Microsoft

AZURE

Fiabilidad en Azure

Alta disponibilidad y recuperación para construir soluciones confiables en Microsoft Azure.

ALTA DISPONIBILIDAD

RESILIENCIA

RECUPERACIÓN

ARQUITECTURAS CONFIABLES

Microsoft Azure

Documentación técnica · Diseño original de portada

© Microsoft Corporation.

Azure Firewall

Funcionalidades nativas de firewall con alta disponibilidad integrada, escalabilidad en la nube sin restricciones y cero...

Azure Kubernetes Service (AKS)

Simplifique la implementación, la administración y las operaciones de Kubernetes

puerta de enlace de Azure ExpressRoute

Conexiones de fibra de red privada dedicadas a Azure

características de confiabilidad de Azure

Obtenga información sobre las características de la plataforma de Azure que le ayudan a crear aplicaciones confiables, incluidas las zonas de disponibilidad y las regiones.

zonas de disponibilidad de Azure

- ¿Qué son las zonas de disponibilidad Azure?
- Habilitación de la resistencia de zona para cargas de trabajo de Azure
- Recursos zonales y resistencia de zona
- Servicios con soporte con zonas de disponibilidad

regiones de Azure

- ¿Qué son las regiones de Azure?
- Regiones emparejadas y no emparejadas
- Servicios con compatibilidad con varias regiones
- Servicios de Azure no regionales
- lista de regiones de Azure

Descripción de los aspectos básicos de confiabilidad

Comprenda conceptos como confiabilidad, resistencia y recuperación en la nube.

Conceptos de confiabilidad

- Responsabilidad compartida para la confiabilidad
- Alta disponibilidad, continuidad empresarial y recuperación ante desastres
- Redundancia, replicación y copia de seguridad

Confiabilidad en la plataforma Azure

- Confiabilidad y soberanía en Azure
- respuesta a incidentes del servicio Azure

- Conmutación por error y conmutación por recuperación
- Cómo leer un acuerdo de nivel de servicio (SLA)

Guía de confiabilidad para cada fase del recorrido en la nube

Planee, diseñe, implemente y funcione cargas de trabajo confiables abordando las consideraciones de estrategia, prácticas operativas y confiabilidad a lo largo del ciclo de vida de la solución.

Estrategia y planificación

Definir la justificación comercial y los resultados esperados de la adopción

- Consideraciones de resistencia para la estrategia en la nube

Implementación y operación

Migración y modernización de la resiliencia de la carga de trabajo

- Preparar su entorno de nube segura para garantizar la disponibilidad
- Adopción del principio de disponibilidad

Arquitectura en la nube

Diseño, compilación y mejora continua de una arquitectura de nube resistente

- Pilar de confiabilidad de Azure Well-Architected Framework
- Principios de diseño para aplicaciones de Azure

Servicios y funcionalidades de confiabilidad

Azure proporciona una gama de servicios que le ayudan a mejorar la confiabilidad de las aplicaciones a través de la conmutación por error, la copia de seguridad, las pruebas, la supervisión y las instrucciones dirigidas.

- Funcionalidades de resistencia en agentes (versión preliminar)
- Resistencia en Azure
- Azure Service Health
- Azure Monitor

Azure Advisor

agente de SRE de Azure (versión preliminar)

Azure Backup

Azure Site Recovery

Azure Chaos Studio

Pruebas de aplicaciones de Azure

Azure Front Door

Azure Traffic Manager

Azure Load Balancer

Confiabilidad en Azure

En entornos de nube, los errores son inevitables. Pueden producirse errores de hardware, defectos de software, errores de configuración, picos de tráfico, interrupciones del centro de datos e incluso interrupciones en toda la región. *La confiabilidad* es la capacidad de una carga de trabajo para mantener las expectativas empresariales, incluso durante errores o interrupciones. Con la arquitectura y las operaciones correctas, los errores no necesitan dar lugar a tiempo de inactividad.

Un tiempo de inactividad no planeado puede tener un gran impacto. Conlleva costos financieros y daña la confianza del cliente y del usuario. Además de estos impactos inmediatos, la confiabilidad deficiente afecta a las obligaciones de cumplimiento y al posicionamiento competitivo. Los equipos dedican más tiempo a la lucha contra incendios y menos tiempo entregando valor. Incluso podría desencadenar penalizaciones contractuales de nivel de servicio.

Azure proporciona servicios administrados y de infraestructura resistentes, pero la confiabilidad de la carga de trabajo depende del diseño y las decisiones operativas que tome: cómo diseñar la arquitectura, configurar servicios, administrar dependencias, automatizar respuestas a problemas y probar software y procesos. El diseño para la confiabilidad le ayuda a minimizar los errores y garantiza que, cuando se produzcan, la carga de trabajo se degrada de maneras predecibles y controladas que se alinean con las prioridades empresariales.

Una carga de trabajo confiable tiene dos propiedades esenciales:

- Es *resistente*, lo que significa que puede absorber errores y cambios en el entorno mientras continúa funcionando en un nivel de servicio aceptable.
- También es *recuperable*, lo que significa que, cuando se produce una interrupción, la carga de trabajo puede restaurar las operaciones normales dentro de los límites de tiempo y pérdida de datos definidos.

Necesita ambas propiedades para satisfacer las expectativas de disponibilidad del mundo real y mantener la continuidad empresarial.

La confiabilidad en Azure funciona en tres capas interconectadas:

- **La aplicación**: las opciones de arquitectura, los procedimientos y los procesos, incluida la administración de dependencias, los runbooks, la automatización y las pruebas
- **Características de confiabilidad**, entre las que se incluyen:

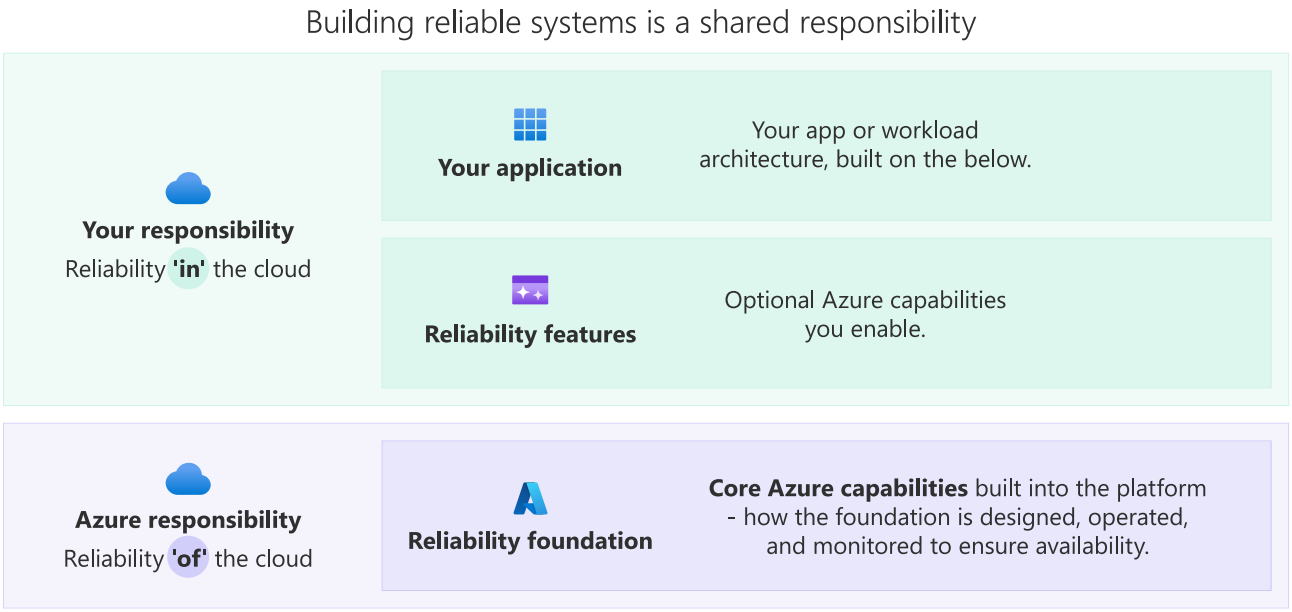
- **Servicios de carga de trabajo y configuración**: configuración de los servicios de Azure que ejecutan la carga de trabajo y las funcionalidades de confiabilidad que configure dentro de esos servicios
  - **Azure servicios de plataforma:** servicios Azure que admiten específicamente la confiabilidad de la carga de trabajo, como equilibrio de carga, DNS, enrutamiento de tráfico y supervisión
- **Fundamentos de confiabilidad**: Azure resistencia integrada, como zonas de disponibilidad, regiones y prácticas de implementación seguras

Estas capas funcionan conjuntamente para determinar la confiabilidad general de la carga de trabajo. Comprenderlos le ayuda a distinguir entre lo que Azure proporciona de forma predeterminada y lo que necesita para diseñar, configurar y operar.

Este modelo le ayuda a aprovechar lo que Azure proporciona al tiempo que se hace responsable de lo que solo se puede diseñar. Este artículo se centra en lo que Azure proporciona en estas capas. Para obtener instrucciones completas sobre el diseño de cargas de trabajo y la capa de operaciones (diseño de soluciones resistentes y patrones arquitectónicos), consulte el marco de Azure Well-Architected y, especialmente, el pilar de confiabilidad.

Responsabilidad compartida para la confiabilidad

La confiabilidad en Azure sigue un modelo de responsabilidad compartida. Microsoft proporciona la plataforma resistente a través de Azure. Diseñe la carga de trabajo resistente.



- **Microsoft** posee los fundamentos de la plataforma y los servicios de confiabilidad de la plataforma. Esta propiedad incluye la infraestructura resistente (redundancia física,

aislamiento de errores y recuperación), zonas de disponibilidad, distribución regional, prácticas de implementación seguras y servicios de nivel de plataforma, como el equilibrio de carga, el enrutamiento del tráfico y la supervisión. Microsoft es responsable de la confiabilidad de la plataforma Azure y de la confiabilidad de los servicios según lo definido por el Acuerdo de Nivel de Servicio publicado y la documentación de cada servicio, como su guía de confiabilidad.

- **Posee** el diseño y las operaciones de la carga de trabajo. Es responsable de traducir las funcionalidades de Azure al comportamiento confiable de la carga de trabajo a través de decisiones de arquitectura, opciones de configuración, prácticas operativas y pruebas.

  Azure no puede conocer los objetivos de disponibilidad, los inconvenientes aceptables, el contexto empresarial o los requisitos específicos de la aplicación. Solo puede definir esos requisitos y diseñarlos en consecuencia. Azure proporciona funcionalidades, al seleccionarlas y configurarlas. Por ejemplo, se configura el comportamiento de conmutación por error para servicios como bases de datos y se definen sondeos de estado del equilibrador de carga que representan con precisión el estado de la aplicación para que las decisiones de enrutamiento del tráfico sean correctas, incluso durante los errores.

  > **Importante**
  >
  > La plataforma proporciona bloques de creación, no garantías de un extremo a otro. Las opciones operativas y de diseño determinan si esas funcionalidades se traducen en una carga de trabajo confiable. Para obtener un desglose completo del modelo de responsabilidad compartida, consulte **Responsabilidad compartida en la nube**.

Los acuerdos de nivel de servicio (SLA) definen los compromisos y expectativas de confiabilidad en los servicios de Azure. Comprender los Acuerdos de Nivel de Servicio es esencial al evaluar los servicios y diseñar para destinos de disponibilidad específicos. Azure los servicios publican compromisos de Acuerdo de Nivel de Servicio que pueden variar según la configuración y el nivel de redundancia. Otros servicios que use de otros proveedores también pueden publicar acuerdos de nivel de servicio. Para obtener instrucciones completas sobre cómo funcionan los SLA y se relacionan con las expectativas de disponibilidad de la carga de trabajo, consulte Acuerdos de nivel de servicio.

Fundamentos resistentes en Azure

Azure está diseñado con varias capas de resistencia integradas en la propia plataforma. Estas bases proporcionan las funcionalidades base sobre las que se crean cargas de trabajo confiables:

- **Resistencia de la infraestructura física:** Azure centros de datos están diseñados con infraestructura redundante, como la alimentación, la refrigeración y la conectividad de red. El controlador de tejido Azure aísla y administra automáticamente los errores de hardware. Detecta errores y organiza la migración de cargas de trabajo a hardware correcto sin intervención del cliente.

- **Regiones:** Azure opera en más de 70 regiones de todo el mundo. Esta distribución geográfica permite que las cargas de trabajo resistan interrupciones y interrupciones en toda la región a través de funcionalidades de conmutación por error planeadas, al tiempo que siguen los requisitos de residencia de datos. Para obtener más información, consulte introducción a las regiones de Azure.

- **Zonas de disponibilidad:** Muchas Azure regiones incluyen centros de datos físicamente independientes dentro de la misma región. Estas zonas de disponibilidad están conectadas por redes de alta velocidad y baja latencia. Cada zona tiene energía, refrigeración y redes independientes para protegerse frente a errores de nivel de centro de datos, a la vez que mantiene funcionalidades de replicación sincrónicas para muchos servicios de Azure. Para obtener más información, consulte ¿Qué son las zonas de disponibilidad?

- **Implementaciones seguras:** Azure implementa procesos de implementación controlados y escalonados para actualizaciones de plataforma y cambios en el servicio para minimizar el riesgo de interrupción generalizada del servicio. La plataforma implementa actualizaciones gradualmente, como en dominios de error, zonas de disponibilidad y regiones, con funcionalidades de reversión automatizadas cuando detecta problemas. Este enfoque garantiza que los cambios en la plataforma no introduzcan riesgos de confiabilidad en las cargas de trabajo de los clientes.

Azure proporciona automáticamente estas bases de nivel de plataforma. Forman la capa base en la que se crean cargas de trabajo confiables. Sin embargo, aún debe configurar los servicios correctamente y combinar estas funcionalidades de maneras que cumplan sus requisitos empresariales específicos.

Azure servicios que admiten la confiabilidad

Azure proporciona funcionalidades de plataforma que convierten los conceptos de confiabilidad en bloques de creación accionables para la arquitectura. Estas funcionalidades funcionan

conjuntamente para habilitar arquitecturas resistentes y muchos servicios Azure incluyen implementaciones integradas de estos patrones:

Expandir tabla

Capability	Description
tecnología de **inteligencia artificial Optimización de confiabilidad con tecnología de inteligencia artificial**	Usa funcionalidades controladas por ia para evaluar y mejorar la confiabilidad. Estas funcionalidades analizan patrones de carga de trabajo, identifican posibles riesgos y proporcionan recomendaciones que ayudan a los equipos a pasar de la solución de problemas reactiva a la administración proactiva de confiabilidad. Estas funcionalidades exponen señales de confiabilidad y las convierten en recomendaciones que ayudan a los equipos a priorizar y mejorar la confiabilidad a lo largo del tiempo. Azure proporciona varios servicios de confiabilidad controlados por ia, entre los que se incluyen: - Azure funcionalidades de resistencia - Funcionalidades de resistencia en agentes (versión preliminar) en Azure Copilot - Agente de SRE de Azure Estos servicios analizan los patrones de carga de trabajo y sugieren mejoras para optimizar la posición de confiabilidad en todo el panel.
Equilibrio de carga y administración del tráfico	Permite la confiabilidad mediante el enrutamiento del tráfico entre instancias redundantes y el control de la conmutación por error automáticamente. Esta funcionalidad es esencial para mantener la disponibilidad del servicio cuando se produce un error en los componentes individuales. Azure proporciona equilibrio de carga en varias capas, entre las que se incluyen: - Azure Load Balancer para la distribución de tráfico TCP/UDP de nivel 4 - Azure Application Gateway para el enrutamiento HTTP de nivel 7 con comprobaciones de estado compatibles con la aplicación - Azure Front Door para el equilibrio de carga HTTP global con conmutación por error rápida - Azure Traffic Manager para la distribución del punto de conexión global basado en DNS Si necesita ayuda para decidir qué equilibrador de carga usar para su escenario, consulte Opciones de equilibrio de carga.
y protección de datos	Protege contra la pérdida de datos y los daños, y habilita la recuperación después de que se produzca un problema. Azure proporciona varias funcionalidades de copia de seguridad y recuperación, entre las que se incluyen: - Azure Backup para la copia de seguridad centralizada con retención

Capability	Description
	configurable para máquinas virtuales, contenedores de blobs, archivos y algunas bases de datos - Características nativas de copia de seguridad y restauración en muchos servicios de Azure, incluidos la mayoría de los servicios de base de datos - Bicep y otra infraestructura como herramientas de código para la copia de seguridad de configuración, protección contra desfase y recreación rápida del entorno Revise cada guía de confiabilidad del servicio Azure para comprender los enfoques de copia de seguridad que admite el servicio.
Geográfica y conmutación por error	Habilita la recuperación de errores regionales a través de la replicación de datos entre regiones y las funcionalidades de conmutación por error automatizadas. Azure Site Recovery orquesta la recuperación ante desastres para cargas de trabajo de máquina virtual con la replicación automatizada y la secuenciación de conmutación por error. Muchos servicios Azure también proporcionan características integradas de replicación geográfica, entre las que se incluyen: - Azure Cosmos DB con funcionalidades nativas de replicación de datos entre regiones y conmutación por error - Azure API Management con funcionalidades nativas de conmutación por error entre regiones Las guías de confiabilidad del servicio detallan las opciones de replicación geográfica disponibles para cada servicio.
y observabilidad	Proporciona visibilidad completa de la posición de confiabilidad y permite una respuesta proactiva a los problemas. Azure proporciona varios servicios de observabilidad, entre los que se incluyen: - Azure Monitor y Application Insights para la supervisión, alertas y seguimiento de dependencias en tiempo real - Modelos de mantenimiento en Azure Monitor para supervisar el estado de una carga de trabajo completa - Azure Service Health para alertas personalizadas e instrucciones para problemas de servicio de Azure que pueden afectar a las cargas de trabajo Juntas, estas funcionalidades le ayudan a comprender si cumple los requisitos de confiabilidad y responde rápidamente cuando surgen problemas.
Pruebas de confiabilidad y validación	Ayuda a comprobar que las cargas de trabajo se comportan según lo esperado en condiciones de error y ayuda a garantizar que la solución cumpla los requisitos de confiabilidad antes de que los problemas afecten a los usuarios. Azure proporciona servicios de pruebas de confiabilidad, entre los que se

Capability	Description
	incluyen: - Azure Chaos Studio para experimentos controlados de inyección de errores para validar el comportamiento de recuperación automática y la resistencia a errores reales - Pruebas de aplicaciones de Azure para pruebas funcionales y de rendimiento para comprender cómo se comportan las aplicaciones, incluido el estrés

Habilitación de la confiabilidad en servicios de Azure

La plataforma proporciona funcionalidades de confiabilidad de nivel de servicio a través de docenas de servicios Azure, cada uno con puntos fuertes y casos de uso específicos. La guía de confiabilidad de cada servicio proporciona detalles sobre cómo el servicio puede permanecer disponible durante diferentes escenarios, por ejemplo:

- **Errores transitorios**, que son errores intermitentes cortos. Las guías de servicio proporcionan recomendaciones para los procedimientos recomendados para minimizar su impacto, como reintentar y usar SDK proporcionados por Microsoft.
- **Errores de zona** de disponibilidad, por lo que el servicio puede redirigir automáticamente las solicitudes para mantener la alta disponibilidad.
- **Errores en toda** la región, por lo que el servicio puede conmutar por error a una región secundaria y seguir funcionando durante una interrupción de la región.

Para ver las guías de confiabilidad de muchos servicios Azure, consulte Guías de confiabilidad por servicio.

Diseño de cargas de trabajo confiables

La confiabilidad surge de un ciclo continuo de definición de destinos, diseño para errores y recuperación rápida, suposiciones de prueba y mejora a través del aprendizaje operativo. Use las guías de confiabilidad de Azure Well-Architected Framework y servicio para comprender lo que proporciona cada servicio de forma predeterminada y lo que requiere una configuración explícita. Para obtener un enfoque estructurado para implementar la confiabilidad, consulte el modelo de madurez de confiabilidad de Azure Well-Architected Framework.

A medida que diseña, conecte *conceptos fundamentales* con *conceptos técnicos*. Los conceptos fundamentales, como la continuidad empresarial y la responsabilidad compartida , definen los resultados que necesita lograr. Los conceptos técnicos como redundancia, replicación, copia de seguridad, conmutación por error y conmutación por recuperación definen cómo implementar

esos resultados en la arquitectura y las operaciones. Los contratos de nivel de servicio le ayudan a comprender las garantías que recibe de sus proveedores de servicios.

Soberanía y residencia de datos

Al diseñar la confiabilidad, incluya los requisitos de soberanía y residencia de datos al principio porque afectan a la selección de regiones, la estrategia de replicación y las rutas de conmutación por error. Una arquitectura resiliente puede seguir incumpliendo los requisitos de cumplimiento normativo si la conmutación por error o el movimiento de datos atraviesan límites restringidos. Para obtener más información, consulte Confiabilidad y soberanía.

La confiabilidad en Azure se logra mediante la combinación de bases de plataforma resistentes con un diseño y operaciones de carga de trabajo cuidadosas. Al comprender qué Azure proporciona y dónde necesita tomar decisiones de diseño y configuración, puede crear sistemas que sigan cumpliendo las expectativas empresariales, incluso en presencia de errores.

ⓘ**Nota:** El autor ha creado este artículo con ayuda de inteligencia artificial. Más información

---

Last updated on 14/08/2026

Guías de confiabilidad por servicio

En este artículo se proporcionan vínculos a instrucciones de confiabilidad para muchos servicios de Azure. La mayoría de las guías de confiabilidad contienen la siguiente información:

- *Las recomendaciones de implementación de producción* proporcionan instrucciones sobre cómo implementar el servicio para satisfacer los requisitos de confiabilidad en entornos de producción.
- *La resistencia a errores transitorios* describe cómo el servicio controla los errores transitorios diarios que pueden producirse en la nube. También se describe cómo controlar estos errores en la aplicación, incluida la información sobre las directivas de reintento, los tiempos de espera y otros procedimientos recomendados.
- *Información general sobre la arquitectura de confiabilidad* es una sinopsis de cómo el servicio admite la confiabilidad. Incluye información sobre qué componentes Microsoft administra y qué componentes administra, características de redundancia integradas y cómo aprovisionar y administrar varios recursos, si procede.
- *La resistencia a los errores de zona de disponibilidad* describe cómo el servicio admite zonas de disponibilidad, requisitos que debe cumplir para usar zonas de disponibilidad, cómo se enruta el tráfico y los datos se replican entre zonas, qué ocurre cuando una zona experimenta una interrupción, recuperación de zona y cómo configurar los recursos para la compatibilidad con zonas de disponibilidad.
- *La resiliencia a los fallos a nivel de región* describe si el servicio proporciona capacidades multirregionales, los requisitos para utilizarlas, cómo se enruta el tráfico y se replican los datos entre regiones, la experiencia cuando una región no está disponible, la compatibilidad con conmutación por error y recuperación, y cómo implementar soluciones personalizadas multirregionales.
- *La resistencia al mantenimiento del servicio* describe cómo el servicio controla los eventos de mantenimiento planeado, incluido cómo minimizar el tiempo de inactividad y la pérdida de datos durante estos eventos. También se muestra cómo configurar el servicio para mejorar la resistencia durante los tiempos de mantenimiento.
- *Contratos de nivel de servicio (SLA),* que definen y describen el tiempo de actividad esperado y cómo cambia el tiempo de actividad esperado en función de la configuración que use.

- *Copia de seguridad y recuperación* para los servicios admitidos, incluidos quién controla y administra las copias de seguridad, donde se almacenan y replican, cómo se pueden recuperar y si solo son accesibles dentro de una región o entre regiones.

Guías de confiabilidad por servicio

En la tabla siguiente se proporcionan vínculos a instrucciones de confiabilidad para los servicios de Azure. Cada guía contiene información sobre cómo el servicio admite características de confiabilidad.

> **Nota:**
>
> Algunos documentos no siguen un único formato de guía de confiabilidad. Estos servicios pueden enumerar más de un artículo que hace referencia a instrucciones de confiabilidad.

Expandir tabla

Service	Guía de confiabilidad	Otra documentación de confiabilidad
	Confiabilidad en la búsqueda de IA	
centro de API Azure CENTRO de API	Confiabilidad en el Centro de API de Azure	
	Confiabilidad en API Management	
	Confiabilidad en Azure App Configuration	
	Confiabilidad en App Service	
App Service Environment Azure App Service App Service Environment	Relibilidad en App Service Environment	
Azure Application Gateway for Containers	Confiabilidad en Puerta de enlace de aplicaciones para contenedores	
Azure Application Gateway v2	Confiabilidad en Azure Application Gateway	

Service	Guía de confiabilidad	Otra documentación de confiabilidad
	Confiabilidad en la Copia de seguridad	
	Relibilidad en Azure Bastion	
	Confiabilidad en Batch	
	Fiabilidad en el Almacenamiento de Blobs	
	Relibilidad en Bot Service	
	Reliability en Chaos Studio	
	Confiabilidad en Container Apps	
	Relibilidad en Container Instances	
	Confiabilidad en Container Registry	
	Confiabilidad en Azure Cosmos DB	
		Recuperación de datos si se produce un error en toda una región
	Confiabilidad en Azure Data Explorer	
	Reliability en Azure Data Factory	
Azure Data Manager for Energy		Confiabilidad en Azure Data Manager for Energy
		recuperación Disaster para Data Share
	Confiabilidad en Azure Database for MySQL	
	Confiabilidad en Azure Database for PostgreSQL	

Service	Guía de confiabilidad	Otra documentación de confiabilidad
	Relibilidad en Azure Databricks	
Azure DDoS Protection	Confiabilidad en DDoS Protection	
Azure Registro de dispositivos	Confiabilidad en el Registro de dispositivos	
		Información general sobre la protección de datos
	Confiabilidad en Azure Disk Storage	
	Relibilidad en Azure DNS	
Azure DocumentDB	Confiabilidad en Azure DocumentDB	
	Confiabilidad en Elastic SAN	
	Confiabilidad en Event Grid	
	Relibilidad en Azure Event Hubs	
	Confiabilidad en Azure ExpressRoute	
	Confiabilidad en Azure Files	
	Relibilidad en Azure Firewall	
	Confiabilidad de Azure Functions	
		Recuperación ante desastres para Health Data Services
Azure Health Data Services: Servicio de desidentificación		Confiabilidad en el servicio de desidentificación de Health Data Services
Azure Health Data Services: Servicios de área de trabajo (FHIR®, DICOM®, medtech)		Consideraciones sobre continuidad empresarial y recuperación ante desastres

Service	Guía de confiabilidad	Otra documentación de confiabilidad
	Confiabilidad en HDInsight	
	Relibilidad en IoT Hub	
	Confiabilidad en Key Vault	
Azure Key Vault HSM administrado	Confiabilidad en HSM administrado de Azure Key Vault	
	Confiabilidad en AKS	
	Relibilidad en Load Balancer	
	Confiabilidad en Logic Apps	
	Relibilidad en Azure Managed Grafana	
		Conmutación por error para la continuidad empresarial y recuperación ante desastres
Azure Managed Redis	Confiabilidad en Azure Managed Redis	
		Azure Migrate y copia de seguridad y recuperación ante desastres
registros de Azure Monitor	Confiabilidad en los registros de Azure Monitor	
	Relibilidad en Azure NAT Gateway	
	Relibilidad en Azure NetApp Files	
		Network Watcher disponibilidad y redundancia del servicio
	Confiabilidad en Notification Hubs	
servicio Azure Private Link	Confiabilidad en el servicio Azure Private Link	

Service	Guía de confiabilidad	Otra documentación de confiabilidad
Azure direcciones IP públicas		Azure zona de disponibilidad de direcciones IP públicas
	Confiabilidad en Queue Storage	
		Preguntas frecuentes sobre Route Server
	Reliability en Service Bus	
		Desplegar un clúster de Service Fabric a través de zonas de disponibilidad recuperación de Disaster en Service Fabric
	Relibilidad en Azure SignalR Service	
	Reliability en Azure Site Recovery	
	Confiabilidad en Azure SQL Database	
	Reliability en Azure SQL Managed Instance	
Azure Storage	Confiabilidad en Acciones de almacenamiento	
de Azure Storage	Confiabilidad en la detección de almacenamiento	
Mover	Confiabilidad en Storage Mover	
	Reliability en Azure Stream Analytics	
	Confiabilidad en Table Storage	
	Confiabilidad en Traffic Manager	
	Relibilidad en Virtual Machines	

Service	Guía de confiabilidad	Otra documentación de confiabilidad
Azure VM Image Builder	Confiabilidad en Azure VM Image Builder	
	Confiabilidad en Conjuntos de Escalado de Máquinas Virtuales	
	Reliability en Virtual Network	
	Confiabilidad en Azure Virtual Network Manager	
		Zonas y resistencia de disponibilidad en Virtual WAN Diseño de recuperación ante desastres
	Confiabilidad en Azure VMware Solution	
	Confiabilidad en VPN Gateway	
	Confiabilidad en Azure Web PubSub Service	
	Reliability en Microsoft Fabric	

Contenido relacionado

- Azure servicios que admiten zonas de disponibilidad
- Lista de regiones de Azure
- Compilación de soluciones para alta disponibilidad mediante zonas de disponibilidad

ⓘ**Nota:** El autor ha creado este artículo con ayuda de inteligencia artificial. Más información

Last updated on 25/06/2026

Seguridad en Azure AI Search

Azure AI Search es una infraestructura de búsqueda escalable que indexa contenido heterogéneo y permite la recuperación a través de API, aplicaciones y agentes de IA. Es adecuado para escenarios de búsqueda empresarial y experiencias de clientes con tecnología de inteligencia artificial que requieren generación dinámica de contenido a través de modelos de finalización de chat. Como servicio de Azure, AI Search proporciona una variedad de funcionalidades para satisfacer sus requisitos de confiabilidad.

Cuando se usa Azure, la confiabilidad es una responsabilidad compartida. Microsoft proporciona una variedad de funcionalidades para admitir resistencia y recuperación. Es responsable de comprender cómo funcionan esas funcionalidades dentro de todos los servicios que usa y de seleccionar las funcionalidades que necesita para cumplir los objetivos empresariales y los objetivos de tiempo de actividad.

En este artículo se describe cómo hacer que Azure AI Search sea resistente a una variedad de posibles interrupciones y problemas, incluidos errores transitorios, interrupciones de zona de disponibilidad, interrupciones de regiones y mantenimiento del servicio. También se describe cómo puede usar copias de seguridad para recuperarse de otros tipos de problemas y se resalta cierta información clave sobre el acuerdo de nivel de servicio (SLA) de Azure AI Search.

Recomendaciones de implementación de producción para la confiabilidad

En el caso de las cargas de trabajo de producción, se recomienda que usted:

- ✓ Use un nivel facturable que tenga al menos dos réplicas. Esta configuración hace que su servicio de búsqueda sea más resistente a errores transitorios y operaciones de mantenimiento. También cumple el acuerdo de nivel de servicio (SLA) para la búsqueda de IA. El Acuerdo de Nivel de Servicio requiere dos réplicas para cargas de trabajo de solo lectura y tres o más réplicas para cargas de trabajo de lectura y escritura.
- ✓ No use el nivel Gratis para su uso en producción. Ai Search no proporciona un Acuerdo de Nivel de Servicio para el nivel Gratis, que está limitado a una réplica.

Introducción a la arquitectura de confiabilidad

Cuando se usa AI Search, se crea un *servicio de búsqueda*. Cada servicio de búsqueda admite muchos *índices de búsqueda* que almacenan el contenido que se puede buscar.

La búsqueda de IA no está diseñada como un almacén de datos principal. En su lugar, se usan *indexadores* para conectar el servicio de búsqueda a orígenes de datos externos. Un indexador

rastrea los datos de origen, invoca *aptitudes* que realizan el procesamiento y el enriquecimiento, y rellena el índice con las salidas de aptitud.

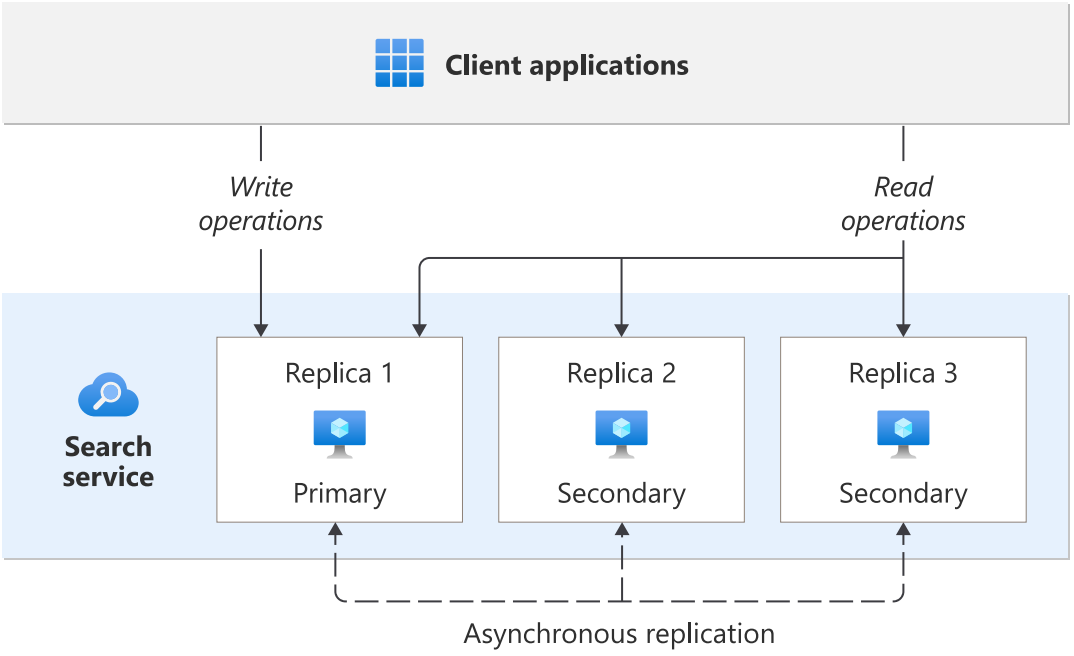
También configurará el número de *réplicas* para el servicio. En AI Search, una réplica es una copia del motor de búsqueda de tu servicio. Puede pensar en una réplica como si representara una sola máquina virtual (VM). Cada servicio Search puede tener entre 1 y 12 réplicas.

La adición de varias réplicas permite que AI Search pueda:

- Aumente la disponibilidad del servicio Search.
- Realice el mantenimiento en una réplica mientras las consultas continúan ejecutándose en otras réplicas.
- Controle cargas de trabajo de consulta e indexación más altas.
- Mejore la resistencia intentando aprovisionar réplicas en diferentes zonas de disponibilidad, si la región las admite.

Ai Search asigna automáticamente una réplica para que sea la *réplica principal*. Todas las operaciones de escritura se realizan en esa réplica. Las otras réplicas se usan para las operaciones de lectura.

En el diagrama siguiente se muestra cómo un servicio Search con tres réplicas se puede distribuir entre tres zonas de disponibilidad:



También puede configurar el número de *particiones*, que representan el almacenamiento que usan los índices de búsqueda.

Es importante comprender el impacto de agregar réplicas y particiones, ya que cada uno afecta al rendimiento de lectura y escritura de maneras diferentes. Para obtener más información sobre las réplicas y las particiones, consulte Estimación y administración de la capacidad de un servicio Search.

Resistencia a errores transitorios

Los errores transitorios son errores breves e intermitentes en los componentes. Se producen con frecuencia en un entorno distribuido como la nube y son una parte normal de las operaciones. Los errores transitorios se corrigen después de un breve período de tiempo. Es importante que las aplicaciones puedan controlar errores transitorios, normalmente mediante el reintento de solicitudes afectadas.

Todas las aplicaciones hospedadas en la nube deben seguir las instrucciones de control de errores transitorios de Azure cuando se comunican con cualquier API, bases de datos y otros componentes hospedados en la nube. Para obtener más información, consulte Recomendaciones para controlar errores transitorios.

Los indexadores de AI Search tienen un control de errores transitorios integrado. Si un origen de datos no está disponible brevemente, el indexador está diseñado para recuperarse y volver a intentarlo. Usa el seguimiento de cambios para reanudar la indexación del último documento indizado correctamente.

Los servicios de búsqueda pueden experimentar errores transitorios durante las operaciones de mantenimiento estándar y no programadas. Búsqueda de Azure AI no proporciona notificaciones anticipadas ni permite la programación de mantenimiento en momentos específicos. Aunque se hace todo lo posible por minimizar el tiempo de inactividad, incluso en los servicios de réplica única, pueden producirse breves interrupciones. Para mejorar la resistencia frente a estos errores transitorios, se recomienda usar dos o más réplicas.

Si construyes aplicaciones que interactúan con AI Search, estas deben manejar errores transitorios. Use una estrategia de reintento con retrocesos exponenciales para las operaciones de lectura y escritura.

Resistencia a errores de zona de disponibilidad

Las zonas de disponibilidad son grupos físicamente independientes de centros de datos dentro de una región de Azure. Cuando una zona falla, los servicios pueden transferirse a una de las zonas restantes.

AI Search tiene redundancia de zona, lo que significa que las réplicas se distribuyen a través de varias zonas de disponibilidad dentro de la región del servicio de búsqueda.

Al agregar dos o más réplicas al servicio, AI Search intenta colocar cada réplica en una zona de disponibilidad diferente. En el caso de los servicios que tienen más réplicas que las zonas disponibles, las réplicas se distribuyen entre zonas lo más uniformemente posible.

En el diagrama siguiente se muestra cómo se puede implementar un servicio de búsqueda de ejemplo con cuatro réplicas en tres zonas de disponibilidad:

> ⓘ **Importante**
>
> La búsqueda de IA no garantiza la ubicación exacta de las réplicas. La selección de ubicación está sujeta a restricciones de capacidad, operaciones de escalado y otros factores.

Requisitos

La redundancia de zona se habilita automáticamente cuando el servicio Search cumple todos los criterios siguientes:

- **Compatibilidad con regiones:** La compatibilidad con zonas de disponibilidad depende de la infraestructura y el almacenamiento. Para obtener una lista de las regiones admitidas, consulte Elección de una región para la búsqueda de IA.
- **Nivel:** El servicio debe estar en el nivel Básico o superior.
- **Número de réplicas:** El servicio debe tener al menos dos réplicas.

  > ⓘ **Nota:**
  >
  > Ai Search intenta distribuir réplicas entre varias zonas cuando tiene dos o más réplicas. Sin embargo, para las cargas de trabajo de lectura y escritura, debería usar tres o más réplicas para lograr el SLA de disponibilidad más alto posible.

Distribución de instancias a través de zonas

AI Search intenta colocar réplicas en diferentes zonas de disponibilidad. Sin embargo, en ocasiones hay situaciones en las que todas las réplicas de un servicio de búsqueda pueden colocarse en la misma zona de disponibilidad. Esta situación puede ocurrir cuando las réplicas se quitan del servicio, como cuando se *reduce horizontalmente* configurando el servicio para que use menos réplicas. La eliminación de réplicas no hace que las réplicas restantes se redistribuyan automáticamente entre las zonas de disponibilidad.

Para reducir la probabilidad de que todas las réplicas se coloquen en una sola zona de disponibilidad, puede desencadenar manualmente una operación de escalabilidad horizontal inmediatamente después de una operación de reducción horizontal. Por ejemplo, supongamos que el servicio Search tiene 10 réplicas y desea reducir horizontalmente a 7 réplicas. En lugar de realizar una sola operación de escalado, puede escalar temporalmente a 6 instancias y, a continuación, escalar inmediatamente a 7 instancias para desencadenar el reequilibrio de zona.

Cost

Cada servicio de búsqueda comienza con una réplica. La redundancia de zona requiere dos o más réplicas, lo que aumenta el costo de ejecutar el servicio. Para comprender las implicaciones de facturación de las réplicas, utilice la calculadora de precios.

Configurar soporte de zonas de disponibilidad

Si su servicio de búsqueda cumple los requisitos de redundancia de zona, no es necesaria ninguna configuración adicional. Siempre que sea posible, AI Search procura colocar las réplicas en diferentes zonas de disponibilidad.

Planeamiento y administración de capacidad

Para prepararse para un fallo en la zona de disponibilidad, considere sobreaprovisionar el número de réplicas. El aprovisionamiento excesivo permite al servicio de búsqueda tolerar cierta pérdida de capacidad y seguir funcionando sin un rendimiento degradado. Agregar réplicas durante una interrupción es difícil, por lo que el aprovisionamiento excesivo ayuda a garantizar que el servicio de búsqueda pueda controlar los volúmenes de solicitudes normales, incluso con una capacidad reducida. Para más información, consulte Administración de la capacidad mediante el sobreaprovisionamiento.

Comportamiento cuando todas las zonas están en buen estado

En esta sección se describe qué esperar cuando los servicios Search están configurados para la redundancia de zona y todas las zonas de disponibilidad están operativas.

- **Enrutamiento de tráfico entre zonas:** AI Search ejecuta el equilibrio de carga automático de todas las consultas y escrituras a través de todas las réplicas disponibles. AI Search puede enviar operaciones de lectura a cualquier réplica de cualquier zona de disponibilidad. Envía operaciones de escritura a una sola réplica principal que el servicio AI Search selecciona.

- **Replicación de datos entre zonas:** Los cambios en los datos se replican automáticamente entre réplicas entre zonas de disponibilidad. La replicación se produce de forma asincrónica, lo que significa que las escrituras se confirman en una réplica principal antes de que se repliquen en otras réplicas.

Comportamiento durante un fallo de zona

En esta sección se describe qué esperar cuando los servicios de búsqueda están configurados para la redundancia de zona y se produce una interrupción de zona de disponibilidad.

- **Detección y respuesta:** Ai Search es responsable de detectar un error en una zona de disponibilidad. No es necesario hacer nada para iniciar una conmutación por error de la zona.

- **Notificación**: Microsoft no le notifica automáticamente cuando una zona está inactiva. Sin embargo, puede usar Azure Resource Health para supervisar el estado de un recurso individual y puede configurar alertas de Resource Health para notificarle problemas. También puede usar Azure Service Health para comprender el estado general del servicio, incluidos los errores de zona, y puede configurar alertas de Service Health para notificarle problemas.

- **Solicitudes activas:** Se finalizan las solicitudes que las réplicas procesan en la zona con errores. Los clientes deben reintentar las solicitudes siguiendo las instrucciones para controlar errores transitorios.

- **Pérdida de datos esperada:** Si la zona de disponibilidad afectada solo contiene réplicas de lectura, no se espera ninguna pérdida de datos.

  Si se pierde la réplica principal porque estaba en la zona afectada, es posible que se pierdan las operaciones de escritura que aún no se hayan replicado.

- **Tiempo de inactividad esperado:** En la mayoría de las situaciones, no se espera que un error de zona provoque tiempo de inactividad en el servicio de búsqueda para las

operaciones de lectura porque las réplicas de lectura de otras zonas de disponibilidad siguen atendiendo solicitudes.

Si se pierde la réplica principal porque estaba en la zona afectada, AI Search promueve automáticamente otra réplica para convertirse en la nueva principal para que las operaciones de escritura se puedan reanudar. Normalmente, la promoción de la réplica tarda unos segundos en producirse. Durante este tiempo, es posible que las operaciones de escritura no se realicen correctamente. Asegúrese de que las aplicaciones están preparadas siguiendo las instrucciones de control de errores transitorios.

Sin embargo, hay algunas situaciones poco probables en las que todas las réplicas de tu servicio de búsqueda podrían encontrarse en una única zona de disponibilidad. En este escenario, puede experimentar una interrupción del servicio hasta que se recupere la zona. Para obtener más información y comprender una solución alternativa, consulte Distribución de instancias.

- **Reenrutamiento del tráfico:** Cuando se produce un error en una zona, AI Search detecta el error y enruta las solicitudes a réplicas activas en las zonas supervivientes. Si se pierde la réplica principal, se promueve otra réplica para que sea la nueva principal.

Recuperación de zona

Cuando se recupera la zona de disponibilidad, AI Search restaura automáticamente las operaciones normales y comienza a enrutar el tráfico a las réplicas disponibles en todas las zonas, incluida la zona recuperada.

Prueba de fallos de zona

AI Search administra el enrutamiento del tráfico para los servicios con redundancia de zona. No es necesario iniciar ni validar ningún proceso de error de zona.

Resistencia a errores en toda la región

AI Search es un servicio de una sola región. Si la región deja de estar disponible, su servicio de búsqueda también dejará de estar disponible.

Soluciones personalizadas de varias regiones para la resistencia

Opcionalmente, puede implementar varios servicios de búsqueda de IA en diferentes regiones. Es responsable de implementar y configurar servicios independientes en cada región. Si crea

una implementación idéntica en una región secundaria de Azure que usa una arquitectura de varias regiones, la aplicación será menos susceptible a un desastre de una sola región.

Cuando se sigue este enfoque, es necesario sincronizar los índices entre regiones para recuperar el último estado de la aplicación. También debe configurar las directivas de equilibrio de carga y conmutación por error.

Para optimizar el rendimiento de la solución general, busque oportunidades para realizar la indexación en réplicas de solo lectura de los orígenes de datos. Por ejemplo, algunos indexadores admiten la lectura desde las réplicas de lectura de un origen de datos distribuido geográficamente.

Para más información, consulte Implementaciones de varias regiones en Búsqueda de Azure AI.

Copias de seguridad y restauración

Dado que AI Search no es una solución de almacenamiento de datos principal, no proporciona opciones de copia de seguridad y restauración de autoservicio. Sin embargo, puede usar la muestra `index-backup-restore` para .NET o Python para hacer una copia de seguridad de la definición del índice y sus documentos en una serie de archivos JSON, que luego se usan para restaurar el índice.

Sin embargo, si elimina accidentalmente el índice y no tiene una copia de seguridad, puede volver a generar el índice. La regeneración implica volver a crear el índice en su servicio de búsqueda y luego volver a cargarlo recuperando los datos de su almacén de datos principal.

Acuerdo de nivel de servicio

El contrato de nivel de servicio (SLA) para los servicios de Azure describe la disponibilidad esperada de cada servicio y las condiciones que la solución deberá cumplir para lograr esa expectativa de disponibilidad. Para obtener más información, consulte Acuerdos de Nivel de Servicio para servicios en línea.

En AI Search, el Acuerdo de Nivel de Servicio de disponibilidad se aplica a los servicios de búsqueda que:

- Están configurados para usar un nivel facturable.
- Tener al menos dos réplicas para cargas de trabajo de solo lectura (consultas).
- Tener al menos tres réplicas para cargas de trabajo de lectura y escritura (consultas e indexación).

Contenido relacionado

- Confiabilidad en Azure
- Límites de servicio en AI Search
- Elección de una región para la búsqueda de IA
- Elección de un plan de tarifa para AI Search
- Planeamiento o adición de capacidad en la búsqueda de IA

ⓘ **Nota:** El autor ha creado este artículo con ayuda de inteligencia artificial. Más información

---

Last updated on 22/01/2026

Confiabilidad en Azure Bot Service

En este artículo se describe el soporte para la confiabilidad en Azure Bot Service. Cubre tanto la confiabilidad regional con zonas de disponibilidad como la resiliencia entre regiones con recuperación ante desastres para bots con residencia local de datos. Para obtener información general más detallada sobre la confiabilidad de Azure, consulte Confiabilidad de Azure.

Al crear una aplicación (bot) en Bot Service, puede elegir la residencia de datos global o local. La residencia de datos locales garantiza que los datos personales del bot se conserven, almacenen y procesen dentro de límites geográficos específicos, como la Unión Europea. Este enfoque le ayuda a cumplir los requisitos de cumplimiento y confiabilidad regional.

> **Importante**
>
> La compatibilidad con la zona de disponibilidad no está habilitada para ningún canal estándar en el servicio de bot regional.

Para más información sobre la implementación de bots con residencia de datos local y cumplimiento regional, consulte Regionalización en Azure Bot Service.

Soporte para zonas de disponibilidad

Las zonas de disponibilidad son grupos físicamente independientes de centros de datos dentro de una región de Azure. Cuando una zona falla, los servicios pueden transferirse a una de las zonas restantes.

En el caso de los bots regionales, Azure Bot Service admite la redundancia de zona de manera predeterminada. No es necesario configurarlo ni volver a configurarlo para la compatibilidad con las zonas de disponibilidad.

Prerrequisitos

- El bot debe ser regional (no global).
- Actualmente, solo la región "westeurope" admite zonas de disponibilidad.

Experiencia a nivel de zona

Durante una interrupción en toda la zona, el cliente debe esperar una breve degradación del rendimiento hasta que la recuperación automática del servicio vuelva a equilibrar la capacidad subyacente para ajustarse a las zonas en estado correcto. Esto no depende de la restauración

de la zona. Se espera que el estado de recuperación automática del servicio administrado por Microsoft compense una zona perdida, lo que usaría la capacidad de otras zonas.

Recuperación ante desastres entre regiones en una geografía de varias regiones

La recuperación ante desastres (DR) hace referencia a las prácticas que las organizaciones usan para recuperarse de eventos de alto impacto, como desastres naturales o implementaciones con errores que producen tiempo de inactividad y pérdida de datos. Independientemente de la causa, el mejor remedio para un desastre es un plan de recuperación ante desastres bien definido y probado y un diseño de aplicaciones que apoye activamente la recuperación ante desastres. Antes de empezar a crear el plan de recuperación ante desastres, consulte Recomendaciones para diseñar una estrategia de recuperación ante desastres.

Para la recuperación ante desastres, Microsoft usa el modelo de responsabilidad compartida. En este modelo, Microsoft garantiza que la infraestructura de línea base y los servicios de plataforma estén disponibles. Sin embargo, muchos servicios de Azure no replican automáticamente datos ni se revierten de una región con errores para realizar la replicación cruzada en otra región habilitada. En esos servicios, es responsable de configurar un plan de recuperación ante desastres válido para su carga de trabajo. La mayoría de los servicios que se ejecutan en ofertas de plataforma como servicio (PaaS) de Azure proporcionan características e instrucciones para admitir la recuperación ante desastres. Puede usar características específicas del servicio para admitir la recuperación rápida con el fin de contribuir al desarrollo del plan de recuperación ante desastres.

Azure Bot Service se ejecuta en modo activo-activo para los servicios globales y regionales. Cuando se produce una interrupción, no es necesario detectar errores ni administrar el servicio. Azure Bot Service realiza automáticamente la conmutación por error automática y la recuperación automática en una arquitectura geográfica de varias regiones. Para el servicio regional de bots de la UE, Azure Bot Service proporciona dos regiones completas dentro de Europa con replicación activa/activa para garantizar la redundancia. Para el servicio de bot global, todas las regiones o zonas geográficas disponibles se pueden servir como superficie global.

Pasos siguientes

Confiabilidad en Azure

Last updated on 22/01/2026

Confiabilidad en Azure Databricks

Azure Databricks es una plataforma de inteligencia artificial y datos basados en Apache Spark colaborativa optimizada para Microsoft Azure. Proporciona un entorno unificado para cargas de trabajo de macrodatos e inteligencia artificial y combina lo mejor de Databricks y Azure para simplificar la ingeniería de datos, la ciencia de datos y el aprendizaje automático.

Cuando se usa Azure, la confiabilidad es una responsabilidad compartida. Microsoft proporciona una variedad de funcionalidades para admitir resistencia y recuperación. Es responsable de comprender cómo funcionan esas funcionalidades dentro de todos los servicios que usa y de seleccionar las funcionalidades que necesita para cumplir los objetivos empresariales y los objetivos de tiempo de actividad.

En este artículo se describe cómo Azure Databricks mantiene la resistencia frente a varias posibles interrupciones y problemas y cómo puede configurar la resistencia para satisfacer sus requisitos. En las instrucciones se describen los errores transitorios, las interrupciones de zona de disponibilidad, las interrupciones de región y el mantenimiento del servicio. En este artículo también se describe cómo usar copias de seguridad para recuperarse de otros problemas y se resalta información clave sobre el contrato de nivel de servicio (SLA) de Azure Databricks.

Recomendaciones de implementación de producción

Para obtener información sobre cómo implementar Azure Databricks para admitir los requisitos de confiabilidad de la solución y cómo afecta la confiabilidad a otros aspectos de la arquitectura, consulte Procedimientos recomendados de arquitectura para Azure Databricks.

Introducción a la arquitectura de confiabilidad

Debe comprender la confiabilidad de cada componente principal de Azure Databricks:

- **El plano de control** es una colección de servicios sin estado que administra los metadatos del área de trabajo, el acceso de usuario, la programación de trabajos y la administración de clústeres. Estos servicios están respaldados por bases de datos que se replican en múltiples zonas de disponibilidad dentro de las regiones admitidas.
- **La raíz del sistema de archivos de Databricks (DBFS)** es una cuenta de almacenamiento que Azure Databricks aprovisiona automáticamente al crear un área de trabajo de Azure Databricks en la cuenta en la nube. Se recomienda no almacenar datos en la raíz de DBFS y deshabilitar esta cuenta de almacenamiento si es posible.

- **El almacenamiento del catálogo de Unity** incluye una o varias cuentas de almacenamiento que almacenan los datos del catálogo de Unity en la cuenta en la nube. Para obtener más información, consulte Introducción al catálogo de Unity.

- **El plano de proceso** ejecuta cargas de trabajo de procesamiento de datos mediante clústeres de máquinas virtuales (VM). El plano de proceso controla los errores transitorios y reemplaza automáticamente los nodos con errores sin intervención del usuario. Puede elegir entre varios tipos de recursos de proceso. Para obtener más información, consulte Computación.

  La disponibilidad del área de trabajo depende de la disponibilidad del plano de control, pero los clústeres de proceso pueden seguir procesando trabajos incluso durante las interrupciones del plano de control.

Resistencia a errores transitorios

Los errores transitorios son errores breves e intermitentes en los componentes. Se producen con frecuencia en un entorno distribuido como la nube y son una parte normal de las operaciones. Los errores transitorios se corrigen después de un breve período de tiempo. Es importante que las aplicaciones puedan controlar errores transitorios, normalmente mediante el reintento de solicitudes afectadas.

Todas las aplicaciones hospedadas en la nube deben seguir las instrucciones de control de errores transitorios de Azure cuando se comunican con cualquier API, bases de datos y otros componentes hospedados en la nube. Para obtener más información, consulte Recomendaciones para controlar errores transitorios.

Puede controlar los reintentos de las tareas dentro de trabajos de Lakeflow para ayudar a recuperarse de errores transitorios.

En el caso de las aplicaciones que se ejecutan en Azure Databricks, implemente la lógica de reintento con retroceso exponencial al conectarse a servicios externos o servicios de Azure, como Storage, Azure SQL Database o Azure Event Hubs. Databricks Runtime incluye resistencia integrada para muchos servicios de Azure, pero el código de la aplicación debe controlar errores transitorios específicos del servicio.

Resistencia a errores de zona de disponibilidad

Las zonas de disponibilidad son grupos físicamente independientes de centros de datos dentro de una región de Azure. Cuando una zona falla, los servicios pueden transferirse a una de las zonas restantes.

Azure Databricks admite *redundancia de zona* para cada componente:

- **Plano de control:** En regiones que admiten zonas de disponibilidad, el plano de control se ejecuta en varias zonas de disponibilidad. El plano de control controla los errores de zona automáticamente, con un impacto mínimo y ninguna intervención del usuario necesaria.

  Los datos del área de trabajo del plano de control se almacenan en bases de datos. En las regiones que admiten zonas de disponibilidad, las bases de datos se replican en varias zonas de la región. Las cuentas de almacenamiento que atienden imágenes en tiempo de ejecución de Databricks también son redundantes dentro de la región. Todas las regiones tienen cuentas de almacenamiento secundarias que se usan cuando la cuenta de almacenamiento principal está inactiva.

- **Raíz de DBFS:** En las regiones que admiten zonas de disponibilidad, puede configurar la cuenta de almacenamiento para que la raíz de DBFS use el almacenamiento con redundancia de zona (ZRS). En regiones emparejadas que admiten zonas de disponibilidad, puede usar opcionalmente almacenamiento con redundancia de zona geográfica (GZRS).

- **Plano de proceso:** Databricks admite la *distribución automática de zonas* para los recursos de proceso, lo que significa que los recursos se distribuyen entre varias zonas de disponibilidad. Esta distribución ayuda a las cargas de trabajo de producción a lograr resistencia a las interrupciones de zona.

  Cuando se usa el proceso sin servidor, no se seleccionan explícitamente zonas para el proceso. Databricks administra la selección de zona de las máquinas virtuales y el reemplazo de máquinas virtuales que podrían perderse debido a interrupciones de zona.

Requisitos

Para usar la compatibilidad con zonas de disponibilidad en Azure Databricks, necesita los siguientes requisitos:

- **Compatibilidad con regiones:** La compatibilidad con la zona de disponibilidad de Azure Databricks está disponible en todas las regiones de Azure que admiten Azure Databricks y proporcionan zonas de disponibilidad. Para obtener una lista de las regiones que admiten Azure Databricks, consulte Productos disponibles por región. Para obtener una lista completa de las regiones que admiten zonas de disponibilidad, consulte Regiones de Azure que admiten zonas de disponibilidad.

- **Replicación de almacenamiento:** Configure las cuentas de almacenamiento del área de trabajo para usar ZRS o GZRS (cuando esté disponible).

- **Capacidad de proceso:** Asegúrese de que existe suficiente capacidad de proceso en varias zonas de la región de destino. Azure Databricks distribuye automáticamente los nodos de clúster entre zonas, pero debe comprobar que los tipos de instancia seleccionados están disponibles en todas las zonas de destino.

Consideraciones

Azure Databricks distribuye automáticamente los nodos de clúster entre zonas de disponibilidad. La distribución depende de la capacidad disponible en cada zona. Durante períodos de alta demanda, los nodos de un clúster se pueden concentrar en menos zonas. Cuando usas computación sin servidor, Azure Databricks gestiona la selección de zona de las máquinas virtuales (VMs) y el reemplazo de VMs que podrían perderse debido a fallos de zonas.

Cost

La distribución de zona no afecta a los costos de proceso porque paga por el mismo número de máquinas virtuales, independientemente de su ubicación de zona de disponibilidad. Para obtener más información, consulte Precios de computación de Azure Databricks.

La redundancia predeterminada para la cuenta de almacenamiento administrada o la raíz de DBFS es el almacenamiento con redundancia geográfica (GRS). Cambiar a ZRS o GZRS puede afectar a los costos de almacenamiento. Para obtener más información, consulte Precios de Azure Blob Storage.

Configurar soporte de zonas de disponibilidad

- **Plano de control:** El plano de control admite automáticamente la redundancia de zona en regiones que tienen zonas de disponibilidad. No es necesario configurar nada.

- **Raíz de DBFS:** Puede configurar la redundancia de zona para el almacenamiento raíz de DBFS al crear un área de trabajo o modificar una área de trabajo existente:

  - **Cree un nuevo área de trabajo con almacenamiento raíz de DBFS redundante de zona:** Cuando cree un nuevo área de trabajo de Azure Databricks, puede configurar opcionalmente la cuenta de almacenamiento asociada para utilizar ZRS o GZRS en lugar del GRS predeterminado. Para más información, consulte Cambio de las opciones de redundancia de almacenamiento del área de trabajo.

  - **Habilite la redundancia de zona en el almacenamiento raíz de DBFS:** En el caso de las áreas de trabajo existentes, puede cambiar la configuración de redundancia de la cuenta de almacenamiento del área de trabajo a ZRS o GZRS. Para obtener más

información sobre cómo habilitar la redundancia de zona, consulte Cambio de la configuración de replicación para una cuenta de almacenamiento.

- **Plano de proceso:** Los nodos de clúster se distribuyen automáticamente entre zonas de disponibilidad. No se requiere ninguna configuración de cliente para la distribución de zona.

Comportamiento cuando todas las zonas están en buen estado

En esta sección se describe qué esperar cuando un área de trabajo está configurada con compatibilidad con zonas de disponibilidad y todas las zonas de disponibilidad están operativas.

- **Replicación de datos entre zonas:** La replicación de datos para el almacenamiento del área de trabajo se produce sincrónicamente entre zonas cuando la raíz de DBFS usa una cuenta ZRS o GZRS. Este enfoque garantiza una coherencia sólida con un impacto mínimo en el rendimiento.

- **Enrutamiento de tráfico entre zonas:** Azure Databricks distribuye automáticamente los nodos de clúster entre zonas durante la creación del clúster. El servicio equilibra la carga de proceso entre zonas mientras mantiene la localidad de datos para obtener un rendimiento óptimo.

Comportamiento durante un fallo de zona

En esta sección se describe qué esperar cuando un área de trabajo está configurada con soporte para zona de disponibilidad y hay una interrupción en la zona de disponibilidad.

- **Detección y respuesta:** Microsoft detecta automáticamente errores de zona e inicia procedimientos de respuesta. No es necesario realizar ninguna acción para la conmutación por error de nivel de zona.

- **Notificación:** Microsoft no le notifica automáticamente cuando una zona está inactiva. Pero puede usar la página de estado de Azure Databricks para ver una introducción a todos los servicios principales de Azure Databricks. También puede suscribirse a actualizaciones de estado de componentes individuales del servicio y recibir una alerta cuando cambie el estado del servicio al que está suscrito.

- **Solicitudes activas:** La ejecución de clústeres podría perder nodos en la zona afectada. El administrador de clústeres solicita automáticamente nodos de reemplazo de zonas

restantes. Si se pierde el nodo del controlador, el clúster y el trabajo se reinician completamente.

- **Pérdida de datos esperada:**
  - **Plano de control:** No se espera ninguna pérdida de datos durante una interrupción de zona.
  - **Raíz de DBFS:** Los datos del área de trabajo permanecen disponibles si usa configuraciones de almacenamiento de ZRS o GZRS.
  - **Plano de proceso:** Los datos almacenados en caché en máquinas virtuales son efímeros. Los datos perdidos de las máquinas virtuales durante un error de zona se recuperan del almacenamiento. Si se pierde el nodo del controlador, el trabajo se reinicia y vuelve a calcular los resultados.
- **Tiempo de inactividad esperado:**
  - **Plano de control:** El plano de control de Databricks realiza la conmutación automática en caso de fallo a zonas operativas en unos 15 minutos.
  - **Raíz de DBFS:** No se espera ningún tiempo de inactividad para las cuentas de almacenamiento que usan ZRS o GZRS.
  - **Plano de proceso:** Si se pierden nodos porque sus máquinas virtuales residen en la zona de disponibilidad afectada, el administrador de clústeres de Azure solicita nodos de reemplazo del proveedor de proceso de Azure. Si las zonas correctas restantes tienen capacidad suficiente para cumplir la solicitud, el proveedor de proceso extrae nodos de las zonas correctas para reemplazar los nodos perdidos. Este proceso puede tardar varios minutos.

    Si el nodo del controlador se pierde debido al error de zona, se reinicia todo el clúster, lo que puede provocar tiempos de recuperación más largos en comparación con la pérdida de nodos de trabajo. Planee este comportamiento en las estrategias de programación y supervisión del trabajo.

    Puede usar grupos de instancias o sin servidor para reducir este tiempo.
- **Reenrutamiento del tráfico:**
  - **Plano de control:** El plano de control de Databricks realiza la conmutación automática en caso de fallo a zonas operativas en unos 15 minutos.
  - **Raíz de DBFS:** Azure Storage redirige automáticamente las solicitudes a los clústeres de almacenamiento en zonas correctas.

- **Plano de proceso:** El administrador de clústeres cambia automáticamente a los nodos de zonas correctas.

Recuperación de zona

Cuando la zona de disponibilidad fallida se recupera, Azure Databricks reanuda las operaciones normales automáticamente en todas las zonas. El administrador de clústeres puede reequilibrar la distribución de nodos durante las posteriores creaciones de nodos, pero los nodos existentes continúan ejecutándose en sus zonas actuales hasta que finalizan.

No es necesario realizar ninguna acción para las operaciones de conmutación por recuperación. La distribución de zona normal se reanuda para las nuevas implementaciones de clúster.

Prueba de fallos de zona

Azure Databricks es un servicio administrado en el que Microsoft controla la conmutación por error de zona automáticamente y realiza pruebas periódicas de zona inactiva. No es necesario probar los escenarios de fallo de zona para el servicio en sí.

Para las aplicaciones que se ejecutan en Azure Databricks, pruebe la resistencia del trabajo simulando errores de nodo de controlador y supervisando el comportamiento de reinicio del clúster. Verifique que los trabajos de procesamiento de datos pueden manejar los reinicios del cluster y reanudarse desde los puntos de control adecuados.

Resistencia a errores en toda la región

Azure Databricks es un servicio de una sola región. Si la región no está disponible, el área de trabajo tampoco está disponible. Si necesita implementaciones en varias regiones, consulte Recuperación ante desastres de Azure Databricks.

Soluciones personalizadas de varias regiones para la resistencia

Azure Databricks no proporciona funcionalidades integradas en varias regiones. Para una protección multirregional completa de sus cargas de trabajo analíticas, debe implementar su propio enfoque.

Las soluciones típicas de varias regiones implican dos o más áreas de trabajo. Puede elegir entre varias estrategias, incluidas las arquitecturas activa-pasiva y activa-activa.

Para elegir una arquitectura, tenga en cuenta los siguientes factores:

- La importancia de la carga de trabajo para su empresa
- La posible duración de una interrupción (horas o posiblemente un día completo)
- El esfuerzo necesario para hacer que el área de trabajo esté totalmente operativa
- El esfuerzo necesario para restaurar o recuperar a la región primaria

Para ver las cargas de trabajo que requieren protección en varias regiones, consulte Recuperación ante desastres de Azure Databricks.

Copia de seguridad y recuperación

Azure Databricks realiza automáticamente copias de seguridad de bases de datos como parte de las operaciones administradas del servicio. Este proceso incluye contenido de cuadernos, definiciones de trabajos, configuraciones de clúster y configuración de control de acceso.

> **Nota:**
>
> Si se produce un error de zona, Azure Databricks no espera ninguna pérdida de datos.

Se recomienda almacenar los datos en el almacenamiento del catálogo de Unity. Puede replicar datos a través de la replicación de almacenamiento o la clonación diferencial.

Las funcionalidades de copia de seguridad y restauración de nivel de área de trabajo no están disponibles directamente. Planee los procedimientos de recreación del área de trabajo que incluyen la restauración de configuraciones, usuarios y controles de acceso desde los procesos de sincronización.

Resistencia al mantenimiento del servicio

Azure Databricks realiza el mantenimiento automático de la plataforma para aplicar actualizaciones de seguridad, implementar nuevas características y mejorar la confiabilidad del servicio. Puede configurar las ventanas de mantenimiento del clúster para reducir la probabilidad de mantenimiento que afecte a las cargas de trabajo de producción. Para más información, consulte Actualización automática del clúster.

Acuerdo de nivel de servicio

El contrato de nivel de servicio (SLA) para los servicios de Azure describe la disponibilidad esperada de cada servicio y las condiciones que la solución deberá cumplir para lograr esa

expectativa de disponibilidad. Para obtener más información, consulte Acuerdos de Nivel de Servicio para servicios en línea.

Contenido relacionado

- Recuperación ante desastres de Azure Databricks
- Procedimientos recomendados de confiabilidad para Azure Databricks
- Introducción a la confiabilidad de Azure

Last updated on 22/01/2026

Failover para la continuidad del negocio y la recuperación ante desastres

Para maximizar el tiempo de actividad, planee con antelación la continuidad empresarial y prepare todo lo necesario para la recuperación ante desastres con Azure Machine Learning.

Microsoft se esfuerza por garantizar que los servicios de Azure siempre estén disponibles. Sin embargo, es posible que se produzcan interrupciones no planeadas en el servicio. Tener implementado un plan de recuperación ante desastres para controlar las interrupciones del servicio regional. En este artículo aprenderá a:

- Planee una implementación multirregional de Azure Machine Learning y recursos asociados.
- Maximice las posibilidades de recuperar registros, cuadernos, imágenes de Docker y otros metadatos.
- Diseñe su solución para una alta disponibilidad.
- Iniciar una conmutación por error a otra región.

> **Importante**
>
> Azure Machine Learning no proporciona conmutación automática por error ni recuperación ante desastres. La copia de seguridad y restauración de metadatos del área de trabajo, como el historial de ejecución, no está disponible.

Si eliminó accidentalmente el área de trabajo o los componentes correspondientes, en este artículo también se proporcionan opciones de recuperación admitidas actualmente.

Descripción de los servicios de Azure para Azure Machine Learning

Azure Machine Learning depende de varios servicios de Azure. Algunos de estos servicios se aprovisionan en la suscripción. Es responsable de la configuración de alta disponibilidad de estos servicios. Otros servicios se crean en una suscripción de Microsoft y son administrados por Microsoft.

Los servicios de Azure incluyen:

- **Infraestructura de Azure Machine Learning**: Un entorno administrado por Microsoft para el área de trabajo de Azure Machine Learning.

- **Recursos asociados**: recursos aprovisionados en la suscripción durante la creación del área de trabajo de Azure Machine Learning. Estos recursos incluyen Azure Storage, Azure Key Vault, Azure Container Registry y Application Insights.
  - El almacenamiento predeterminado tiene datos como el modelo, los datos del registro de entrenamiento y las referencias a los recursos de datos.
  - Key Vault tiene credenciales para Azure Storage, Container Registry y almacenes de datos.
  - Container Registry tiene una imagen de Docker para entornos de aprendizaje e inferencia.
  - Application Insights se usa para la supervisión de Azure Machine Learning.

- **Recursos de proceso**: recursos creados después de la implementación del área de trabajo. Por ejemplo, puede crear una instancia de proceso o un clúster de proceso para entrenar un modelo de Machine Learning.
  - Instancia de proceso y clúster de proceso: Entornos de desarrollo de modelos administrado por Microsoft.
  - Otros recursos: recursos de proceso de Microsoft que puede adjuntar a Azure Machine Learning, como Azure Kubernetes Service (AKS), Azure Databricks, Azure Container Instances y Azure HDInsight. Eres responsable de configurar la alta disponibilidad de estos recursos.

- **Otros almacenes de datos**: Azure Machine Learning puede montar otros almacenes de datos, como Azure Storage y Azure Data Lake Storage para los datos de entrenamiento. Estos almacenes de datos se crean dentro de su suscripción. Usted es responsable de configurar su configuración de alta disponibilidad. Para ver otras opciones del almacén de datos, consulte Creación de almacenes de datos.

En la tabla siguiente se muestran los servicios Azure administrados por Microsoft y que administra usted. También se indican los servicios de alta disponibilidad, de forma predeterminada.

Expandir tabla

Servicio	Administrado por	Alta disponibilidad de forma predeterminada
Infraestructura de Azure Machine Learning	Microsoft	

Servicio	Administrado por	Alta disponibilidad de forma predeterminada
Recursos asociados		
Azure Storage	Tú	
Bóveda de claves	Tú	✓
Registro de contenedores	Tú	
Application Insights	Tú	N/D
Recursos informáticos		
Instancia de cómputo	Microsoft	
Clúster de proceso	Microsoft	
Otros recursos de proceso, como AKS, Azure Databricks, Container Instances, HDInsight	Tú	
Otros almacenes de datos, como Azure Storage, SQL Database, Bases de datos de Azure para PostgreSQL, Bases de datos de Azure para MySQL. Sistema de archivos de Azure Databricks	Tú	

En el resto de este artículo se describen las acciones que el usuario tiene que realizar para que cada uno de estos servicios sea de alta disponibilidad.

Planeamiento de la implementación multirregional

Una implementación multirregional se basa en la creación de Azure Machine Learning y otros recursos (infraestructura) en dos regiones de Azure. Si se produce una interrupción en una región, se puede cambiar a la otra. Cuando se plantee dónde implementar sus recursos, tenga en cuenta lo siguiente:

- **Disponibilidad regional**: si es posible, use una región en la misma área geográfica, no necesariamente la más cercana. Para comprobar la disponibilidad por región para Azure Machine Learning, consulte los productos de Azure disponibles por región.
- **Regiones emparejadas de Azure**: coordinan las actualizaciones de la plataforma y priorizan las iniciativas de recuperación según las necesidades. Sin embargo, no todas las regiones admiten regiones emparejadas. Para más información, consulte Regiones emparejadas de Azure.

- **Disponibilidad del servicio**: determine si la disponibilidad de los recursos utilizados por la solución debe ser activa/activa, activa/en espera o activa/pasiva.
  - **Hot/hot**: ambas regiones están activas al mismo tiempo, y una de ellas está lista para usarse de inmediato.
  - **activa/en espera**: la región primaria está activa y la secundaria tiene recursos críticos (por ejemplo, modelos implementados) preparados para iniciarse. Los recursos no críticos tendrían que implementarse manualmente en la región secundaria.
  - **activa/pasiva**: la región primaria está activa y la región secundaria tiene el servicio Azure Machine Learning y otros recursos implementados, junto con los datos necesarios. Recursos como los modelos, las implementaciones de modelos o las canalizaciones tendrían que implementarse manualmente.

> **Sugerencia**
>
> En función de los requisitos empresariales, puede decidir tratar diferentes recursos de Azure Machine Learning de forma diferente. Por ejemplo, puede que quiera usar activo/activo para los modelos desplegados (inferencia) y activo/inactivo para los experimentos (entrenamiento).

Azure Machine Learning se basa en otros servicios. Algunos servicios se pueden configurar para la replicación en otras regiones. Otros los debes crear manualmente en varias regiones. En la tabla siguiente se proporciona una lista de servicios responsables de la replicación y una descripción de la configuración:

Expandir tabla

Servicio de Azure	Georreplicado por	Configuración
Área de trabajo de Machine Learning	Tú	Cree un área de trabajo en las regiones seleccionadas.
Cómputo de aprendizaje automático	Tú	Cree los recursos informáticos en las regiones seleccionadas. Para los recursos de computación que pueden escalar dinámicamente, asegúrese de que ambas regiones dispongan de una cuota de computación suficiente para sus necesidades.
Registro de Machine Learning	Tú	Cree el registro en varias regiones.
Bóveda de claves	Microsoft	Use la misma instancia de Key Vault con el área de trabajo y los recursos de Azure Machine Learning en ambas regiones. Key Vault

Servicio de Azure	Georreplicado por	Configuración
		conmuta por error de forma automática en una región secundaria. Para más información, consulte Redundancia y disponibilidad de Azure Key Vault.
Registro de contenedores	Microsoft	Configure la instancia de Container Registry para que replique geográficamente los registros en la región emparejada para Azure Machine Learning. Use la misma instancia para ambas instancias del espacio de trabajo. Para más información, consulte Replicación geográfica en Azure Container Registry.
Cuenta de almacenamiento	Tú	Azure Machine Learning no admite la conmutación por error de la **cuenta de almacenamiento predeterminada** mediante almacenamiento con redundancia geográfica (GRS), almacenamiento con redundancia de zona geográfica (GZRS), almacenamiento con redundancia geográfica con acceso de lectura (RA-GRS) ni almacenamiento con redundancia de zona geográfica con acceso de lectura (RA-GZRS). Cree una cuenta de almacenamiento separada para el almacenamiento predeterminado de cada área de trabajo. Cree servicios o cuentas de almacenamiento separadas para otro almacenamiento de datos. Para más información, vea Redundancia de Azure Storage.
Application Insights	Tú	Cree un recurso de Application Insights para el espacio de trabajo en ambas regiones. Para ajustar el período de retención de datos y los detalles consulte Recopilación, retención y almacenamiento de datos en Application Insights.

Para acelerar la recuperación y el reinicio en la región secundaria, se recomiendan los siguientes procedimientos de desarrollo:

- Utilice plantillas de Azure Resource Manager. Las plantillas generan "infraestructura como código" y permiten implementar rápidamente servicios en ambas regiones.
- Para evitar el desvío entre las dos regiones, actualice las canalizaciones de integración e implementación continuas para que se implementen en ambas regiones.
- Cuando automatice las implementaciones, incluya la configuración de los recursos de proceso asociados al área de trabajo, como Azure Kubernetes Service.
- Cree asignaciones de rol para los usuarios de ambas regiones.
- Cree recursos de red, como las redes virtuales y los puntos de conexión privados de Azure, para ambas regiones. Asegúrese de que los usuarios tengan acceso a ambos entornos de red. Por ejemplo, las configuraciones de VPN y DNS para ambas redes virtuales.

Servicios de proceso y de datos

En función de cuáles sean sus necesidades, podrá tener más servicios de proceso o de datos para que Azure Machine Learning los utilice. Por ejemplo, puede usar Azure Kubernetes Services o Azure SQL Database. Use la siguiente información para aprender a configurar estos servicios con fines de alta disponibilidad.

Recursos informáticos

- **Azure Kubernetes Service**: consulte Procedimientos recomendados para continuidad empresarial y recuperación ante desastres en Azure Kubernetes Service (AKS) y Creación de un clúster de Azure Kubernetes Service (AKS) que use zonas de disponibilidad. Si el clúster de AKS se creó mediante Azure Machine Learning Studio, el SDK o la CLI, no se admite la alta disponibilidad entre regiones.
- **Azure Databricks**: consulte Recuperación ante desastres regional para clústeres de Azure Databricks.
- **Container Instances**: Un orquestador se encarga de la conmutación por error. Consulte Azure Container Instances y orquestadores de contenedores.
- **HDInsight**: consulte Servicios de alta disponibilidad admitidos en Azure HDInsight.

Servicios de datos

- **Contenedor de Azure Blob Storage, Azure Files y Data Lake Storage Gen2**: consulte Redundancia de Azure Storage.
- **Data Lake Storage Gen1**: Consulte Guía de alta disponibilidad y recuperación ante desastres para Data Lake Storage Gen1.

> **Sugerencia**
>
> Si proporciona su propia clave administrada por el cliente para implementar el área de trabajo de Azure Machine Learning, Azure Cosmos DB también se aprovisiona dentro de la suscripción. En ese caso, eres responsable de configurar sus ajustes de alta disponibilidad. Consulte **Alta disponibilidad con Azure Cosmos DB**.

Diseño para lograr alta disponibilidad

Zonas de disponibilidad

Algunos servicios de Azure admiten zonas de disponibilidad. En las regiones que admiten zonas de disponibilidad, si una zona falla, cualquier carga de trabajo se pausará y los datos deberían

guardarse. Sin embargo, los datos no están disponibles para actualizarse hasta que la zona vuelva a estar en línea.

Para más información, consulte Soporte técnico para el servicio de zonas de disponibilidad.

Implementación de componentes críticos en varias regiones

Determine el nivel de continuidad empresarial que tiene como objetivo. El nivel puede diferir entre los componentes de la solución. Por ejemplo, tal vez desee usar una configuración activa/activa para las canalizaciones de producción o las implementaciones de modelos, y una configuración activa/inactiva con fines de experimentación.

Administración de datos de entrenamiento en un almacenamiento aislado

Al mantener el almacenamiento de datos aislado del almacenamiento predeterminado que el área de trabajo usa para los registros, podrá hacer lo siguiente:

- Adjunte las mismas instancias de almacenamiento como almacenes de datos a los espacios de trabajo principal y secundario.
- Utilice la geo-replicación para las cuentas de almacenamiento de datos y maximice su tiempo de actividad.

Administración de recursos de aprendizaje automático como código

> **Nota:**
>
> La copia de seguridad y la restauración de metadatos del área de trabajo, como el historial de ejecución, los modelos y los entornos no están disponibles. La especificación de recursos y configuraciones como código mediante especificaciones de YAML le ayuda a volver a crear recursos entre áreas de trabajo en caso de desastre.

Los trabajos de Azure Machine Learning se definen mediante una especificación de trabajo. Esta especificación incluye dependencias de los artefactos de entrada que se administran a nivel de instancia del espacio de trabajo, incluidos los entornos y los recursos de cómputo. Para el envío de trabajos y los despliegues multirregión, recomendamos las siguientes prácticas:

- Administre localmente el código base, con ayuda de un repositorio de Git.

    - Exporte cuadernos importantes de Estudio de Azure Machine Learning.
    - Exportar canalizaciones creadas en Studio como código.

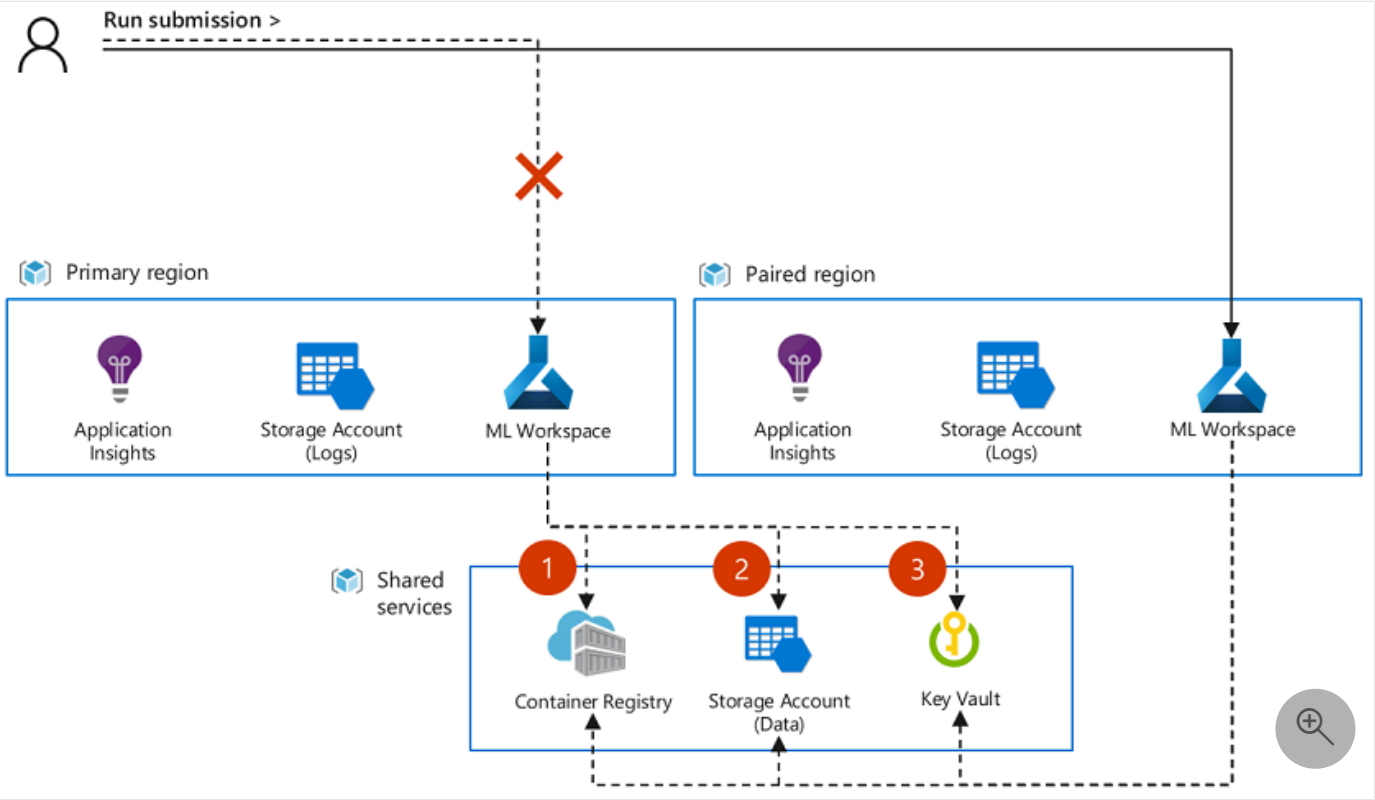
- Administre las configuraciones como código.
    - Evite las referencias codificadas de forma rígida al área de trabajo. En su lugar, configure una referencia a la instancia del área de trabajo mediante un archivo de configuración y utilice MLClient.from_config() para inicializar el área de trabajo.
    - Use Dockerfile si tiene imágenes personalizadas de Docker.

Iniciar una conmutación por error

Continuar el trabajo en el espacio de trabajo de conmutación por error

Cuando el área de trabajo principal deja de estar disponible, puede cambiar al área de trabajo secundaria con fines de experimentación y desarrollo. Azure Machine Learning no envía automáticamente trabajos al área de trabajo secundaria si se produce una interrupción. Actualice la configuración de su código para que apunte al nuevo recurso del espacio de trabajo. Se recomienda evitar la codificación rígida de las referencias al área de trabajo. En su lugar, use un archivo de configuración del área de trabajo para minimizar los pasos manuales del usuario para cambiar las áreas de trabajo. Asegúrese de actualizar también cualquier automatización, como las canalizaciones de integración e implementación continuas, en la nueva área de trabajo.

Azure Machine Learning no puede sincronizar ni recuperar artefactos o metadatos entre instancias de área de trabajo. En función de la estrategia de implementación de la aplicación, es posible que tenga que mover artefactos o volver a crear insumos para la experimentación, como los activos de datos, en el espacio de trabajo de conmutación por error para poder seguir enviando trabajos. En caso de que haya configurado los recursos de su espacio de trabajo principal y de su espacio de trabajo secundario para compartir recursos asociados con la georreplicación habilitada, es posible que algunos objetos estén disponibles directamente en el espacio de trabajo de conmutación por error. Este es el caso, por ejemplo, si ambas áreas de trabajo comparten las mismas imágenes de Docker, almacenes de datos configurados y recursos de Azure Key Vault. En el diagrama siguiente se muestra una configuración en la que dos áreas de trabajo comparten las mismas imágenes (1), almacenes de datos (2) y recursos de Key Vault (3).



> **Nota:**
>
> Los trabajos que se ejecutan cuando se produce una interrupción del servicio no pasarán automáticamente al área de trabajo secundaria. También es poco probable que los trabajos se reanuden y finalicen correctamente en el área de trabajo principal una vez que se resuelva la interrupción. En su lugar, los trabajos se deben volver a enviar, ya sea en el área de trabajo secundaria o en la principal (una vez resuelta la interrupción).

Movilidad de artefactos entre áreas de trabajo

En función del enfoque de recuperación, es posible que tenga que copiar artefactos entre las áreas de trabajo para continuar con el trabajo. Actualmente, la movilidad de artefactos entre áreas de trabajo es limitada. Se recomienda administrar los artefactos como código siempre que sea posible para que se puedan volver a crear en la instancia de conmutación por error.

Los artefactos siguientes se pueden exportar e importar entre áreas de trabajo mediante la extensión de la CLI de Azure para el aprendizaje automático:

Expandir tabla

Artefacto	Exportación	Importar
Modelos	az ml model download --name {NAME} --version {VERSION}	az ml model create

Artefacto	Exportación	Importar
Entornos	az ml environment share --name my-environment --version {VERSION} --resource-group {RESOURCE_GROUP} --workspace-name {WORKSPACE} --share-with-name {NEW_NAME_IN_REGISTRY} --share-with-version {NEW_VERSION_IN_REGISTRY} --registry-name {REGISTRY_NAME}	az ml environment create
Trabajos de Azure Machine Learning	az ml job download -n {NAME} -g {RESOURCE_GROUP} -w {WORKSPACE_NAME}	az ml job create -f {FILE} -g {RESOURCE_GROUP} -w {WORKSPACE_NAME}
Activos de datos	az ml data share --name {DATA_NAME} --version {VERSION} --resource-group {RESOURCE_GROUP} --workspace-name {WORKSPACE} --share-with-name {NEW_NAME_IN_REGISTRY} --share-with-version {NEW_VERSION_IN_REGISTRY} --registry-name {REGISTRY_NAME}	az ml data create -f {FILE} -g {RESOURCE_GROUP} --registry-name {REGISTRY_NAME}

Sugerencia

- Las **salidas de trabajos** se almacenan en la cuenta de almacenamiento predeterminada asociada a un área de trabajo. Aunque las salidas de trabajo pueden quedar inaccesibles desde la interfaz de usuario de Studio en caso de una interrupción del servicio, se puede acceder directamente a los datos a través de la cuenta de almacenamiento. Para obtener más información sobre cómo trabajar con datos almacenados en blobs, consulte **Creación, descarga y enumeración de blobs mediante la CLI de Azure**.

Opciones de recuperación

Eliminación del área de trabajo

Si eliminaste accidentalmente tu área de trabajo, es posible que puedas recuperarla. Para ver los pasos de recuperación, consulta Recuperación de datos del área de trabajo tras una eliminación accidental con la eliminación temporal.

Incluso si el área de trabajo no se puede recuperar, es posible que pueda recuperar los cuadernos del recurso de almacenamiento de Azure asociado al área de trabajo si sigue estos pasos:

- En Azure Portal, vaya a la cuenta de almacenamiento vinculada al área de trabajo de Azure Machine Learning que se eliminó.
- En la sección Almacenamiento de datos que aparece a la izquierda, seleccione **Recursos compartidos de archivos**.
- Sus cuadernos de notas se encuentran en el recurso compartido de archivos cuyo nombre contiene el identificador del área de trabajo.

Pasos siguientes

Para obtener información sobre las implementaciones de infraestructura repetibles con Azure Machine Learning, use una plantilla de Bicep o una plantilla de Terraform.

Last updated on 21/07/2026

Confiabilidad en Azure Chaos Studio

En este artículo se describe el soporte de confiabilidad ofrecido por Azure Chaos Studio. Trata cómo configurar zonas de disponibilidad y qué esperar durante una interrupción de toda la zona. Para obtener información general más detallada sobre la confiabilidad en Azure, consulte Azure confiabilidad.

Soporte para zonas de disponibilidad

Availability zones son grupos de centros de datos físicamente independientes dentro de una región de Azure. Cuando una zona falla, los servicios pueden transferirse a una de las zonas restantes.

Azure Chaos Studio admite la redundancia de zona como la configuración predeterminada dentro de una región. Los recursos de Chaos Studio se duplican o distribuyen automáticamente entre diferentes zonas.

Prerrequisitos

Las siguientes regiones admiten zonas de disponibilidad para Chaos Studio:

Expandir tabla

Americas	Europa	Asia Pacífico
Sur de Brasil	Centro de Francia	Australia East
Canada Central	Centro-oeste de Alemania	Japón Oriental
Central US	Norte de Italia	Sudeste Asiático
East US	Norte de Europa	Norte de Emiratos Árabes Unidos
Este de EE. UU. 2	Centro de Suecia	
Oeste de EE. UU. 2	UK South	
Oeste de EE. UU. 3	Oeste de Europa	

Para obtener información detallada sobre el modelo de disponibilidad regional para Azure Chaos Studio vea Disponibilidad regional de Azure Chaos Studio.

Experiencia de reducción de zona

En caso de una interrupción en toda la zona, debe prever una breve degradación del rendimiento y la disponibilidad a medida que el servicio pasa a una zona funcional. Esta interrupción no depende de la restauración de la zona afectada, ya que los servicios administrados por Microsoft mitigan las pérdidas de zona mediante el uso de la capacidad de las zonas alternativas. En caso de interrupción de una zona de disponibilidad, es posible que un experimento de caos encuentre errores o interrupciones, pero los metadatos cruciales del experimento, los datos históricos y los detalles específicos deben permanecer accesibles y el servicio no debe experimentar una interrupción completa.

Recuperación ante desastres entre regiones y continuidad empresarial

Chaos Studio es un servicio de una sola región y no admite la conmutación por error entre regiones habilitada por el servicio.

Pasos siguientes

- Crear y ejecutar el primer experimento.
- Más información sobre la ingeniería de caos.
- Confiabilidad en Azure

Last updated on 14/04/2026

Confiabilidad en Azure Data Explorer

Azure Data Explorer es un servicio de análisis para ingerir, almacenar y consultar grandes volúmenes de datos con baja latencia. Normalmente se usa para cargas de trabajo de análisis de registros, telemetría y series temporales que requieren consultas rápidas sobre grandes conjuntos de datos.

Cuando se usa Azure, la confiabilidad es una responsabilidad compartida. Microsoft proporciona una variedad de funcionalidades para admitir resistencia y recuperación. Es responsable de comprender cómo funcionan esas funcionalidades dentro de todos los servicios que usa y de seleccionar las funcionalidades que necesita para cumplir los objetivos empresariales y los objetivos de tiempo de actividad.

En este artículo se describe cómo hacer que Azure Data Explorer sea resistente a varias posibles interrupciones y problemas, incluidos errores transitorios, errores de zona de disponibilidad y errores en toda la región. También se describen las opciones de copia de seguridad y restauración y la resistencia al mantenimiento del servicio y se resalta la información clave sobre el contrato de nivel de servicio (SLA) de Azure Data Explorer.

Recomendaciones de implementación de producción para la confiabilidad

En el caso de las cargas de trabajo de producción, se recomienda realizar los pasos siguientes para mejorar la confiabilidad del clúster de Azure Data Explorer:

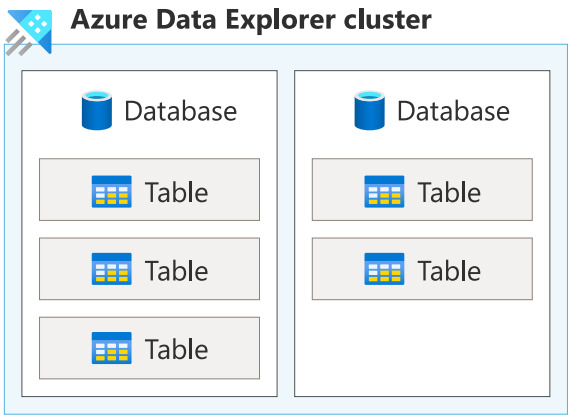
- ✓ **Implemente un clúster completo.** Azure Data Explorer proporciona clústeres gratuitos con fines de prueba. En el caso de las cargas de trabajo de producción, implemente un clúster completo.
- ✓ **Habilite la compatibilidad con la zona de disponibilidad.** Azure Data Explorer admite zonas de disponibilidad. Al habilitar la compatibilidad con zonas de disponibilidad, el servicio distribuye los nodos de proceso entre varias zonas de disponibilidad y almacena datos mediante el almacenamiento con redundancia de zona (ZRS). Esta configuración mejora la resistencia a los errores de zona de disponibilidad.

Introducción a la arquitectura de confiabilidad

En esta sección se describen algunos de los aspectos importantes de cómo funciona el servicio que es más relevante desde una perspectiva de confiabilidad. En la sección se presenta la arquitectura lógica, que incluye algunos de los recursos y características que se implementan y usan. También se describe la arquitectura física, que proporciona detalles sobre cómo funciona el servicio en segundo plano.

Arquitectura lógica

El recurso principal que implementa es un *clúster*, que representa la infraestructura que necesita para ingerir, almacenar y consultar los datos. Con un clúster, se crean *bases de datos* y esas bases de datos contienen *tablas*.



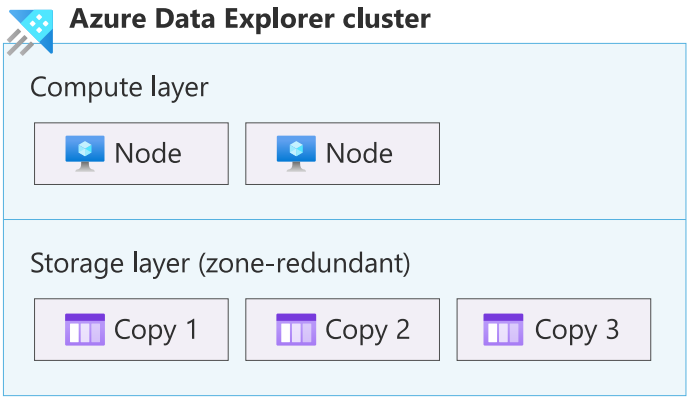
Los clústeres realizan la ingesta para recuperar datos de otros orígenes de datos y cargarlos en una tabla del clúster. A continuación, puede consultar datos mediante la sintaxis del lenguaje de consulta kusto (KQL). Los clústeres también tienen un conjunto de operaciones de administración que puede realizar.

Arquitectura física

Un clúster de Azure Data Explorer tiene dos capas principales que son aplicables a su configuración de confiabilidad:

- **Capa de proceso:** Azure Data Explorer es una plataforma informática distribuida y puede tener de dos a muchos nodos de máquina virtual (VM), según la escala y el tipo de rol de nodo. Los nodos controlan el trabajo de ingesta y procesamiento de consultas de datos. No gestionas directamente ni ves las máquinas virtuales del nodo. La plataforma administra automáticamente la creación de instancias, la supervisión del estado y el reemplazo de nodos incorrectos. Al configurar el clúster para que use varias zonas de disponibilidad, los nodos se distribuyen entre distintos centros de datos.

- **Capa de almacenamiento:** Azure Data Explorer usa Azure Storage como su capa de persistencia duradera. El almacenamiento proporciona automáticamente tolerancia a errores, con la configuración predeterminada que ofrece almacenamiento con redundancia local (LRS) dentro de un centro de datos. Se conservan tres réplicas. Si se pierde una réplica mientras está en uso, se implementa otra sin interrupción. Al configurar el clúster para que use varias zonas de disponibilidad, las réplicas se distribuyen entre distintos centros de datos.



Para más información, consulte Funcionamiento de Azure Data Explorer.

Resistencia a errores transitorios

Los errores transitorios son errores breves e intermitentes en los componentes. Se producen con frecuencia en un entorno distribuido como la nube y son una parte normal de las operaciones. Los errores transitorios se corrigen después de un breve período de tiempo. Es importante que las aplicaciones puedan controlar errores transitorios, normalmente mediante el reintento de solicitudes afectadas.

Todas las aplicaciones hospedadas en la nube deben seguir las instrucciones de control de errores transitorios de Azure cuando se comunican con cualquier API, bases de datos y otros componentes hospedados en la nube. Para obtener más información, consulte Recomendaciones para controlar errores transitorios.

Para crear resistencia a errores transitorios al usar Azure Data Explorer, siga estos procedimientos:

- Confíe en el comportamiento de reintento integrado cuando use la ingesta en cola.
- Use bibliotecas y SDK de cliente proporcionados por Microsoft, que reintentan automáticamente cuando se producen errores transitorios.

- Si usa Azure Data Explorer API REST directamente, vuelva a intentar las consultas y las operaciones de administración que produzcan errores debido a un error transitorio.

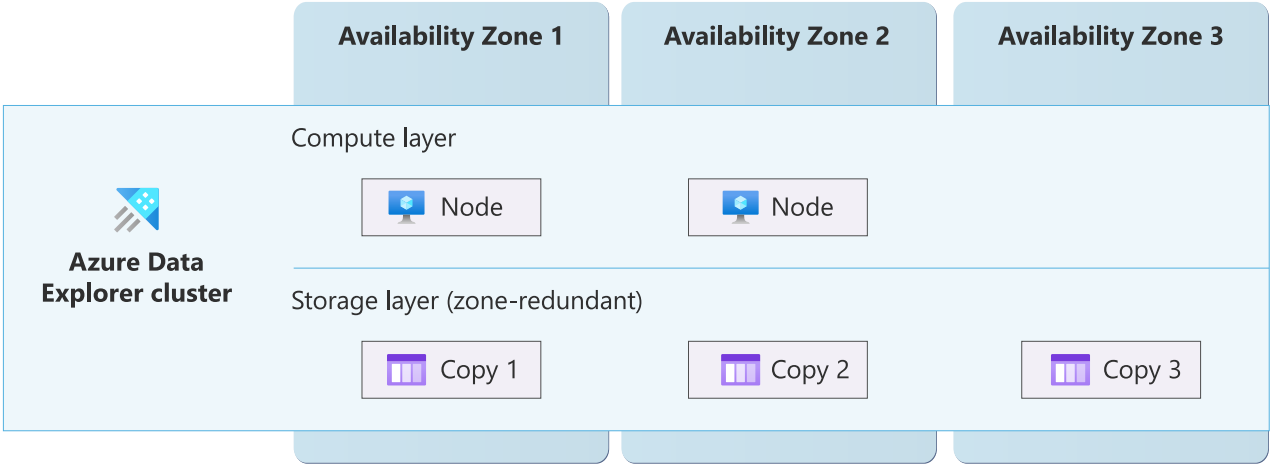
Resistencia a errores de zona de disponibilidad

Las zonas de disponibilidad son grupos físicamente independientes de centros de datos dentro de una región de Azure. Cuando una zona falla, los servicios pueden transferirse a una de las zonas restantes.

Azure Data Explorer admite dos tipos de configuración de zona de disponibilidad:

- **Con redundancia de zona (recomendado):** Al habilitar zonas de disponibilidad en el clúster, los nodos del clúster se distribuyen entre varias zonas. Microsoft administra la distribución de nodos en las zonas de disponibilidad seleccionadas y controla la detección y respuesta a los errores de zona de disponibilidad. Un clúster con redundancia de zona es resistente a una interrupción de zona de disponibilidad.

  Al configurar el clúster para que sea con redundancia de zona, Storage ZRS replica de forma sincrónica al menos tres copias de los datos en varias zonas de disponibilidad.

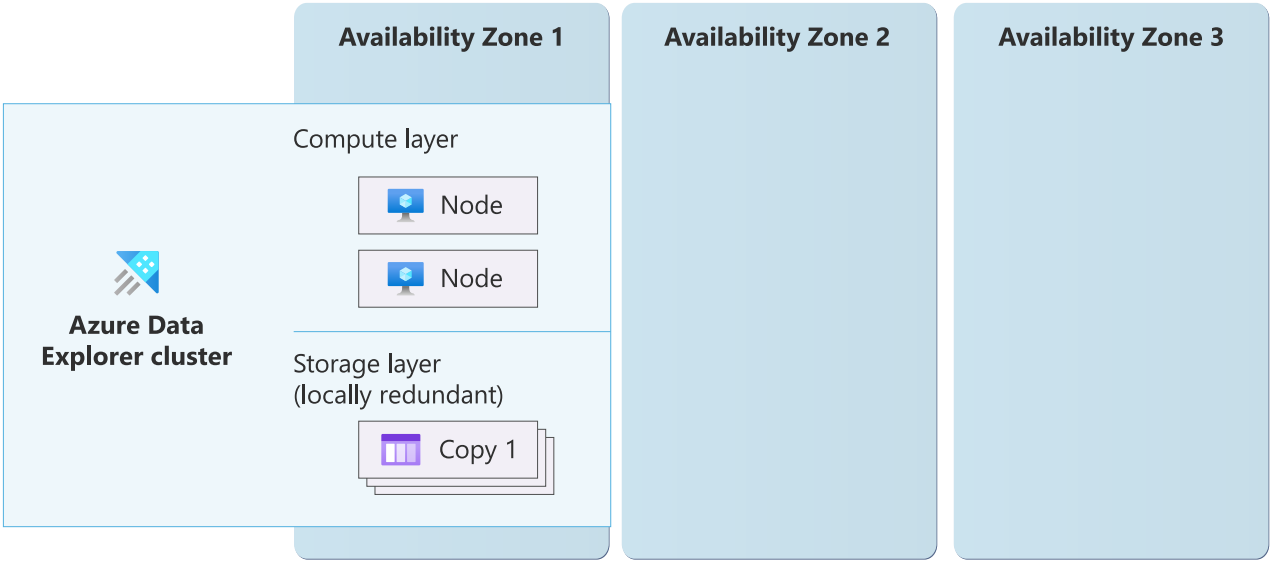


- **Zonal:** Opcionalmente, puede seleccionar una sola zona al habilitar zonas de disponibilidad en el clúster. Microsoft coloca todos los nodos de proceso en esa zona. Esta configuración es un clúster *zonal* (de una sola zona). Un clúster zonal puede reducir la latencia de cargas de trabajo poco sensibles a la latencia porque todos los nodos de proceso se ejecutan en la misma zona, pero no proporciona resistencia a interrupciones de zona.

  > **Importante**
  >
  > La asignación a una sola zona de disponibilidad solo se recomienda cuando la latencia entre zonas es demasiado alta para sus necesidades y después de comprobar que la

latencia no cumple sus requisitos. Por sí mismo, un recurso zonal no proporciona resistencia a una interrupción de zona de disponibilidad. Para mejorar la resiliencia de un recurso zonal, debe desplegar explícitamente recursos independientes en múltiples zonas de disponibilidad y configurar el enrutamiento del tráfico y la conmutación por error. Para obtener más información, consulte [Recursos zonales y resistencia de zona](#).

La selección de zona solo se aplica a los nodos de proceso. En el caso de un clúster zonal, los datos de almacenamiento siguen usando LRS y pueden almacenarse en una zona diferente de los nodos de proceso.



Si no habilita zonas de disponibilidad, el clúster es *no zonal*, lo que significa que Azure selecciona la zona de disponibilidad para cada nodo y los datos. Si alguna zona de disponibilidad de la región tiene una interrupción, puede afectar a los nodos, datos o ambos del clúster. No se recomienda una configuración no zonal porque no proporciona protección contra interrupciones de zona de disponibilidad.

Requisitos

- **Compatibilidad con regiones:** La compatibilidad con zonas de disponibilidad está disponible en regiones de Azure que admiten zonas de disponibilidad.

  Sin embargo, algunos tamaños y tipos de nodo de proceso solo están disponibles en regiones específicas o zonas específicas dentro de una región.

- **Clústeres completos:** La compatibilidad con zonas de disponibilidad está disponible con clústeres completos. No está disponible con clústeres gratuitos.

Consideraciones

Selección de zona: En el caso de los nodos de proceso, elija las zonas de disponibilidad que se van a usar. Microsoft administra la ubicación de la zona de almacenamiento y las réplicas de almacenamiento pueden estar en zonas diferentes de los nodos de proceso.

Costo

Habilitar el soporte para zonas de disponibilidad conlleva costes adicionales para ZRS, que se factura a una tarifa superior a la de LRS. Para más información, consulte Precios de Azure Storage.

Los nodos de cálculo se cobran a la misma tarifa independientemente de si se utiliza soporte para zonas de disponibilidad. Para más información, consulte los Precios de Azure Data Explorer.

Configurar soporte de zonas de disponibilidad

- **Cree un nuevo clúster con soporte para zonas de disponibilidad.** Puede habilitar la compatibilidad con zonas de disponibilidad al crear un nuevo clúster de Azure Data Explorer. Para más información, consulte Creación de un clúster y una base de datos.

  Al crear un clúster con zonas de disponibilidad habilitadas mediante el portal de Azure, este es automáticamente redundante entre zonas, y Microsoft selecciona las zonas.

  Para seleccionar zonas usted mismo o para crear un clúster zonal, use otro enfoque de implementación como las API de Azure Resource Manager o Bicep. En la mayoría de las situaciones, cree un clúster con redundancia de zona y use todas las zonas de la región.

  > **Nota:**
  >
  > Cuando selecciona qué zonas de disponibilidad usar, en realidad está seleccionando la *zona de disponibilidad lógica*. Si implementa otros componentes de carga de trabajo en una suscripción de Azure diferente, pueden usar un número de zona de disponibilidad lógica diferente para acceder a la misma zona de disponibilidad física. Para obtener más información, consulte zonas de disponibilidad física y lógica.

- **Habilite las zonas de disponibilidad en un clúster existente (versión preliminar).** Puede migrar un clúster no zonal existente para usar zonas de disponibilidad. Esta funcionalidad

está en versión preliminar. Para más información, consulte Migración del clúster para admitir varias zonas de disponibilidad.

- **Vuelva a configurar las zonas de disponibilidad en un clúster existente (versión preliminar).** Puede cambiar las zonas usadas para un clúster. Esta funcionalidad está en versión preliminar. Para más información, consulte Migración del clúster para admitir varias zonas de disponibilidad.

- **Deshabilite la compatibilidad con zonas de disponibilidad en un clúster existente.** Después de configurar un clúster con zonas de disponibilidad, no se puede cambiar el clúster para que no use zonas de disponibilidad.

- **Compruebe la configuración de la zona de disponibilidad para los clústeres.** Use la propiedad *de estado de zona* del clúster (la `zoneStatus` propiedad de la API REST) para comprobar la configuración de zona de disponibilidad de un clúster. Un valor de `Zonal` indica que el clúster usa zonas de disponibilidad, pero no significa que el clúster se ejecute en una sola zona.

  Para determinar si un clúster es zonal o con redundancia de zona, use la `zones` propiedad . Si la lista de zonas tiene una zona enumerada, el clúster es zonal (zona única). Si aparecen varias zonas, tiene redundancia de zona.

Planeamiento y administración de capacidad

Cuando una zona de disponibilidad no está disponible, es posible que los nodos de esa zona no estén disponibles temporalmente, lo que reduce la capacidad de proceso del clúster hasta que se recupere la zona.

Si el clúster no puede tolerar la pérdida de capacidad, considere la posibilidad de sobreaprovisionar el clúster. Este enfoque permite que la solución tolere cierta pérdida de capacidad y siga funcionando sin que se vea afectado su rendimiento. Sin embargo, al sobreaprovisionar el clúster, es posible que el clúster tenga un número de nodos desequilibrado entre zonas.

Distribución de instancias a través de zonas

La capa de proceso del clúster usa un enfoque de mejor esfuerzo para distribuir uniformemente las instancias entre las zonas que seleccione.

Comportamiento cuando todas las zonas están en buen estado

En esta sección se describe qué esperar al configurar un clúster para la compatibilidad con zonas de disponibilidad y todas las zonas están operativas.

- **Operación entre zonas:** Durante el funcionamiento normal, Azure Data Explorer usa todos los nodos de proceso disponibles para la ingesta, el procesamiento de consultas y otras operaciones. El trabajo se distribuye entre nodos independientemente de su zona de disponibilidad.
- **Replicación de datos entre zonas:** El comportamiento de replicación de datos entre zonas depende de la configuración de zona de disponibilidad que usa el clúster.
  - *Con redundancia entre zonas:* Los datos se replican de forma sincrónica entre zonas de disponibilidad mediante Storage ZRS, que proporciona un alto nivel de coherencia de los datos y minimiza el riesgo de pérdida de datos durante un fallo de zona.
  - *Zonal:* Los datos se almacenan con Storage LRS, lo que significa que las tres copias podrían estar en una sola zona de disponibilidad.

Comportamiento durante un fallo de zona

En esta sección se describe qué esperar al configurar un clúster para la compatibilidad con zonas de disponibilidad y se produce una interrupción en una de las zonas.

- **Detección y respuesta:** La responsabilidad de la detección y la respuesta depende de la configuración de zona de disponibilidad que usa el clúster.
  - *Con redundancia de zona:* Microsoft detecta fallos en las zonas de disponibilidad y gestiona la respuesta de Azure Data Explorer. No es necesario hacer nada para iniciar una conmutación por error de la zona.
  - *Zonal:* Usted es responsable de detectar fallos en las zonas de disponibilidad que utiliza su clúster. También es responsable de las acciones que decida iniciar, como cambiar a un segundo clúster que creó anteriormente en una zona de disponibilidad diferente.
- **Notificación:** Microsoft no le notifica automáticamente cuando una zona está inactiva. Sin embargo, puede usar Azure Service Health para comprender el estado general del servicio, incluidos los errores de zona, y puede configurar alertas de Service Health para notificarle problemas.
- **Solicitudes activas:** Las solicitudes activas que dependen de los recursos de proceso o almacenamiento de la zona con errores podrían finalizarse y el cliente debería reintentarlas.

Asegúrese de que las aplicaciones están preparadas siguiendo las instrucciones de control de errores transitorios.

- **Pérdida de datos esperada:** La pérdida de datos esperada depende de la configuración de zona de disponibilidad que usa el clúster.
  - *Con redundancia de zona:* No se espera ninguna pérdida de datos durante una interrupción de la zona de disponibilidad porque los datos se replican sincrónicamente entre zonas.
  - *Zonal:* Los datos no están disponibles hasta que se recupere la zona. En el improbable caso de una pérdida permanente de una zona que contiene todas las réplicas de almacenamiento, es posible que los datos se pierdan permanentemente.
- **Tiempo de inactividad esperado:** El tiempo de inactividad esperado depende de la configuración de zona de disponibilidad que usa el clúster.
  - *Zona-redundante:* Una breve interrupción del servicio podría ocurrir mientras el tráfico se redirige a zonas de disponibilidad saludables. Asegúrese de que las aplicaciones están preparadas siguiendo las instrucciones de control de errores transitorios.
  - *Zonal:* Los nodos de cómputo del clúster no están disponibles hasta que se restablezca la zona de disponibilidad. Es posible que tampoco pueda acceder a los datos del clúster durante un fallo de zona.
- **Redistribución:** El comportamiento de reenrutamiento del tráfico depende de la configuración de zona de disponibilidad que usa el clúster.
  - *Redundancia de zona:* Azure Data Explorer enruta nuevas solicitudes a los recursos de proceso y almacenamiento en las zonas saludables restantes.
  - *Zonal:* El clúster no está disponible hasta que se recupere la zona de disponibilidad.

Recuperación de zona

Cuando se recupera la zona de disponibilidad que ha fallado, Microsoft recrea los nodos del clúster y las réplicas de almacenamiento en esa zona y restaura la distribución normal del tráfico en todas las zonas. No es necesario realizar ninguna acción.

Prueba de fallos de zona

Las opciones para probar errores de zona dependen de la configuración de zona de disponibilidad que usa el clúster.

- *Zone-redundant:* Microsoft administra completamente la conmutación por error de zona de disponibilidad y la recuperación para Azure Data Explorer. No es necesario iniciar ni validar los procesos de fallo de una zona de disponibilidad.
- *Zonal:* Para simular parcialmente la pérdida de todos los nodos de cómputo durante una interrupción de zona, puede detener su clúster. Utilice este enfoque para validar partes de sus propios procesos de detección de caída de zona y conmutación por error.

Resistencia a errores en toda la región

Un clúster de Azure Data Explorer se implementa en una sola región de Azure. Si esa región deja de estar disponible, el clúster y sus datos no están disponibles.

Soluciones de varias regiones personalizadas para la resistencia

Para minimizar el impacto empresarial de una interrupción de una región, implemente clústeres de Azure Data Explorer independientes en varias regiones. Cada clúster es independiente. Eres responsable de administrar cada clúster y coordinar la replicación de datos, el enrutamiento del tráfico y la conmutación por error entre regiones.

Puede decidir entre diferentes tipos de configuraciones de clúster de varias regiones, que admiten distintos niveles de tiempo de recuperación, posible pérdida de datos, esfuerzo y costo. Seleccione las regiones de Azure para cada clúster que cumplan los requisitos de latencia y residencia de datos. Para obtener más información sobre las configuraciones y los patrones de clúster de varias regiones que puede seguir, consulte Información general sobre la continuidad empresarial y la recuperación ante desastres.

Copias de seguridad y restauración

Para la mayoría de las soluciones, no debe confiar exclusivamente en copias de seguridad. En su lugar, utilice las otras capacidades descritas en esta guía para apoyar los requisitos de resiliencia. Sin embargo, las copias de seguridad protegen contra algunos riesgos que otros enfoques no. Para más información, consulte ¿Qué son la redundancia, la replicación y la copia de seguridad?.

Azure Data Explorer no proporciona una funcionalidad nativa de copia de seguridad y restauración. Si necesita realizar una copia de seguridad de los datos, tenga en cuenta los siguientes enfoques:

- La exportación continua exporta periódicamente los datos al almacenamiento externo y proporciona *una exportación exactamente una vez* para los tipos de datos admitidos.
- La exportación de datos al almacenamiento en la nube admite la exportación manual de datos al almacenamiento externo.
- Ingerir datos sin procesar en Azure Data Explorer desde una fuente de entrada, como un lago de datos, del que puede hacer una copia de seguridad por separado.

Resistencia a la eliminación accidental

Azure Data Explorer incluye varios mecanismos para ayudarle a protegerse frente a la eliminación accidental de clústeres, bases de datos, tablas y tablas externas:

- **Eliminación accidental de clústeres o bases de datos:** La eliminación accidental de clústeres o bases de datos es una acción irrecuperable. Evite la pérdida de datos habilitando un bloqueo de eliminación en el clúster o el recurso de base de datos.
- **Eliminación accidental de tablas:** Los usuarios que tienen permisos de administrador de tablas o superiores pueden quitar tablas. Si uno de esos usuarios elimina accidentalmente una tabla, puede recuperarla si usa el comando undo drop table. Para que este comando se realice correctamente, primero debe habilitar la propiedad *de capacidad de recuperación* en la directiva de retención.
- **Eliminación accidental de tablas externas:**las tablas externas son entidades de esquema de consulta de Kusto que hacen referencia a datos almacenados fuera de la base de datos. La eliminación de una tabla externa solo elimina los metadatos de la tabla. Puede recuperarlo ejecutando de nuevo el comando de creación de tablas.

  En el caso de las tablas externas de Azure Blob Storage y Azure Data Lake, use la funcionalidad de eliminación temporal para protegerse frente a la eliminación accidental o sobrescritura de un blob durante un período de tiempo configurado por el usuario.

Resistencia al mantenimiento del servicio

Azure Data Explorer aplica periódicamente las actualizaciones del servicio y realiza un mantenimiento rutinario. La plataforma Azure controla estas actividades automáticamente mientras permanece dentro de los niveles de disponibilidad especificados en el Acuerdo de Nivel de Servicio. Asegúrese de que las aplicaciones están preparadas para la pérdida ocasional de

conectividad durante el mantenimiento del servicio siguiendo las instrucciones de control de errores transitorios.

Para obtener información sobre el próximo mantenimiento planeado, use Azure Service Health.

Acuerdo de nivel de servicio

El acuerdo de nivel de servicio (SLA) para Azure servicios describe la disponibilidad esperada de cada servicio y las condiciones que la solución debe cumplir para lograr esa expectativa de disponibilidad. Para obtener más información, consulte Acuerdos de Nivel de Servicio para servicios en línea.

Para que sea apto para el Acuerdo de Nivel de Servicio de disponibilidad de Azure Data Explorer, la aplicación debe controlar los errores transitorios mediante el reintento de solicitudes con errores.

Contenido relacionado

- Confiabilidad en Azure
- Introducción a Azure Data Explorer
- Introducción a BCDR

ⓘ**Nota:** El autor ha creado este artículo con ayuda de inteligencia artificial. Más información

---

Last updated on 11/06/2026

Confiabilidad en Azure Data Factory

Azure Data Factory le permite crear canalizaciones de datos flexibles y eficaces para la integración de datos sin servidor y la transformación de datos. Como servicio de Azure, Data Factory proporciona una variedad de funcionalidades para satisfacer los requisitos de fiabilidad.

Cuando se usa Azure, la confiabilidad es una responsabilidad compartida. Microsoft proporciona una variedad de funcionalidades para admitir resistencia y recuperación. Es responsable de comprender cómo funcionan esas funcionalidades dentro de todos los servicios que usa y de seleccionar las funcionalidades que necesita para cumplir los objetivos empresariales y los objetivos de tiempo de actividad.

En este artículo se describe cómo hacer que Data Factory sea resistente a una variedad de posibles interrupciones y problemas, incluidos errores transitorios, interrupciones de zona de disponibilidad y interrupciones de regiones. También se describe cómo puede usar copias de seguridad para recuperarse de otros tipos de problemas y se resalta cierta información clave sobre el acuerdo de nivel de servicio (SLA) de Data Factory.

> ⓘ **Nota:**
>
> Al considerar la confiabilidad de la factoría de datos, también debe tener en cuenta la confiabilidad de los almacenes de datos a los que se conecta. La mejora de la resistencia de la factoría de datos por sí sola podría tener un impacto limitado si los almacenes de datos no son igualmente resistentes. En función de los requisitos de resistencia, es posible que tenga que realizar cambios de configuración en varias áreas. Para ayudar a garantizar que los almacenes de datos cumplan los requisitos de continuidad empresarial, consulte la documentación y las instrucciones de confiabilidad del producto.

Introducción a la arquitectura de confiabilidad

Data Factory consta de varios componentes de infraestructura. Cada componente admite la confiabilidad de la infraestructura de maneras diferentes.

Los componentes de Data Factory incluyen:

- **El servicio principal de Data Factory**, que administra los desencadenadores de canalización y supervisa la coordinación de las actividades de canalización. El servicio principal también administra los metadatos de cada componente de la factoría de datos. Microsoft administra el servicio principal.

- **Entornos de ejecución de integración (IR)**, que se conectan a almacenes de datos y realizan actividades definidas en la canalización. Hay diferentes tipos de infrarrojos.
    - *IR administrados por Microsoft*, que incluyen Azure IR y Azure-SQL Server Integration Services (Azure-SSIS) IR. Microsoft administra los componentes que componen estos entornos de ejecución. En algunos escenarios, se configuran las opciones que afectan a la resiliencia del IR.
    - *Entornos de ejecución de integración autohospedados (SHIR)*. Microsoft proporciona software que puede ejecutar en su propia infraestructura informática para realizar algunas partes de sus propias canalizaciones de Data Factory. Es responsable de la implementación y administración de los recursos de computación y de la resiliencia de esos recursos de computación.

Resistencia a errores transitorios

Los errores transitorios son errores breves e intermitentes en los componentes. Se producen con frecuencia en un entorno distribuido como la nube y son una parte normal de las operaciones. Los errores transitorios se corrigen después de un breve período de tiempo. Es importante que las aplicaciones puedan controlar errores transitorios, normalmente mediante el reintento de solicitudes afectadas.

Todas las aplicaciones hospedadas en la nube deben seguir las instrucciones de control de errores transitorios de Azure cuando se comunican con cualquier API, bases de datos y otros componentes hospedados en la nube. Para obtener más información, consulte Recomendaciones para controlar errores transitorios.

Cuando se usa Data Factory, es importante prepararse para errores transitorios, especialmente al diseñar canalizaciones y actividades.

Idempotencia

Las actividades de canalización deben ser *idempotentes*, lo cual significa que se puedan volver a ejecutar varias veces sin causar efectos adversos. Si se produce un error transitorio como un error de red o una interrupción de zona de disponibilidad, Data Factory podría volver a ejecutar actividades de canalización. Esta repetición puede crear registros duplicados.

Para evitar la inserción de registros duplicados debido a un error transitorio, implemente los procedimientos recomendados siguientes:

- *Use identificadores únicos* para cada registro antes de escribir en la base de datos. Este enfoque puede ayudarle a encontrar y eliminar registros duplicados.

- *Use una estrategia upsert* para los conectores que admiten upsert. Antes de que se produzca la inserción de registros duplicados, use este enfoque para comprobar si ya existe un registro. Si existe, actualícela. Si no existe, insértelo. Por ejemplo, comandos SQL como `MERGE` o `ON DUPLICATE KEY UPDATE` usan este enfoque de upsert.

- *Use estrategias de acción de copia.* Para obtener más información, consulte Comprobación de coherencia de datos en la actividad de copia.

Directivas de reintentos

Puede usar directivas de reintento para configurar partes de la canalización para reintentar si hay un problema, como errores transitorios en los recursos conectados. En Data Factory, puede configurar directivas de reintento en los siguientes tipos de objetos de canalización:

- Desencadenadores periódicos
- Actividades de ejecución

Para obtener más información sobre cómo cambiar o deshabilitar las directivas de reintento para los desencadenadores y actividades de Data Factory, consulte Ejecuciones y desencadenadores de canalización.

Resistencia a errores de zona de disponibilidad

Las zonas de disponibilidad son grupos físicamente independientes de centros de datos dentro de una región de Azure. Cuando una zona falla, los servicios pueden transferirse a una de las zonas restantes.

Data Factory admite *redundancia de zona*, lo que proporciona resistencia a errores en las zonas de disponibilidad.

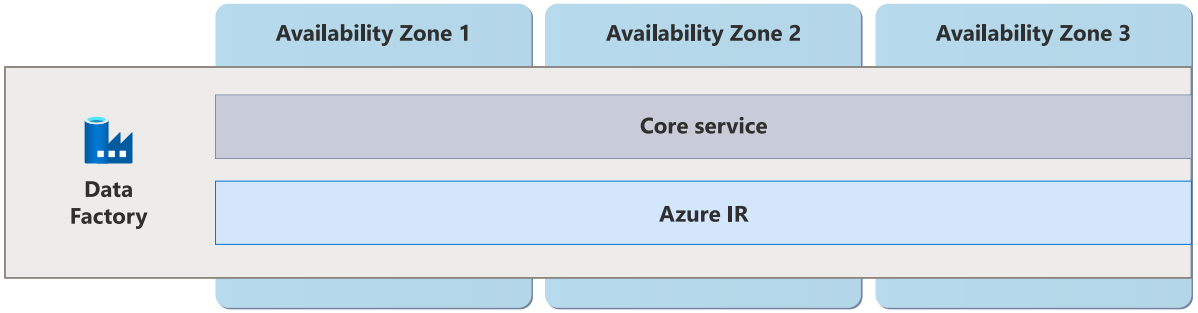
Cada parte de Data Factory admite redundancia de zona:

- **Servicio principal:** Microsoft administra los componentes en el servicio principal de Data Factory y los distribuye entre zonas de disponibilidad.

  Pero después de un error de zona, Microsoft no garantiza el estado de los desencadenadores de ventana de saltos de tamaño constante.

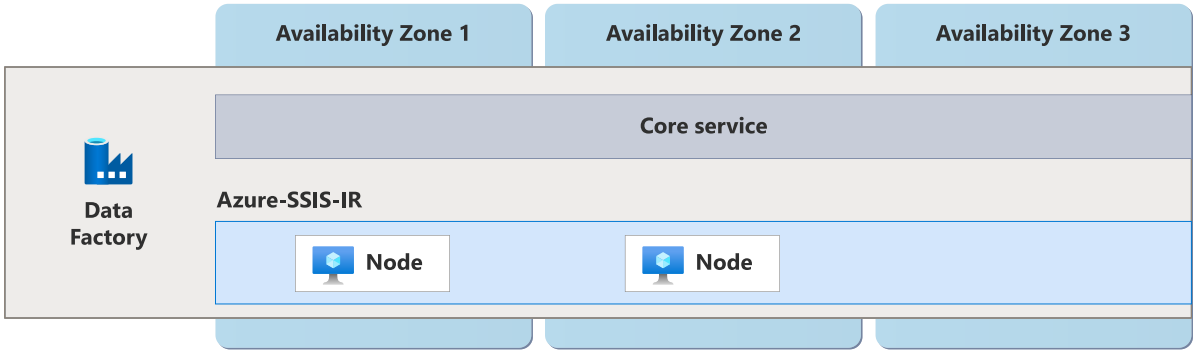
- **IRs:** La compatibilidad con la redundancia de zona depende del tipo de IR que se utilice.

  - *Azure IR* admite redundancia de zona y Microsoft administra esta funcionalidad.



- *Un Azure-SSIS IR* requiere la implementación de al menos dos nodos. Estos nodos se asignan automáticamente a diferentes zonas de disponibilidad.

  En el diagrama siguiente se muestra una canalización con redundancia de zona y un entorno de ejecución de integración de Azure-SSIS con dos nodos que se implementan en zonas diferentes:



- *Un SHIR* le ofrece la responsabilidad de implementar la infraestructura de proceso para hospedar el entorno de ejecución. Puede implementar varios nodos, como máquinas virtuales individuales y configurarlos para alta disponibilidad. Después, puede distribuir esos nodos en varias zonas de disponibilidad. Para más información, consulte Alta disponibilidad y escalabilidad.

Requisitos

Los recursos de Data Factory con redundancia de zona se pueden implementar en cualquier región que admita zonas de disponibilidad.

Cost

- **Servicio principal:** No se aplica ningún costo adicional para la redundancia de zona.
- **IR:** El costo de la redundancia de zona varía en función del tipo de IR que use.
  - *Azure IR* incluye redundancia de zona sin costo adicional.

- *Un Azure-SSIS IR* requiere que despliegue al menos dos nodos para lograr redundancia de zona. Para obtener más información sobre cómo se factura cada nodo, vea Ejemplo de precios: Ejecutar paquetes SSIS en un IR de Azure-SSIS.
- *Un SHIR* requiere que se implemente y administre la infraestructura de proceso. Para lograr la resiliencia de la zona, debe distribuir los recursos informáticos entre varias zonas. Dependiendo del número de nodos que implemente y cómo los configure, puede incurrir en costos adicionales de los servicios de proceso subyacentes y otros servicios auxiliares. No hay ningún cargo adicional para ejecutar el SHIR en varios nodos.

Configurar soporte de zonas de disponibilidad

- **Servicio principal:** No se requiere ninguna configuración. El servicio principal de Data Factory admite automáticamente la redundancia de zona.
- **IRs:**
  - *Una instancia de Azure IR:* No se requiere ninguna configuración. Azure IR habilita automáticamente la redundancia de zona.
  - *Un Azure-SSIS IR:* no se requiere configuración. Un IR de nivel Azure-SSIS habilita automáticamente la redundancia de zona cuando se implementa con dos o más nodos.
  - *Una SHIR* requiere que configure su propia resistencia, lo que incluye la propagación de los nodos entre varias zonas de disponibilidad.

Planeamiento y administración de capacidad

- **Servicio principal:** El servicio principal de Data Factory se escala automáticamente en función de la demanda y no es necesario planear ni administrar la capacidad.
- **IRs:**
  - *Una instancia de Azure IR* se escala automáticamente en función de la demanda y no es necesario planear ni administrar la capacidad.
  - *Un Azure-SSIS IR* requiere que configure específicamente el número de nodos que use. Para prepararse para errores de zona de disponibilidad, considere la posibilidad de sobreaprovisionar la capacidad del IR. El sobreaprovisionamiento permite a la solución tolerar cierto grado de pérdida de capacidad y seguir funcionando sin degradar el

rendimiento. Para obtener más información, consulte Administración de la capacidad mediante el aprovisionamiento excesivo.

- *Un SHIR* requiere que configure su propia capacidad y escalamiento. Considere la posibilidad de sobreaprovisionar al implementar una SHIR.

Comportamiento cuando todas las zonas están en buen estado

En esta sección se describe qué esperar cuando los recursos de Data Factory están configurados para la redundancia de zona y todas las zonas de disponibilidad están operativas.

- **Enrutamiento de tráfico entre zonas:** Durante las operaciones normales, Data Factory distribuye automáticamente las actividades de canalización, los desencadenadores y otros trabajos entre instancias correctas en cada zona de disponibilidad.

- **Replicación de datos entre zonas:** En general, Data Factory es un servicio sin estado, por lo que no es necesario replicar ningún estado entre zonas.

  Sin embargo, los desencadenadores de ventanas deslizantes contienen estado, que podría no estar completamente replicado entre zonas.

Comportamiento durante un fallo de zona

En esta sección se describe qué puede ocurrir cuando los recursos de Data Factory están configurados para la redundancia de zona y se produce una interrupción en la zona de disponibilidad.

- **Detección y respuesta:** La plataforma data Factory es responsable de detectar un error en una zona de disponibilidad y responder. No es necesario hacer nada para iniciar una conmutación por error de zona en las canalizaciones u otros componentes.

- **Notificación**: Microsoft no le notifica automáticamente cuando una zona está inactiva. Sin embargo, puede usar Azure Resource Health para supervisar el estado de un recurso individual y puede configurar alertas de Resource Health para notificarle problemas. También puede usar Azure Service Health para comprender el estado general del servicio, incluidos los errores de zona, y puede configurar alertas de Service Health para notificarle problemas.

- **Solicitudes activas:** cualquier canalización y desencadenador en proceso sigue ejecutándose, y no experimentará ninguna interrupción inmediata debido a un error en la zona. Sin embargo, las actividades en curso durante un error de zona pueden producir un error y reiniciarse. Es importante diseñar actividades para que sean idempotentes, lo que

les ayuda a recuperarse de errores de zona y de otros errores. Para obtener más información, consulte Resistencia a errores transitorios.

- **Tiempo de inactividad esperado:** No se espera ningún tiempo de inactividad durante un error de zona.
- **Pérdida de datos esperada:** En general, Data Factory es un servicio sin estado, por lo que no se espera ninguna pérdida de datos durante un error de zona.

  Pero si usa un desencadenador periódico, el estado del desencadenador podría perderse después de un error de zona. Debe reiniciar o volver a ejecutar los triggers que estaban en ejecución en el momento del fallo de zona.

Recuperación de zona

Cuando se recupera la zona de disponibilidad, Data Factory vuelve automáticamente a la zona original. No es necesario hacer nada para iniciar una conmutación por recuperación de zona en las canalizaciones u otros componentes.

Sin embargo, si usa una SHIR, es posible que tenga que reiniciar los recursos de cómputo si se han detenido.

Prueba de fallos de zona

Para el servicio principal, así como para los Azure y los Azure-SSIS IR, Data Factory administra el enrutamiento del tráfico, la conmutación por error y la conmutación por recuperación para los recursos con redundancia de zona. Dado que esta característica está totalmente administrada, no es necesario iniciar ni validar los procesos de error de zona de disponibilidad.

En el caso de los SHIR, puede usar Azure Chaos Studio para simular un error de zona de disponibilidad en una máquina virtual de Azure.

Resistencia a errores en toda la región

Los recursos de Data Factory se implementan en una sola región de Azure. Si la región deja de estar disponible, la factoría de datos tampoco está disponible. Sin embargo, hay enfoques que puede usar para ayudar a garantizar la resistencia a las interrupciones de la región. Estos enfoques dependen de si la factoría de datos está en una región emparejada o no emparejada y en sus requisitos y configuración específicos.

Conmutación por error administrada por Microsoft en una región emparejada

Data Factory admite la conmutación por error administrada por Microsoft para factorías de datos en regiones emparejadas, excepto Sur de Brasil y Sudeste de Asia. En el improbable caso de un error prolongado en la región, Microsoft podría iniciar una conmutación por error regional de la instancia de Data Factory.

Debido a los requisitos de residencia de datos en el Sur de Brasil y sudeste de Asia, los datos de Data Factory solo se almacenan en la región local mediante el almacenamiento con redundancia de zona de Azure Storage. Para sudeste asiático, todos los datos se almacenan en Singapur. Para el Sur de Brasil, todos los datos se almacenan en Brasil.

En el caso de las factorías de datos en regiones no emparejadas o en el Sur de Brasil o Sudeste de Asia, Microsoft no realiza la conmutación por error regional en su nombre.

> **ⓘ Importante**
>
> Microsoft inicia la conmutación por error administrada por Microsoft. Es probable que se produzca después de un retraso significativo y se realice con el mejor esfuerzo. También hay algunas excepciones a este proceso. Es posible que experimente alguna pérdida de los metadatos de la factoría de datos. La conmutación por error de los recursos de Data Factory puede producirse en un momento diferente del tiempo de conmutación por error de otros servicios de Azure.
>
> Si necesita ser resistente a interrupciones de regiones, considere la posibilidad de usar una de las **soluciones personalizadas de varias regiones para lograr resistencia**.

Conmutación por error de los IR

Para prepararse para una conmutación por error, puede haber algunas consideraciones adicionales, en función del entorno de ejecución de integración utilizado.

- Puede configurar *Azure IR* para resolver automáticamente la región que usa. Si la región está establecida para *resolverse automáticamente* y hay una interrupción en la región principal, el Azure IR cambia automáticamente a la región emparejada. Esta conmutación por error está sujeta a limitaciones. Para configurar la región de Azure IR para la implementación de la actividad o el envío en la configuración de IR, establezca la región para *resolverse automáticamente*.
- La conmutación por error de *Azure-SSIS IR* se administra por separado en la conmutación por error administrada por Microsoft de Data Factory. Para obtener más información,

consulte Soluciones personalizadas de varias regiones para lograr resistencia.

- *Una SHIR* se ejecuta en la infraestructura por la que es responsable, por lo que una conmutación por error administrada por Microsoft no se aplica a las SHIR. Para obtener más información, consulte Soluciones personalizadas de varias regiones para lograr resistencia.

Reconfiguración posterior a la conmutación por error

Una vez completada una conmutación por error administrada por Microsoft, puede acceder a la canalización de Data Factory en la región emparejada. Sin embargo, una vez completada la conmutación por error, es posible que tenga que realizar alguna reconfiguración para los IR u otros componentes. Este proceso incluye volver a establecer la configuración de red.

Soluciones personalizadas de varias regiones para la resistencia

Si necesita que las canalizaciones sean resistentes a las interrupciones regionales y necesita controlar el proceso de conmutación por error, considere la posibilidad de usar una canalización controlada por metadatos.

- **Configure el control de código fuente para que Data Factory** realice un seguimiento y audite los cambios realizados en los metadatos. Puede usar este enfoque para acceder a los archivos JSON de metadatos para canalizaciones, conjuntos de datos, servicios vinculados y desencadenadores. Data Factory admite diferentes tipos de repositorio de Git, como Azure DevOps y GitHub. Para más información, consulte Control de código fuente en Data Factory.

- **Utilice un sistema de integración continua y entrega continua (CI/CD),** como Azure DevOps, para gestionar los metadatos de la canalización y las implementaciones. Puede usar CI/CD para restaurar rápidamente las operaciones en una instancia de otra región. Si una región no está disponible, puede aprovisionar una nueva factoría de datos manualmente o a través de la automatización. Una vez creada la nueva factoría de datos, puede restaurar las canalizaciones, los conjuntos de datos y los servicios vinculados JSON desde el repositorio de Git existente. Para más información, consulte Continuidad empresarial y recuperación ante desastres (BCDR) para canalizaciones de Data Factory y Azure Synapse Analytics.

Dependiendo del IR que utilice, puede haber otras consideraciones.

- Un *Azure-SSIS IR* usa una base de datos almacenada en Azure SQL Database o Azure SQL Managed Instance. Puede configurar la replicación geográfica o un grupo de

conmutación por error para esta base de datos. La base de datos Azure-SSIS se encuentra en una región primaria de Azure que tiene acceso de lectura y escritura. La base de datos se replica continuamente en una región secundaria que tiene acceso de solo lectura. Si la región primaria no está disponible, se desencadena una conmutación por error, lo que hace que las bases de datos principales y secundarias intercambien roles.

También puede configurar un par de Azure-SSIS IR en espera dual que funcione en sincronización con el grupo de conmutación por error de Azure SQL Database o SQL Managed Instance.

Para obtener más información, consulte Configurar un IR de Azure-SSIS para BCDR.

- *Una SHIR* se ejecuta en la infraestructura que administra. Si el SHIR se implementa en una máquina virtual de Azure, puede usar Azure Site Recovery para desencadenar la conmutación por error de la máquina virtual en otra región.

Copias de seguridad y restauración

Data Factory admite CI/CD a través de la integración del control de código fuente, de modo que pueda realizar una copia de seguridad de los metadatos de una instancia de Data Factory. Las canalizaciones de CI/CD implementan estos metadatos sin problemas en un nuevo entorno. Para más información, consulte CI/CD en Data Factory.

Acuerdo de nivel de servicio

El contrato de nivel de servicio (SLA) para los servicios de Azure describe la disponibilidad esperada de cada servicio y las condiciones que la solución deberá cumplir para lograr esa expectativa de disponibilidad. Para obtener más información, consulte Acuerdos de Nivel de Servicio para servicios en línea.

Data Factory proporciona acuerdos de nivel de servicio de disponibilidad independientes para:

- La tasa de éxito de las llamadas API que hagas, como aquellas para gestionar tu fábrica de datos.
- Número de ejecuciones de actividad que comienzan a ejecutarse.

Las ejecuciones de actividad se pueden retrasar brevemente y requieren que se cumplan todas las dependencias para ejecutar el trabajo.

Contenido relacionado

- Confiabilidad en Azure

Last updated on 22/01/2026

Recuperación ante desastres para Azure Data Share

En este artículo se explica cómo configurar un entorno de recuperación ante desastres para Azure Data Share. Las interrupciones del centro de datos de Azure son poco frecuentes, pero pueden durar desde unos minutos hasta algunas horas. Las interrupciones del centro de datos pueden provocar interrupciones en los entornos que se basan en los datos compartidos por el proveedor de datos. Al seguir los pasos que se detallan en este artículo, los proveedores de datos pueden seguir compartiendo los datos con los consumidores de datos si se produce una interrupción del centro de datos en la región primaria que hospeda el recurso compartido de datos.

Consecución de la continuidad empresarial para Azure Data Share

Para estar preparado para una interrupción del centro de datos, el proveedor de datos puede tener un entorno de recurso compartido de datos aprovisionado en una región secundaria. Se pueden tomar medidas para garantizar una conmutación por error fluida si se produce una interrupción en el centro de datos.

Los proveedores de datos pueden crear recursos secundarios de Azure Data Share en otra región. Estos recursos de Data Share se pueden configurar para incluir recursos compartidos y conjuntos de datos que existan en el recurso principal de Azure Data Share. Pueden invitar a los consumidores de datos a los recursos compartidos secundarios al configurar el entorno de recuperación ante desastres, o en un momento posterior (por ejemplo, como parte de los pasos de una conmutación por error manual).

Si los consumidores de datos tienen una suscripción activa al recurso compartido en un entorno secundario aprovisionado con fines de recuperación ante desastres, estos pueden habilitar la programación de instantáneas como parte de una conmutación por error. Si los consumidores de datos no quieren suscribirse a una región secundaria para fines de recuperación ante desastres, se pueden invitar al recurso compartido de datos secundario en un momento posterior.

Los consumidores de datos pueden tener una suscripción activa al recurso compartido que esté inactiva para propósitos de recuperación ante desastres, o bien los proveedores de datos pueden agregarlos en un momento posterior como parte de los procedimientos de conmutación por error manual.

Contenido relacionado

- Continuidad empresarial y recuperación ante desastres
- Creación de alta disponibilidad a su estrategia de continuidad empresarial y recuperación ante desastres

Para obtener información acerca de cómo empezar a compartir datos, vaya al tutorial que cubre cómo compartir sus datos.

Last updated on 27/01/2026

Confiabilidad en Azure Databricks

Azure Databricks es una plataforma de inteligencia artificial y datos basados en Apache Spark colaborativa optimizada para Microsoft Azure. Proporciona un entorno unificado para cargas de trabajo de macrodatos e inteligencia artificial y combina lo mejor de Databricks y Azure para simplificar la ingeniería de datos, la ciencia de datos y el aprendizaje automático.

Cuando se usa Azure, la confiabilidad es una responsabilidad compartida. Microsoft proporciona una variedad de funcionalidades para admitir resistencia y recuperación. Es responsable de comprender cómo funcionan esas funcionalidades dentro de todos los servicios que usa y de seleccionar las funcionalidades que necesita para cumplir los objetivos empresariales y los objetivos de tiempo de actividad.

En este artículo se describe cómo Azure Databricks mantiene la resistencia frente a varias posibles interrupciones y problemas y cómo puede configurar la resistencia para satisfacer sus requisitos. En las instrucciones se describen los errores transitorios, las interrupciones de zona de disponibilidad, las interrupciones de región y el mantenimiento del servicio. En este artículo también se describe cómo usar copias de seguridad para recuperarse de otros problemas y se resalta información clave sobre el contrato de nivel de servicio (SLA) de Azure Databricks.

Recomendaciones de implementación de producción

Para obtener información sobre cómo implementar Azure Databricks para admitir los requisitos de confiabilidad de la solución y cómo afecta la confiabilidad a otros aspectos de la arquitectura, consulte Procedimientos recomendados de arquitectura para Azure Databricks.

Introducción a la arquitectura de confiabilidad

Debe comprender la confiabilidad de cada componente principal de Azure Databricks:

- **El plano de control** es una colección de servicios sin estado que administra los metadatos del área de trabajo, el acceso de usuario, la programación de trabajos y la administración de clústeres. Estos servicios están respaldados por bases de datos que se replican en múltiples zonas de disponibilidad dentro de las regiones admitidas.
- **La raíz del sistema de archivos de Databricks (DBFS)** es una cuenta de almacenamiento que Azure Databricks aprovisiona automáticamente al crear un área de trabajo de Azure Databricks en la cuenta en la nube. Se recomienda no almacenar datos en la raíz de DBFS y deshabilitar esta cuenta de almacenamiento si es posible.

- **El almacenamiento del catálogo de Unity** incluye una o varias cuentas de almacenamiento que almacenan los datos del catálogo de Unity en la cuenta en la nube. Para obtener más información, consulte Introducción al catálogo de Unity.

- **El plano de proceso** ejecuta cargas de trabajo de procesamiento de datos mediante clústeres de máquinas virtuales (VM). El plano de proceso controla los errores transitorios y reemplaza automáticamente los nodos con errores sin intervención del usuario. Puede elegir entre varios tipos de recursos de proceso. Para obtener más información, consulte Computación.

  La disponibilidad del área de trabajo depende de la disponibilidad del plano de control, pero los clústeres de proceso pueden seguir procesando trabajos incluso durante las interrupciones del plano de control.

Resistencia a errores transitorios

Los errores transitorios son errores breves e intermitentes en los componentes. Se producen con frecuencia en un entorno distribuido como la nube y son una parte normal de las operaciones. Los errores transitorios se corrigen después de un breve período de tiempo. Es importante que las aplicaciones puedan controlar errores transitorios, normalmente mediante el reintento de solicitudes afectadas.

Todas las aplicaciones hospedadas en la nube deben seguir las instrucciones de control de errores transitorios de Azure cuando se comunican con cualquier API, bases de datos y otros componentes hospedados en la nube. Para obtener más información, consulte Recomendaciones para controlar errores transitorios.

Puede controlar los reintentos de las tareas dentro de trabajos de Lakeflow para ayudar a recuperarse de errores transitorios.

En el caso de las aplicaciones que se ejecutan en Azure Databricks, implemente la lógica de reintento con retroceso exponencial al conectarse a servicios externos o servicios de Azure, como Storage, Azure SQL Database o Azure Event Hubs. Databricks Runtime incluye resistencia integrada para muchos servicios de Azure, pero el código de la aplicación debe controlar errores transitorios específicos del servicio.

Resistencia a errores de zona de disponibilidad

Las zonas de disponibilidad son grupos físicamente independientes de centros de datos dentro de una región de Azure. Cuando una zona falla, los servicios pueden transferirse a una de las zonas restantes.

Azure Databricks admite *redundancia de zona* para cada componente:

- **Plano de control:** En regiones que admiten zonas de disponibilidad, el plano de control se ejecuta en varias zonas de disponibilidad. El plano de control controla los errores de zona automáticamente, con un impacto mínimo y ninguna intervención del usuario necesaria.

  Los datos del área de trabajo del plano de control se almacenan en bases de datos. En las regiones que admiten zonas de disponibilidad, las bases de datos se replican en varias zonas de la región. Las cuentas de almacenamiento que atienden imágenes en tiempo de ejecución de Databricks también son redundantes dentro de la región. Todas las regiones tienen cuentas de almacenamiento secundarias que se usan cuando la cuenta de almacenamiento principal está inactiva.

- **Raíz de DBFS:** En las regiones que admiten zonas de disponibilidad, puede configurar la cuenta de almacenamiento para que la raíz de DBFS use el almacenamiento con redundancia de zona (ZRS). En regiones emparejadas que admiten zonas de disponibilidad, puede usar opcionalmente almacenamiento con redundancia de zona geográfica (GZRS).

- **Plano de proceso:** Databricks admite la *distribución automática de zonas* para los recursos de proceso, lo que significa que los recursos se distribuyen entre varias zonas de disponibilidad. Esta distribución ayuda a las cargas de trabajo de producción a lograr resistencia a las interrupciones de zona.

  Cuando se usa el proceso sin servidor, no se seleccionan explícitamente zonas para el proceso. Databricks administra la selección de zona de las máquinas virtuales y el reemplazo de máquinas virtuales que podrían perderse debido a interrupciones de zona.

Requisitos

Para usar la compatibilidad con zonas de disponibilidad en Azure Databricks, necesita los siguientes requisitos:

- **Compatibilidad con regiones:** La compatibilidad con la zona de disponibilidad de Azure Databricks está disponible en todas las regiones de Azure que admiten Azure Databricks y proporcionan zonas de disponibilidad. Para obtener una lista de las regiones que admiten Azure Databricks, consulte Productos disponibles por región. Para obtener una lista completa de las regiones que admiten zonas de disponibilidad, consulte Regiones de Azure que admiten zonas de disponibilidad.

- **Replicación de almacenamiento:** Configure las cuentas de almacenamiento del área de trabajo para usar ZRS o GZRS (cuando esté disponible).

- **Capacidad de proceso:** Asegúrese de que existe suficiente capacidad de proceso en varias zonas de la región de destino. Azure Databricks distribuye automáticamente los nodos de clúster entre zonas, pero debe comprobar que los tipos de instancia seleccionados están disponibles en todas las zonas de destino.

Consideraciones

Azure Databricks distribuye automáticamente los nodos de clúster entre zonas de disponibilidad. La distribución depende de la capacidad disponible en cada zona. Durante períodos de alta demanda, los nodos de un clúster se pueden concentrar en menos zonas. Cuando usas computación sin servidor, Azure Databricks gestiona la selección de zona de las máquinas virtuales (VMs) y el reemplazo de VMs que podrían perderse debido a fallos de zonas.

Cost

La distribución de zona no afecta a los costos de proceso porque paga por el mismo número de máquinas virtuales, independientemente de su ubicación de zona de disponibilidad. Para obtener más información, consulte Precios de computación de Azure Databricks.

La redundancia predeterminada para la cuenta de almacenamiento administrada o la raíz de DBFS es el almacenamiento con redundancia geográfica (GRS). Cambiar a ZRS o GZRS puede afectar a los costos de almacenamiento. Para obtener más información, consulte Precios de Azure Blob Storage.

Configurar soporte de zonas de disponibilidad

- **Plano de control:** El plano de control admite automáticamente la redundancia de zona en regiones que tienen zonas de disponibilidad. No es necesario configurar nada.

- **Raíz de DBFS:** Puede configurar la redundancia de zona para el almacenamiento raíz de DBFS al crear un área de trabajo o modificar una área de trabajo existente:

  - **Cree un nuevo área de trabajo con almacenamiento raíz de DBFS redundante de zona:** Cuando cree un nuevo área de trabajo de Azure Databricks, puede configurar opcionalmente la cuenta de almacenamiento asociada para utilizar ZRS o GZRS en lugar del GRS predeterminado. Para más información, consulte Cambio de las opciones de redundancia de almacenamiento del área de trabajo.

  - **Habilite la redundancia de zona en el almacenamiento raíz de DBFS:** En el caso de las áreas de trabajo existentes, puede cambiar la configuración de redundancia de la cuenta de almacenamiento del área de trabajo a ZRS o GZRS. Para obtener más

información sobre cómo habilitar la redundancia de zona, consulte Cambio de la configuración de replicación para una cuenta de almacenamiento.

- **Plano de proceso:** Los nodos de clúster se distribuyen automáticamente entre zonas de disponibilidad. No se requiere ninguna configuración de cliente para la distribución de zona.

Comportamiento cuando todas las zonas están en buen estado

En esta sección se describe qué esperar cuando un área de trabajo está configurada con compatibilidad con zonas de disponibilidad y todas las zonas de disponibilidad están operativas.

- **Replicación de datos entre zonas:** La replicación de datos para el almacenamiento del área de trabajo se produce sincrónicamente entre zonas cuando la raíz de DBFS usa una cuenta ZRS o GZRS. Este enfoque garantiza una coherencia sólida con un impacto mínimo en el rendimiento.

- **Enrutamiento de tráfico entre zonas:** Azure Databricks distribuye automáticamente los nodos de clúster entre zonas durante la creación del clúster. El servicio equilibra la carga de proceso entre zonas mientras mantiene la localidad de datos para obtener un rendimiento óptimo.

Comportamiento durante un fallo de zona

En esta sección se describe qué esperar cuando un área de trabajo está configurada con soporte para zona de disponibilidad y hay una interrupción en la zona de disponibilidad.

- **Detección y respuesta:** Microsoft detecta automáticamente errores de zona e inicia procedimientos de respuesta. No es necesario realizar ninguna acción para la conmutación por error de nivel de zona.

- **Notificación:** Microsoft no le notifica automáticamente cuando una zona está inactiva. Pero puede usar la página de estado de Azure Databricks para ver una introducción a todos los servicios principales de Azure Databricks. También puede suscribirse a actualizaciones de estado de componentes individuales del servicio y recibir una alerta cuando cambie el estado del servicio al que está suscrito.

- **Solicitudes activas:** La ejecución de clústeres podría perder nodos en la zona afectada. El administrador de clústeres solicita automáticamente nodos de reemplazo de zonas

restantes. Si se pierde el nodo del controlador, el clúster y el trabajo se reinician completamente.

- **Pérdida de datos esperada:**
  - **Plano de control:** No se espera ninguna pérdida de datos durante una interrupción de zona.
  - **Raíz de DBFS:** Los datos del área de trabajo permanecen disponibles si usa configuraciones de almacenamiento de ZRS o GZRS.
  - **Plano de proceso:** Los datos almacenados en caché en máquinas virtuales son efímeros. Los datos perdidos de las máquinas virtuales durante un error de zona se recuperan del almacenamiento. Si se pierde el nodo del controlador, el trabajo se reinicia y vuelve a calcular los resultados.
- **Tiempo de inactividad esperado:**
  - **Plano de control:** El plano de control de Databricks realiza la conmutación automática en caso de fallo a zonas operativas en unos 15 minutos.
  - **Raíz de DBFS:** No se espera ningún tiempo de inactividad para las cuentas de almacenamiento que usan ZRS o GZRS.
  - **Plano de proceso:** Si se pierden nodos porque sus máquinas virtuales residen en la zona de disponibilidad afectada, el administrador de clústeres de Azure solicita nodos de reemplazo del proveedor de proceso de Azure. Si las zonas correctas restantes tienen capacidad suficiente para cumplir la solicitud, el proveedor de proceso extrae nodos de las zonas correctas para reemplazar los nodos perdidos. Este proceso puede tardar varios minutos.

    Si el nodo del controlador se pierde debido al error de zona, se reinicia todo el clúster, lo que puede provocar tiempos de recuperación más largos en comparación con la pérdida de nodos de trabajo. Planee este comportamiento en las estrategias de programación y supervisión del trabajo.

    Puede usar grupos de instancias o sin servidor para reducir este tiempo.
- **Reenrutamiento del tráfico:**
  - **Plano de control:** El plano de control de Databricks realiza la conmutación automática en caso de fallo a zonas operativas en unos 15 minutos.
  - **Raíz de DBFS:** Azure Storage redirige automáticamente las solicitudes a los clústeres de almacenamiento en zonas correctas.

- **Plano de proceso:** El administrador de clústeres cambia automáticamente a los nodos de zonas correctas.

Recuperación de zona

Cuando la zona de disponibilidad fallida se recupera, Azure Databricks reanuda las operaciones normales automáticamente en todas las zonas. El administrador de clústeres puede reequilibrar la distribución de nodos durante las posteriores creaciones de nodos, pero los nodos existentes continúan ejecutándose en sus zonas actuales hasta que finalizan.

No es necesario realizar ninguna acción para las operaciones de conmutación por recuperación. La distribución de zona normal se reanuda para las nuevas implementaciones de clúster.

Prueba de fallos de zona

Azure Databricks es un servicio administrado en el que Microsoft controla la conmutación por error de zona automáticamente y realiza pruebas periódicas de zona inactiva. No es necesario probar los escenarios de fallo de zona para el servicio en sí.

Para las aplicaciones que se ejecutan en Azure Databricks, pruebe la resistencia del trabajo simulando errores de nodo de controlador y supervisando el comportamiento de reinicio del clúster. Verifique que los trabajos de procesamiento de datos pueden manejar los reinicios del cluster y reanudarse desde los puntos de control adecuados.

Resistencia a errores en toda la región

Azure Databricks es un servicio de una sola región. Si la región no está disponible, el área de trabajo tampoco está disponible. Si necesita implementaciones en varias regiones, consulte Recuperación ante desastres de Azure Databricks.

Soluciones personalizadas de varias regiones para la resistencia

Azure Databricks no proporciona funcionalidades integradas en varias regiones. Para una protección multirregional completa de sus cargas de trabajo analíticas, debe implementar su propio enfoque.

Las soluciones típicas de varias regiones implican dos o más áreas de trabajo. Puede elegir entre varias estrategias, incluidas las arquitecturas activa-pasiva y activa-activa.

Para elegir una arquitectura, tenga en cuenta los siguientes factores:

- La importancia de la carga de trabajo para su empresa
- La posible duración de una interrupción (horas o posiblemente un día completo)
- El esfuerzo necesario para hacer que el área de trabajo esté totalmente operativa
- El esfuerzo necesario para restaurar o recuperar a la región primaria

Para ver las cargas de trabajo que requieren protección en varias regiones, consulte Recuperación ante desastres de Azure Databricks.

Copia de seguridad y recuperación

Azure Databricks realiza automáticamente copias de seguridad de bases de datos como parte de las operaciones administradas del servicio. Este proceso incluye contenido de cuadernos, definiciones de trabajos, configuraciones de clúster y configuración de control de acceso.

> **Nota:**
>
> Si se produce un error de zona, Azure Databricks no espera ninguna pérdida de datos.

Se recomienda almacenar los datos en el almacenamiento del catálogo de Unity. Puede replicar datos a través de la replicación de almacenamiento o la clonación diferencial.

Las funcionalidades de copia de seguridad y restauración de nivel de área de trabajo no están disponibles directamente. Planee los procedimientos de recreación del área de trabajo que incluyen la restauración de configuraciones, usuarios y controles de acceso desde los procesos de sincronización.

Resistencia al mantenimiento del servicio

Azure Databricks realiza el mantenimiento automático de la plataforma para aplicar actualizaciones de seguridad, implementar nuevas características y mejorar la confiabilidad del servicio. Puede configurar las ventanas de mantenimiento del clúster para reducir la probabilidad de mantenimiento que afecte a las cargas de trabajo de producción. Para más información, consulte Actualización automática del clúster.

Acuerdo de nivel de servicio

El contrato de nivel de servicio (SLA) para los servicios de Azure describe la disponibilidad esperada de cada servicio y las condiciones que la solución deberá cumplir para lograr esa

expectativa de disponibilidad. Para obtener más información, consulte Acuerdos de Nivel de Servicio para servicios en línea.

Contenido relacionado

- Recuperación ante desastres de Azure Databricks
- Procedimientos recomendados de confiabilidad para Azure Databricks
- Introducción a la confiabilidad de Azure

Last updated on 22/01/2026

Confiabilidad en Azure Event Hubs

Azure Event Hubs es un servicio en la nube nativo que puede transmitir millones de eventos por segundo con baja latencia, desde cualquier origen a cualquier destino. Use Event Hubs para ingerir y almacenar datos de streaming e integrarlos con aplicaciones cliente creadas para Apache Kafka o aplicaciones que usan los SDK de cliente de Event Hubs.

Cuando se usa Azure, la confiabilidad es una responsabilidad compartida. Microsoft proporciona una variedad de funcionalidades para admitir resistencia y recuperación. Es responsable de comprender cómo funcionan esas funcionalidades dentro de todos los servicios que usa y de seleccionar las funcionalidades que necesita para cumplir los objetivos empresariales y los objetivos de tiempo de actividad.

En este artículo se describe cómo Event Hubs es resistente a una variedad de posibles interrupciones y problemas, y cómo puede configurarlo para que sea resistente a otros usuarios, incluidos errores transitorios, interrupciones de zona de disponibilidad y interrupciones de región. También se describen las opciones de copia de seguridad y recuperación, y se resalta cierta información clave sobre el contrato de nivel de servicio (SLA) de Azure Event Hubs.

Recomendaciones de implementación de producción

Para obtener información sobre cómo implementar Event Hubs para admitir los requisitos de confiabilidad de la solución y comprender cómo afecta la confiabilidad a otros aspectos de la arquitectura, consulte Procedimientos recomendados de arquitectura para Event Hubs en Azure Well-Architected Framework.

Introducción a la arquitectura de confiabilidad

En esta sección se describen aspectos importantes sobre cómo funciona Event Hubs desde una perspectiva de confiabilidad. Presenta la arquitectura lógica, que incluye recursos y características que se implementan y usan. También se explica la arquitectura física, que proporciona detalles sobre cómo el servicio administra las operaciones internamente.

Arquitectura lógica

Un *espacio de nombres* de Event Hubs actúa como contenedor de administración para uno o varios centros de eventos. Configure el servicio, como asignar capacidad de streaming,

configurar la seguridad de red y habilitar la resistencia geográfica y la recuperación ante desastres geográficas, en el nivel de espacio de nombres.

Dentro de un espacio de nombres, puede organizar eventos en un *centro de eventos*. El ecosistema de Apache® Kafka hace referencia a este tipo de entidad como *tema*. El centro de eventos o tema es un registro distribuido de solo adición de tus eventos.

Cada centro de eventos contiene una o varias *particiones*, que son registros de eventos secuenciales. Un centro de eventos puede usar varias particiones para realizar el procesamiento paralelo y el escalado horizontal. Event Hubs solo garantiza la ordenación dentro de una sola partición. La creación de particiones desempeña un papel clave en el diseño de confiabilidad de la aplicación. Al diseñar la aplicación, realice un equilibrio entre maximizar la disponibilidad y la coherencia. Para maximizar el tiempo de actividad de la mayoría de las aplicaciones, evite direccionar particiones directamente desde las aplicaciones cliente. Para más información, consulte Disponibilidad y coherencia en Event Hubs.

Los consumidores que leen desde el centro de eventos pueden leer sus eventos secuencialmente manteniendo su propio *punto de control*, que identifica el último evento que reciben.

Para más información sobre las particiones y otros conceptos fundamentales de Event Hubs, consulte Características y terminología en Event Hubs.

Arquitectura física

En la arquitectura física, un espacio de nombres de Event Hubs se ejecuta dentro de un *clúster*. Un clúster proporciona los recursos de proceso y almacenamiento subyacentes. La mayoría de los espacios de nombres se ejecutan en clústeres compartidos por otros clientes de Azure. Cuando se usa el nivel Premium, el espacio de nombres se asigna a recursos dedicados dentro de un clúster compartido. Cuando usas el nivel Dedicado, un clúster se asigna exclusivamente a tus espacios de nombres. Para más información sobre los clústeres dedicados, consulte Introducción al nivel dedicado de Event Hubs. Independientemente del tipo de clúster y del nivel, Microsoft administra los clústeres y sus máquinas virtuales y almacenamiento subyacentes.

Para lograr redundancia, cada clúster tiene varias réplicas que procesan solicitudes de lectura y escritura. Para optimizar la alta disponibilidad y el rendimiento, todos los datos se almacenan en tres réplicas de almacenamiento. Para escalar los recursos de proceso del espacio de nombres, implemente unidades de rendimiento (TUs), unidades de procesamiento (PUs) o unidades de capacidad (CUs), en función del nivel. Para más información, consulte Escalado con Event Hubs.

Los clústeres abarcan varias máquinas físicas y bastidores, lo que reduce el riesgo de errores catastróficos que afectan al espacio de nombres. En las regiones que tienen zonas de disponibilidad, los clústeres se extienden entre centros de datos físicos independientes. Para obtener más información, consulte Resilience to availability zone failures (Resistencia a errores de zona de disponibilidad).

Resistencia a errores transitorios

Los errores transitorios son errores breves e intermitentes en los componentes. Se producen con frecuencia en un entorno distribuido como la nube y son una parte normal de las operaciones. Los errores transitorios se corrigen después de un breve período de tiempo. Es importante que las aplicaciones puedan controlar errores transitorios, normalmente mediante el reintento de solicitudes afectadas.

Todas las aplicaciones hospedadas en la nube deben seguir las instrucciones de control de errores transitorios de Azure cuando se comunican con cualquier API, bases de datos y otros componentes hospedados en la nube. Para obtener más información, consulte Recomendaciones para controlar errores transitorios.

Event Hubs implementa mecanismos transparentes de detección de errores y conmutación por error para que el servicio siga funcionando dentro de los niveles de servicio garantizados, normalmente sin interrupciones perceptibles cuando se produzcan errores.

Al diseñar aplicaciones cliente para trabajar con Event Hubs, siga estas instrucciones:

- **Utilice políticas de reintento integradas.** Los SDK de Event Hubs y Apache Kafka reintentan automáticamente las operaciones para errores que se pueden reintentar, como tiempos de espera en la red, respuestas de limitación, o cuando el servidor está ocupado. Para evitar sobrecargar innecesariamente el servicio, implementan el retroceso exponencial de forma predeterminada.

- **Configure los valores de tiempo de espera adecuados** en función de los requisitos de la aplicación. El tiempo de espera predeterminado suele ser de 60 segundos, pero puede ajustarlo en función de su escenario.

- **Implemente puntos de control** en el procesador de eventos para realizar un seguimiento del progreso y habilitar la recuperación desde la última posición procesada después de errores transitorios.

- **Use el procesamiento por lotes para las operaciones de envío** para mejorar el rendimiento y reducir el impacto de los problemas de red transitorios en mensajes individuales.

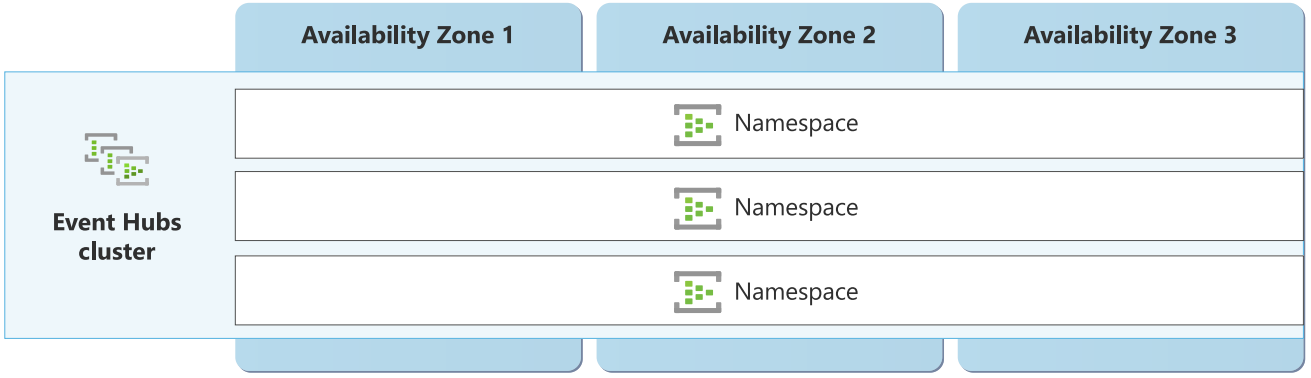
- **Use los SDK de Apache Kafka** si trabaja con el protocolo Kafka. Los SDK de Kafka también implementan directivas de reintento y otros procedimientos recomendados que ayudan a controlar errores transitorios.

Resistencia a errores de zona de disponibilidad

Las zonas de disponibilidad son grupos físicamente independientes de centros de datos dentro de una región de Azure. Cuando una zona falla, los servicios pueden transferirse a una de las zonas restantes.

Event Hubs admite implementaciones con redundancia de zona en todos los niveles de servicio. Al crear un espacio de nombres de Event Hubs en una región admitida, la redundancia de zona se habilita automáticamente sin costo adicional. Pero con el nivel Dedicado, las zonas de disponibilidad solo son compatibles con un mínimo de tres CUs. El modelo de implementación con redundancia de zona se aplica a todas las características de Event Hubs, incluida la compatibilidad con el protocolo Capture, Schema Registry y Kafka.

Event Hubs replica de forma transparente los datos de configuración, metadatos y eventos en tres zonas de disponibilidad de la región. La redundancia de zona proporciona conmutación automática por error sin requerir intervención de su parte. Todos los componentes de Event Hubs, incluidos el proceso, las redes y el almacenamiento, se replican entre zonas. Event Hubs tiene suficientes reservas de capacidad para controlar instantáneamente la pérdida completa de una zona. Incluso si una zona de disponibilidad completa deja de estar disponible, Event Hubs sigue funcionando sin pérdida de datos ni interrupción en las aplicaciones de streaming.



Soporte para regiones

Los espacios de nombres de Event Hubs con redundancia de zona se pueden implementar en cualquier región de Azure que admita zonas de disponibilidad.

Requisitos

- Los niveles Estándar y Premium admiten zonas de disponibilidad sin ninguna configuración adicional necesaria.
- En el nivel dedicado, las zonas de disponibilidad requieren un mínimo de tres CUs.

Cost

La redundancia de zona en Event Hubs no agrega ningún costo adicional.

Configurar soporte de zonas de disponibilidad

Los espacios de nombres de Event Hubs admiten automáticamente la redundancia de zona cuando se implementan en regiones admitidas. No es necesario realizar ninguna otra configuración.

Comportamiento cuando todas las zonas están en buen estado

Cuando los espacios de nombres de Event Hubs usan redundancia de zona y todas las zonas de disponibilidad funcionan normalmente, espere el siguiente comportamiento:

- **Enrutamiento de tráfico entre zonas:** Event Hubs funciona en un modelo activo-activo donde la infraestructura de tres zonas de disponibilidad procesa simultáneamente eventos entrantes.
- **Replicación de datos entre zonas:** Event Hubs usa la replicación sincrónica entre zonas de disponibilidad. Cuando un productor de eventos envía un evento, Event Hubs lo escribe en réplicas en varias zonas antes de confirmar al cliente que la operación de escritura se ha completado. Este enfoque garantiza una pérdida de datos cero, incluso si una zona completa deja de estar disponible. El enfoque de replicación sincrónica proporciona garantías de coherencia sólidas al tiempo que mantiene una baja latencia a través de protocolos de replicación optimizados.

Comportamiento durante un fallo de zona

Cuando los espacios de nombres de Event Hubs utilizan redundancia de zona y ocurre una interrupción en una zona de disponibilidad, se debe esperar el siguiente comportamiento:

- **Detección y respuesta:** Event Hubs es responsable de detectar automáticamente un error en una zona de disponibilidad. No es necesario iniciar una conmutación por error de zona.

- **Notificación:** Microsoft no le notifica automáticamente cuando una zona está inactiva. Sin embargo, puede usar Azure Service Health para comprender el estado general del servicio, incluidos los errores de zona, y puede configurar alertas de Service Health para notificarle problemas.
- **Solicitudes activas:** Durante un error de zona, Event Hubs podría quitar las solicitudes activas. Si los clientes controlan correctamente los errores transitorios mediante el reintento después de un breve período de tiempo, normalmente evitan un impacto significativo.
- **Pérdida de datos esperada:** No se produce ninguna pérdida de datos durante un error de zona porque Event Hubs replica de forma sincrónica eventos entre zonas antes de la confirmación.
- **Tiempo de inactividad esperado:** Un error de zona puede provocar unos segundos de tiempo de inactividad. Si los clientes controlan correctamente los errores transitorios mediante el reintento después de un breve período de tiempo, normalmente evitan un impacto significativo.
- **Reenrutamiento del tráfico:** Event Hubs detecta la pérdida de la zona y redirige automáticamente las nuevas solicitudes a otra réplica en una de las zonas de disponibilidad correctas.

  Los SDK de cliente de Event Hubs normalmente controlan la administración de conexiones y la lógica de reintento de forma transparente.

Recuperación de zona

Cuando se recupera una zona de disponibilidad, Event Hubs reintegra automáticamente la zona a la topología activa del servicio. La zona recuperada comienza a aceptar nuevas conexiones y procesar eventos junto con las demás zonas. Los datos que se han replicado en zonas supervivientes durante la interrupción permanecen intactos y la replicación sincrónica normal se reanuda en todas las zonas. No es necesario tomar medidas para la recuperación y la reintegración de zonas.

Prueba de fallos de zona

Event Hubs administra el enrutamiento del tráfico, la conmutación por error y la recuperación de zonas en caso de fallos de zona, por lo que no es necesario validar los procesos en caso de fallos de zona de disponibilidad ni proporcionar más intervención.

Resistencia a errores en toda la región

Event Hubs proporciona dos tipos de compatibilidad con varias regiones:

- La replicación geográfica (niveles Premium y Dedicado) proporciona replicación activa-activa de metadatos y datos de eventos entre una región primaria y una o varias regiones secundarias. Use la replicación geográfica para la mayoría de las aplicaciones que necesitan permanecer resistentes a las interrupciones de la región y tener una tolerancia baja para la pérdida de datos de eventos.

- La recuperación ante desastres geográfica de metadatos (nivel Estándar y superior) proporciona la replicación activa-pasiva de la configuración y los metadatos entre una región primaria y secundaria, pero no replica los datos de eventos. Use la recuperación ante desastres geográficas para las aplicaciones que pueden tolerar la pérdida de datos de eventos durante un escenario de desastre y que necesite reanudar rápidamente las operaciones en otra región.

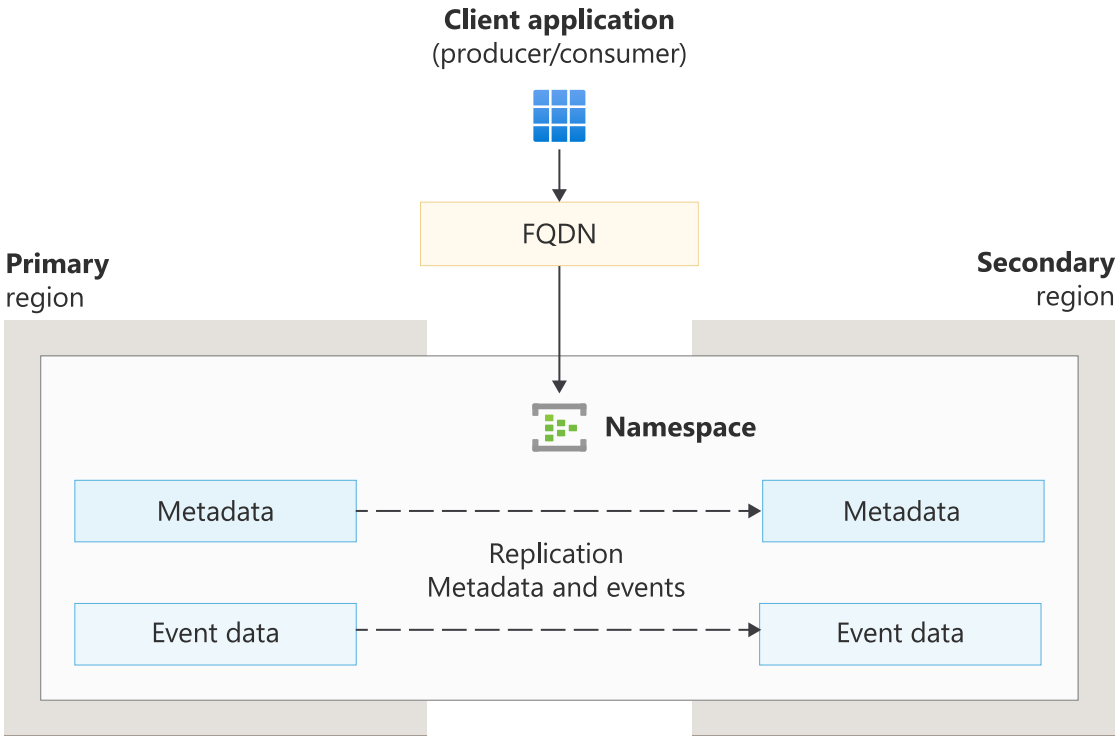
Tanto la replicación geográfica como la recuperación geográfica de desastres de metadatos requieren que usted inicie manualmente la conmutación por error o la promoción de una región secundaria para que se convierta en la nueva región primaria. Microsoft no realiza automáticamente la conmutación por error ni la promoción, incluso si la región primaria está inactiva.

Replicación geográfica

Los niveles Premium y Dedicado admiten la replicación geográfica. Esta característica replica ambos metadatos (como entidades, configuración y propiedades) y datos (como cargas de eventos) para el espacio de nombres. Tú configuras el enfoque de replicación para la configuración del espacio de nombres y los datos de eventos. Esta característica garantiza que los eventos permanecen disponibles en otra región y le permite cambiar a la región secundaria cuando sea necesario. También replica los metadatos y los datos del Registro de esquemas.

Use la replicación geográfica para escenarios que requieran resistencia a interrupciones de regiones y que tengan una tolerancia baja para la pérdida de datos de eventos.

El espacio de nombres se extiende esencialmente entre regiones. Una región actúa como principal y las demás regiones sirven como secundarias. La suscripción de Azure muestra un único espacio de nombres, independientemente de cuántas regiones secundarias configure para la replicación geográfica.



En cualquier momento, puede *promover* una región secundaria a una región primaria. Al promover una región secundaria, Event Hubs redirige el nombre de dominio completo (FQDN) del espacio de nombres a la región secundaria seleccionada y degrada la región primaria anterior a una región secundaria. Decide si desea realizar una *promoción planeada*, lo que significa que espera a que se complete la replicación de datos o una *promoción forzada*, lo que podría provocar la pérdida de datos.

> **Nota:**
>
> La replicación geográfica de Event Hubs usa el término *promoción* porque representa mejor el proceso de promoción de una región secundaria a una región primaria (y, posteriormente, degradación de una región primaria a una región secundaria). También podría ver el término *conmutación por error* utilizado para describir el proceso general.

En esta sección se resumen aspectos importantes de la replicación geográfica. Revise la documentación completa para comprender exactamente cómo funciona. Para más información, consulte Replicación geográfica de Event Hubs.

Soporte para regiones

Puede elegir cualquier región de Azure que admita Event Hubs como región primaria o regiones secundarias. No es necesario usar regiones emparejadas de Azure, por lo que puede elegir regiones secundarias en función de los requisitos de latencia, cumplimiento o residencia de datos.

Requisitos

Para habilitar la replicación geográfica, el espacio de nombres debe usar el nivel Premium o Dedicado.

Consideraciones

Al habilitar la replicación geográfica, tenga en cuenta los siguientes factores:

- **Formato de punto de comprobación:** El formato de los puntos de control cambia. Para más información, consulte Replicación geográfica: Consumo de datos.
- **Puntos de conexión privados:** Si usa puntos de conexión privados para conectarse al espacio de nombres, también debe configurar las redes en las regiones primarias y secundarias. Para obtener más información, consulte Puntos de conexión privados.

Cost

Para comprender cómo funcionan los precios para la replicación geográfica, consulte Precios.

Configuración de la compatibilidad con varias regiones

- **Habilite la replicación geográfica en un espacio de nombres nuevo o existente.** Para configurar la replicación activa-activa para un espacio de nombres recién creado, consulte Habilitación de la replicación geográfica en un nuevo espacio de nombres. Para configurar la replicación activa-activa en un espacio de nombres existente, consulte Habilitación de la replicación geográfica en un espacio de nombres existente.
- **Cambie el enfoque de replicación.** Para cambiar entre modos de replicación sincrónicos y asincrónicos, consulte Cambiar modo de replicación.
- **Deshabilitar la replicación geográfica.** Para deshabilitar la replicación geográfica en una región secundaria, consulte Eliminación de una región secundaria.

Comportamiento cuando todas las regiones están en buen estado

En esta sección se describe qué esperar cuando se configura un espacio de nombres de Event Hubs para la replicación geográfica y la región primaria está operativa.

- **Enrutamiento del tráfico entre regiones:** Las aplicaciones cliente se conectan a través del FQDN de su espacio de nombres, y su tráfico se dirige a la región primaria.

Solo la región primaria procesa activamente los eventos de los clientes durante las operaciones normales. La región secundaria recibe eventos replicados, pero de lo contrario permanece pasiva en modo de espera.

- **Replicación de datos entre regiones:** El comportamiento de replicación de datos entre las regiones principal y secundaria depende de si configura el emparejamiento de replicación para usar la replicación sincrónica o asincrónica.

  - *Síncrono:* Los eventos se replican en la región secundaria antes de que se complete la operación de escritura.

    Este modo proporciona la mayor seguridad de que los datos del evento están seguros, ya que deben confirmarse tanto en la región primaria como en la secundaria. Sin embargo, la replicación sincrónica aumenta considerablemente la latencia de escritura para los eventos entrantes. También requiere que la región secundaria esté disponible para aceptar la operación de escritura, por lo que una interrupción en cualquier región secundaria hace que se produzca un error en la operación de escritura.
    - *Asíncrono:* Los eventos se escriben en la región primaria y, a continuación, se completa la operación de escritura. Poco tiempo después, replica los eventos en la región secundaria.

    Este modo proporciona un mayor rendimiento de escritura que la replicación sincrónica porque no hay ninguna latencia de replicación entre regiones durante las operaciones de escritura. Además, el modo de replicación asincrónica puede tolerar la pérdida de una región secundaria, a la vez que permite operaciones de escritura en la región primaria. Sin embargo, si la región primaria tiene una interrupción, es posible que los datos que aún no se hayan replicado en la región secundaria no estén disponibles o se pierdan.

    Al configurar la replicación asincrónica, establezca el retardo máximo aceptable que puede tener la replicación. En cualquier momento, puede comprobar el retraso de replicación actual mediante métricas de Azure Monitor.

    Si el retraso de replicación asincrónico aumenta más allá del máximo especificado, la región primaria comienza a limitar las solicitudes entrantes para que la replicación pueda alcanzar el nivel deseado. Para evitar esta situación, es importante seleccionar regiones secundarias que no estén demasiado distantes geográficamente y asegurarse de que la capacidad sea suficiente para el rendimiento.

    Para obtener más información, consulte Modos de replicación.

Comportamiento durante una falla de región

En esta sección se describe qué esperar cuando se configura un espacio de nombres de Event Hubs para la replicación geográfica y se produce una interrupción en la región principal o secundaria.

- **Detección y respuesta:** Usted es responsable de decidir cuándo promover la región secundaria del espacio de nombres para convertirse en la nueva región primaria. Microsoft no toma esta decisión ni inicia el proceso por ti, incluso si hay una interrupción de la región. Para obtener más información sobre cómo promover una región secundaria a la nueva principal, consulte Promover secundaria.

  Al promover una región secundaria, elija si desea realizar una *promoción planeada* o una *promoción forzada*. Una promoción planeada espera a que la región secundaria se ponga al día antes de aceptar nuevo tráfico. Este enfoque elimina la pérdida de datos, pero presenta tiempo de inactividad.

  Durante una interrupción en la región principal, típicamente usted necesita realizar una promoción forzada. Si la región primaria está disponible y activa una promoción por otro motivo específico, podría elegir una promoción planeada.

- **Notificación:** Microsoft no le notifica automáticamente cuando una región está inactiva. Sin embargo, puede usar Azure Service Health para comprender el estado general del servicio, incluidos los errores de región, y puede configurar alertas de Service Health para notificarle problemas.

  Use esa información y otras métricas para decidir cuándo promover una región secundaria a una región primaria.

- **Solicitudes activas:** El comportamiento depende de si la interrupción de la región se produce en la región primaria o en una región secundaria:

  - *Interrupción de la región primaria:* Si la región primaria no está disponible, se finalizan todas las solicitudes activas. Las aplicaciones cliente deben reintentar las operaciones una vez completada la promoción.

  - *Interrupción de la región secundaria:* Una interrupción en la región secundaria puede provocar problemas para las solicitudes activas en las situaciones siguientes:

    - Si usa el modo de replicación sincrónica, la región primaria no puede completar las operaciones de escritura si alguna región secundaria no está disponible.

    - Si usa el modo de replicación asincrónica, el espacio de nombres limita su capacidad y no acepta nuevos eventos después de que el retraso de la replicación alcance el máximo que ha configurado.

Para seguir usando el espacio de nombres en la región primaria, quite el espacio de nombres secundario de la configuración de replicación geográfica.

- **Pérdida de datos esperada:** La cantidad de pérdida de datos depende del tipo de promoción que realice (planeado o forzado) y del modo de replicación (sincrónico o asincrónico):

  - *Promoción planeada:* No se espera ninguna pérdida de datos. Sin embargo, durante una interrupción de la región, es posible que una promoción planeada no sea posible porque requiere que todas las regiones primarias y secundarias estén disponibles.

  - *Promoción forzada, replicación sincrónica:* No se espera ninguna pérdida de datos.

  - *Promoción forzada, replicación asincrónica:* Es posible que experimente alguna pérdida de datos para eventos recientes que no se replican en la región secundaria. La cantidad depende del retraso de replicación. Para comprobar el retraso de replicación actual, use las métricas de Azure Monitor.

  Si realiza una promoción forzada, no puede recuperar los datos perdidos, incluso después de que la región primaria esté disponible.

- **Tiempo de inactividad esperado:** La cantidad de tiempo de inactividad esperado depende de si realiza una promoción planeada o forzada:

  - *Promoción planeada:* El primer paso de una promoción planeada replica los datos en la región secundaria. Ese proceso suele completarse rápidamente, pero en algunas situaciones, puede tardar hasta la longitud del retraso de replicación. Una vez completada la replicación, el proceso de promoción suele tardar entre 5 y 10 minutos. A veces, los servidores del sistema de nombres de dominio (DNS) pueden tardar más tiempo en actualizar las entradas y replicar completamente sus registros en los clientes.

    La región primaria no admite operaciones de escritura durante el proceso completo de promoción.

    Es posible que esta opción no sea posible durante una interrupción de la región porque requiere que todas las regiones primarias y secundarias estén disponibles.

  - *Promoción forzada:* Durante una promoción forzada, Event Hubs no espera a que se complete la replicación de datos e inicia la promoción inmediatamente. El proceso de promoción suele tardar entre 5 y 10 minutos. A veces, las entradas DNS pueden tardar más tiempo en replicarse y actualizarse por completo en todos los clientes.

    La región primaria no admite operaciones de escritura durante el proceso completo de promoción.

- **Reenrutamiento del tráfico:** Una vez completada la promoción, el FQDN del espacio de nombres apunta a la nueva región primaria. Pero este redireccionamiento depende de la rapidez con la que se actualizan los registros DNS de los clientes, incluido para que sus servidores DNS respeten el período de vida (TTL) de los registros DNS del espacio de nombres.

  En algunas situaciones, debe configurar las aplicaciones de consumo para que se comporten de forma coherente después de que se produzca la promoción de la región. Para más información, consulte Replicación geográfica: Consumo de datos.

Recuperación de regiones

Después de que la región primaria original se recupere, si desea devolver el namespace a su región primaria original, siga el mismo proceso de promoción de la región.

Si realizó una promoción forzada durante la interrupción de la región, no podrá recuperar los datos perdidos, incluso después de que la región primaria esté disponible.

Prueba de fallos de región

Para probar la replicación geográfica, promueva temporalmente la región secundaria a principal y compruebe que las aplicaciones cliente pueden cambiar entre regiones con una interrupción mínima.

Supervise la duración de la promoción y compruebe que los runbooks y la automatización funcionan correctamente. Después de realizar las pruebas, puede revertir a la configuración original.

Comprenda el posible tiempo de inactividad y pérdida de datos que puede experimentar durante y después del proceso de promoción. Pruebe la replicación geográfica en un entorno que no sea de producción que refleje la configuración del espacio de nombres de producción.

Recuperación ante desastres geográfica de metadatos

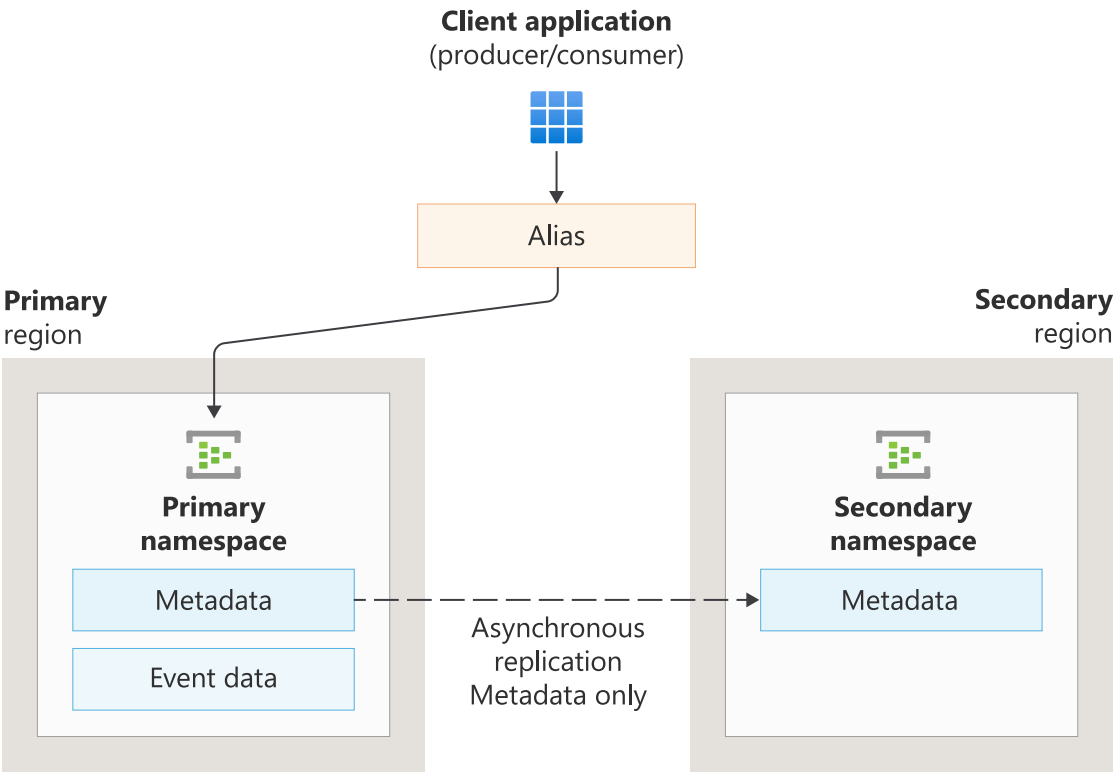
El nivel Estándar y versiones posteriores admiten la recuperación ante desastres geográfica de metadatos. Esta característica mejora la recuperación de escenarios de desastres, incluida la pérdida grave de una región. La recuperación ante desastres geográfica solo replica la configuración y los metadatos del espacio de nombres. Sin embargo, no replica los datos de eventos. Para admitir la recuperación ante desastres, esta característica garantiza que un espacio de nombres de otra región esté preconfigurado y listo para aceptar inmediatamente eventos de los clientes. La recuperación ante desastres geográfica sirve como una solución de

recuperación unidireccional y no admite la conmutación por recuperación en la región primaria anterior.

La recuperación ante desastres geográfica de metadatos funciona mejor para las aplicaciones que no necesitan mantener todos los eventos y pueden tolerar cierta pérdida de datos durante un escenario de desastre. Por ejemplo, si sus eventos representan lecturas de sensores que luego usted agrega, podría decidir que puede permitirse perder algunos eventos de una región fallida si puede reanudar rápidamente el procesamiento de nuevos eventos en otra región.

> **Importante**
>
> La recuperación ante desastres geográfica permite la continuidad de las operaciones que tienen la misma configuración, pero **no replica los datos de eventos**. Si necesita replicar datos de eventos, considere la posibilidad de usar la replicación geográfica.

Al configurar la recuperación ante desastres geográfica de metadatos, se crea un *alias* al que se conectan las aplicaciones cliente. El alias es un FQDN que dirige todo el tráfico al espacio de nombres principal de forma predeterminada.



Si se produce un error en la región primaria u otro tipo de desastre, puede iniciar manualmente una conmutación por error unidireccional desde la región primaria a la región secundaria en cualquier momento. La conmutación por error se completa casi al instante. Durante el proceso de conmutación por error, el alias de recuperación ante desastres geográficos se redirige al espacio de nombres secundario.

En esta sección se resumen aspectos importantes de la recuperación ante desastres geográfica. Revise la documentación completa para comprender exactamente cómo funciona. Para más información, consulte Recuperación ante desastres geográfica de Event Hubs.

Soporte para regiones

Puede seleccionar cualquier región de Azure que admita Event Hubs como espacio de nombres principal o secundario. No es necesario usar regiones emparejadas de Azure, por lo que puede elegir regiones secundarias en función de los requisitos de latencia, cumplimiento o residencia de datos.

Requisitos

- **Nivel de espacio de nombres principal:** El espacio de nombres principal debe estar en el nivel Estándar o superior para usar la recuperación ante desastres geográfica de metadatos.
- **Nivel de espacio de nombres secundario:** La recuperación ante desastres geográfica de metadatos admite combinaciones específicas de niveles para los espacios de nombres principal y secundario. Para obtener más información, vea Pares de espacios de nombres admitidos.

Consideraciones

- **Asignaciones de roles:** las asignaciones de control de acceso basado en roles (RBAC) de Microsoft Entra a entidades en el espacio de nombres principal no se replican en el espacio de nombres secundario. Cree asignaciones de roles manualmente en el espacio de nombres secundario para proteger el acceso a estas.
- **Registro de Esquema:** Los metadatos del Registro de Esquema se replican al usar la recuperación ante desastres geográfica de metadatos, pero los esquemas inscritos en el registro de esquema no se replican.
- **Diseño de aplicaciones:** La recuperación ante desastres geográfica requiere consideraciones específicas al diseñar las aplicaciones cliente. Para más información, consulte Consideraciones.
- **Puntos de conexión privados:** Si usa puntos de conexión privados para conectarse al espacio de nombres, configure las redes en la región primaria y secundaria. Para obtener más información, consulte Puntos de conexión privados.

Cost

Al habilitar la recuperación ante desastres geográficos de metadatos, usted paga por los espacios de nombres principal y secundario.

Configuración de la compatibilidad con varias regiones

- **Cree un emparejamiento de recuperación ante desastres geográficos para metadatos.** Para configurar la recuperación ante desastres entre los espacios de nombres principal y secundario, consulte Configuración y proceso de conmutación por error.

- **Deshabilite la recuperación ante desastres geográfica de metadatos.** Para interrumpir el emparejamiento entre espacios de nombres, consulte Configuración y flujo de conmutación por error.

Planeamiento y administración de capacidad

Cuando planee implementaciones de varias regiones, asegúrese de que ambas regiones tienen capacidad suficiente para controlar la carga completa si se produce un error en una región. La región secundaria permanece pasiva durante las operaciones normales, pero debe controlar inmediatamente el tráfico después de la conmutación por error. Planee cómo escalar la capacidad del espacio de nombres secundario para que pueda recibir tráfico de producción sin demora. Si puede tolerar un tiempo de inactividad adicional durante el proceso de conmutación por error, puede optar por escalar la capacidad del espacio de nombres secundario durante o después de la conmutación por error. Para reducir el tiempo de inactividad, aprovisione la capacidad con antelación en el espacio de nombres secundario para que permanezca lista para recibir la carga de producción.

Comportamiento cuando todas las regiones están en buen estado

En esta sección se describe qué esperar cuando se configura un espacio de nombres de Event Hubs para la recuperación ante desastres geográfica y la región primaria está operativa.

- Enrutamiento del tráfico entre regiones: Las aplicaciones cliente se conectan a través del alias de recuperación ante desastres geográfico de su espacio de nombres, y su tráfico se enruta al espacio de nombres principal en la región primaria.

  Solo el espacio de nombres principal procesa activamente eventos de los clientes durante las operaciones normales. El espacio de nombres secundario permanece pasivo en modo de espera y se produce un error en las solicitudes de acceso a los datos.

- **Replicación de datos entre regiones:** Solo los metadatos de configuración se replican entre los espacios de nombres. La replicación de la configuración se produce de forma continua y asincrónica.

  Todos los datos de eventos permanecen solo en el espacio de nombres principal y no se replican en el espacio de nombres secundario.

Comportamiento durante una falla de región

En esta sección se describe qué esperar cuando se configura un espacio de nombres de Event Hubs para la recuperación ante desastres geográficos y se produce una interrupción en la región principal.

- **Detección y respuesta:** usted es responsable de monitorear la salud de la región e iniciar la conmutación por error manualmente. Microsoft no realiza la conmutación por error ni promueve automáticamente una región secundaria, incluso si la región primaria está inactiva.

  Para obtener más información sobre cómo iniciar la conmutación por error, consulte Conmutación por error manual.

  La conmutación por error es una operación unidireccional, por lo que deberá restablecer la pareja de recuperación ante desastres geográficos más adelante. Para obtener más información, consulte Recuperación de regiones.

- **Notificación:** Microsoft no le notifica automáticamente cuando una región está inactiva. Sin embargo, puede usar Azure Service Health para comprender el estado general del servicio, incluidos los errores de región, y puede configurar alertas de Service Health para notificarle problemas.

  Use esa información y otras métricas para decidir cuándo conmutar a la región secundaria.

- **Solicitudes activas:** Las solicitudes activas terminan cuando se inicia la conmutación por error. Las aplicaciones cliente deben reintentar las operaciones una vez completada la conmutación por error.

- **Pérdida de datos esperada:**

  - *Metadatos:* Normalmente, la configuración y los metadatos se replican en el espacio de nombres secundario. Pero la replicación de metadatos se produce de forma asincrónica, por lo que es posible que los cambios recientes no se repliquen, especialmente los cambios complejos. Compruebe la configuración del espacio de nombres secundario antes de que los clientes accedan a él.

- *Datos de eventos:* Los datos del evento no se replican entre regiones. Si la región primaria deja de funcionar, los eventos del espacio de nombres principal no estarán disponibles.

  Los eventos no se pierden permanentemente a menos que un desastre catastrófico cause una pérdida total de la región primaria. Si la región se recupera, puede recuperar eventos del espacio de nombres principal más adelante.

- **Tiempo de inactividad esperado:** Normalmente, la conmutación ocurre en un plazo de 5 a 10 minutos. Los clientes pueden tardar más tiempo en replicar y actualizar completamente las entradas DNS.

- **Reenrutamiento del tráfico:** Los clientes que usan el alias de recuperación ante desastres geográfica para conectarse al espacio de nombres se redirigen automáticamente al espacio de nombres secundario después de la conmutación por error. Pero este redireccionamiento depende de los servidores DNS que respetan el TTL de los registros DNS del espacio de nombres y los clientes que reciben esos registros DNS actualizados.

Recuperación de regiones

Una vez recuperada la región primaria original, debe reestablecer manualmente el emparejamiento y, opcionalmente, realizar una conmutación de retorno. Cree un nuevo emparejamiento de recuperación ante desastres geográfica con la región recuperada como secundaria y, a continuación, realice otra conmutación por error si desea volver a la región original. Este proceso implica la posible pérdida de datos de eventos enviados al servidor principal temporal.

Si el desastre provoca la pérdida de todas las zonas de la región primaria, los datos podrían ser irrecuperables. En otros escenarios, los datos del evento permanecen en el espacio de nombres principal de antes de que se pueda recuperar la conmutación por error. Puede obtener eventos históricos del espacio de nombres principal anterior después de restaurar el acceso. Usted es responsable de configurar las aplicaciones para recibir y procesar esos eventos. Microsoft no los restaura automáticamente en tu región secundaria.

Prueba de fallos de región

Para probar los procesos de respuesta y recuperación ante desastres, realice una conmutación por error planeada durante una ventana de mantenimiento. Inicie la conmutación por error desde el espacio de nombres principal al espacio de nombres secundario y compruebe que las aplicaciones pueden conectarse y procesar eventos desde el nuevo espacio de nombres principal.

Supervise la duración de la conmutación por error y compruebe que los runbooks y la automatización funcionan correctamente. Después de realizar las pruebas, puede revertir a la configuración original.

Comprenda el posible tiempo de inactividad y la pérdida de datos que puede experimentar durante y después del proceso de conmutación por error. Pruebe la replicación geográfica en un entorno que no sea de producción que refleje la configuración del espacio de nombres de producción.

Soluciones personalizadas de varias regiones para la resistencia

La replicación geográfica y la recuperación ante desastres geográficos de metadatos proporcionan resiliencia frente a interrupciones regionales y otros problemas, y son compatibles con la mayoría de las cargas de trabajo. Algunos niveles de Event Hubs no admiten estas características, o es posible que necesite la replicación personalizada o necesite mantener varias regiones activas simultáneamente.

Varios patrones de diseño pueden lograr diferentes tipos de compatibilidad con varias regiones en Event Hubs. Muchos de los patrones requieren la implementación de varios espacios de nombres y el uso de servicios como Azure Functions para replicar eventos entre ellos. Para obtener más información, consulte Federación multisitio y de varias regiones.

Copias de seguridad y restauración

Event Hubs no está diseñado como una ubicación de almacenamiento a largo plazo para los datos. Normalmente, los datos se almacenan en un centro de eventos durante un breve período de tiempo y, a continuación, se procesan o se conservan en otro sistema de almacenamiento de datos. Puede configurar el período de retención de datos para su centro de eventos según sus requisitos y el nivel que utiliza su espacio de nombres. Para más información, consulte Retención de eventos.

Si necesita conservar una copia de los eventos, considere la posibilidad de usar Event Hubs Capture, que guarda copias de eventos en una cuenta de Azure Blob Storage.

Acuerdo de nivel de servicio

El contrato de nivel de servicio (SLA) para los servicios de Azure describe la disponibilidad esperada de cada servicio y las condiciones que la solución deberá cumplir para lograr esa expectativa de disponibilidad. Para obtener más información, consulte Acuerdos de Nivel de Servicio para servicios en línea.

El SLA de disponibilidad del namespace es superior cuando se utilizan los niveles Premium o Dedicado.

Contenido relacionado

- Documentación de Event Hubs
- Introducción a la confiabilidad de Azure

Last updated on 22/01/2026

Confiabilidad en Azure HDInsight

En este artículo se describe la compatibilidad con la confiabilidad en Azure HDInsighty se tratan zonas de disponibilidad y recuperación entre regiones y continuidad empresarial. Para obtener información general más detallada sobre la confiabilidad de Azure, consulte Confiabilidad de Azure.

Soporte para zonas de disponibilidad

Las zonas de disponibilidad son grupos físicamente independientes de centros de datos dentro de una región de Azure. Cuando una zona falla, los servicios pueden transferirse a una de las zonas restantes.

Azure HDInsight admite una configuración de implementación zonal. Los nodos de clúster de Azure HDInsight se colocan en una sola zona que se selecciona en la región seleccionada. Un clúster de HDInsight zonal está aislado de las interrupciones que se producen en otras zonas. Sin embargo, si una interrupción afecta a la zona específica elegida para el clúster de HDInsight, el clúster no estará disponible. Este modelo de implementación proporciona conectividad de red de baja latencia y económica dentro del clúster. La replicación de este modelo de implementación en varias zonas de disponibilidad puede proporcionar un mayor nivel de disponibilidad para protegerse frente a errores de hardware.

> **ⓘ Importante**
>
> En el caso de las implementaciones en las que los usuarios no especifican una zona específica, los tipos de nodo no son resistentes a la zona y pueden experimentar tiempo de inactividad durante una interrupción en cualquier zona de esa región.

Prerequisites

- Las zonas de disponibilidad solo se admiten para los clústeres creados después del 15 de junio de 2023. La configuración de la zona de disponibilidad no se puede actualizar después de crear el clúster. Tampoco puede actualizar un clúster de zona sin disponibilidad para usar zonas de disponibilidad.
- Los clústeres se deben crear en una red virtual personalizada.

- Debe traer su propia base de datos SQL para la base de datos de Ambari y el metastore externo, como metastore de Hive, para que pueda configurar estas bases de datos en la misma zona de disponibilidad.

- Los clústeres de HDInsight deben crearse con la opción de zona de disponibilidad en una de las siguientes regiones:
  - Australia East
  - Sur de Brasil
  - Canada Central
  - Central US
  - East US
  - Este de EE. UU. 2
  - France Central
  - Centro-oeste de Alemania
  - Japan East
  - Korea Central
  - Norte de Europa
  - Centro de Qatar
  - Sudeste asiático
  - Centro-sur de EE. UU.
  - UK South
  - Gobierno de EE. UU. - Virginia
  - West Europe
  - Oeste de EE. UU. 2

Creación de un clúster de HDInsight con zona de disponibilidad

Puede usar la plantilla de Azure Resource Manager (ARM) para iniciar un clúster de HDInsight en una zona de disponibilidad especificada.

En la sección de recursos, debe agregar una sección de "zonas" y proporcionar en qué zona de disponibilidad desea que se implemente este clúster.

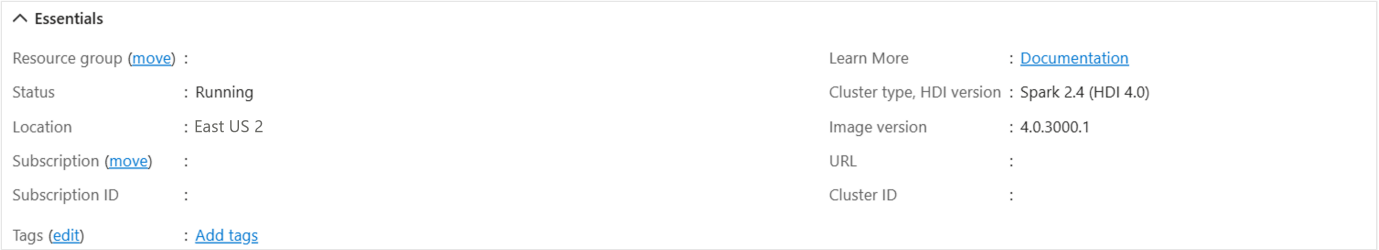
JSON

```
"resources": [
    {
        "type": "Microsoft.HDInsight/clusters",
        "apiVersion": "2021-06-01",
```

```
            "name": "[parameters('cluster name')]",
            "location": "East US 2",
            "zones": [
                "1"
            ],
        }
    ]
```

Comprobación de nodos dentro de una zona de disponibilidad entre zonas

Cuando el clúster de HDInsight esté listo, puede comprobar la ubicación para ver en qué zona de disponibilidad se implementan.



Obtener la respuesta de la API:

JSON

```
[
        {
            "location": "East US 2",
            "zones": [
                "1"
            ],
        }
]
```

Escalado vertical del clúster

Puede escalar verticalmente un clúster de HDInsight con más nodos de trabajo. Los nodos de trabajo recién agregados se colocarán en la misma zona de disponibilidad de este clúster.

Migración de zonas de disponibilidad

Actualmente, los clústeres de Azure HDInsight no admiten la migración local de instancias de clúster existentes a compatibilidad con zonas de disponibilidad. Sin embargo, puede elegir volver a crear el clústery elegir una región o zona de disponibilidad diferente durante la creación del

clúster. Un clúster en espera secundario en una región diferente y una zona de disponibilidad diferente se puede usar en escenarios de recuperación ante desastres.

Experiencia de reducción de zona

Cuando una zona de disponibilidad deja de funcionar:

- No se puede usar ssh en este clúster.
- No se puede eliminar, escalar verticalmente o reducir verticalmente este clúster.
- No se pueden enviar trabajos ni ver el historial de trabajos.
- Todavía puede enviar una nueva solicitud de creación de clústeres en otra región.

Recuperación ante desastres entre regiones y continuidad empresarial

La recuperación ante desastres (DR) hace referencia a las prácticas que las organizaciones usan para recuperarse de eventos de alto impacto, como desastres naturales o implementaciones con errores que producen tiempo de inactividad y pérdida de datos. Independientemente de la causa, el mejor remedio para un desastre es un plan de recuperación ante desastres bien definido y probado y un diseño de aplicaciones que apoye activamente la recuperación ante desastres. Antes de empezar a crear el plan de recuperación ante desastres, consulte Recomendaciones para diseñar una estrategia de recuperación ante desastres.

Para la recuperación ante desastres, Microsoft usa el modelo de responsabilidad compartida. En este modelo, Microsoft garantiza que la infraestructura de línea base y los servicios de plataforma estén disponibles. Sin embargo, muchos servicios de Azure no replican automáticamente datos ni se revierten de una región con errores para realizar la replicación cruzada en otra región habilitada. En esos servicios, es responsable de configurar un plan de recuperación ante desastres válido para su carga de trabajo. La mayoría de los servicios que se ejecutan en ofertas de plataforma como servicio (PaaS) de Azure proporcionan características e instrucciones para admitir la DR. Puede usar características específicas del servicio para admitir la recuperación rápida con el fin de contribuir al desarrollo del plan de recuperación ante desastres.

Los clústeres de Azure HDInsight dependen de muchos servicios de Azure, como almacenamiento, bases de datos, Active Directory, Active Directory Domain Services, redes y Key Vault. Una aplicación de análisis bien diseñada, de alta disponibilidad y tolerante a errores debe diseñarse con suficiente redundancia para resistir las interrupciones regionales o locales en uno o varios de estos servicios. En esta sección se proporciona información general sobre los

procedimientos recomendados, la disponibilidad de una sola región y varias regiones, y las opciones de optimización para el planeamiento de la continuidad empresarial.

Recuperación ante desastres en la geografía de varias regiones

La mejora de la continuidad empresarial mediante la recuperación ante desastres de alta disponibilidad entre regiones requiere diseños arquitectónicos de mayor complejidad y un costo más alto. En las siguientes tablas se detallan algunas áreas técnicas que pueden aumentar el costo total de propiedad.

Optimizaciones de costos

Expandir tabla

Área	Causa de la escalación de costo	Estrategias de optimización
Almacenamiento de datos	Duplicación de tablas o datos principales en una región secundaria.	Replique solo los datos mantenidos.
Salida de datos	Las transferencias de datos entre regiones salientes tienen un precio. Revise las directrices de precios de ancho de banda.	Replique solo los datos mantenidos para reducir la superficie de salida de la región.
Cálculo del clúster	Clústeres de HDInsight adicionales en la región secundaria.	Use scripts automatizados para implementar un proceso secundario después del error principal. Use el escalado automático para mantener el tamaño del clúster secundario al mínimo. Use SKU de máquinas virtuales más baratas. Cree los clústeres secundarios en las regiones donde se pueden descontar las SKU de máquinas virtuales.
Autenticación	Los escenarios multiusuario de la región secundaria incurren en configuraciones adicionales de Microsoft Entra Domain Services	Evite las configuraciones multiusuario en la región secundaria.

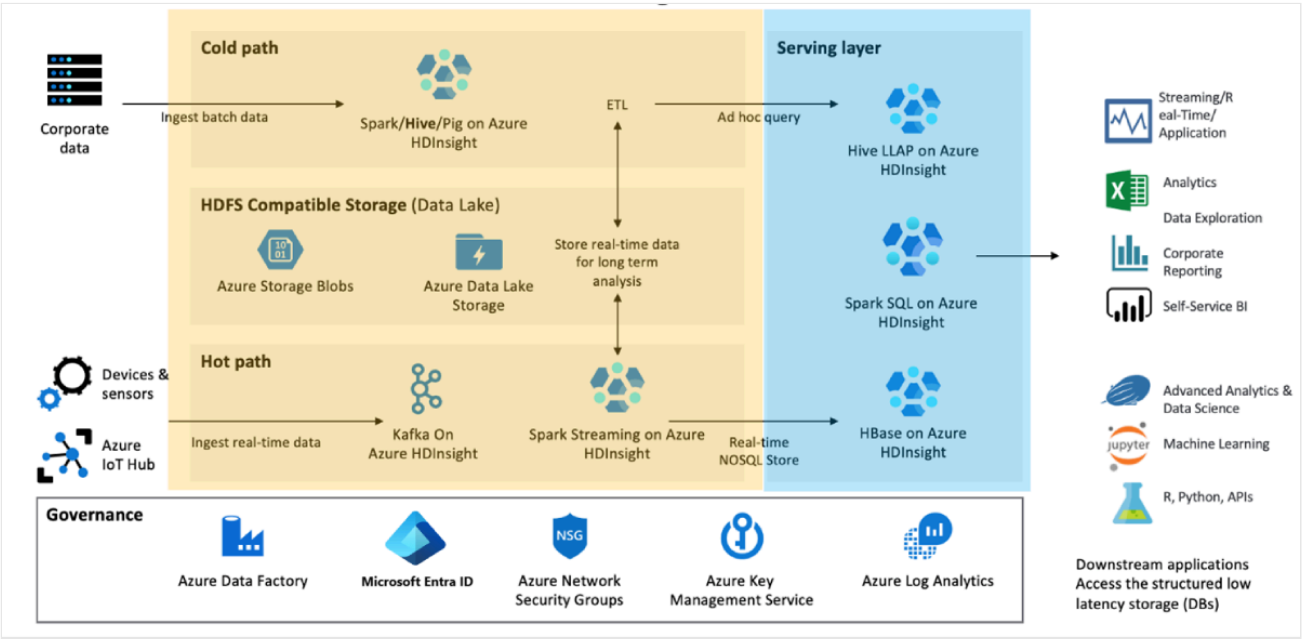
Optimizaciones de complejidad

Expandir tabla

Área	Causa de la escalación de complejidad	Estrategias de optimización
Patrones de lectura y escritura	Requerir que los elementos primarios y secundarios estén habilitados para lectura y escritura.	Diseñe el elemento secundario para que sea de solo lectura.
RTO y RPO cero	Requerir pérdida de datos cero (RPO = 0) y tiempo de inactividad cero (RTO = 0).	Diseñe RPO y RTO de maneras para reducir el número de componentes que necesitan conmutar por error. Para obtener más información sobre RTO y RPO, consulte Azure HDInsight arquitecturas de continuidad empresarial.
Funciones empresariales	Requerir toda la funcionalidad empresarial de la primaria en la secundaria.	Evalúe si puede realizar la ejecución con un subconjunto crítico mínimo básico de la funcionalidad empresarial en la secundaria.
Connectivity	Requerir que todos los sistemas ascendentes y descendentes de la conectividad primaria se conecten también a la secundaria.	Limite la conectividad secundaria a un subconjunto crítico mínimo básico.

Cuando el plan de recuperación ante desastres de varias regiones, tenga en cuenta las siguientes recomendaciones:

- Determine la funcionalidad de negocio mínima que necesitará si se produce un desastre y por qué. Por ejemplo, evalúe si necesita funcionalidades de conmutación por error para la capa de transformación de datos (que se muestra en amarillo) *y* la capa de servicio de datos (que se muestra en azul) o si solo necesita conmutación por error para la capa de servicio de datos.



- Segmente los clústeres en función de la carga de trabajo, el ciclo de vida de desarrollo y los departamentos. El hecho de tener más clústeres reduce las posibilidades de que un único error grande afecte a varios procesos empresariales diferentes.
- Haga que sus regiones secundarias sean de solo lectura. Las regiones de conmutación por error con capacidades de lectura y escritura pueden conducir a arquitecturas complejas.
- Los clústeres transitorios son más fáciles de administrar cuando se produce un desastre. Diseñe sus cargas de trabajo de forma que los clústeres se puedan reciclar y no se mantenga ningún estado en ellos.
- A menudo, las cargas de trabajo se dejan sin terminar si se produce un desastre y es necesario reiniciar en la nueva región. Diseñe las cargas de trabajo para que sean idempotentes por naturaleza.
- Use la automatización durante las implementaciones de clústeres y asegúrese de que los valores de configuración del clúster se incluyen en el script en la medida de lo posible para garantizar una implementación rápida y totalmente automatizada si se produce un desastre.

Detección, notificación y administración de interrupciones

- Use las herramientas de supervisión de Azure en HDInsight para detectar un comportamiento anómalo en el clúster y establecer las notificaciones de alerta correspondientes. Puede implementar las soluciones de administración específicas del clúster de HDInsight preconfiguradas que recopilan las métricas de rendimiento importantes del tipo de clúster específico. Para más información, consulte Supervisión de Azure para HDInsight.
- Suscríbase a las alertas de estado de Azure para recibir notificaciones sobre problemas de servicio, mantenimiento planeado, avisos de estado y seguridad para una suscripción, un servicio o una región. Las notificaciones de estado que incluyen la causa del problema y la hora de llegada estimada resuelta le ayudan a ejecutar mejor la conmutación por error y la conmutación por recuperación. Para obtener más información, consulte Documentación de Azure Service Health.

Recuperación ante desastres en una sola región geográfica

Cada componente de un sistema HDInsight básico tiene sus propios mecanismos de tolerancia a errores de una sola región. Tenga en cuenta que no siempre toma un evento catastrófico para afectar a la funcionalidad empresarial. Los incidentes de servicio en uno o varios de los servicios

siguientes en una sola región también pueden provocar la pérdida de la funcionalidad empresarial esperada.

- **Proceso (máquinas virtuales): clúster de Azure HDInsight**. HDInsight ofrece un Acuerdo de Nivel de Servicio de disponibilidad del 99,9 %. Para proporcionar alta disponibilidad en una sola implementación, HDInsight viene acompañado de muchos servicios que están en modo de alta disponibilidad de manera predeterminada. Los mecanismos de tolerancia a errores de HDInsight los proporcionan los servicios de alta disponibilidad de los ecosistemas de Microsoft y Apache OSS.

  Los siguientes componentes de infraestructura están diseñados para ser de alta disponibilidad:
  - Nodos principales activos y en espera
  - Varios nodos de puerta de enlace
  - Tres nodos de Quorum de Zookeeper
  - Nodos de trabajo distribuidos por dominios de actualización y de error

  Los siguientes servicios también están diseñados para ser de alta disponibilidad:
  - Servidor Apache Ambari
  - Servidores de escala de tiempo de la aplicación para YARN
  - Servidor de Historial de trabajos para MapReduce de Hadoop
  - Apache Livy
  - HDFS
  - Administrador de Recursos YARN
  - HBase Master

  Para más información, consulte servicios de alta disponibilidad admitidos con Azure HDInsight.

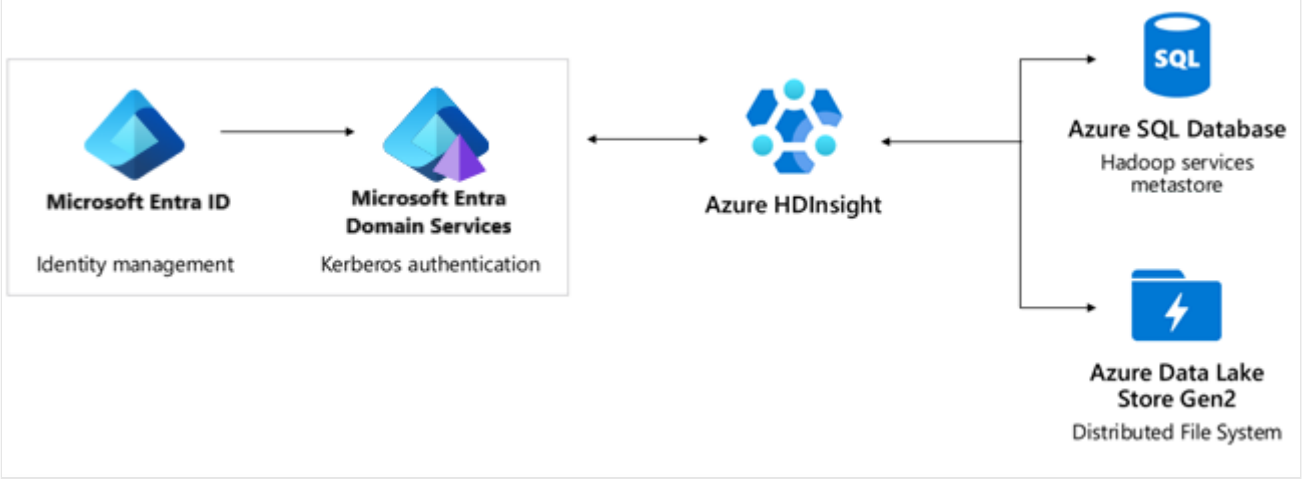
- **Metastores: Azure SQL Database**. HDInsight usa Azure SQL Database como metastore, lo que proporciona un Acuerdo de Nivel de Servicio del 99,99 %. Se conservan tres réplicas de datos dentro de un centro de datos con replicación sincrónica. Si se produce una pérdida de réplica, se proporciona una réplica alternativa sin problemas. Replicación geográfica activa es compatible de forma integrada con un máximo de cuatro centros de datos. Cuando hay una conmutación por error, ya sea manual o del centro de datos, la primera réplica de la jerarquía se convierte automáticamente en compatible con lectura y escritura. Para más información, consulte Continuidad empresarial de Azure SQL Database.

- **Storage: Azure Data Lake Gen2 o Blob Storage**. HDInsight recomienda Azure Data Lake Storage Gen2 como capa de almacenamiento subyacente. Azure Storage, incluido Azure

Data Lake Storage Gen2, proporciona un Acuerdo de Nivel de Servicio del 99,9 %. HDInsight usa el servicio LRS en el que se conservan tres réplicas de datos dentro de un centro de datos y la replicación es sincrónica. Si se produce una pérdida de réplica, se proporciona una réplica sin problemas.

- **Autenticación: Microsoft Entra ID, Microsoft Entra Domain Services, Enterprise Security Package**.
  - Microsoft Entra ID proporciona un Acuerdo de Nivel de Servicio del 99,9 %. Active Directory es un servicio global con varios niveles de redundancia interna y capacidad de recuperación automática. Para obtener más información, consulte cómo Microsoft en la mejora continua de la confiabilidad del Microsoft Entra ID.
  - Microsoft Entra Domain Services proporciona un Acuerdo de Nivel de Servicio del 99,9 %. Microsoft Entra Domain Services es un servicio de alta disponibilidad hospedado en centros de datos distribuidos globalmente. Los conjuntos de réplicas son una característica en vista previa de Microsoft Entra Domain Services que habilita la recuperación geográfica ante desastres si una región de Azure se queda sin conexión. Para obtener más información, consulte Conceptos y características de los conjuntos de réplicas de Microsoft Entra Domain Services para obtener más información.
  - Azure DNS proporciona un Acuerdo de Nivel de Servicio del 100 %. HDInsight usa Azure DNS en varios lugares para la resolución de nombres de dominio.

- **Servicios opcionales**, como Azure Key Vault y Azure Data Factory.



Contenido relacionado

- Arquitecturas de continuidad empresarial de Azure HDInsight
- Caso práctico de arquitectura de solución de alta disponibilidad de Azure HDInsight
- ¿Qué son Apache Hive y HiveQL en Azure HDInsight?
- confiabilidad Azure

Last updated on 01/05/2026

Failover para la continuidad del negocio y la recuperación ante desastres

Para maximizar el tiempo de actividad, planee con antelación la continuidad empresarial y prepare todo lo necesario para la recuperación ante desastres con Azure Machine Learning.

Microsoft se esfuerza por garantizar que los servicios de Azure siempre estén disponibles. Sin embargo, es posible que se produzcan interrupciones no planeadas en el servicio. Tener implementado un plan de recuperación ante desastres para controlar las interrupciones del servicio regional. En este artículo aprenderá a:

- Planee una implementación multirregional de Azure Machine Learning y recursos asociados.
- Maximice las posibilidades de recuperar registros, cuadernos, imágenes de Docker y otros metadatos.
- Diseñe su solución para una alta disponibilidad.
- Iniciar una conmutación por error a otra región.

> **Importante**
>
> Azure Machine Learning no proporciona conmutación automática por error ni recuperación ante desastres. La copia de seguridad y restauración de metadatos del área de trabajo, como el historial de ejecución, no está disponible.

Si eliminó accidentalmente el área de trabajo o los componentes correspondientes, en este artículo también se proporcionan opciones de recuperación admitidas actualmente.

Descripción de los servicios de Azure para Azure Machine Learning

Azure Machine Learning depende de varios servicios de Azure. Algunos de estos servicios se aprovisionan en la suscripción. Es responsable de la configuración de alta disponibilidad de estos servicios. Otros servicios se crean en una suscripción de Microsoft y son administrados por Microsoft.

Los servicios de Azure incluyen:

- **Infraestructura de Azure Machine Learning**: Un entorno administrado por Microsoft para el área de trabajo de Azure Machine Learning.

- **Recursos asociados**: recursos aprovisionados en la suscripción durante la creación del área de trabajo de Azure Machine Learning. Estos recursos incluyen Azure Storage, Azure Key Vault, Azure Container Registry y Application Insights.
  - El almacenamiento predeterminado tiene datos como el modelo, los datos del registro de entrenamiento y las referencias a los recursos de datos.
  - Key Vault tiene credenciales para Azure Storage, Container Registry y almacenes de datos.
  - Container Registry tiene una imagen de Docker para entornos de aprendizaje e inferencia.
  - Application Insights se usa para la supervisión de Azure Machine Learning.

- **Recursos de proceso**: recursos creados después de la implementación del área de trabajo. Por ejemplo, puede crear una instancia de proceso o un clúster de proceso para entrenar un modelo de Machine Learning.
  - Instancia de proceso y clúster de proceso: Entornos de desarrollo de modelos administrado por Microsoft.
  - Otros recursos: recursos de proceso de Microsoft que puede adjuntar a Azure Machine Learning, como Azure Kubernetes Service (AKS), Azure Databricks, Azure Container Instances y Azure HDInsight. Eres responsable de configurar la alta disponibilidad de estos recursos.

- **Otros almacenes de datos**: Azure Machine Learning puede montar otros almacenes de datos, como Azure Storage y Azure Data Lake Storage para los datos de entrenamiento. Estos almacenes de datos se crean dentro de su suscripción. Usted es responsable de configurar su configuración de alta disponibilidad. Para ver otras opciones del almacén de datos, consulte Creación de almacenes de datos.

En la tabla siguiente se muestran los servicios Azure administrados por Microsoft y que administra usted. También se indican los servicios de alta disponibilidad, de forma predeterminada.

Expandir tabla

Servicio	Administrado por	Alta disponibilidad de forma predeterminada
Infraestructura de Azure Machine Learning	Microsoft	

Servicio	Administrado por	Alta disponibilidad de forma predeterminada
Recursos asociados		
Azure Storage	Tú	
Bóveda de claves	Tú	✓
Registro de contenedores	Tú	
Application Insights	Tú	N/D
Recursos informáticos		
Instancia de cómputo	Microsoft	
Clúster de proceso	Microsoft	
Otros recursos de proceso, como AKS, Azure Databricks, Container Instances, HDInsight	Tú	
Otros almacenes de datos, como Azure Storage, SQL Database, Bases de datos de Azure para PostgreSQL, Bases de datos de Azure para MySQL. Sistema de archivos de Azure Databricks	Tú	

En el resto de este artículo se describen las acciones que el usuario tiene que realizar para que cada uno de estos servicios sea de alta disponibilidad.

Planeamiento de la implementación multirregional

Una implementación multirregional se basa en la creación de Azure Machine Learning y otros recursos (infraestructura) en dos regiones de Azure. Si se produce una interrupción en una región, se puede cambiar a la otra. Cuando se plantee dónde implementar sus recursos, tenga en cuenta lo siguiente:

- **Disponibilidad regional**: si es posible, use una región en la misma área geográfica, no necesariamente la más cercana. Para comprobar la disponibilidad por región para Azure Machine Learning, consulte los productos de Azure disponibles por región.
- **Regiones emparejadas de Azure**: coordinan las actualizaciones de la plataforma y priorizan las iniciativas de recuperación según las necesidades. Sin embargo, no todas las regiones admiten regiones emparejadas. Para más información, consulte Regiones emparejadas de Azure.

- **Disponibilidad del servicio**: determine si la disponibilidad de los recursos utilizados por la solución debe ser activa/activa, activa/en espera o activa/pasiva.
  - **Hot/hot**: ambas regiones están activas al mismo tiempo, y una de ellas está lista para usarse de inmediato.
  - **activa/en espera**: la región primaria está activa y la secundaria tiene recursos críticos (por ejemplo, modelos implementados) preparados para iniciarse. Los recursos no críticos tendrían que implementarse manualmente en la región secundaria.
  - **activa/pasiva**: la región primaria está activa y la región secundaria tiene el servicio Azure Machine Learning y otros recursos implementados, junto con los datos necesarios. Recursos como los modelos, las implementaciones de modelos o las canalizaciones tendrían que implementarse manualmente.

> **Sugerencia**
>
> En función de los requisitos empresariales, puede decidir tratar diferentes recursos de Azure Machine Learning de forma diferente. Por ejemplo, puede que quiera usar activo/activo para los modelos desplegados (inferencia) y activo/inactivo para los experimentos (entrenamiento).

Azure Machine Learning se basa en otros servicios. Algunos servicios se pueden configurar para la replicación en otras regiones. Otros los debes crear manualmente en varias regiones. En la tabla siguiente se proporciona una lista de servicios responsables de la replicación y una descripción de la configuración:

Expandir tabla

Servicio de Azure	Georreplicado por	Configuración
Área de trabajo de Machine Learning	Tú	Cree un área de trabajo en las regiones seleccionadas.
Cómputo de aprendizaje automático	Tú	Cree los recursos informáticos en las regiones seleccionadas. Para los recursos de computación que pueden escalar dinámicamente, asegúrese de que ambas regiones dispongan de una cuota de computación suficiente para sus necesidades.
Registro de Machine Learning	Tú	Cree el registro en varias regiones.
Bóveda de claves	Microsoft	Use la misma instancia de Key Vault con el área de trabajo y los recursos de Azure Machine Learning en ambas regiones. Key Vault

Servicio de Azure	Georreplicado por	Configuración
		conmuta por error de forma automática en una región secundaria. Para más información, consulte Redundancia y disponibilidad de Azure Key Vault.
Registro de contenedores	Microsoft	Configure la instancia de Container Registry para que replique geográficamente los registros en la región emparejada para Azure Machine Learning. Use la misma instancia para ambas instancias del espacio de trabajo. Para más información, consulte Replicación geográfica en Azure Container Registry.
Cuenta de almacenamiento	Tú	Azure Machine Learning no admite la conmutación por error de la **cuenta de almacenamiento predeterminada** mediante almacenamiento con redundancia geográfica (GRS), almacenamiento con redundancia de zona geográfica (GZRS), almacenamiento con redundancia geográfica con acceso de lectura (RA-GRS) ni almacenamiento con redundancia de zona geográfica con acceso de lectura (RA-GZRS). Cree una cuenta de almacenamiento separada para el almacenamiento predeterminado de cada área de trabajo. Cree servicios o cuentas de almacenamiento separadas para otro almacenamiento de datos. Para más información, vea Redundancia de Azure Storage.
Application Insights	Tú	Cree un recurso de Application Insights para el espacio de trabajo en ambas regiones. Para ajustar el período de retención de datos y los detalles consulte Recopilación, retención y almacenamiento de datos en Application Insights.

Para acelerar la recuperación y el reinicio en la región secundaria, se recomiendan los siguientes procedimientos de desarrollo:

- Utilice plantillas de Azure Resource Manager. Las plantillas generan "infraestructura como código" y permiten implementar rápidamente servicios en ambas regiones.
- Para evitar el desvío entre las dos regiones, actualice las canalizaciones de integración e implementación continuas para que se implementen en ambas regiones.
- Cuando automatice las implementaciones, incluya la configuración de los recursos de proceso asociados al área de trabajo, como Azure Kubernetes Service.
- Cree asignaciones de rol para los usuarios de ambas regiones.
- Cree recursos de red, como las redes virtuales y los puntos de conexión privados de Azure, para ambas regiones. Asegúrese de que los usuarios tengan acceso a ambos entornos de red. Por ejemplo, las configuraciones de VPN y DNS para ambas redes virtuales.

Servicios de proceso y de datos

En función de cuáles sean sus necesidades, podrá tener más servicios de proceso o de datos para que Azure Machine Learning los utilice. Por ejemplo, puede usar Azure Kubernetes Services o Azure SQL Database. Use la siguiente información para aprender a configurar estos servicios con fines de alta disponibilidad.

Recursos informáticos

- **Azure Kubernetes Service**: consulte Procedimientos recomendados para continuidad empresarial y recuperación ante desastres en Azure Kubernetes Service (AKS) y Creación de un clúster de Azure Kubernetes Service (AKS) que use zonas de disponibilidad. Si el clúster de AKS se creó mediante Azure Machine Learning Studio, el SDK o la CLI, no se admite la alta disponibilidad entre regiones.
- **Azure Databricks**: consulte Recuperación ante desastres regional para clústeres de Azure Databricks.
- **Container Instances**: Un orquestador se encarga de la conmutación por error. Consulte Azure Container Instances y orquestadores de contenedores.
- **HDInsight**: consulte Servicios de alta disponibilidad admitidos en Azure HDInsight.

Servicios de datos

- **Contenedor de Azure Blob Storage, Azure Files y Data Lake Storage Gen2**: consulte Redundancia de Azure Storage.
- **Data Lake Storage Gen1**: Consulte Guía de alta disponibilidad y recuperación ante desastres para Data Lake Storage Gen1.

> **Sugerencia**
>
> Si proporciona su propia clave administrada por el cliente para implementar el área de trabajo de Azure Machine Learning, Azure Cosmos DB también se aprovisiona dentro de la suscripción. En ese caso, eres responsable de configurar sus ajustes de alta disponibilidad. Consulte **Alta disponibilidad con Azure Cosmos DB**.

Diseño para lograr alta disponibilidad

Zonas de disponibilidad

Algunos servicios de Azure admiten zonas de disponibilidad. En las regiones que admiten zonas de disponibilidad, si una zona falla, cualquier carga de trabajo se pausará y los datos deberían

guardarse. Sin embargo, los datos no están disponibles para actualizarse hasta que la zona vuelva a estar en línea.

Para más información, consulte Soporte técnico para el servicio de zonas de disponibilidad.

Implementación de componentes críticos en varias regiones

Determine el nivel de continuidad empresarial que tiene como objetivo. El nivel puede diferir entre los componentes de la solución. Por ejemplo, tal vez desee usar una configuración activa/activa para las canalizaciones de producción o las implementaciones de modelos, y una configuración activa/inactiva con fines de experimentación.

Administración de datos de entrenamiento en un almacenamiento aislado

Al mantener el almacenamiento de datos aislado del almacenamiento predeterminado que el área de trabajo usa para los registros, podrá hacer lo siguiente:

- Adjunte las mismas instancias de almacenamiento como almacenes de datos a los espacios de trabajo principal y secundario.
- Utilice la geo-replicación para las cuentas de almacenamiento de datos y maximice su tiempo de actividad.

Administración de recursos de aprendizaje automático como código

> **Nota:**
>
> La copia de seguridad y la restauración de metadatos del área de trabajo, como el historial de ejecución, los modelos y los entornos no están disponibles. La especificación de recursos y configuraciones como código mediante especificaciones de YAML le ayuda a volver a crear recursos entre áreas de trabajo en caso de desastre.

Los trabajos de Azure Machine Learning se definen mediante una especificación de trabajo. Esta especificación incluye dependencias de los artefactos de entrada que se administran a nivel de instancia del espacio de trabajo, incluidos los entornos y los recursos de cómputo. Para el envío de trabajos y los despliegues multirregión, recomendamos las siguientes prácticas:

- Administre localmente el código base, con ayuda de un repositorio de Git.

    - Exporte cuadernos importantes de Estudio de Azure Machine Learning.
    - Exportar canalizaciones creadas en Studio como código.

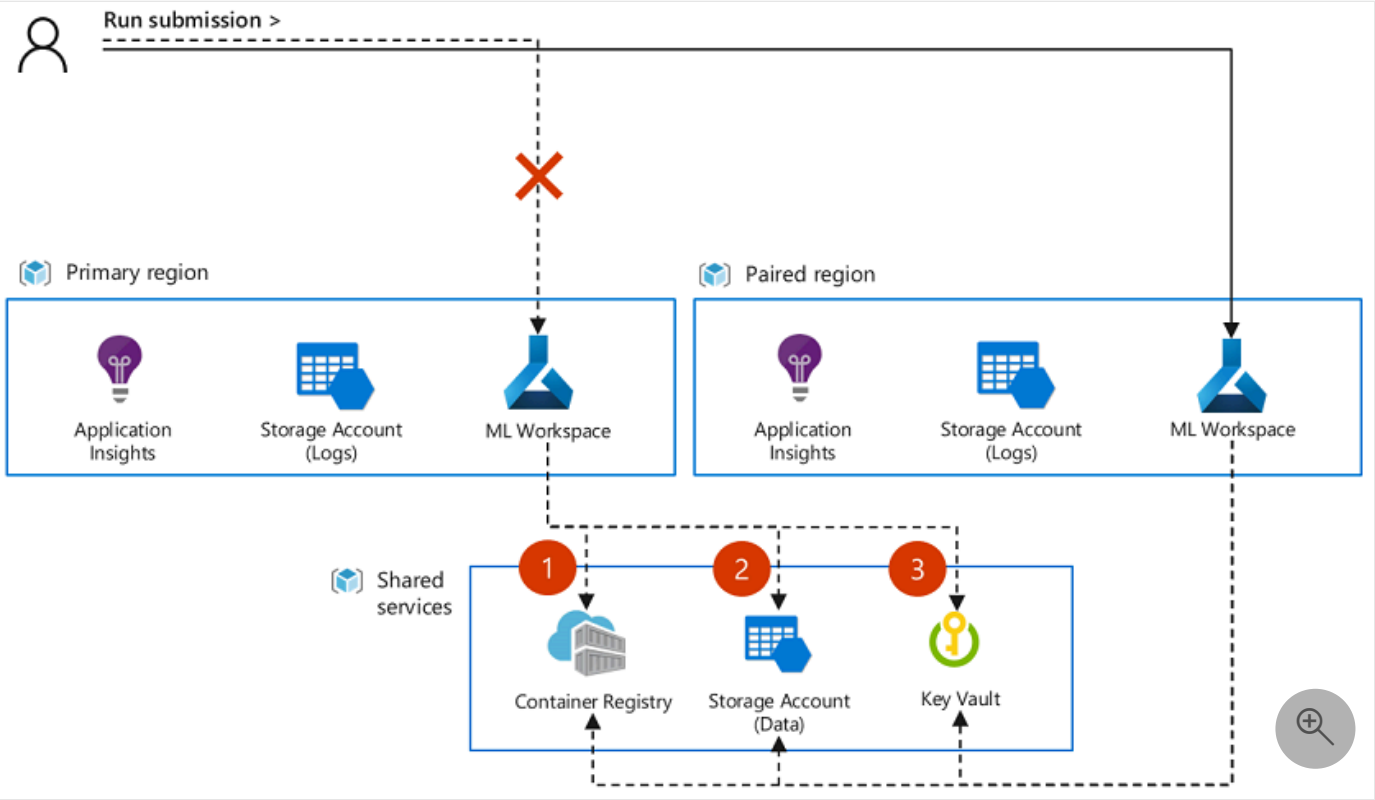
- Administre las configuraciones como código.
    - Evite las referencias codificadas de forma rígida al área de trabajo. En su lugar, configure una referencia a la instancia del área de trabajo mediante un archivo de configuración y utilice MLClient.from_config() para inicializar el área de trabajo.
    - Use Dockerfile si tiene imágenes personalizadas de Docker.

Iniciar una conmutación por error

Continuar el trabajo en el espacio de trabajo de conmutación por error

Cuando el área de trabajo principal deja de estar disponible, puede cambiar al área de trabajo secundaria con fines de experimentación y desarrollo. Azure Machine Learning no envía automáticamente trabajos al área de trabajo secundaria si se produce una interrupción. Actualice la configuración de su código para que apunte al nuevo recurso del espacio de trabajo. Se recomienda evitar la codificación rígida de las referencias al área de trabajo. En su lugar, use un archivo de configuración del área de trabajo para minimizar los pasos manuales del usuario para cambiar las áreas de trabajo. Asegúrese de actualizar también cualquier automatización, como las canalizaciones de integración e implementación continuas, en la nueva área de trabajo.

Azure Machine Learning no puede sincronizar ni recuperar artefactos o metadatos entre instancias de área de trabajo. En función de la estrategia de implementación de la aplicación, es posible que tenga que mover artefactos o volver a crear insumos para la experimentación, como los activos de datos, en el espacio de trabajo de conmutación por error para poder seguir enviando trabajos. En caso de que haya configurado los recursos de su espacio de trabajo principal y de su espacio de trabajo secundario para compartir recursos asociados con la georreplicación habilitada, es posible que algunos objetos estén disponibles directamente en el espacio de trabajo de conmutación por error. Este es el caso, por ejemplo, si ambas áreas de trabajo comparten las mismas imágenes de Docker, almacenes de datos configurados y recursos de Azure Key Vault. En el diagrama siguiente se muestra una configuración en la que dos áreas de trabajo comparten las mismas imágenes (1), almacenes de datos (2) y recursos de Key Vault (3).



> **Nota:**
>
> Los trabajos que se ejecutan cuando se produce una interrupción del servicio no pasarán automáticamente al área de trabajo secundaria. También es poco probable que los trabajos se reanuden y finalicen correctamente en el área de trabajo principal una vez que se resuelva la interrupción. En su lugar, los trabajos se deben volver a enviar, ya sea en el área de trabajo secundaria o en la principal (una vez resuelta la interrupción).

Movilidad de artefactos entre áreas de trabajo

En función del enfoque de recuperación, es posible que tenga que copiar artefactos entre las áreas de trabajo para continuar con el trabajo. Actualmente, la movilidad de artefactos entre áreas de trabajo es limitada. Se recomienda administrar los artefactos como código siempre que sea posible para que se puedan volver a crear en la instancia de conmutación por error.

Los artefactos siguientes se pueden exportar e importar entre áreas de trabajo mediante la extensión de la CLI de Azure para el aprendizaje automático:

Expandir tabla

Artefacto	Exportación	Importar
Modelos	az ml model download --name {NAME} --version {VERSION}	az ml model create

Artefacto	Exportación	Importar
Entornos	az ml environment share --name my-environment --version {VERSION} --resource-group {RESOURCE_GROUP} --workspace-name {WORKSPACE} --share-with-name {NEW_NAME_IN_REGISTRY} --share-with-version {NEW_VERSION_IN_REGISTRY} --registry-name {REGISTRY_NAME}	az ml environment create
Trabajos de Azure Machine Learning	az ml job download -n {NAME} -g {RESOURCE_GROUP} -w {WORKSPACE_NAME}	az ml job create -f {FILE} -g {RESOURCE_GROUP} -w {WORKSPACE_NAME}
Activos de datos	az ml data share --name {DATA_NAME} --version {VERSION} --resource-group {RESOURCE_GROUP} --workspace-name {WORKSPACE} --share-with-name {NEW_NAME_IN_REGISTRY} --share-with-version {NEW_VERSION_IN_REGISTRY} --registry-name {REGISTRY_NAME}	az ml data create -f {FILE} -g {RESOURCE_GROUP} --registry-name {REGISTRY_NAME}

Sugerencia

- Las **salidas de trabajos** se almacenan en la cuenta de almacenamiento predeterminada asociada a un área de trabajo. Aunque las salidas de trabajo pueden quedar inaccesibles desde la interfaz de usuario de Studio en caso de una interrupción del servicio, se puede acceder directamente a los datos a través de la cuenta de almacenamiento. Para obtener más información sobre cómo trabajar con datos almacenados en blobs, consulte **Creación, descarga y enumeración de blobs mediante la CLI de Azure**.

Opciones de recuperación

Eliminación del área de trabajo

Si eliminaste accidentalmente tu área de trabajo, es posible que puedas recuperarla. Para ver los pasos de recuperación, consulta Recuperación de datos del área de trabajo tras una eliminación accidental con la eliminación temporal.

Incluso si el área de trabajo no se puede recuperar, es posible que pueda recuperar los cuadernos del recurso de almacenamiento de Azure asociado al área de trabajo si sigue estos pasos:

- En Azure Portal, vaya a la cuenta de almacenamiento vinculada al área de trabajo de Azure Machine Learning que se eliminó.
- En la sección Almacenamiento de datos que aparece a la izquierda, seleccione **Recursos compartidos de archivos**.
- Sus cuadernos de notas se encuentran en el recurso compartido de archivos cuyo nombre contiene el identificador del área de trabajo.

Pasos siguientes

Para obtener información sobre las implementaciones de infraestructura repetibles con Azure Machine Learning, use una plantilla de Bicep o una plantilla de Terraform.

Last updated on 21/07/2026

Reliability in Azure Stream Analytics

Azure Stream Analytics is a highly resilient service that processes and analyzes streaming data from multiple sources simultaneously. Stream Analytics provides the information that you can use to build complex event processing pipelines by using SQL-like queries.

When you use Azure, reliability is a shared responsibility. Microsoft provides a range of capabilities to support resiliency and recovery. You're responsible for understanding how those capabilities work within all of the services you use, and selecting the capabilities you need to meet your business objectives and uptime goals.

This article describes how Stream Analytics is resilient to potential problems like transient faults and availability zone outages. It provides guidance about how to protect mission-critical jobs against region outages and service maintenance. It also highlights key information about the Stream Analytics service-level agreement (SLA).

> **Importante**
>
> Improving the resiliency of Stream Analytics alone might have limited effect if the other components aren't equally resilient. Consider the reliability of your data sources, including inputs and outputs. Depending on your resiliency requirements, you might need to make configuration changes across multiple areas.

Production deployment recommendations

To ensure high reliability in production environments that use Stream Analytics, we recommend that you take the following actions:

- **Use regions that support availability zones.** Deploy your streaming jobs and other resources in regions that support availability zones.
- **Deploy sufficient capacity.** Set your streaming units (SUs) based on your expected throughput and additional capacity to handle peak loads. Add a buffer above your baseline requirements to handle sudden increases.
- **Monitor health.** Implement comprehensive monitoring by using Azure Monitor metrics and diagnostic logs to track job health, input and output events, and resource utilization. Set up alerts for critical metrics like watermark delay and runtime errors to detect problems before they affect data processing. For more information, see Monitor Stream Analytics.
- **Implement multi-region redundancy for mission-critical workloads.** Deploy identical Stream Analytics jobs across multiple regions. Replicate their configurations and ensure appropriate data routing to achieve regional resiliency. Stream Analytics doesn't provide

native multi-region replication, but this approach enables failover and continuity. For more information, see Custom multi-region solutions for resiliency.

Reliability architecture overview

This section describes some of the important aspects of how the service works that are most relevant from a reliability perspective. The section introduces the logical architecture, which includes some of the resources and features that you deploy and use. It also discusses the physical architecture, which provides details on how the service works under the covers.

Logical architecture

A *job* is the fundamental unit in Stream Analytics that lets you define and run your stream processing logic. A job consists of the following major components:

- *Inputs* that read streaming data from data sources like Azure Event Hubs, Azure IoT Hub, or Azure Storage.
- A *query* that processes and transforms the data.
- *Outputs* that continuously write results to different destinations, like Azure SQL Database, Azure Data Lake Storage, Azure Cosmos DB, and Power BI.

For more information, see Stream Analytics resource model.

Physical architecture

Stream Analytics achieves high reliability by applying multiple layers of resiliency to mitigate problems in the underlying infrastructure and input and output data sources. The following components help ensure that your jobs run robustly:

- **Worker nodes:** Stream Analytics jobs run within a *cluster*. The virtual machines (VMs) within the cluster are known as *worker nodes*. When you use the Standard or Standard V2 SKUs, your jobs run on shared clusters. When you use the Dedicated SKU, your jobs run on their own dedicated cluster.

  The platform automatically manages worker node creation, job placement across worker nodes, health monitoring, and the replacement of unhealthy worker nodes, so you don't see or manage the VMs directly.

- **SUs:** SUs represent the compute resources that run a job. The higher the number of SUs, the more compute resources are allocated for the job. The platform manages worker nodes and job distribution across worker nodes, but you're responsible for allocating SUs to jobs. For more information, see Understand and adjust Stream Analytics SUs.

- **Checkpoints:** Stream Analytics maintains job state through regular *checkpointing* of state. Checkpoints help failed jobs recover quickly with minimal data reprocessing, even for jobs that use stateful query logic.

  When processing failures occur, Stream Analytics automatically restarts from the last checkpoint and reprocesses events that fail during processing. This guarantee applies to all built-in functions and user-defined functions within the job. But achieving end-to-end, exactly-once delivery depends on your output destination's capabilities. For more information, see Checkpoint and replay concepts in Stream Analytics jobs.

> **Nota:**
>
> By using Stream Analytics on IoT Edge, you can run jobs on your own infrastructure. When you use Stream Analytics on IoT Edge, you're responsible for setting it up to meet your reliability requirements. Stream Analytics on IoT Edge is outside the scope of this article.

Resilience to transient faults

Transient faults are short, intermittent failures in components. They occur frequently in a distributed environment like the cloud, and they're a normal part of operations. Transient faults correct themselves after a short period of time. It's important that your applications can handle transient faults, usually by retrying affected requests.

All cloud-hosted applications should follow the Azure transient fault handling guidance when they communicate with any cloud-hosted APIs, databases, and other components. For more information, see Recommendations for handling transient faults.

Stream Analytics automatically handles many transient faults when it ingests data from inputs and writes data to outputs by using built-in retry mechanisms. After a worker node restart or job reassignment, the job uses checkpoints to replay any events that weren't fully processed and continues processing until it reaches the current input stream.

It's a good practice to set up output error policies. But these policies only apply to data conversion errors, and they don't change how Stream Analytics handles transient faults.

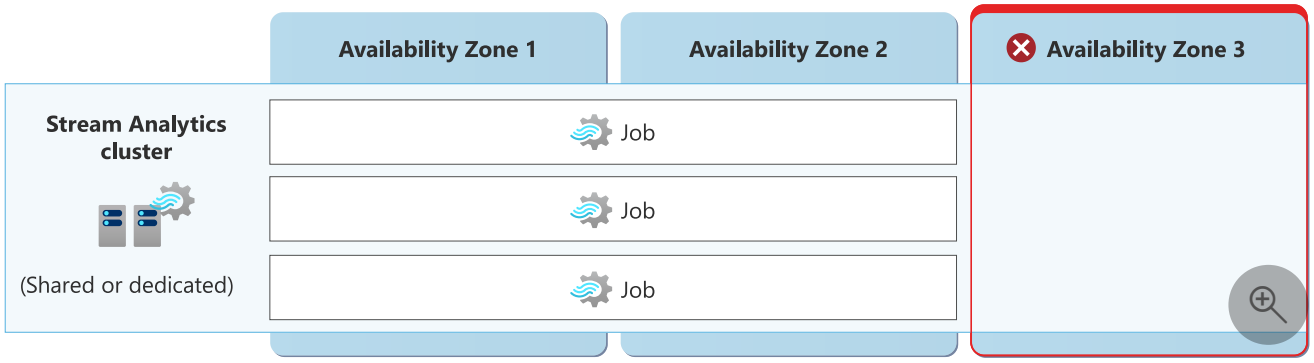
Resilience to availability zone failures

Availability zones are physically separate groups of datacenters within an Azure region. When one zone fails, services can fail over to one of the remaining zones.

Stream Analytics is automatically zone-redundant in regions that support availability zones, which means that jobs use multiple availability zones. Zone redundancy ensures that your job remains resilient to a wide range of failures, including catastrophic datacenter outages, without any changes to the application logic.

When you create a Stream Analytics job in a zone-enabled region, the service distributes your job's compute resources across multiple availability zones.

This zone-redundant deployment model ensures that your streaming jobs continue to process data even if an entire availability zone becomes unavailable. For example, the following diagram shows how jobs continue to run if zone 3 experiences an outage.



Zone redundancy applies to all Stream Analytics features, including query processing, checkpointing, and job management tasks. Stream Analytics automatically replicates your job's state and checkpoint data across availability zones, which prevents data loss and reduces downtime during zone failures.

Requirements

- **Region support:** Zone redundancy for Stream Analytics resources is supported in any region that supports availability zones. For the complete list of regions that support availability zones, see Azure regions list.
- **SKU requirements:** Zone redundancy is available in all Stream Analytics SKUs.

Cost

Zone redundancy on Stream Analytics doesn't incur extra charges. You pay the same rate for SUs whether your job runs in a zone-redundant configuration or in a non-zone-redundant configuration. For more information, see Stream Analytics pricing.

Configure availability zone support

- **Create a zone-redundant Stream Analytics job.** Stream Analytics jobs are automatically zone redundant when you create them in a supported region. No configuration is required.

  For deployment instructions, see Quickstart: Create a Stream Analytics job by using the Azure portal and Quickstart: Create a dedicated Stream Analytics cluster by using the Azure portal.

- **Enable zone redundancy.** All jobs and dedicated clusters are automatically zone-redundant in regions that have availability zones. You don't need to enable zone redundancy.

- **Turn off zone redundancy.** You can't turn off zone redundancy.

Behavior when all zones are healthy

This section describes what to expect when Stream Analytics jobs are zone-redundant and all availability zones are operational.

- **Traffic routing between zones:** Stream Analytics runs each job on worker nodes. Workers in any zone might process incoming streaming data. The service uses internal load balancing to distribute processing tasks across zones.

- **Data replication between zones:** Stream Analytics replicates job state and checkpoint data synchronously across availability zones. As the job processes events and updates its state, Stream Analytics writes those changes to multiple availability zones before it acknowledges them. This synchronous replication ensures zero data loss even if an entire zone becomes unavailable. The replication process is transparent to your application and doesn't affect processing latency under normal conditions.

Behavior during a zone failure

This section describes what to expect when Stream Analytics jobs are zone-redundant and there's an availability zone outage.

- **Detection and response:** The Stream Analytics platform is responsible for detecting a failure in an availability zone and responding to it. Stream Analytics marks workers in the failed zone as unhealthy and automatically redistributes jobs to workers in the remaining healthy zones. You don't need to do anything to initiate a zone failover.
- **Notification**: Microsoft doesn't automatically notify you when a zone is down. However, you can use Azure Resource Health to monitor for the health of an individual resource, and you can set up Resource Health alerts to notify you of problems. You can also use Azure Service Health to understand the overall health of the service, including any zone failures, and you can set up Service Health alerts to notify you of problems.
- **Active requests:** Stream Analytics shifts running jobs to another worker in a healthy availability zone.

  Stream Analytics uses checkpointing to maintain processing state. During a zone failure, workers in healthy zones automatically reprocess jobs running in the failed zone from the last checkpoint.
- **Expected data loss:** The job checkpointing system ensures no data loss.
- **Expected downtime:** Jobs in progress automatically resume after the platform moves them to a healthy worker.
- **Traffic rerouting:** The service automatically redirects all new input data to workers in healthy zones. Existing connections from input sources are reestablished with workers in operational zones. Output connections are similarly reestablished, which ensures continuous data flow through your streaming pipeline.

Zone recovery

When the failed availability zone recovers, Stream Analytics automatically reintegrates it into the active processing pool. Jobs begin to use the recovered infrastructure.

You don't take any action for zone recovery. The platform handles all zone recovery tasks, including state replication and workload redistribution.

Test for zone failures

The Stream Analytics platform manages traffic routing, failover, and zone recovery, so you don't need to initiate or validate availability zone failure processes.

Resilience to region-wide failures

Stream Analytics deploys resources into a single Azure region. If the region becomes unavailable, your jobs (and dedicated clusters, if applicable) are also unavailable.

Custom multi-region solutions for resiliency

To achieve multi-region resilience for your streaming workloads, consider deploying separate jobs in multiple regions. When you take this approach, you're responsible for deploying and managing the jobs and for setting up the appropriate data routing and replication strategies. The Stream Analytics jobs are two separate entities. It's your application's responsibility to send input data into both regional inputs and to reconcile the regional outputs. For more information, see Achieve geo-redundancy for Stream Analytics jobs.

Backup and recovery

Stream Analytics doesn't have a built-in backup and restore feature.

If you want to move, copy, or back up the definition and configuration of your jobs, you can use the Stream Analytics extension for Visual Studio Code to export an existing job in the Azure cloud to your local computer. After you save the entire configuration of your Stream Analytics jobs locally, you can deploy it to the same region or another Azure region. For more information, see Copy, back up, and move your Stream Analytics jobs.

Resilience to service maintenance

Stream Analytics does automatic platform maintenance to apply security updates, deploy new features, and improve service reliability. It might deploy service updates weekly or more frequently. Stream Analytics tests all new updates to ensure high quality.

Consider the following points to ensure that your jobs are resilient to service maintenance activities:

- **Set up jobs to be resilient to replays.** Stream Analytics typically uses checkpoints to restore data after service maintenance. But sometimes it needs to use a replay technique instead. You need to set up your input data sources so that replays don't cause incorrect or partial results in your output. For more information, see Job recovery from a service upgrade.
- **Consider mitigating the risk of bugs by deploying identical jobs.** The service proactively looks for signals after deployment to each batch to check whether the deployment introduces bugs. But no matter how much you test, an existing in-progress job might fail

when maintenance introduces a problem. If you run mission-critical jobs, take steps to mitigate this risk.

You can reduce the risk of a bug affecting your workload by deploying identical jobs to two Azure regions. Monitor these jobs to receive notifications when something unexpected occurs. If one of these jobs enters a failed state after a Stream Analytics service update, take the following actions:

- Contact Azure support to help identify the cause and fix the problem.
- Fail over any downstream consumers to use the healthy job output.

When you select Azure regions to use for your secondary job, consider whether your region has a paired region. The Azure regions list has the most up-to-date information about which regions are paired. Stream Analytics guarantees that infrastructure in paired regions is updated at different times. Updates to Stream Analytics don't deploy at the same time in a set of paired regions. A sufficient time gap exists between the updates to identify and fix potential problems.

Service-level agreement

The service-level agreement (SLA) for Azure services describes the expected availability of each service and the conditions that your solution must meet to achieve that availability expectation. For more information, see SLAs for online services.

Stream Analytics provides separate availability SLAs for API calls to manage jobs and for the operations of the jobs.

Related content

- Stream Analytics documentation
- Stream Analytics monitoring
- Stream processing architectures

ⓘ **Nota:** El autor ha creado este artículo con ayuda de inteligencia artificial. Más información

---

Last updated on 22/01/2026

Confiabilidad en Microsoft Fabric

Este artículo describe el soporte de fiabilidad en Microsoft Fabric, incluyendo tanto la resiliencia regional con zonas de disponibilidad como la recuperación y continuidad del negocio entre regiones. Para obtener información general más detallada sobre la confiabilidad de Azure, consulte Confiabilidad de Azure.

Soporte para zonas de disponibilidad

Las zonas de disponibilidad son grupos físicamente independientes de centros de datos dentro de una región de Azure. Cuando una zona falla, los servicios pueden transferirse a una de las zonas restantes.

Fabric usa zonas de disponibilidad de Azure para proteger elementos y datos de Fabric y Power BI ante fallos en el centro de datos. El servicio distribuye automáticamente los recursos Fabric entre varias zonas sin requerir ninguna configuración del cliente.

- La ingeniería de datos admite zonas de disponibilidad si usa OneLake. Si usa otros orígenes de datos como ADLS Gen2, debe asegurarse de que el almacenamiento con redundancia de zona (ZRS) esté habilitado.

Experiencia de reducción de zona

Durante una interrupción en toda la zona, no se requiere ninguna acción del cliente. Capacidades del tejido que se auto-sanarse y se reequilibran automáticamente para aprovechar la zona saludable. En algunos casos, las operaciones en curso pueden necesitar reiniciarse. Por ejemplo, ejecutar Spark Jobs podría fallar si el nodo principal está en la zona fallida. En ese caso, tienes que volver a enviar los trabajos. Las consultas de endpoint de almacén de datos y análisis SQL pueden fallar si el nodo frontal está en la zona fallida. En tal caso, debes reiniciar la consulta de forma segura.

> **ⓘ Importante**
>
> Aunque Microsoft se esfuerza por proporcionar compatibilidad uniforme y coherente con las zonas de disponibilidad, en algunos casos de error de zona de disponibilidad, las capacidades de Fabric ubicadas en regiones de Azure con mayores fluctuaciones de la demanda de los clientes podrían experimentar una latencia superior a la normal.

Recuperación ante desastres entre regiones y continuidad empresarial

La recuperación ante desastres (DR) hace referencia a las prácticas que las organizaciones usan para recuperarse de eventos de alto impacto, como desastres naturales o implementaciones con errores que producen tiempo de inactividad y pérdida de datos. Independientemente de la causa, el mejor remedio para un desastre es un plan de recuperación ante desastres bien definido y probado y un diseño de aplicaciones que apoye activamente la recuperación ante desastres. Antes de empezar a crear el plan de recuperación ante desastres, consulte Recomendaciones para diseñar una estrategia de recuperación ante desastres.

Para la recuperación ante desastres, Microsoft usa el modelo de responsabilidad compartida. En este modelo, Microsoft garantiza que la infraestructura de línea base y los servicios de plataforma estén disponibles. Sin embargo, muchos servicios de Azure no replican automáticamente datos ni se revierten de una región con errores para realizar la replicación cruzada en otra región habilitada. En esos servicios, es responsable de configurar un plan de recuperación ante desastres válido para su carga de trabajo. La mayoría de los servicios que se ejecutan en las ofertas de plataforma como servicio (PaaS) de Azure proporcionan funciones y orientación para dar soporte a la recuperación ante desastres. Puede usar características específicas del servicio para admitir la recuperación rápida con el fin de contribuir al desarrollo del plan de recuperación ante desastres.

En esta sección se describe un plan de recuperación ante desastres para Fabric diseñado para ayudar a su organización a mantener sus datos seguros y accesibles cuando se produce un desastre regional no planeado. El plan trata los temas siguientes:

- **Replicación entre regiones**: Fabric ofrece replicación entre regiones para los datos almacenados en OneLake. Puede participar o no en esta característica en función de sus requisitos.
- **Acceso a los datos después del desastre**: en un escenario de desastre regional, Fabric garantiza el acceso a los datos, con ciertas limitaciones. Aunque la creación o modificación de nuevos elementos está restringida después de la conmutación por error, el enfoque principal sigue siendo asegurarse de que los datos existentes permanecen accesibles e intactos.
- **Guía para la recuperación**: Fabric proporciona un conjunto estructurado de instrucciones para guiarle a través del proceso de recuperación. La guía estructurada facilita la transición a las operaciones normales.

Power BI, ahora una parte de Fabric, tiene un sistema de recuperación ante desastres sólido y ofrece las siguientes características:

- **BCDR como valor predeterminado**: si una región es emparejada con un región que admite Power BI, las funcionalidades de recuperación ante desastres se incluyen de forma predeterminada. No es necesario participar ni activar esta característica por separado.
- **Replicación entre regiones**: Power BI utiliza la replicación con redundancia geográfica del almacenamiento de Azure y la replicación con redundancia geográfica de Azure SQL para garantizar que existen instancias de copia de seguridad en otras regiones y que se pueden utilizar. Esto significa que los datos se duplican en distintas regiones, lo que mejora su disponibilidad y reduce los riesgos asociados a las interrupciones regionales.
- **Servicios continuos y acceso después del desastre**: incluso durante eventos disruptivos, los elementos de Power BI permanecen accesibles en modo de solo lectura. Los elementos incluyen modelos semánticos, informes y paneles de información, garantizando que las empresas puedan continuar sus procesos de análisis y toma de decisiones sin obstáculos significativos.

Para más información, consulte las preguntas frecuentes sobre alta disponibilidad, conmutación por error y recuperación ante desastres de Power BI.

> **Importante**
>
> Para los clientes afectados por un desastre y cuyas regiones de origen no tienen una región emparejada con Azure que soporte Fabric, la capacidad de utilizar capacidades de Fabric podría verse comprometida, incluso si los datos dentro de esas capacidades se replican. Esta limitación está ligada a la infraestructura de la región de origen, que es esencial para el funcionamiento de las capacidades. Para ver la lista de regiones que admiten Fabric, vaya a Disponibilidad de la región de Fabric.

Región principal y funcionalidad de capacidad

Para un planeamiento eficaz de la recuperación ante desastres, es fundamental que comprenda la relación entre la región principal y las ubicaciones de capacidad. Comprender la región principal y las ubicaciones de capacidad le ayuda a realizar selecciones estratégicas de regiones de capacidad, así como los procesos de replicación y recuperación correspondientes.

La **región principal** para el espacio empresarial y el almacenamiento de datos de su organización se establece en la ubicación de la dirección de facturación del primer usuario que se registra. Para más información sobre la configuración del inquilino, vaya a Planeamiento de implementación de Power BI: Configuración de inquilino. Al crear nuevas capacidades, el almacenamiento de datos se establece en la región principal de forma predeterminada. Si deseas cambiar tu región de almacenamiento de datos a otra región, necesitas activar Multi-Geo, una función de Fabric Premium.

> **Importante**
>
> Elegir una región diferente para la capacidad no reubica completamente todos los datos en esa región. Algunos elementos de datos siguen estando almacenados en la región principal. Para ver qué datos permanecen en la región principal y qué datos se almacenan en la región habilitada para Multi-Geo, consulte **Configuración de la compatibilidad de Multi-Geo con Fabric Premium**.
>
> En el caso de una región de origen que no tenga una región emparejada, las capacidades en cualquier región habilitada para Multi-Geo podrían enfrentarse a problemas operativos si la región de origen sufre un desastre, ya que la funcionalidad principal del servicio está vinculada a la región de origen.
>
> Si selecciona una región habilitada para Multi-Geo dentro de la UE, se garantiza que los datos se almacenan dentro del límite de datos de la UE.

Para obtener información sobre cómo identificar la región principal, consulte Búsqueda de la región principal de Fabric.

Configuración de la capacidad de recuperación ante desastres

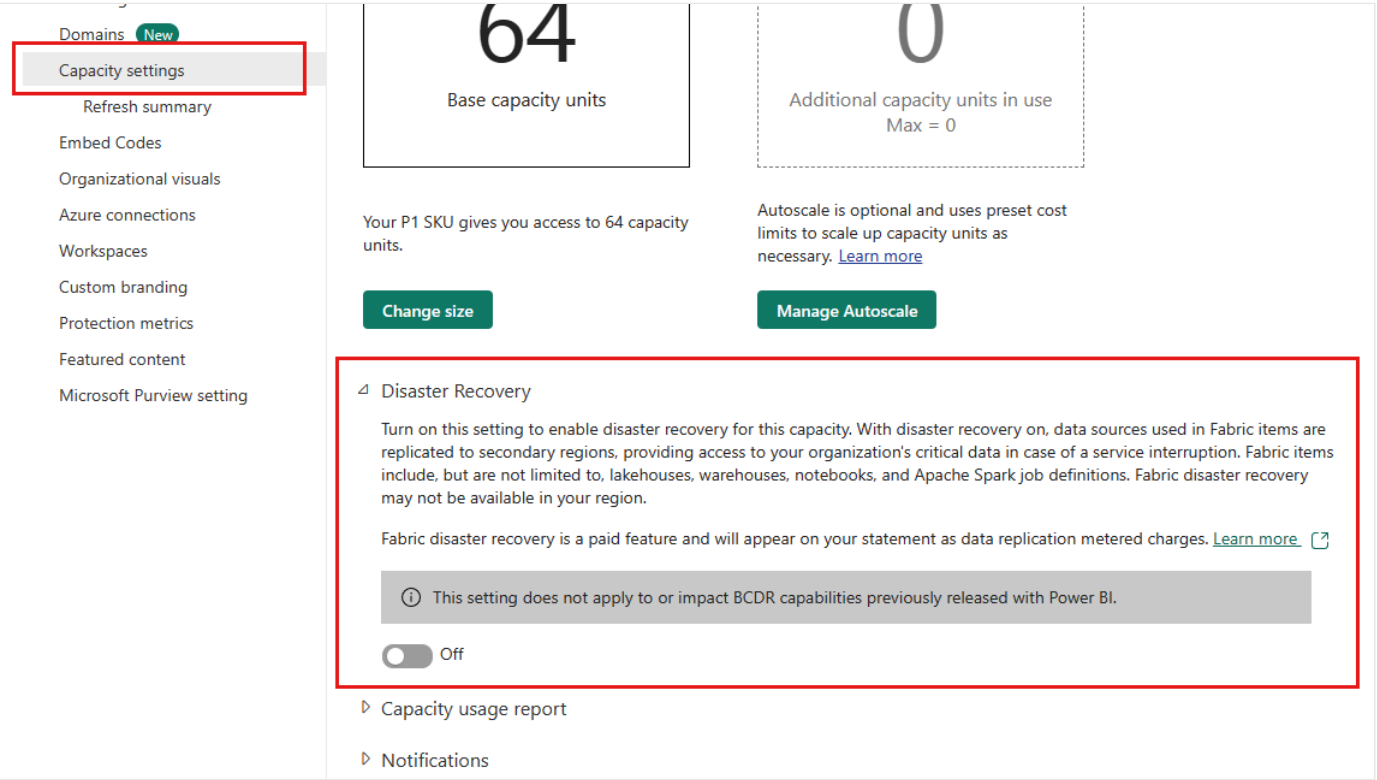
Fabric proporciona un conmutador de recuperación ante desastres en la página de configuración de capacidad. Está disponible donde los emparejamientos regionales de Azure coinciden con la presencia de servicios de Fabric. Estos son los detalles de este cambio:

- **Acceso a rol**: Solo los usuarios con el rol de administrador de capacidad o superior pueden utilizar este conmutador.
- **Granularidad**: La granularidad del conmutador es el nivel de capacidad. Está disponible para las capacidades Premium y Fabric.

- **Ámbito de datos**: El conmutador de recuperación de desastres se refiere específicamente a los datos de OneLake, que incluye los datos de Lakehouse y Warehouse. **El switch no influye en tus datos almacenados fuera de OneLake.**

- **Continuidad BCDR para Power BI**: Aunque puedes activar o desactivar la recuperación ante desastres para datos de OneLake, BCDR para Power BI siempre está soportado, independientemente de si el interruptor está encendido o apagado.

- **Frecuencia**: Una vez que cambias la configuración de capacidad de recuperación ante desastres, debes esperar 30 días antes de poder modificarla de nuevo. El tiempo de espera mantiene la estabilidad y evita cambios constantes.



> **Nota:**
>
> Tras activar la configuración de capacidad de recuperación ante desastres o crear nuevos espacios de trabajo dentro de la capacidad, la replicación de datos puede tardar un tiempo en iniciarse. Puede comprobar el estado de cada área de trabajo en la página de configuración de capacidad en Áreas de trabajo **asignadas a esta capacidad**. La columna **Replicación geográfica de OneLake** muestra el estado para habilitar la replicación geográfica.

Replicación de datos

Cuando active la configuración de capacidad de recuperación ante desastres, se habilita la replicación entre regiones como capacidad de recuperación ante desastres para los datos de OneLake. La plataforma Fabric se alinea con las regiones Azure para aprovisionar los pares de redundancia geográfica. Sin embargo, algunas regiones no tienen una región de par de Azure, o la región de par no admite Fabric. Para estas regiones, la replicación de datos no está disponible. Para obtener más información, consulte Regiones con zonas de disponibilidad y ningún par de regiones y Disponibilidad de regiones de Fabric.

> **Nota:**
>
> Aunque Fabric ofrece una solución de replicación de datos en OneLake para admitir la recuperación ante desastres, existen notables limitaciones. Por ejemplo, los datos de las bases de datos KQL y los conjuntos de consultas se almacenan externamente a OneLake, lo que significa que es necesario un enfoque separado de recuperación de desastres. Consulte el resto de este documento para obtener más información sobre el enfoque de recuperación ante desastres para cada elemento de Fabric.

Billing

La función de recuperación ante desastres de Fabric permite la replicación geográfica de sus datos para mejorar la seguridad y la fiabilidad. Esta característica consume más almacenamiento y transacciones, que se facturan como almacenamiento BCDR y operaciones BCDR, respectivamente. Puede supervisar y administrar estos costes en la aplicación Microsoft Fabric Capacity Metrics, donde aparecen como partidas independientes.

Para obtener un desglose exhaustivo de todos los costos de recuperación ante desastres asociados para ayudarle a planear y presupuestar en consecuencia, consulte Consumo de almacenamiento y proceso de OneLake.

Configurar recuperación ante desastres

Aunque Fabric proporciona características de recuperación ante desastres para admitir la resistencia de los datos, **debe** seguir ciertos pasos manuales para restaurar el servicio durante las interrupciones. En esta sección se detallan las acciones que debe realizar para prepararse para posibles interrupciones.

Fase 1: Preparación

- **Active la configuración de la capacidad de recuperación ante desastres**: Revise y configure periódicamente los configuración de capacidad de recuperación ante desastres para asegurarse de que satisfacen sus necesidades de protección y rendimiento.
- **Cree copias de seguridad de los datos**: Copie los datos almacenados críticos fuera de OneLake a otra región de forma que se ajuste a su plan de recuperación ante desastres.

Fase 2: Conmutación por error en caso de desastre

Cuando un desastre grave deja irrecuperable la región principal, Microsoft Fabric inicia una conmutación por error regional. No puedes acceder al portal de Fabric hasta que se complete el failover. Se publica una notificación en la página de soporte de Microsoft Fabric.

El tiempo que tarda en completarse la conmutación por error puede variar, aunque normalmente tarda menos de una hora. Una vez completada la conmutación por error, esto es lo que puede esperar:

- **Portal de Fabric**: puede acceder al portal y las operaciones de lectura, como la navegación por las áreas de trabajo, los flujos de tareas en áreas de trabajo y elementos, seguir funcionando. Todas las operaciones de escritura, como la creación o modificación de un área de trabajo, se ponen en pausa.
- **Power BI**: Puede realizar operaciones de lectura, como la visualización de cuadros de mando e informes. No se admiten actualizaciones, operaciones de publicación de informes, modificaciones de cuadros de mando e informes ni otras operaciones que requieran cambios en los metadatos.
- **Lakehouse/Almacén:** No puedes abrir estos elementos, pero sí acceder a archivos a través de APIs o herramientas de OneLake.
- **Definición de trabajo en Spark**: No puedes abrir definiciones de trabajos en Spark, pero sí puedes acceder a archivos de código a través de APIs o herramientas de OneLake. Cualquier metadato o configuración se guarda tras el failover.
- **Cuaderno**: No puedes abrir cuadernos, y el contenido del código no se guarda después del desastre.
- **Modelo/Experimento ML**: No se pueden abrir modelos o experimentos ML. El contenido del código y los metadatos, como métricas de ejecución y configuraciones, no se guardan después del desastre.

- **Flujo de datos Gen2/Canalización/Eventstream**: no puede abrir estos elementos, pero puede utilizar los destinos de recuperación ante desastres compatibles (almacenes de lago o almacenes) para proteger los datos.
- **Base de datos o conjunto de consultas de KQL**: No puedes acceder a las bases de datos ni a los conjuntos de consultas de KQL después de una conmutación por error. Se requieren más pasos de requisitos previos para proteger los datos en bases de datos de KQL y conjuntos de consultas.

En un escenario de desastre, el portal de Fabric y Power BI están en modo de solo lectura, y otros elementos de Fabric no están disponibles. Puedes acceder a sus datos almacenados en OneLake usando APIs o herramientas de terceros. Tanto el portal como Power BI conservan la capacidad de realizar operaciones de lectura y escritura en esos datos. Esta capacidad garantiza que los datos críticos sigan siendo accesibles y modificables, y mitiga la posible interrupción de las operaciones de su empresa.

Puedes acceder a los datos de OneLake a través de varios canales:

- API OneLake ADLS Gen2: Consulte Conexión a Microsoft OneLake
- Ejemplos de herramientas que pueden conectarse a los datos de OneLake:
  - Explorador de Azure Storage: Consulte Integrar OneLake con Explorador de Azure Storage
  - Explorador de archivos OneLake: Consulte Utilizar el explorador de archivos de OneLake para acceder a los datos de Fabric
- En un escenario de desastre, el catálogo de OneLake está en modo de solo lectura:
  - Pestaña Explorar: puede acceder a la pestaña Explorar para ver todos los elementos y áreas de trabajo, incluidos sus metadatos y detalles relacionados.
  - Pestaña Gobernanza: puede acceder a la pestaña Gobernanza para ver información, acciones recomendadas y herramientas de gobernanza, en función de la actualización correcta del modelo más reciente antes de la conmutación por error.

Fase 3: Plan de recuperación

Aunque Fabric garantiza que los datos permanecen accesibles después de un desastre, también puede actuar para restaurar completamente sus servicios al estado antes del incidente. En esta sección se proporciona una guía paso a paso para ayudarle a través del proceso de recuperación.

Pasos de la recuperación

1. Cree una nueva capacidad de Fabric en cualquier región después de un desastre. Dada la alta demanda en este tipo de eventos, selecciona una región fuera de tu región geográfica principal para aumentar la probabilidad de disponibilidad de capacidad de proceso. Para obtener información sobre cómo crear una capacidad, consulte Comprar una suscripción de Microsoft Fabric.

2. Cree áreas de trabajo en la capacidad recién creada. Si es necesario, use los mismos nombres que las áreas de trabajo antiguas.

3. Cree elementos con los mismos nombres que los que desea recuperar. Este paso es importante si usa el script personalizado para recuperar almacenes y almacenes de lagos.

4. Restauración de los elementos. Para cada elemento, siga la sección correspondiente de la guía de recuperación de desastres específica de la experiencia para restaurar el elemento.

Pasos siguientes

- Guía de recuperación ante desastres específica de la experiencia
- Confiabilidad en Azure

Last updated on 23/07/2026

Confiabilidad en Azure App Service

Azure App Service es un servicio basado en HTTP para hospedar aplicaciones web, API REST y back-end móvil. App Service se integra con Microsoft Azure para proporcionar seguridad, equilibrio de carga, escalado automático y administración automatizada para aplicaciones. Como servicio de Azure, App Service proporciona una variedad de funcionalidades para satisfacer los requisitos de confiabilidad.

Cuando se usa Azure, la confiabilidad es una responsabilidad compartida. Microsoft proporciona una variedad de funcionalidades para admitir resistencia y recuperación. Es responsable de comprender cómo funcionan esas funcionalidades dentro de todos los servicios que usa y de seleccionar las funcionalidades que necesita para cumplir los objetivos empresariales y los objetivos de tiempo de actividad.

En este artículo se describe cómo hacer que App Service sea resistente a una variedad de posibles interrupciones y problemas, incluidos errores transitorios, interrupciones de zona de disponibilidad, interrupciones de regiones y mantenimiento del servicio. También se describe cómo puede usar copias de seguridad para recuperarse de otros tipos de problemas y se resalta cierta información clave sobre el acuerdo de nivel de servicio (SLA) de App Service.

> **Nota:**
>
> Si busca información sobre el soporte de confiabilidad en App Service Environment, consulte Confiabilidad en App Service Environment.

Recomendaciones de implementación de producción

Azure Well-Architected Framework proporciona recomendaciones sobre confiabilidad, rendimiento, seguridad, costo y operaciones. Para comprender cómo estas áreas influyen entre sí y contribuir a una solución confiable de App Service, consulte Procedimientos recomendados de arquitectura para App Service (Web Apps) en Azure Well-Architected Framework.

Introducción a la arquitectura de confiabilidad

Al crear una aplicación web de App Service, se especifica el plan de App Service que ejecuta la aplicación.

Un plan de App Service define un conjunto de recursos de proceso que ejecutan las aplicaciones web. Todas las Web Apps deben ejecutarse dentro de un plan. Puede escalar un

plan para ejecutarse en varias *instancias* de máquina virtual, también denominadas *trabajadores*. Estas instancias proporcionan los recursos de proceso que ejecutan el código de su aplicación. Un único plan de App Service puede hospedar varias aplicaciones. Todas las aplicaciones se ejecutan en el mismo conjunto compartido de instancias de máquina virtual.

App Service ofrece las siguientes características de redundancia:

- **Distribución entre dominios de error:** En el nivel de plataforma, Azure distribuye automáticamente las instancias de máquina virtual del plan de App Service entre dominios de error dentro de la región de Azure. Esta distribución minimiza el riesgo de errores de hardware localizados mediante la agrupación de máquinas virtuales que comparten una fuente de alimentación común y un conmutador de red.

- **Distribución entre zonas de disponibilidad:** Si habilita la redundancia de zona en un plan de App Service compatible, Azure distribuye las instancias entre zonas de disponibilidad dentro de la región. Esta configuración proporciona una mayor resistencia si se produce una interrupción de zona. Para obtener más información sobre la redundancia de zonas, consulte Compatibilidad con zonas de disponibilidad.

- **Escalado de aplicaciones:** Al configurar el plan de App Service para ejecutar varias instancias de máquina virtual, todas las aplicaciones del plan se ejecutan en todas las instancias de forma predeterminada. Si configura su plan para el escalado automático, todas las aplicaciones se escalarán juntas según la configuración de escalado automático. Sin embargo, puede personalizar el número de instancias de plan que ejecutan una aplicación específica mediante el escalado por aplicación.

- **Unidades de escalado:** Internamente, App Service se ejecuta en una infraestructura de plataforma denominada *unidades de escalado*, también conocida como *stamps* o *espacios web*. Una unidad de escalado incluye todos los componentes necesarios para hospedar y ejecutar App Service, incluido el proceso, el almacenamiento, las redes y el equilibrio de carga. Azure administra las unidades de escalado para garantizar una distribución equilibrada de cargas de trabajo, realizar un mantenimiento rutinario y mantener la confiabilidad general de la plataforma.

  Es posible que algunas funcionalidades solo se apliquen a unidades de escalado específicas. Por ejemplo, algunas unidades de escalado de App Service pueden admitir la redundancia de zona, mientras que otras unidades de escala de la misma región no.

Resistencia a errores transitorios

Los errores transitorios son errores breves e intermitentes en los componentes. Se producen con frecuencia en un entorno distribuido como la nube y son una parte normal de las

operaciones. Los errores transitorios se corrigen después de un breve período de tiempo. Es importante que las aplicaciones puedan controlar errores transitorios, normalmente mediante el reintento de solicitudes afectadas.

Todas las aplicaciones hospedadas en la nube deben seguir las instrucciones de control de errores transitorios de Azure cuando se comunican con cualquier API, bases de datos y otros componentes hospedados en la nube. Para obtener más información, consulte Recomendaciones para controlar errores transitorios.

Los SDK proporcionados por Microsoft suelen controlar errores transitorios. Dado que hospeda sus propias aplicaciones en App Service, siga los pasos necesarios para reducir la posibilidad de errores transitorios:

- **Implemente varias instancias en el plan.** App Service realiza actualizaciones automatizadas y otras formas de mantenimiento en instancias del plan. Si una instancia se vuelve incorrecta, el servicio puede reemplazar automáticamente esa instancia por una nueva instancia correcta. Durante el proceso de reemplazo, puede haber un breve período cuando la instancia anterior no está disponible y una nueva instancia no está lista para atender el tráfico. Para mitigar estos efectos, implemente varias instancias del plan de App Service.

- **Uso de ranuras de implementación.** Las ranuras de implementación de App Service permiten implementaciones sin tiempo de inactividad de las aplicaciones. Use ranuras de implementación para minimizar el efecto de las implementaciones y los cambios de configuración para los usuarios. Las ranuras de implementación también reducen la probabilidad de que se reinicie la aplicación. Reiniciar la aplicación produce un error transitorio.

- **Evite escalar o reducir verticalmente.** Estas operaciones modifican la CPU, la memoria y otros recursos asignados a cada instancia, y pueden provocar el reinicio de la aplicación empresarial. En su lugar, seleccione un nivel y un tamaño de instancia que cumplan los requisitos de rendimiento bajo una carga típica. Para escalar y reducir horizontalmente, agregue y elimine instancias de forma dinámica para administrar los cambios en el volumen de tráfico.

Resistencia a errores de zona de disponibilidad

Las zonas de disponibilidad son grupos físicamente independientes de centros de datos dentro de una región de Azure. Cuando una zona falla, los servicios pueden transferirse a una de las zonas restantes.

En el caso de los **niveles Premium v2 a v4**, puede configurar App Service como *redundante de zona*, lo que significa que los recursos se distribuyen entre varias zonas de disponibilidad. La distribución entre varias zonas ayuda a las cargas de trabajo de producción a lograr resistencia y confiabilidad. Al configurar la redundancia de zona en los planes de App Service, todas las aplicaciones que utilizan el plan pasan a ser redundantes en la zona.

Requisitos

Para habilitar la redundancia de zona, debe cumplir los siguientes requisitos:

- **Compatibilidad con regiones:** En el caso de los planes de App Service **Premium v2 y v3** , la redundancia de zona se admite en cualquier región que admita zonas de disponibilidad.
- **Tipo de plan:** Use tipos de plan Premium v2 a v4.

  > ⓘ **Importante**
  >
  > Para habilitar la redundancia de zona para los planes **de App Service Premium v4** , debe **confirmar que la región deseada admite planes v4** y que **admite zonas de disponibilidad**.

- **Número mínimo de instancias:** Implemente un mínimo de dos instancias en el plan.
- **Unidad de escalado:** La aplicación debe implementarse en una unidad de escalado que admita zonas de disponibilidad. No controla directamente la unidad de escalado que utiliza el plan. En su lugar, al crear un plan de App Service, el plan se asigna a una unidad de escalado en función del grupo de recursos del propio plan. Para determinar si la unidad de escalado del plan de App Service admite redundancia de zona, consulte Comprobar la compatibilidad con redundancia de zona para un plan de App Service.

  Si tu plan de App Service está en una unidad de escalado que no admite redundancia de zona, no puedes habilitar la redundancia de zona en tu plan. En su lugar, debe reimplementar tus aplicaciones en un nuevo plan en una unidad de escala diferente.

Distribución de instancias a través de zonas

Al crear un plan de App Service con redundancia de zona, Azure distribuye las instancias del plan entre zonas de disponibilidad de la región. Esta distribución garantiza que sus aplicaciones sigan estando disponibles incluso si una zona sufre una interrupción del servicio.

La distribución de instancias en una implementación con redundancia de zona sigue reglas específicas. Estas reglas también se aplican a medida que la aplicación se escala y se reduce horizontalmente:

- **Instancias mínimas:** El plan de App Service debe tener un mínimo de dos instancias para la redundancia de zona.
- **Zonas de disponibilidad máxima compatibles con el plan:** Azure determina el número de zonas de disponibilidad que puede usar el plan, lo que se conoce como *maximumNumberOfZones*. Para ver el número de zonas de disponibilidad que puede usar su plan específico, consulte Comprobar la compatibilidad con la redundancia de zonas para un plan de App Service.
- **Distribución de instancias:** Cuando se habilita la redundancia de zona, Azure distribuye instancias de plan entre varias zonas de disponibilidad automáticamente. La distribución se basa en las reglas siguientes:
  - Si el número de instancias supera el *maximumNumberOfZones* y divide uniformemente, Azure distribuye las instancias uniformemente entre zonas.
  - Si el número de instancias no se divide de manera uniforme, Azure distribuye las instancias restantes entre las zonas restantes.
  - Cuando la plataforma de App Service asigna instancias para un plan de App Service con redundancia de zona, usa el equilibrio de zona de mejor esfuerzo que proporcionan los conjuntos de escalado de máquinas virtuales de Azure subyacentes. Un plan se equilibra si cada zona tiene el mismo número de máquinas virtuales o difiere en una instancia de todas las demás zonas. Para obtener más información, consulte Equilibrio de zona.
- **Ubicación de zona física:** Puede ver la zona de disponibilidad física que se usa para cada una de las instancias del plan de App Service. Para más información, consulte Visualización de zonas físicas para un plan de App Service.

Consideraciones

En el caso de los planes **Premium v2 a v4**, una interrupción de zona de disponibilidad podría afectar a algunos aspectos de Azure App Service, aunque la aplicación siga sirviendo tráfico. Estos comportamientos incluyen el escalado del plan de App Service, la creación de aplicaciones, la configuración de la aplicación y la publicación de aplicaciones.

Al habilitar la redundancia de zona en su plan App Service **Premium v2 a v4**, también mejora la resistencia durante las actualizaciones de la plataforma. Para obtener más información,

consulte Resistencia al mantenimiento del servicio.

En los planes de App Service que no están configurados como redundantes por zona, las instancias de máquina virtual (VM) subyacentes no son resistentes a los errores de la zona de disponibilidad. Pueden experimentar tiempo de inactividad durante una interrupción en cualquier zona de esa región.

Cost

Cuando use los **planes Premium v2 a v4 de App Service**, habilitar las zonas de disponibilidad no supone ningún coste adicional si tiene dos o más instancias. Los cargos se basan en la SKU de su plan de App Service, la capacidad que especifique y las instancias a las que escale según sus criterios de escalado automático.

Si habilita zonas de disponibilidad pero especifica una capacidad de menos de dos, la plataforma aplica un recuento mínimo de instancias de dos. La plataforma le cobra por esas dos instancias.

Configurar soporte de zonas de disponibilidad

- **Cree un nuevo plan de App Service con redundancia de zona.** Para más información, consulte Crear un nuevo plan de App Service que incluya redundancia de zona.
- **Habilite o deshabilite la redundancia de zona en un plan de App Service existente.** Para obtener más información, consulte Configurar la redundancia de zona para un plan de App Service existente.

Planeamiento y administración de capacidad

Para prepararse para errores de zona de disponibilidad, considere la posibilidad de *sobreaprovisionar* la capacidad del plan de App Service. Este enfoque permite que la solución tolere cierta pérdida de capacidad y siga funcionando sin que se vea afectado su rendimiento. Para obtener más información, consulte Administración de la capacidad mediante el aprovisionamiento excesivo.

Comportamiento cuando todas las zonas están en buen estado

En la siguiente lista se describe qué esperar cuando los planes de App Service están configurados para la redundancia de zona y todas las zonas de disponibilidad están operativas:

- **Enrutamiento del tráfico entre zonas:** Durante el funcionamiento normal, el tráfico se enruta entre todas las instancias disponibles del plan de App Service en todas las zonas de disponibilidad.

- **Replicación de datos entre zonas:** Durante las operaciones normales, cualquier estado almacenado en el sistema de archivos de la aplicación se almacena en el almacenamiento con redundancia de zona y se replica sincrónicamente entre zonas de disponibilidad.

Comportamiento durante un fallo de zona

Una interrupción de la zona de disponibilidad puede afectar a algunos aspectos de App Service, aunque la aplicación siga sirviendo al tráfico. Estos comportamientos incluyen el escalado del plan de App Service, la creación de aplicaciones, la configuración de la aplicación y la publicación de aplicaciones.

La siguiente lista describe qué puede ocurrir cuando los planes de App Service están configurados para la redundancia de zona y una o varias zonas de disponibilidad no están disponibles:

- **Detección y respuesta:** La plataforma de App Service detecta automáticamente errores en una zona de disponibilidad e inicia una respuesta. No se requiere ninguna intervención manual para iniciar una conmutación por error de zona.

- **Notificación**: Microsoft no le notifica automáticamente cuando una zona está inactiva. Sin embargo, puede usar Azure Resource Health para supervisar el estado de un recurso individual y puede configurar alertas de Resource Health para notificarle problemas. También puede usar Azure Service Health para comprender el estado general del servicio, incluidos los errores de zona, y puede configurar alertas de Service Health para notificarle problemas.

- **Solicitudes activas:** Las solicitudes en curso que se conectan a una instancia del plan de App Service en la zona de disponibilidad defectuosa se finalizan. Reintentar esas solicitudes.

- **Enrutamiento del tráfico:**App Service detecta las instancias perdidas de esa zona e intenta encontrar nuevas instancias de reemplazo Después de que App Service encuentre reemplazos, distribuye el tráfico entre las nuevas instancias según sea necesario.

  Si se configura el ajuste automático y se determina que se necesitan más instancias, se solicitan instancias a App Service. El comportamiento de escalabilidad automática funciona independientemente del comportamiento de la plataforma de App Service. Por lo tanto, la especificación del número de instancias no tiene por qué ser un múltiplo de

dos. Para más información, consulte Escalado vertical de una aplicación en App Service y Información general sobre el escalado automático.

> **Importante**
>
> Azure no garantiza que las solicitudes de más instancias se realicen correctamente en un escenario de zona descendente. La plataforma intenta reponer instancias perdidas de forma óptima. Si necesita capacidad garantizada durante un error de zona de disponibilidad, cree y configure sus planes de App Service para tener en cuenta la pérdida de zona mediante el **sobreaprovisionamiento de la capacidad**.

- **Comportamientos fuera del tiempo de ejecución:** Las aplicaciones en un plan de App Service con redundancia de zona continúan ejecutándose y prestando servicio al tráfico incluso si una zona de disponibilidad sufre una interrupción. Sin embargo, los comportamientos que no son de ejecución pueden verse afectados durante una interrupción de la zona de disponibilidad. Estos comportamientos incluyen el escalado del plan de App Service, la creación de aplicaciones, la configuración de la aplicación y la publicación de aplicaciones.

Recuperación de zona

Cuando se recupera la zona de disponibilidad, App Service crea automáticamente instancias en la zona de disponibilidad recuperada, quita las instancias temporales creadas en las otras zonas de disponibilidad y enruta el tráfico entre las instancias como de costumbre.

Prueba de fallos de zona

La plataforma de App Service administra el enrutamiento del tráfico, la conmutación por error y la conmutación por recuperación para planes de App Service con redundancia de zona. Esta característica está totalmente administrada, por lo que no es necesario iniciar ni validar los procesos de error de zona de disponibilidad.

Resistencia a errores en toda la región

App Service es un servicio de una sola región. Si la región deja de estar disponible, la aplicación tampoco está disponible.

Soluciones personalizadas de varias regiones para la resistencia

Para reducir el riesgo de un error de una sola región que afecte a la aplicación, puede implementar planes en varias regiones. Los pasos siguientes ayudan a reforzar la resistencia:

- Despliegue su aplicación en los planes de cada región.
- Configure directivas de equilibrio de carga y conmutación por error.
- Replique los datos entre regiones para que pueda recuperar el último estado de la aplicación.

Considere los siguientes recursos relacionados:

- Arquitectura de referencia: aplicación web de alta disponibilidad en múltiples regiones
- Enfoques que se deben tener en cuenta
- Tutorial: Crear una aplicación de alta disponibilidad en varias regiones en App Service

Copias de seguridad y restauración

Cuando use el nivel Básico o superior, puede hacer una copia de seguridad de sus aplicaciones de App Service en un archivo usando las funciones de copia de seguridad y restauración de App Service.

Estas funcionalidades resultan útiles cuando resulta complicado volver a implementar el código o cuando se almacena el estado en el disco. La mayoría de las soluciones no deberían depender exclusivamente de las copias de seguridad. En su lugar, use las otras capacidades de esta guía para satisfacer sus requisitos de resistencia. Sin embargo, las copias de seguridad protegen contra algunos riesgos que otros enfoques no.

> ⓘ **Importante**
>
> A partir del **31 de marzo de 2028**, las copias de seguridad personalizadas de Azure App Service **ya no admiten la copia de seguridad de bases de datos vinculadas**. Consulte Desuso de copias de seguridad de bases de datos vinculadas para obtener más información.
>
> En su lugar, use las herramientas nativas de copia de seguridad y restauración de la base de datos vinculada. Para más información, consulte Copia de seguridad y restauración de la aplicación en App Service.

Resistencia al mantenimiento del servicio

App Service realiza actualizaciones de servicio periódicas y otras tareas de mantenimiento. Para mantener la capacidad esperada durante una actualización, la plataforma agrega

automáticamente instancias adicionales del plan de App Service durante el proceso de actualización.

Habilitación de la redundancia de zona. Al habilitar la redundancia de zona en su plan de App Service, también mejora la resiliencia durante las actualizaciones de la plataforma. *Los dominios de actualización* constan de colecciones de máquinas virtuales que se desconectan durante una actualización y se asignan a zonas de disponibilidad. La implementación de varias instancias en su plan de App Service y la habilitación de la redundancia de zona para su plan agregan una capa adicional de resistencia en caso de que una instancia o zona sea incorrecta durante una actualización.

Para obtener más información, consulte Mantenimiento planeado rutinario para App Service y Mantenimiento rutinario para App Service, reinicios y tiempo de inactividad.

Acuerdo de nivel de servicio

El contrato de nivel de servicio (SLA) para los servicios de Azure describe la disponibilidad esperada de cada servicio y las condiciones que la solución deberá cumplir para lograr esa expectativa de disponibilidad. Para obtener más información, consulte Acuerdos de Nivel de Servicio para servicios en línea.

Al implementar un plan de App Service con redundancia de zona, aumenta el porcentaje de tiempo de actividad definido en el Acuerdo de Nivel de Servicio.

Contenido relacionado

- Confiabilidad en Azure
- Confiabilidad en App Service Environment

Last updated on 22/01/2026

Confiabilidad en Azure App Service Environment

App Service Environment es una característica de Azure App Service que proporciona un entorno totalmente aislado y dedicado para ejecutar aplicaciones de App Service de forma segura a gran escala. Como servicio de Azure, App Service Environment proporciona una variedad de funcionalidades para satisfacer sus requisitos de confiabilidad.

Cuando se usa Azure, la confiabilidad es una responsabilidad compartida. Microsoft proporciona una variedad de funcionalidades para admitir resistencia y recuperación. Es responsable de comprender cómo funcionan esas funcionalidades dentro de todos los servicios que usa y de seleccionar las funcionalidades que necesita para cumplir los objetivos empresariales y los objetivos de tiempo de actividad.

En este artículo se describe el soporte de confiabilidad en App Service Environment, incluida la resistencia a errores transitorios, fallos en zonas de disponibilidad e interrupciones a nivel regional. También se describen las estrategias de copia de seguridad y la resistencia al mantenimiento del servicio y se resalta cierta información clave sobre el acuerdo de nivel de servicio (SLA) de App Service Environment.

> ⓘ **Nota:**
>
> A diferencia de la oferta multiinquilino pública de App Service que comparte la infraestructura de soporte, una instancia de App Service Environment proporciona recursos de proceso dedicados y aislamiento de red mejorado para un único cliente.
>
> Un entorno ofrece las siguientes ventajas clave en materia de confiabilidad:
>
> - Recursos de proceso dedicados que no se comparten con otros clientes
> - Aislamiento de red mejorado para mejorar la seguridad y la estabilidad
> - La capacidad de implementar en su propia red virtual para un mayor control sobre el enrutamiento del tráfico y las directivas de seguridad.
>
> Para más información sobre la compatibilidad con la confiabilidad en App Service, consulte Confiabilidad en App Service.

Recomendaciones de implementación de producción para la confiabilidad

Se recomienda habilitar la redundancia de zona en el entorno y los planes de App Service, lo que requiere que los planes usen un mínimo de dos instancias.

Introducción a la arquitectura de confiabilidad

Cuando implementa un App Service Environment, implemente el entorno como contenedor para sus planes de App Service y aplicaciones web. Durante la configuración, configure los ajustes básicos de red y el aislamiento opcional del hardware. Elija si desea admitir la redundancia de zona en el entorno si la región admite zonas de disponibilidad.

Después de crear su entorno, puede crear uno o varios planes de App Service.

Un plan de App Service define un conjunto de recursos de proceso que ejecutan las aplicaciones web. Todas las Web Apps deben ejecutarse dentro de un plan. Puede escalar un plan para ejecutarse en varias *instancias* de máquina virtual, también denominadas *trabajadores*. Estas instancias proporcionan los recursos de proceso que ejecutan el código de su aplicación. Un único plan de App Service puede hospedar varias aplicaciones. Todas las aplicaciones se ejecutan en el mismo conjunto compartido de instancias de máquina virtual.

Para usar App Service Environment, sus planes deben usar el nivel de precios Aislado v2. Este nivel admite redundancia de zona y aplicaciones críticas a gran escala.

App Service ofrece las siguientes características de redundancia:

- **Distribución entre dominios de error:** En el nivel de plataforma, Azure distribuye automáticamente las instancias de máquina virtual del plan de App Service entre dominios de error dentro de la región de Azure. Esta distribución minimiza el riesgo de errores de hardware localizados mediante la agrupación de máquinas virtuales que comparten una fuente de alimentación común y un conmutador de red.

- **Distribución entre zonas de disponibilidad:** Si habilita la redundancia de zona en un plan de App Service compatible, Azure distribuye las instancias entre zonas de disponibilidad dentro de la región. Esta configuración proporciona una mayor resistencia si se produce una interrupción de zona. Para obtener más información sobre la redundancia de zonas, consulte Compatibilidad con zonas de disponibilidad.

- **Escalado de aplicaciones:** Al configurar el plan de App Service para ejecutar varias instancias de máquina virtual, todas las aplicaciones del plan se ejecutan en todas las instancias de forma predeterminada. Si configura su plan para el escalado automático, todas las aplicaciones se escalarán juntas según la configuración de escalado automático. Sin embargo, puede personalizar el número de instancias de plan que ejecutan una aplicación específica mediante el escalado por aplicación.

- **Unidades de escalado:** Internamente, App Service se ejecuta en una infraestructura de plataforma denominada *unidades de escalado*, también conocida como *stamps* o *espacios web*. Una unidad de escalado incluye todos los componentes necesarios para hospedar y ejecutar App Service, incluido el proceso, el almacenamiento, las redes y el equilibrio de carga. Azure administra las unidades de escalado para garantizar una distribución equilibrada de cargas de trabajo, realizar un mantenimiento rutinario y mantener la confiabilidad general de la plataforma.

  Es posible que algunas funcionalidades solo se apliquen a unidades de escalado específicas. Por ejemplo, algunas unidades de escalado de App Service pueden admitir la redundancia de zona, mientras que otras unidades de escala de la misma región no.

Resistencia a errores transitorios

Los errores transitorios son errores breves e intermitentes en los componentes. Se producen con frecuencia en un entorno distribuido como la nube y son una parte normal de las operaciones. Los errores transitorios se corrigen después de un breve período de tiempo. Es importante que las aplicaciones puedan controlar errores transitorios, normalmente mediante el reintento de solicitudes afectadas.

Todas las aplicaciones hospedadas en la nube deben seguir las instrucciones de control de errores transitorios de Azure cuando se comunican con cualquier API, bases de datos y otros componentes hospedados en la nube. Para obtener más información, consulte Recomendaciones para controlar errores transitorios.

Los SDK proporcionados por Microsoft suelen controlar errores transitorios. Dado que hospeda sus propias aplicaciones en App Service, siga los pasos necesarios para reducir la posibilidad de errores transitorios:

- **Implemente varias instancias en el plan.** App Service realiza actualizaciones automatizadas y otras formas de mantenimiento en instancias del plan. Si una instancia se vuelve incorrecta, el servicio puede reemplazar automáticamente esa instancia por una nueva instancia correcta. Durante el proceso de reemplazo, puede haber un breve período cuando la instancia anterior no está disponible y una nueva instancia no está lista para atender el tráfico. Para mitigar estos efectos, implemente varias instancias del plan de App Service.

- **Uso de ranuras de implementación.** Las ranuras de implementación de App Service permiten implementaciones sin tiempo de inactividad de las aplicaciones. Use ranuras de implementación para minimizar el efecto de las implementaciones y los cambios de configuración para los usuarios. Las ranuras de implementación también reducen la

probabilidad de que se reinicie la aplicación. Reiniciar la aplicación produce un error transitorio.

- **Evite escalar o reducir verticalmente.** Estas operaciones modifican la CPU, la memoria y otros recursos asignados a cada instancia, y pueden provocar el reinicio de la aplicación empresarial. En su lugar, seleccione un nivel y un tamaño de instancia que cumplan los requisitos de rendimiento bajo una carga típica. Para escalar y reducir horizontalmente, agregue y elimine instancias de forma dinámica para administrar los cambios en el volumen de tráfico.

Resistencia a errores de zona de disponibilidad

Las zonas de disponibilidad son grupos físicamente independientes de centros de datos dentro de una región de Azure. Cuando una zona falla, los servicios pueden transferirse a una de las zonas restantes.

Puede configurar su App Service Environment como *redundante por zona*. También puede configurar sus planes de App Service para que sean redundantes por zona, lo que los distribuye entre varias zonas de disponibilidad.

Sin embargo, puede habilitar o deshabilitar la redundancia de zona en cada plan. Esto significa que puede tener algunos planes en el entorno con redundancia de zona y otros que no.

Al crear un plan de App Service con redundancia de zona en el entorno, las instancias del plan de App Service se distribuyen entre las zonas de disponibilidad de la región. Para obtener más información, consulte Distribución de instancias entre zonas.

Requisitos

Para habilitar la redundancia de zona para su App Service Environment, debe cumplir los siguientes requisitos:

- **Compatibilidad con regiones:** Para ver qué regiones admiten zonas de disponibilidad para App Service Environment v3, consulte Regiones.
- **Tipo de plan:** Utilice tipos de plan Aislado versión 2.
- **Número mínimo de instancias:** Implemente un mínimo de dos instancias en el plan.
- **Unidad de escalado:** El entorno debe implementarse en una unidad de escalado que admita zonas de disponibilidad. No controla directamente la unidad de escalado que usa el entorno. En su lugar, al crear un entorno de App Service, el entorno se asigna a una unidad de escalado en función del grupo de recursos del entorno. Para determinar si la

unidad de escalado de App Service Environment admite redundancia de zona, compruebe si hay compatibilidad con la redundancia de zona para un entorno de App Service.

Si su entorno está en una unidad de escalado que no admite zonas de disponibilidad, no puede habilitar la redundancia de zona en su entorno ni en sus planes. En su lugar, debe crear un nuevo entorno en un nuevo grupo de recursos y volver a implementar las aplicaciones en nuevos planes dentro de ese entorno.

- **Requisitos de configuración:** Configure tanto su entorno del servicio de aplicaciones como sus planes para admitir la redundancia de zona. Puede habilitar la redundancia de zona durante la creación del entorno o actualizando un entorno existente.

Distribución de instancias a través de zonas

Al crear un plan de App Service con redundancia de zona, Azure distribuye las instancias del plan entre zonas de disponibilidad de la región. Esta distribución garantiza que sus aplicaciones sigan estando disponibles incluso si una zona sufre una interrupción del servicio.

La distribución de instancias en una implementación con redundancia de zona sigue reglas específicas. Estas reglas también se aplican a medida que la aplicación se escala y se reduce horizontalmente:

- **Instancias mínimas:** El plan de App Service debe tener un mínimo de dos instancias para la redundancia de zona.

- **Zonas de disponibilidad máxima compatibles con el plan:** Azure determina el número de zonas de disponibilidad que puede usar el plan, lo que se conoce como *maximumNumberOfZones*. Para ver el número de zonas de disponibilidad que puede usar su plan específico, consulte Comprobar la compatibilidad con la redundancia de zonas para un plan de App Service.

- **Distribución de instancias:** Cuando se habilita la redundancia de zona, Azure distribuye instancias de plan entre varias zonas de disponibilidad automáticamente. La distribución se basa en las reglas siguientes:

  - Si el número de instancias supera el *maximumNumberOfZones* y divide uniformemente, Azure distribuye las instancias uniformemente entre zonas.

  - Si el número de instancias no se divide de manera uniforme, Azure distribuye las instancias restantes entre las zonas restantes.

  - Cuando la plataforma de App Service asigna instancias para un plan de App Service con redundancia de zona, usa el equilibrio de zona de mejor esfuerzo que

proporcionan los conjuntos de escalado de máquinas virtuales de Azure subyacentes. Un plan se equilibra si cada zona tiene el mismo número de máquinas virtuales o difiere en una instancia de todas las demás zonas. Para obtener más información, consulte Equilibrio de zona.

- **Ubicación de zona física:** Puede ver la zona de disponibilidad física que se usa para cada una de las instancias del plan de App Service. Para más información, consulte Visualización de zonas físicas para un plan de App Service.

Consideraciones

Una interrupción de la zona de disponibilidad puede afectar a algunos aspectos de App Service, aunque la aplicación siga sirviendo al tráfico. Estos comportamientos incluyen el escalado del plan de App Service, la creación de aplicaciones, la configuración de la aplicación y la publicación de aplicaciones.

Al habilitar la redundancia de zona en su plan de App Service, también mejora la resiliencia durante las actualizaciones de la plataforma. Para obtener más información, consulte Resistencia al mantenimiento del servicio.

En los planes de App Service que no son redundantes por zona, las instancias de máquina virtual subyacentes no son resistentes a los errores de la zona de disponibilidad. Pueden experimentar tiempo de inactividad durante una interrupción en cualquier zona de esa región.

Cost

Puede habilitar la redundancia de zona en un App Service Environment o sus planes sin coste adicional. Sin embargo, la redundancia de zona de un plan requiere que tenga dos o más instancias. Se le cobrará en función de la SKU del plan de App Service, la capacidad que especifique y las instancias a las que escale según los criterios de escalado automático.

Si habilita las zonas de disponibilidad pero especifica una capacidad inferior a dos instancias, la plataforma impone un mínimo de dos instancias. La plataforma le cobra por esas dos instancias.

> **ⓘ Importante**
>
> Al habilitar zonas de disponibilidad para un App Service Environment, todos los planes de App Service con menos de 3 instancias se escalan a 3 instancias. Cualquier plan con 3 o más instancias permanece sin cambios. Una vez completada la operación para habilitar las

zonas de disponibilidad, puede ajustar los planes de App Service según sea necesario, incluyendo menos de 3 instancias.

Configurar soporte de zonas de disponibilidad

Para obtener información sobre cómo crear, habilitar o deshabilitar un nuevo App Service Environment redundante en zonas y nuevos planes de App Service redundantes en zonas, consulte Configurar App Service Environments y planes de App Service aislados v2 para la redundancia en zonas.

> **Nota:**
>
> Un cambio en el estado de redundancia de zona de un App Service Environment tarda entre 12 y 24 horas en completarse. Durante el proceso de actualización, no se produce inactividad ni problemas de rendimiento.

Planeamiento y administración de capacidad

Para prepararse para errores de zona de disponibilidad, considere la posibilidad de *sobreaprovisionar* la capacidad del plan de App Service. Este enfoque permite que la solución tolere cierta pérdida de capacidad y siga funcionando sin que se vea afectado su rendimiento. Para obtener más información, consulte Administración de la capacidad mediante el aprovisionamiento excesivo.

Comportamiento cuando todas las zonas están en buen estado

En la siguiente lista se describe qué esperar cuando los planes de App Service están configurados para la redundancia de zona y todas las zonas de disponibilidad están operativas:

- **Enrutamiento del tráfico entre zonas:** Durante el funcionamiento normal, el tráfico se enruta entre todas las instancias disponibles del plan de App Service en todas las zonas de disponibilidad.
- **Replicación de datos entre zonas:** Durante las operaciones normales, cualquier estado almacenado en el sistema de archivos de la aplicación se almacena en el almacenamiento con redundancia de zona y se replica sincrónicamente entre zonas de disponibilidad.

Comportamiento durante un fallo de zona

Una interrupción de la zona de disponibilidad puede afectar a algunos aspectos de App Service, aunque la aplicación siga sirviendo al tráfico. Estos comportamientos incluyen el escalado del plan de App Service, la creación de aplicaciones, la configuración de la aplicación y la publicación de aplicaciones.

La siguiente lista describe qué puede ocurrir cuando los planes de App Service están configurados para la redundancia de zona y una o varias zonas de disponibilidad no están disponibles:

- **Detección y respuesta:** La plataforma de App Service detecta automáticamente errores en una zona de disponibilidad e inicia una respuesta. No se requiere ninguna intervención manual para iniciar una conmutación por error de zona.

- **Notificación**: Microsoft no le notifica automáticamente cuando una zona está inactiva. Sin embargo, puede usar Azure Resource Health para supervisar el estado de un recurso individual y puede configurar alertas de Resource Health para notificarle problemas. También puede usar Azure Service Health para comprender el estado general del servicio, incluidos los errores de zona, y puede configurar alertas de Service Health para notificarle problemas.

- **Solicitudes activas:** Las solicitudes en curso que se conectan a una instancia del plan de App Service en la zona de disponibilidad defectuosa se finalizan. Reintentar esas solicitudes.

- **Enrutamiento del tráfico:**App Service detecta las instancias perdidas de esa zona e intenta encontrar nuevas instancias de reemplazo Después de que App Service encuentre reemplazos, distribuye el tráfico entre las nuevas instancias según sea necesario.

  Si se configura el ajuste automático y se determina que se necesitan más instancias, se solicitan instancias a App Service. El comportamiento de escalabilidad automática funciona independientemente del comportamiento de la plataforma de App Service. Por lo tanto, la especificación del número de instancias no tiene por qué ser un múltiplo de dos. Para más información, consulte Escalado vertical de una aplicación en App Service y Información general sobre el escalado automático.

  > ⓘ **Importante**
  >
  > Azure no garantiza que las solicitudes de más instancias se realicen correctamente en un escenario de zona descendente. La plataforma intenta reponer instancias perdidas de forma óptima. Si necesita capacidad garantizada durante un error de zona de disponibilidad, cree y configure sus planes de App Service para tener en cuenta la pérdida de zona mediante el **sobreaprovisionamiento de la capacidad**.

- **Comportamientos fuera del tiempo de ejecución:** Las aplicaciones en un plan de App Service con redundancia de zona continúan ejecutándose y prestando servicio al tráfico incluso si una zona de disponibilidad sufre una interrupción. Sin embargo, los comportamientos que no son de ejecución pueden verse afectados durante una interrupción de la zona de disponibilidad. Estos comportamientos incluyen el escalado del plan de App Service, la creación de aplicaciones, la configuración de la aplicación y la publicación de aplicaciones.

Recuperación de zona

Cuando se recupera la zona de disponibilidad, App Service crea automáticamente instancias en la zona de disponibilidad recuperada, quita las instancias temporales creadas en las otras zonas de disponibilidad y enruta el tráfico entre las instancias como de costumbre.

Prueba de fallos de zona

La plataforma de App Service administra el enrutamiento del tráfico, la conmutación por error y la conmutación por recuperación para planes de App Service con redundancia de zona. Esta característica está totalmente administrada, por lo que no es necesario iniciar ni validar los procesos de error de zona de disponibilidad.

Resistencia a errores en toda la región

App Service es un servicio de una sola región. Si la región deja de estar disponible, su entorno y sus planes y aplicaciones también dejarán de estar disponibles.

Soluciones personalizadas de varias regiones para la resistencia

Para reducir el riesgo de un error de una sola región que afecte a la aplicación, implemente varias instancias de App Service Environment en varias regiones. Los pasos siguientes ayudan a reforzar la resistencia:

- Implemente la aplicación en los entornos de App Service en cada región.
- Configure directivas de equilibrio de carga y conmutación por error.
- Replique los datos entre regiones para que pueda recuperar el último estado de la aplicación.

Para obtener un enfoque de ejemplo que ilustra esta arquitectura, consulte Implementación empresarial de alta disponibilidad mediante App Service Environment.

Copias de seguridad y restauración

Para realizar una copia de seguridad de las aplicaciones de App Service en un archivo, use las funcionalidades de copia de seguridad y restauración de App Service.

Estas funcionalidades resultan útiles cuando resulta complicado volver a implementar el código o cuando se almacena el estado en el disco. La mayoría de las soluciones no deberían depender exclusivamente de las copias de seguridad. En su lugar, use las otras capacidades de esta guía para satisfacer sus requisitos de resistencia. Sin embargo, las copias de seguridad protegen contra algunos riesgos que otros enfoques no.

> ⓘ **Importante**
>
> A partir del **31 de marzo de 2028**, las copias de seguridad personalizadas de Azure App Service **ya no admiten la copia de seguridad de bases de datos vinculadas**. Consulte Desuso de copias de seguridad de bases de datos vinculadas para obtener más información.
>
> En su lugar, use las herramientas nativas de copia de seguridad y restauración de la base de datos vinculada. Para más información, consulte Copia de seguridad y restauración de la aplicación en App Service.

Resistencia al mantenimiento del servicio

App Service realiza actualizaciones de servicio periódicas y otras tareas de mantenimiento. Para mantener la capacidad esperada durante una actualización, la plataforma agrega automáticamente instancias adicionales del plan de App Service durante el proceso de actualización.

Habilitación de la redundancia de zona. Al habilitar la redundancia de zona en su plan de App Service, también mejora la resiliencia durante las actualizaciones de la plataforma. *Los dominios de actualización* constan de colecciones de máquinas virtuales que se desconectan durante una actualización y se asignan a zonas de disponibilidad. La implementación de varias instancias en su plan de App Service y la habilitación de la redundancia de zona para su plan agregan una capa adicional de resistencia en caso de que una instancia o zona sea incorrecta durante una actualización.

Personalice el ciclo de actualización. Puede personalizar el ciclo de actualización de un entorno de App Service. Si necesita validar el efecto de las actualizaciones en su carga de trabajo, habilite las actualizaciones manuales. Este enfoque le permite realizar validaciones y

pruebas en una instancia que no es de producción antes de aplicarlas a su instancia de producción.

Para obtener más información sobre las preferencias de mantenimiento, consulte Preferencias de actualización para el mantenimiento planificado de App Service Environment.

Acuerdo de nivel de servicio

El contrato de nivel de servicio (SLA) para los servicios de Azure describe la disponibilidad esperada de cada servicio y las condiciones que la solución deberá cumplir para lograr esa expectativa de disponibilidad. Para obtener más información, consulte Acuerdos de Nivel de Servicio para servicios en línea.

Al implementar un plan de App Service con redundancia de zona, aumenta el porcentaje de tiempo de actividad definido en el Acuerdo de Nivel de Servicio.

Contenido relacionado

- Confiabilidad en Azure
- Confiabilidad en App Service

Last updated on 22/01/2026

Fiabilidad en Azure Batch

En este artículo se describe el soporte de confiabilidad en Azure Batch. Describe cómo mejorar la resistencia intrar regional mediante zonas de disponibilidad, grupos de lotes y nodos de proceso para minimizar el tiempo de inactividad y la pérdida de datos. También se vincula a información sobre la recuperación entre regiones y la continuidad empresarial.

Soporte para zonas de disponibilidad

Las zonas de disponibilidad son grupos físicamente independientes de centros de datos dentro de una región de Azure. Cuando una zona falla, los servicios pueden transferirse a una de las zonas restantes.

Batch mantiene la paridad con Azure con la compatibilidad de las zonas de disponibilidad.

Prerrequisitos

- Para las cuentas de Batch del modo de suscripción de usuario, asegúrese de que la suscripción en la que va a crear el grupo no tenga una restricción de oferta de zona en la SKU de máquina virtual solicitada. Para ver si la suscripción no tiene restricciones, llame a la API de enumeración de SKU de recursos y compruebe el campo `ResourceSkuRestrictions`. Si existe una restricción de zona, puede enviar una incidencia de soporte técnico para quitar la restricción de zona.

- Como InfiniBand no admite la comunicación entre zonas, no puede crear un grupo con una directiva de zona si tiene habilitada la comunicación entre nodos y usa una SKU de máquina virtual que admite InfiniBand.

- Batch mantiene la paridad con Azure con la compatibilidad de las zonas de disponibilidad. Para usar la opción de zona, el grupo se debe crear en una región de Azure compatible con las zonas de disponibilidad.

- Para asignar el grupo de Batch en las zonas de disponibilidad, la región de Azure en la que se creó el grupo debe admitir la SKU de máquina virtual solicitada en más de una zona. Para validar que la región admite la SKU de máquina virtual solicitada en más de una zona, llame a la API de enumeración de SKU de recursos y compruebe el campo `locationInfo` de `resourceSku`. Asegúrese de que se admite más de una zona para la SKU de máquina virtual solicitada. También puede usar la CLI de Azure para enumerar todas las SKU de recursos disponibles con el siguiente comando:

  Azure CLI

```
az vm list-skus
```

Creación de un grupo de Azure Batch en zonas de disponibilidad

Para obtener ejemplos sobre cómo crear un grupo de Batch en zonas de disponibilidad, consulte Creación de un grupo de Azure Batch en zonas de disponibilidad.

Obtenga más información acerca de cómo crear cuentas de Batch con Azure Portal, la CLI de Azure, Powershell o Batch Management API.

Experiencia a nivel de zona

Durante la interrupción de una zona, los nodos de esa zona dejan de estar disponibles. Los nodos de ese mismo grupo de nodos de otras zonas no se ven afectados y siguen estando disponibles.

La cuenta de Azure Batch no reasigna ni crea nodos nuevos para compensar los nodos que han quedado fuera de servicio debido a la interrupción. Los usuarios deben agregar más nodos al grupo de nodos, que luego se asignan desde otras zonas en estado correcto.

Tolerancia a errores

Para prepararse para un posible error de zona de disponibilidad, debe sobreaprovisionar la capacidad del servicio para asegurarse de que la solución puede tolerar un tercio de la pérdida de capacidad y seguir funcionando sin degradar el rendimiento durante interrupciones en toda la zona. Como la plataforma distribuye las máquinas virtuales entre tres zonas y debe tener en cuenta al menos el error de una de ellas, multiplique el número máximo de instancias de carga de trabajo por un factor de zonas/(zonas-1) o 3/2. Por ejemplo, si la carga de trabajo máxima típica necesita cuatro instancias, debería aprovisionar seis instancias: (2/3 * 6 instancias) = 4 instancias.

Migración de zonas de disponibilidad

No se puede migrar un grupo de Batch existente a la compatibilidad con la zona de disponibilidad. Si desea recrear el grupo de Batch entre zonas de disponibilidad, consulte Creación de un grupo de Azure Batch entre zonas de disponibilidad.

Recuperación ante desastres entre regiones y continuidad empresarial

Azure Batch está disponible en todas las regiones de Azure. Sin embargo cuando se crea una cuenta de Batch, se debe asociar con una región específica. Todas las operaciones posteriores para dicha cuenta de Batch solo se aplican a esa región. Por ejemplo, se crean los grupos y las máquinas virtuales (VM) asociadas en la misma región que la cuenta de Batch.

Al diseñar una aplicación que usa Batch, debe considerar la posibilidad de que Batch no esté disponible en una región. Es posible que encuentre una situación poco frecuente en la que haya un problema con la región como un todo, con el servicio completo de Batch en la región o con su cuenta de Batch específica.

Si la aplicación o solución que usa Batch siempre debe estar disponible, se debe diseñar para conmutarse por error a otra región o para que la carga de trabajo siempre se divida en dos o más regiones. Ambos enfoques requieren al menos dos cuentas de Batch y que cada cuenta se encuentre en una región distinta.

La responsabilidad de configurar la recuperación ante desastres entre regiones con Azure Batch es suya. Si ejecuta varias cuentas de Batch en regiones específicas y aprovecha las zonas de disponibilidad, la aplicación puede cumplir los objetivos de recuperación ante desastres cuando una de las cuentas de Batch deje de estar disponible.

Al proporcionar la capacidad de conmutación por error a una región alternativa, se deben tener en cuenta todos los componentes de una solución. No basta con tener simplemente una segunda cuenta de Batch. Por ejemplo, en la mayoría de las aplicaciones de Batch, se requiere una cuenta de almacenamiento de Azure. La cuenta de almacenamiento y la cuenta de Batch deben estar en la misma región para lograr un rendimiento aceptable.

Tenga en cuenta lo siguiente a la hora de diseñar una solución con conmutación por error:

- Cree previamente todos los servicios necesarios en cada región, como la cuenta de Batch y la de almacenamiento. No se suele aplicar ningún cargo por tener cuentas creadas y los cargos se generan solo cuando se usa la cuenta o se almacenan datos.
- Asegúrese con tiempo de que se establecen las cuotas adecuadas para todas las cuentas de Batch de **suscripción de usuario** para asignar el número necesario de núcleos mediante la cuenta de Batch.
- Use scripts o plantillas para automatizar la implementación de la aplicación en una región.

- Mantenga actualizados los archivos binarios de aplicación y los datos de referencia en todas las regiones. Mantenerlos actualizados garantizará que la región se pueda conectar rápidamente sin necesidad de esperar a que los archivos se carguen y se implementen. Por ejemplo, considere el caso en el que se almacena una aplicación personalizada para instalar en los nodos del grupo y se hace referencia a ella mediante paquetes de aplicación de Batch. Cuando se publica una actualización de la aplicación, se debe cargar en cada cuenta de Batch y hacer referencia a ella mediante la configuración del grupo (o convertir la versión más reciente en la versión predeterminada).

- En la aplicación que realice llamadas a Batch, al almacenamiento y a cualquier otro servicio, cambie los clientes o la carga con facilidad a regiones distintas.

- Considere la posibilidad de cambiar con frecuencia a una región alternativa como parte del funcionamiento normal. Por ejemplo, si tiene dos implementaciones en regiones distintas, cambie a la región alternativa cada mes.

La duración del tiempo de recuperación ante un desastre depende de la configuración que elija. Batch es independiente respecto a si usa varias cuentas o una sola cuenta. En las configuraciones de tipo activo-activo, en las que dos instancias de Batch reciben tráfico simultáneamente, la recuperación ante desastres es más rápida que en el caso de una configuración de tipo activo-pasivo. La configuración que elija debe basarse en las necesidades empresariales (regiones diferentes, requisitos de latencia) y las consideraciones técnicas.

Recuperación ante desastres de una sola región

La forma de implementar la recuperación ante desastres en Batch es la misma, tanto si trabaja en una zona geográfica de una sola región o de varias regiones. Las únicas diferencias son las SKU que usa para el almacenamiento y si tiene previsto usar la misma cuenta de almacenamiento o una diferente en todas las regiones.

Pruebas de recuperación ante desastres

Debe realizar sus propias pruebas de recuperación ante desastres de la solución habilitada para Batch. Se considera un procedimiento recomendado para permitir cambiar de forma sencilla entre la carga de cliente y servicio en diferentes regiones.

Probar el plan de recuperación ante desastres para Batch puede ser tan sencillo como alternar cuentas de Batch. Por ejemplo, podría confiar en una sola cuenta de Batch en una región específica durante un día operativo. A continuación, el día siguiente, podría cambiar a una segunda cuenta de Batch en otra región. La recuperación ante desastres se administra principalmente en el lado cliente. Este enfoque de varias cuentas para la recuperación ante

desastres se encarga de las expectativas de RTO y RPO en zonas geográficas de una sola región o de varias regiones.

Capacidad y resistencia proactiva de la recuperación ante desastres

Microsoft y sus clientes operan bajo el modelo de responsabilidad compartida. Microsoft es responsable de la resistencia de la plataforma y la infraestructura. La responsabilidad de abordar la recuperación ante desastres para cualquier servicio específico que implemente y controle es suya. Para asegurarse de que la recuperación sea proactiva:

- Siempre debe implementar previamente recursos secundarios. La implementación previa de recursos secundarios es necesaria porque no hay ninguna garantía de capacidad en el momento del impacto para aquellos que no han asignado previamente dichos recursos.
- Cree previamente todos los servicios necesarios en cada región, como las cuentas de Batch y las cuentas de almacenamiento asociadas. No se aplica ningún cargo por crear nuevas cuentas; solo se generan cargos cuando se usa la cuenta o se almacenan datos.
- Asegúrese de establecer las cuotas adecuadas en todas las suscripciones de antemano, de forma que pueda asignar el número necesario de núcleos mediante la cuenta de Batch. Al igual que en otros servicios de Azure, existen límites en determinados recursos asociados con el servicio Batch. Muchos de ellos son cuotas predeterminadas que Azure aplica en el nivel de cuenta o suscripción. Tenga en cuenta estas cuotas al diseñar y escalar verticalmente sus cargas de trabajo de Batch.

> ⓘ **Nota:**
>
> Si planea ejecutar cargas de trabajo de producción en Batch, es posible que tenga que aumentar el valor predeterminado de una o varias de las cuotas. Para generar una cuota, puede solicitar un aumento de la cuota sin cargo alguno. Para obtener más información, consulte **Solicitud de un aumento de cuota**.

Storage

Debe configurar el almacenamiento de Batch para asegurarse de que se realiza una copia de seguridad de los datos entre regiones; la responsabilidad del cliente es el enfoque predeterminado. La mayoría de las soluciones de Batch usan Azure Storage para almacenar los archivos de recursos y los archivos de salida. Por ejemplo, las tareas de Batch (incluidas las tareas estándar, las de inicio, las de preparación de trabajos y las de liberación de trabajos) especifican normalmente archivos de recursos que residen en cuentas de almacenamiento. Las

cuentas de almacenamiento también almacenan los datos que se procesan y los datos de salida que se generan. Comprender la posible pérdida de datos en las regiones de las operaciones de servicio es una consideración importante. También debe confirmar si los datos se pueden reescribir o si son de solo lectura.

Batch admite los siguientes tipos de cuentas de almacenamiento de Azure:

- Cuentas de uso general v2 (GPv2)
- Cuentas de uso general v1 (GPv1)
- Cuentas de Blob Storage (actualmente admitidas para grupos en la configuración de máquina virtual)

Para más información sobre las cuentas de almacenamiento, vea Introducción a las cuentas de Azure Storage.

Puede asociar una cuenta de almacenamiento con su cuenta de Batch cuando crea la cuenta o llevar a cabo este paso en otro momento.

Si va a configurar una cuenta de almacenamiento independiente para cada región en la que está disponible el servicio, debe usar cuentas de almacenamiento con redundancia de zona (ZRS). Use cuentas de almacenamiento con redundancia de zona geográfica (GZRS) si va a usar la misma cuenta de almacenamiento en varias regiones emparejadas. En el caso de las zonas geográficas que contienen una sola región, debe crear una cuenta de almacenamiento con redundancia de zona (ZRS), ya que el almacenamiento con redundancia de zona geográfica no está disponible.

El planeamiento de capacidad es otra consideración importante respecto al almacenamiento y debe abordarse de forma proactiva. Al elegir una cuenta de almacenamiento, tenga en cuenta los requisitos de costo y rendimiento. Por ejemplo, las opciones de cuenta GPv2 y de Blob Storage admiten mayores límites de capacidad y escalabilidad si se compara con GPv1. (Póngase en contacto con el servicio de soporte técnico de Azure para solicitar un aumento en el límite de almacenamiento). Estas opciones de cuenta pueden mejorar el rendimiento de las soluciones de Batch que contienen un gran número de tareas en paralelo que se leen o escriben en la cuenta de almacenamiento.

Cuando una cuenta de almacenamiento está vinculada a una cuenta de Batch, imagínese que se trata de la cuenta de almacenamiento automático. Se requiere una cuenta de almacenamiento automático si tiene previsto usar la funcionalidad de paquetes de aplicación, ya que se usa para almacenar los archivos ZIP de los paquetes de aplicación. También se puede usar una cuenta de almacenamiento automático para los archivos de recursos de tarea; dado que la cuenta de almacenamiento automático ya está vinculada a la cuenta de Batch, esto evita la necesidad de que las direcciones URL de firma de acceso compartido (SAS) accedan a los archivos de recursos.

Pasos siguientes

- Confiabilidad en Azure

Last updated on 22/01/2026

Confiabilidad en Azure Container Apps

Azure Container Apps es un servicio de hospedaje de contenedores sin servidor totalmente administrado para implementar microservicios y aplicaciones en contenedor.

Cuando se usa Azure, la confiabilidad es una responsabilidad compartida. Microsoft proporciona una variedad de funcionalidades para admitir resistencia y recuperación. Es responsable de comprender cómo funcionan esas funcionalidades dentro de todos los servicios que usa y de seleccionar las funcionalidades que necesita para cumplir los objetivos empresariales y los objetivos de tiempo de actividad.

En este artículo se describe cómo hacer que Las aplicaciones de contenedor sean resistentes a varias posibles interrupciones y problemas, incluidos errores transitorios, interrupciones de zona de disponibilidad, interrupciones de regiones y mantenimiento del servicio. También se describe cómo usar copias de seguridad para recuperarse de otros tipos de problemas y se resalta información clave sobre el acuerdo de nivel de servicio (SLA) de Container Apps.

Recomendaciones de implementación de producción

Para obtener información sobre cómo implementar Container Apps para admitir los requisitos de confiabilidad de la solución y cómo afecta la confiabilidad a otros aspectos de la arquitectura, consulte Procedimientos recomendados de arquitectura para Aplicaciones de contenedor en Azure Well-Architected Framework.

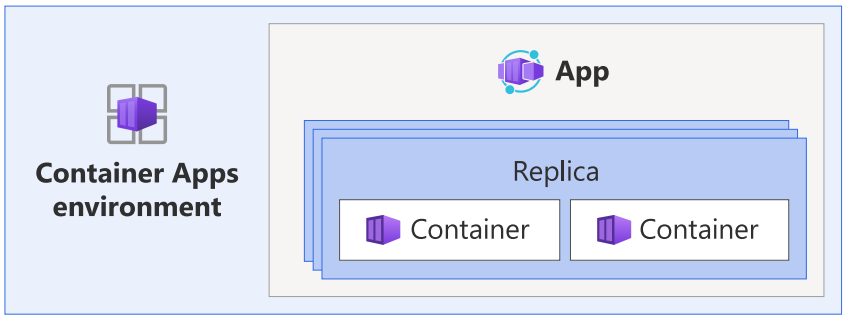
Introducción a la arquitectura de confiabilidad

Al usar Container Apps, implementará un *entorno* que actúa como unidad de implementación básica y define un límite seguro alrededor de un grupo de aplicaciones contenedoras. El entorno es donde se configuran los valores principales, incluida la compatibilidad de la zona de disponibilidad y la configuración de red. Los dos tipos de entornos son entornos de perfiles de carga de trabajo y entornos de solo consumo. Para obtener más información, consulte Estructuras de proceso y facturación en Container Apps.

Puede implementar varias *aplicaciones* en un único entorno. Cada aplicación ejecuta uno o varios *contenedores*. Un entorno también puede ejecutar uno o varios *trabajos*, que representan tareas no interactivas. Para obtener más información, consulte Contenedores en Aplicaciones de contenedor y trabajos en Aplicaciones de contenedor.

Cada aplicación tiene una o varias *réplicas*, que representan las instancias en ejecución de la aplicación. Puedes controlar cómo se escala la aplicación, incluido el número mínimo y máximo

de réplicas y cómo la aplicación agrega y quita réplicas dinámicamente. El planificador de la plataforma garantiza una distribución óptima entre hosts físicos al mismo tiempo que cumple con los requisitos mínimos de cantidad de réplicas. Para obtener más información, consulte Establecimiento de reglas de escalado en Container Apps.



Container Apps apoya la fiabilidad de las aplicaciones mediante diferentes capacidades.

- **Supervisión automática del estado:** El controlador de ingreso integrado equilibra automáticamente la carga del tráfico entre réplicas saludables. Si se produce un error en las comprobaciones de estado de una réplica o su infraestructura subyacente deja de estar disponible durante un tiempo prolongado, el servicio reinicia automáticamente los contenedores con errores o crea réplicas de reemplazo. También redistribuye el tráfico fuera de las réplicas no saludables y administra los reintentos de red en el clúster. Este proceso de recuperación automática no requiere ninguna intervención del cliente y mantiene el recuento de réplicas especificado. Para más información, consulte Sondeos de estado.

- **Resistencia de la aplicación a través de Dapr:** Container Apps proporciona una estrecha integración con Dapr, que es un marco que admite microservicios de nivel de producción y aplicaciones en contenedor. Dapr incluye características que ayudan a mejorar la resistencia, incluido el control de errores en otros servicios. Para obtener más información, consulte Microservicios con Aplicaciones de Contenedores.

- **Resistencia de la infraestructura para los componentes del sistema:** Esta resistencia incluye el plano de control, los controladores de entrada y el entorno de ejecución del contenedor. En las regiones que tienen zonas de disponibilidad, Container Apps proporciona redundancia de zona. Para obtener más información, consulte Resilience to availability zone failures (Resistencia a errores de zona de disponibilidad).

Resistencia a errores transitorios

Los errores transitorios son errores breves e intermitentes en los componentes. Se producen con frecuencia en un entorno distribuido como la nube y son una parte normal de las operaciones. Los errores transitorios se corrigen después de un breve período de tiempo. Es

importante que las aplicaciones puedan controlar errores transitorios, normalmente mediante el reintento de solicitudes afectadas.

Todas las aplicaciones hospedadas en la nube deben seguir las instrucciones de control de errores transitorios de Azure cuando se comunican con cualquier API, bases de datos y otros componentes hospedados en la nube. Para obtener más información, consulte Recomendaciones para controlar errores transitorios.

Container Apps gestiona automáticamente muchos errores transitorios a través de sus mecanismos de reintento a nivel de plataforma y supervisión de salud. Para asegurarse de que las aplicaciones son resistentes a errores transitorios, realice las siguientes acciones:

- **Configure sondeos de estado** que permitan a la plataforma detectar y responder a condiciones de error específicas de la aplicación. Establezca los umbrales de error y los valores de tiempo de espera adecuados en función de las características de inicio de la aplicación. Por ejemplo, para evitar reinicios prematuros del contenedor durante problemas temporales, use un umbral de fallos de 3 con un período de 10 segundos para los probes de vivacidad. Para más información, consulte Sondeos de estado.

- **Utilice directivas de resiliencia para el descubrimiento de servicios (versión preliminar)** para prevenir, detectar y recuperar proactivamente de los fallos en las solicitudes de servicio. Por ejemplo, cuando se usa una directiva de resistencia, cada solicitud entrante a la aplicación se puede reintentar automáticamente si hay un error transitorio que impide que la aplicación responda. Para más información, consulte Resistencia de detección de servicios (versión preliminar).

- **Implemente la lógica de reintento** en las aplicaciones para llamadas de servicio externas, conexiones de base de datos y solicitudes de API.

  Si la aplicación usa Dapr para integrarse con los servicios en la nube, use la resistencia de componentes de Dapr (versión preliminar) para configurar reintentos, tiempos de espera y disyuntores.

  Para otras dependencias, la aplicación debe controlar los errores transitorios. Use estrategias de retroceso exponencial y patrones de cortacircuitos al llamar a servicios externos para evitar fallos en cascada durante interrupciones en servicios posteriores. Las características integradas de detección y equilibrio de carga de Container Apps enrutan automáticamente el tráfico de las instancias con errores, pero las políticas de reintento a nivel de aplicación garantizan una gestión eficaz de manera elegante de los errores transitorios antes de que las comprobaciones de estado a nivel de plataforma desencadenen el reinicio del contenedor.

- **Diseñe trabajos para ser resistentes a errores transitorios**, incluidos los errores durante la ejecución del trabajo o en sus dependencias. Diseñe los trabajos para reanudarlos si se

reinician, o diseñe para la idempotencia de modo que se puedan ejecutar nuevamente de forma segura.

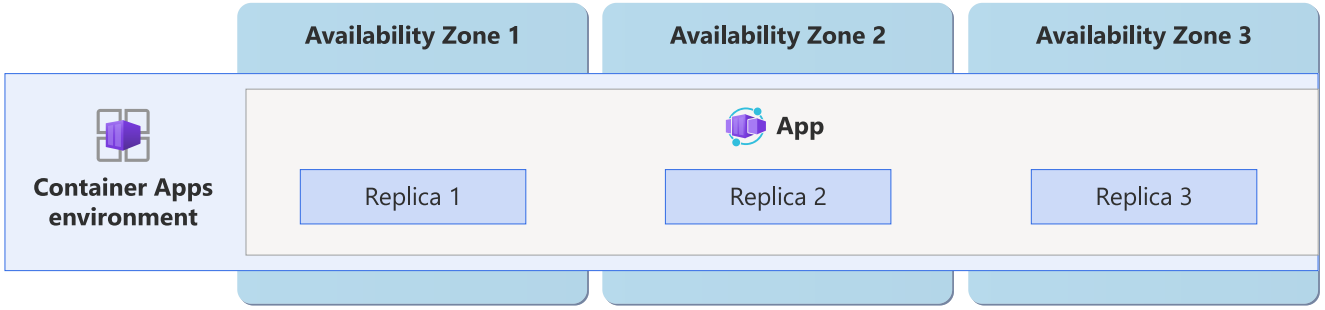
Resistencia a errores de zona de disponibilidad

Las zonas de disponibilidad son grupos físicamente independientes de centros de datos dentro de una región de Azure. Cuando una zona falla, los servicios pueden transferirse a una de las zonas restantes.

Al crear un entorno de Container Apps, puede habilitar la *redundancia de zona* para distribuir la infraestructura subyacente en varias zonas de disponibilidad en la región de Azure elegida. Container Apps programa automáticamente las réplicas de las aplicaciones entre zonas. Esta distribución se produce de forma transparente, lo que significa que no es necesario especificar la ubicación de zona para las réplicas individuales.

La redundancia de zona mejora la resistencia de la aplicación a errores de nivel de zona al garantizar que las réplicas de la aplicación contenedor se reparten entre varias zonas.

En el diagrama siguiente se muestra una aplicación contenedora con redundancia de zona con tres réplicas. Cada réplica se ejecuta en una zona de disponibilidad independiente.



Requisitos

- **Compruebe la compatibilidad con regiones.** La redundancia de zona está disponible en todas las regiones que admiten aplicaciones de contenedores y zonas de disponibilidad.

  Para ver qué regiones admiten zonas de disponibilidad, consulte Regiones de Azure con compatibilidad con zonas de disponibilidad.

  Para ver qué regiones admiten Container Apps, consulte Disponibilidad del producto por región.

- **Use perfiles de carga de trabajo.** La redundancia de zona está disponible para todos los planes de Aplicaciones de Contenedor, incluidos los perfiles de carga de trabajo de consumo y dedicado.

- **Habilite la redundancia de zona durante la creación del entorno.** Esta configuración no se puede cambiar después de crear el entorno.
- **Implemente un entorno de Container Apps en una red virtual.** La red virtual debe estar en una región que admita zonas de disponibilidad. Asegúrese de que la red virtual tiene una subred de tamaño adecuado. Los entornos de solo consumo necesitan una subred con un `/23` intervalo de enrutamiento de Inter-Domain sin clases (CIDR) o mayor, mientras que los entornos de perfiles de carga de trabajo requieren un `/27` intervalo CIDR o superior.
- **Establezca el número mínimo de réplicas en al menos dos para garantizar la distribución entre varias zonas de disponibilidad.** Considere la posibilidad de establecer un número mínimo de réplicas mayor si se aplica alguna de las condiciones siguientes:
  - La carga máxima esperada necesita más de dos réplicas.
  - Debe ser resistente a varias interrupciones de zona simultáneas.
  - Quieres minimizar el tiempo que esperas a que se creen nuevas réplicas en otras zonas durante una interrupción de zona.

Cost

No incurre en cargos adicionales más allá de los precios estándar de Container Apps al habilitar la redundancia de zona. Se pagan las mismas tarifas por los recursos de computación, las solicitudes y los segundos de vCore, independientemente de si la redundancia de zona está habilitada o no. Para más información, consulte Precios de Container Apps y facturación de Container Apps.

Configurar soporte de zonas de disponibilidad

- **Cree un entorno de Container Apps con zonas redundantes.** Para obtener instrucciones de implementación que abarcan Azure Portal, la CLI de Azure y Azure PowerShell, consulte Creación de una aplicación de contenedor con redundancia de zona.
- **Migre a una implementación redundante por zonas.** No se puede habilitar la redundancia de zona en un entorno de Container Apps existente. Para actualizar los entornos existentes que no tienen redundancia de zona, cree un nuevo entorno con redundancia de zona habilitada en una región admitida. A continuación, vuelva a implementar las aplicaciones de contenedor.
- **Desactivar la redundancia de zona.** La redundancia de zona no se puede deshabilitar después de habilitarla durante la creación del entorno. Si necesita una implementación

sin redundancia de zona, debe crear un nuevo entorno sin habilitar la opción de redundancia de zona ni implementarla en una región que no admita zonas de disponibilidad.

- **Compruebe la redundancia de zona.** Puede usar Azure Portal, la CLI de Azure y Azure PowerShell para comprobar el estado de redundancia de zona del entorno.

Planeamiento y administración de capacidad

Si una zona de disponibilidad deja de estar disponible, la plataforma Container Apps usa las reglas de escalado para decidir cuándo reemplazar las réplicas perdidas de esa zona. Es importante configurar correctamente las reglas de escalado para que el programador pueda tomar las decisiones de programación adecuadas.

Para configurar correctamente las reglas de escalado, siga estos principios:

- **Establezca un número mínimo de réplicas** que la aplicación pueda tolerar. Puede tardar un breve período de tiempo en reemplazar las réplicas perdidas porque la plataforma debe detectar que las réplicas antiguas han desaparecido. A continuación, las nuevas réplicas deben iniciarse y devolver un estado de sondeo de preparación correcto para poder recibir solicitudes entrantes. Si no puede tolerar ningún período con menos del número mínimo de réplicas que ha especificado, considere la posibilidad de sobreaprovisionar para mantener el rendimiento de la aplicación incluso si una zona deja de estar disponible.

- **Establezca las solicitudes de recursos y los límites** para ayudar al programador de Container Apps a tomar decisiones de ubicación óptimas entre zonas. Los requisitos de recursos infraespecificados pueden provocar errores de distribución o ubicación desiguales durante una carga alta.

Para obtener más información sobre las opciones de configuración, consulte Establecimiento de reglas de escalado.

Comportamiento cuando todas las zonas están en buen estado

En esta sección se describe qué esperar cuando los recursos de Container Apps están configurados para la redundancia de zona y todas las zonas de disponibilidad están operativas.

- **Enrutamiento de tráfico entre zonas:** Con Container Apps con redundancia de zonas, la plataforma funciona en un modelo activo-activo donde varias réplicas sirven tráfico simultáneamente. El controlador de entrada distribuye las solicitudes entrantes entre

todas las réplicas funcionales, independientemente de su zona, y utiliza el equilibrio de carga 'round robin' de forma predeterminada. Cada zona procesa solicitudes de forma independiente y la plataforma no da prioridad a ninguna zona específica para la distribución del tráfico. Los sondeos de salud se originan en todas las zonas para garantizar una evaluación precisa de la salud de cada réplica desde varias perspectivas.

- **Replicación de datos entre zonas:** Container Apps no replica los datos de la aplicación entre zonas porque está diseñado para cargas de trabajo sin estado. Los datos que la aplicación almacena en almacenamiento efímero, incluido el almacenamiento con ámbito de contenedor y el almacenamiento con ámbito de réplica, se eliminan cuando se cierra el contenedor o la réplica.

  Para los requisitos de datos con estado, monte un recurso compartido de archivos de Azure Files configurado para el almacenamiento con redundancia de zona o use otros servicios de Azure, como Azure Cosmos DB o Azure SQL Database, que proporcionan sus propias funcionalidades de replicación entre zonas.

  La plataforma solo replica los metadatos del plano de control, incluidas las configuraciones de la aplicación, las reglas de escalado y los secretos entre zonas para lograr una alta disponibilidad. Las imágenes de contenedor se extraen del registro de contenedor en cada zona según sea necesario cuando se crean réplicas.

Comportamiento durante un fallo de zona

En esta sección se describe qué esperar cuando los recursos de Container Apps están configurados para redundancia de zona y hay una interrupción en una zona de disponibilidad.

- **Detección y respuesta:** Azure detecta automáticamente errores de zona. Container Apps detiene inmediatamente la programación de nuevas réplicas en la zona fallida y comienza a redistribuir el tráfico a réplicas saludables en las zonas restantes. La plataforma gestiona todas las operaciones de conmutación por error automáticamente, sin necesidad de tu intervención.

- **Notificación:** Microsoft no le notifica automáticamente cuando una zona está inactiva. Sin embargo, puede usar Azure Service Health para comprender el estado general del servicio, incluidos los errores de zona, y puede configurar alertas de Service Health para notificarle problemas.

  También puede supervisar el estado de las aplicaciones a través de las métricas de Container Apps en Azure Monitor. Configure alertas en las caídas del recuento de réplicas y las tasas de error de solicitudes para recibir notificaciones inmediatas cuando se produzcan problemas relacionados con la zona.

- **Solicitudes activas:** Es posible que las solicitudes en vuelo a las réplicas de la zona fallida puedan eliminarse o experimenten tiempos de espera o errores de conexión. Las ejecuciones de trabajos en la zona afectada se anulan y se marcan como fallidas.

- **Pérdida de datos esperada:** No se produce ninguna pérdida de datos en el nivel de plataforma container Apps porque el servicio está diseñado para cargas de trabajo sin estado. Los datos almacenados en el almacenamiento efímero dentro de la zona de disponibilidad se pierden cuando finaliza la réplica y el almacenamiento efímero solo se debe usar para los datos temporales.

- **Tiempo de inactividad esperado:** Las aplicaciones experimentan un tiempo de inactividad mínimo o sin tiempo de inactividad durante los errores de zona. El impacto real depende de la configuración del sondeo de estado de la aplicación y del número de réplicas en zonas correctas. Asegúrese de que los clientes sigan las instrucciones de control de errores transitorios para minimizar los efectos.

  Los trabajos que se ejecutan en la zona afectada son abortados y se marcan como fallidos. Si necesita que un trabajo sea resistente a un error de zona, configure reintentos o configure paralelismo para que el trabajo ejecute varias copias de la misma ejecución. Para obtener más información, consulte Configuración avanzada del trabajo.

- **Reenrutamiento del tráfico:** Las sondas de estado del controlador de ingreso detectan rápidamente las réplicas inaccesibles y las eliminan del pool de equilibrio de carga. En función de la configuración del sondeo de estado de la aplicación, este proceso de conmutación por error suele producirse en unos 30 segundos. El tráfico entrante subsiguiente se distribuye entre las réplicas saludables restantes. Esta redirección de tráfico se produce de forma transparente a los clientes, que siguen usando la misma dirección URL de la aplicación.

  Si la afinidad de sesión está habilitada y una zona deja de funcionar, los clientes que se enrutaron previamente a réplicas de esa zona se enrutan a nuevas réplicas porque las réplicas anteriores ya no están disponibles. Se pierde cualquier estado asociado a las réplicas anteriores.

  No se iniciarán nuevas instancias de trabajos en la zona defectuosa.

- **Administración de instancias:** Es posible que se creen nuevas instancias de réplica en zonas saludables si las reglas de escalado automático se activan debido a un aumento en la carga.

Recuperación de zona

Cuando una zona de disponibilidad se recupera de un error, Container Apps vuelve a reinsertar automáticamente la zona en un servicio activo sin necesidad de intervención. Los sondeos de estado de la plataforma detectan cuándo la infraestructura de la zona recuperada está disponible y Container Apps comienza a programar nuevas réplicas en esa zona en función de la configuración de escalado. Las réplicas existentes en zonas saludables siguen atendiendo el tráfico durante el proceso de reintegración, lo que ayuda a evitar la interrupción del servicio.

Container Apps equilibra gradualmente la distribución de réplicas entre todas las zonas disponibles como parte de las operaciones de escalado estandarizados. Este reajuste automático ocurre cuando se crean réplicas en respuesta a eventos de escalado o cuando se reemplazan las réplicas defectuosas. La plataforma no fuerza la redistribución inmediata de réplicas correctas existentes, lo que evita reinicios innecesarios del contenedor y mantiene la estabilidad de la aplicación durante la recuperación.

Prueba de fallos de zona

La plataforma Container Apps administra el enrutamiento del tráfico, la conmutación por error y la conmutación por recuperación para aplicaciones de contenedor con redundancia de zona. Esta característica está totalmente administrada, por lo que no es necesario iniciar ni validar los procesos de error de zona de disponibilidad.

Para validar la resistencia de la aplicación a errores de zona, simule interrupciones de nivel de zona en el nivel de aplicación mediante enfoques de prueba controlados. Detenga o quite réplicas de zonas específicas mediante el escalado descendente de su aplicación y supervise cómo las réplicas restantes manejan la carga aumentada. Supervise las métricas clave durante las pruebas de resistencia, incluido el recuento de réplicas, las tasas de éxito de las solicitudes, los tiempos de respuesta y el comportamiento de escalado automático. Asegúrese de que el número mínimo de réplicas mantiene la disponibilidad del servicio cuando se eliminan las réplicas y compruebe que las reglas de escalado pueden manejar el aumento de carga en las réplicas restantes. Pruebe las configuraciones de sondeo de estado de salud fallando deliberadamente los puntos de conexión de estado para confirmar que la plataforma quita las instancias no saludables del ciclo de rotación durante los períodos de tiempo esperados.

Resistencia a errores en toda la región

Container Apps es un servicio de una sola región. Si la región deja de estar disponible, el entorno y las aplicaciones tampoco están disponibles.

Soluciones personalizadas de varias regiones para la resistencia

Para reducir el riesgo de un error de una sola región que afecte a la aplicación, puede implementar entornos en varias regiones. Los pasos siguientes ayudan a reforzar la resistencia:

- Implemente las aplicaciones en entornos de cada región. Cada entorno requiere su propia configuración de red virtual y los requisitos de subred se aplican de forma independiente a cada implementación regional. Las imágenes de contenedor deben estar disponibles en todas las regiones, lo que puede lograr mediante Azure Container Registry con la replicación geográfica habilitada.
- Configure las directivas de equilibrio de carga y conmutación por error usando servicios como Azure Front Door o Azure Traffic Manager.
- Replique los datos entre regiones para que pueda recuperar el último estado de la aplicación.

Copias de seguridad y restauración

Container Apps no proporciona funcionalidades de copia de seguridad integradas para las aplicaciones o los datos. Como plataforma de hospedaje de contenedores sin estado, Container Apps espera que las aplicaciones administren sus propias estrategias de persistencia y recuperación de datos a través de servicios externos. Los contenedores de aplicaciones y sus sistemas de archivos locales son efímeros y los datos almacenados localmente se pierden cuando las réplicas se reinician o se mueven.

Resistencia durante las actualizaciones de la aplicación

Use la administración de revisiones para implementar actualizaciones en la aplicación sin tiempo de inactividad. Puede crear nuevas revisiones con imágenes de contenedor actualizadas y realizar una transición mediante una estrategia de implementación azul-verde o cambiar gradualmente el tráfico mediante reglas de división de tráfico. Durante las actualizaciones de aplicaciones, la plataforma mantiene los recuentos mínimos de réplicas mediante la creación de nuevos contenedores antes de retirar los antiguos, lo que ayuda a evitar interrupciones del servicio.

Para obtener más información, consulte Actualización e implementación de cambios en Container Apps.

Resistencia al mantenimiento del servicio

Container Apps realiza el mantenimiento automático de la plataforma para aplicar actualizaciones de seguridad, implementar nuevas características y mejorar la confiabilidad del servicio. La plataforma usa actualizaciones graduales en dominios de error y zonas de disponibilidad para reducir la interrupción de las aplicaciones en ejecución. Durante las ventanas de mantenimiento, los contenedores siguen ejecutándose sin interrupción porque las actualizaciones se aplican a la infraestructura subyacente en fases.

Puede especificar sus propias ventanas de mantenimiento, que son períodos de tiempo que desea realizar el mantenimiento en las aplicaciones. Tenga en cuenta que las actualizaciones críticas pueden producirse fuera de las ventanas de mantenimiento. Para obtener más información, consulte Mantenimiento planeado de Container Apps.

Acuerdo de nivel de servicio

El contrato de nivel de servicio (SLA) para los servicios de Azure describe la disponibilidad esperada de cada servicio y las condiciones que la solución deberá cumplir para lograr esa expectativa de disponibilidad. Para obtener más información, consulte Acuerdos de Nivel de Servicio para servicios en línea.

El Acuerdo de Nivel de Servicio de disponibilidad para las Aplicaciones de Contenedor se basa en las reglas de escalado que establezca en las aplicaciones.

Contenido relacionado

- Introducción a Container Apps
- Sondas de estado en Container Apps
- Escalado automático en aplicaciones de contenedores
- Observabilidad en Aplicaciones de Contenedores
- Documentación sobre confiabilidad de Azure
- Guía de confiabilidad deWell-Architected Framework

ⓘ **Nota:** El autor ha creado este artículo con ayuda de inteligencia artificial. Más información

---

Last updated on 22/01/2026

Confiabilidad en Azure Container Instances

En este artículo se describe la compatibilidad con la confiabilidad en Azure Container Instances, que proporciona una manera sencilla de ejecutar contenedores de Linux o Windows en Azure, sin necesidad de administrar máquinas virtuales (VM) ni adoptar un servicio de nivel superior más complejo.

Al usar Azure, relibilidad es una responsabilidad compartida. Microsoft proporciona una variedad de funcionalidades para admitir resistencia y recuperación. Es responsable de comprender cómo funcionan esas funcionalidades dentro de todos los servicios que usa y de seleccionar las funcionalidades que necesita para cumplir los objetivos empresariales y los objetivos de tiempo de actividad.

En este artículo se describe cómo hacer que Azure Container Instances sean resistentes a una variedad de posibles interrupciones y problemas, incluidos errores transitorios, interrupciones de zona de disponibilidad y interrupciones de regiones. Resalta cierta información clave sobre el acuerdo de nivel de servicio (SLA) de Azure Container Instances.

Recomendaciones de implementación de producción para la confiabilidad

Para aumentar la confiabilidad de las aplicaciones de producción basadas en Container Instances, se recomienda realizar las siguientes acciones:

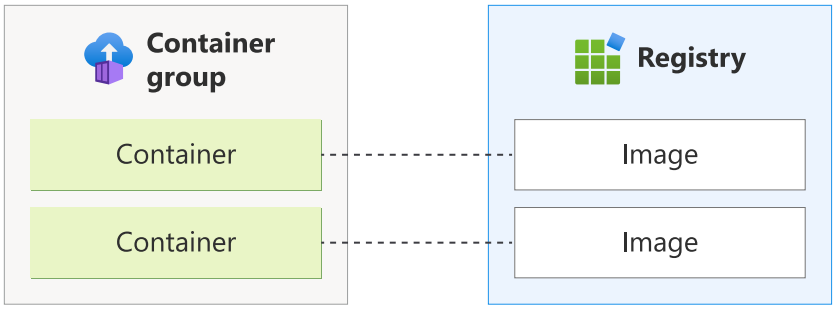
- ✓ Ejecute las aplicaciones en varias zonas de disponibilidad.
- ✓ Considere si también se deben ejecutar grupos de contenedores independientes en varias regiones.
- ✓ Use sondeos de actividad para detectar y automáticamente reiniciar contenedores no saludables.
- ✓ Utilice sondeos de preparación para esperar hasta que los contenedores estén listos antes de recibir tráfico.
- ✓ Use actualizaciones graduales para aplicar cambios progresivamente si usa NGroups. Este enfoque reduce la probabilidad de tiempo de inactividad debido a las actualizaciones.
- ✓ Revise las prácticas recomendadas y consideraciones para Container Instances.

Introducción a la arquitectura de confiabilidad

Para usar Container Instances, implemente un grupo *container*. Un grupo de contenedores contiene uno o varios *contenedores*. Cada contenedor se crea a partir de una imagen *container*, que se almacena en un registro como Azure Container Registry.

Todos los contenedores de un grupo de contenedores se implementan juntos como una sola unidad lógica y comparten la misma infraestructura física.

En el diagrama siguiente se muestra la relación entre grupos de contenedores, contenedores e imágenes.



Container Instances proporciona las siguientes características para administrar grupos de contenedores:

- NGroups (versión preliminar) proporciona un conjunto de funcionalidades para administrar varios grupos de contenedores relacionados. Al crear un NGroup, se define el número de grupos de contenedores para crear. Container Instances proporciona funcionalidades como lanzamientos de actualización automatizados y distribución de grupos de contenedores entre zonas de disponibilidad.
- Los grupos en espera crean un grupo de grupos de contenedores aprovisionados previamente que se pueden usar en respuesta al tráfico entrante. Los grupos en espera están concebidos para optimizar la creación de grupos de contenedores y no para aumentar la resiliencia.

Resistencia a errores transitorios

Los errores transitorios son errores breves e intermitentes en los componentes. Se producen con frecuencia en un entorno distribuido como la nube y son una parte normal de las operaciones. Los errores transitorios se corrigen después de un breve período de tiempo. Es importante que las aplicaciones puedan controlar errores transitorios, normalmente mediante el reintento de solicitudes afectadas.

Todas las aplicaciones hospedadas en la nube deben seguir las instrucciones de control de errores transitorios Azure cuando se comunican con cualquier API, bases de datos y otros componentes hospedados en la nube. Para obtener más información, consulte Recomendaciones para controlar errores transitorios.

Los SDK proporcionados por Microsoft suelen controlar errores transitorios. Dado que hospeda sus propias aplicaciones en Container Instances, siga los pasos necesarios para reducir la posibilidad de errores transitorios:

- Ejecute varios grupos de contenedores para cargas de trabajo importantes para asegurarse de que un error en un contenedor o grupo de contenedores no afecte a toda la aplicación.
- Construye el código de tu aplicación para que sea resistente a errores transitorios en los servicios a los que te conectas, por ejemplo, utilizando políticas de reintentos con estrategias de espera incremental.

Para obtener más información sobre otros errores que pueden producirse en tiempo de ejecución y cómo responder a ellos, consulte Problemas durante el tiempo de ejecución del grupo de contenedores.

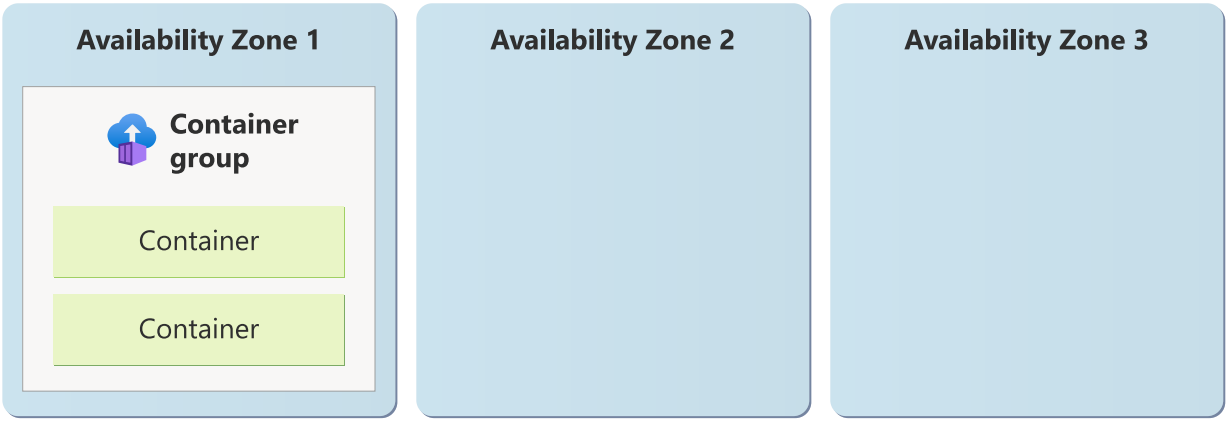
Resistencia a errores de zona de disponibilidad

Availability zones son grupos de centros de datos físicamente independientes dentro de una región de Azure. Cuando una zona falla, los servicios pueden transferirse a una de las zonas restantes.

Container Instances admite zonas de disponibilidad de diferentes maneras, en función de cómo implemente los grupos de contenedores:

- **Grupos de contenedores creados manualmente:** Un grupo de contenedores individual es un recurso *zonal* , lo que significa que se puede implementar en una sola zona de disponibilidad que seleccione. Todos los contenedores del grupo se despliegan en la misma zona de disponibilidad. Si esa zona de disponibilidad tiene una interrupción, el grupo de contenedores y todos sus contenedores podrían experimentar tiempo de inactividad.

  En el diagrama siguiente se muestra un grupo de contenedores que se ha implementado manualmente en la zona de disponibilidad 1:

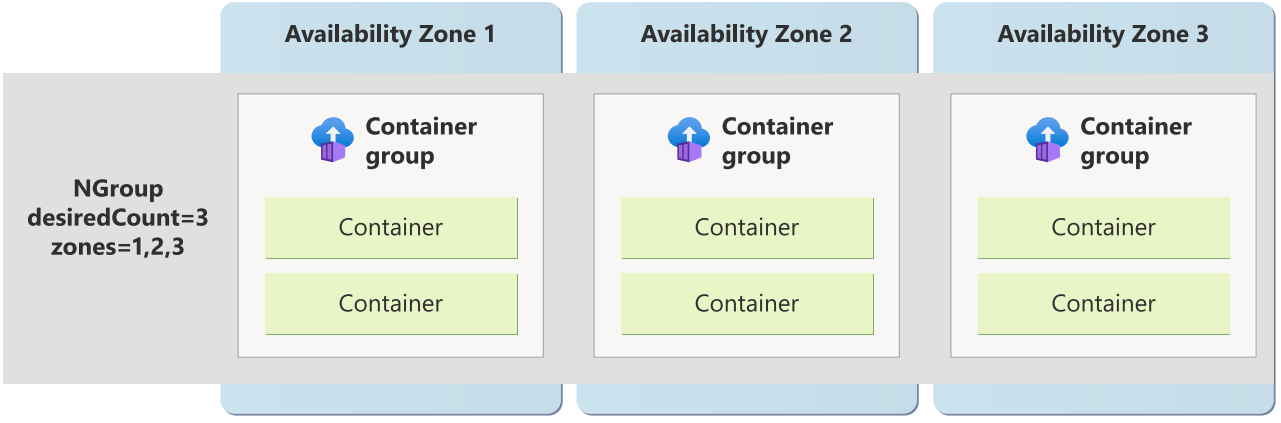


> ⓘ **Nota:**
>
> Para asegurarse de que la aplicación sigue ejecutándose cuando cualquier zona única de la región experimenta una interrupción, se recomienda crear un mínimo de dos grupos de contenedores en dos zonas de disponibilidad diferentes.

Si no especifica zonas de disponibilidad que se van a usar para el grupo de contenedores, es *no zonal* o *regional*, lo que significa que podría colocarse en cualquier zona de disponibilidad dentro de la región o dentro de la misma zona. Si alguna zona de disponibilidad de la región tiene un problema, el grupo de contenedores podría experimentar tiempo de inactividad.

- **NGroups:** Al implementar un NGroup, puede especificar una o varias zonas en las que implementarla. Si implementa un NGroup en dos o más zonas, se trata de un NGroup *con redundancia de zona* y una interrupción de una zona de disponibilidad solo provoca problemas para los grupos de contenedores dentro de la zona afectada.

  En el diagrama siguiente se muestra un NGroup que se implementa en tres zonas de disponibilidad:



Si no especifica zonas de disponibilidad para su NGroup, será no zonal y podría experimentar tiempo de inactividad si alguna zona de disponibilidad en la región presenta un problema.

- **Grupos en espera:** Al implementar un grupo en espera, puede especificar opcionalmente una o varias zonas. La plataforma puede solicitar contenedores en las zonas que seleccione.

  Sin embargo, los grupos en espera no tienen redundancia de zona ni resistencia de zona porque no hay garantía de que los contenedores se creen en varias zonas. Si se produce una interrupción de zona, es posible que todos los contenedores del grupo se coloquen en la zona afectada.

  Debido a que los grupos de respaldo no están diseñados para soportar la resiliencia ante fallos de zona, esta guía no describe el comportamiento detallado de los grupos de respaldo con zonas de disponibilidad.

  > ⓘ **Importante**
  >
  > Los grupos en espera no se diseñaron para tener resistencia de zona. No deben usarse para cargas de trabajo que requieran resistencia a errores de zona.

Requisitos

- **Compatibilidad con regiones:** Las implementaciones de grupos de contenedores zonales se admiten en todas las regiones con zonas de disponibilidad.
- Las implementaciones zonales están disponibles para Linux y Windows Server 2019 grupos de contenedores.
- Para seleccionar una zona de disponibilidad, debe usar la SKU estándar. Los grupos de contenedores zonales no están disponibles con la SKU confidencial.

Consideraciones

Los contenedores de acceso puntual no admiten zonas de disponibilidad y siempre son no zonales.

Cost

No hay ningún costo adicional para configurar zonas de disponibilidad para un grupo de contenedores.

Configurar soporte de zonas de disponibilidad

- **Cree grupos de contenedores con soporte para zonas de disponibilidad.** El enfoque que se usa para configurar zonas de disponibilidad depende de cómo cree grupos de contenedores.
  - *Grupos de contenedores creados manualmente:* Para crear un grupo de contenedores zonales en una zona específica, puede usar uno de los métodos siguientes:
    - El portal Azure
    - El CLI de Azure
    - Azure PowerShell
    - Bicep
    - Una plantilla de Azure Resource Manager (plantilla de ARM)
    - Terraform
  - *NGroups:* Puede implementar un NGroup con redundancia de zona mediante un archivo de Bicep o una plantilla de ARM y especificando varias zonas. Para obtener más información, consulte ejemplo de NGroups con zonas.
  - *Grupos en espera:* Puede implementar un grupo en espera que use zonas de disponibilidad especificando una o varias zonas al crear o actualizar el grupo. Sin embargo, es posible que los contenedores no se creen en varias zonas. Los pools de reserva no deben usarse para cargas de trabajo que requieran tolerancia a fallos de zona. Para obtener más información, consulte Crear un grupo en espera para Container Instances.
- **Habilitar soporte para zonas de disponibilidad en los recursos existentes.** El enfoque que se usa para configurar zonas de disponibilidad depende de cómo cree grupos de contenedores.
  - *Grupos de contenedores creados manualmente:* No se pueden habilitar zonas de disponibilidad en un grupo de contenedores nozonal existente. Debe eliminar el grupo de contenedores y crear un grupo de contenedores zonal.
  - *NGroups:* No se pueden habilitar zonas de disponibilidad en un NGroup no zonal existente. Debe eliminar el NGroup y crear un NGroup con redundancia zonal.

- *Grupos en espera:* No se pueden habilitar zonas de disponibilidad en un grupo de espera no zonal existente. Debe eliminar el grupo de contenedores y crear un nuevo grupo de espera que use varias zonas de disponibilidad.

- **Mueva grupos de contenedores entre zonas o deshabilite el soporte para zonas de disponibilidad.** El enfoque que se usa para modificar zonas de disponibilidad depende de cómo cree grupos de contenedores.

  - *Grupos de contenedores creados manualmente:* Para cambiar la zona de disponibilidad de un grupo de contenedores, debe eliminar el grupo de contenedores y crear otro grupo de contenedores con la nueva zona de disponibilidad. Para obtener información sobre cómo eliminar el grupo de contenedores, consulte los siguientes recursos:
    - El portal Azure
    - El CLI de Azure
    - Azure PowerShell
    - Bicep
    - Una plantilla de ARM
    - Terraform
    - La CLI de Docker

  - *NGroups:* Puede agregar zonas a un NGroup, pero no puede quitar zonas.

  - *Grupos en espera:* Puede agregar zonas a un grupo en espera, pero no puede quitar zonas.

Distribución de grupos de contenedores

La forma en que los grupos de contenedores se distribuyen entre zonas de disponibilidad depende de cómo implemente los grupos de contenedores.

- **Grupos de contenedores creados manualmente:** Es responsable de distribuir los grupos de contenedores creados manualmente en varias zonas de disponibilidad.

- **NGroups:** Durante las operaciones de escalado horizontal, NGroups elimina aleatoriamente instancias, lo que podría no mantener una distribución entre zonas de disponibilidad. Las operaciones de escalabilidad horizontal intentan reequilibrar la distribución entre zonas.

- **Grupos en espera:** Un grupo en espera puede crear contenedores en cualquiera de las zonas de disponibilidad que configure en el grupo. Sin embargo, es posible que los

contenedores no se creen en varias zonas. Los pools de reserva no deben usarse para cargas de trabajo que requieran tolerancia a fallos de zona.

Planeamiento y administración de capacidad

** Para prepararse para un fallo de zona de disponibilidad, considere el sobreaprovisionamiento del número de grupos de contenedores que despliegue. Este enfoque permite que la solución tolere cierta pérdida de capacidad y siga funcionando sin que se vea afectado su rendimiento. Para obtener más información, consulte Administración de la capacidad mediante el sobreaprovisionamiento.

El enfoque que se usa para sobreaprovisionar grupos de contenedores depende de cómo implemente los grupos de contenedores.

- **Grupos de contenedores creados manualmente:** Es responsable de planear la capacidad de los grupos de contenedores creados manualmente, incluido el planeamiento del número de grupos de contenedores que se van a implementar en cada zona.
- **NGroup:** Considere *la posibilidad de sobreaprovisionar* la capacidad de NGroup para tolerar la pérdida de una zona.
- **Grupos en espera:** Los grupos en espera no están diseñados para ser resistentes a los errores de zona. Considere la posibilidad de usar varios grupos en espera en diferentes zonas o usar NGroups.

Comportamiento cuando todas las zonas están en buen estado

En esta sección se describe qué esperar cuando los recursos de Container Instances están configurados para ser compatibles con zonas de disponibilidad y todas las zonas de disponibilidad están operativas.

- **Enrutamiento de tráfico entre zonas:** Usted es responsable de dirigir el tráfico a sus contenedores. Por ejemplo, puede usar Azure Application Gateway como puerta de enlace y equilibrador de carga para los grupos de contenedores.

  Si usa grupos de NGroup o grupos en espera, es responsable del equilibrio de carga en cada contenedor. También debe configurar el sistema de enrutamiento de tráfico para detectar el estado de cada grupo de contenedores y redirigir el tráfico a un grupo alternativo cuando sea necesario.

- **Replicación de datos entre zonas:** Los contenedores y los grupos de contenedores no tienen estado. Puede adjuntar su propio recurso compartido de archivos o conectarse a bases de datos u otros servicios de almacenamiento desde sus aplicaciones. Usted es responsable de asegurar que los recursos compartidos de archivos y los servicios de almacenamiento sean resilientes por zonas. Revise las guías de confiabilidad de cada servicio para comprender cómo hacer que cada zona de componentes sea resistente.

Comportamiento durante un fallo de zona

En esta sección se describe qué esperar cuando los recursos de Container Instances están configurados para admitir la zona de disponibilidad y se produce una interrupción en esta.

- **Detección y respuesta:** La responsabilidad de la detección de errores de zona y la respuesta asociada dependen de cómo implemente los grupos de contenedores.
  - *Grupos de contenedores creados manualmente:* debe detectar la pérdida de una zona de disponibilidad e iniciar una conmutación por error a un grupo de contenedores alternativo que ha creado en otra zona de disponibilidad.
  - *NGroups:* La plataforma Container Instances es responsable de detectar un error en una zona de disponibilidad y responder.

    Sin embargo, es responsable de garantizar que el tráfico se enrute a los contenedores en una zona saludable.
  - *Standby pools:* No se garantiza que la plataforma de Container Instances responda a errores de zona para pools en espera. Los pools de reserva no deben usarse para cargas de trabajo que requieran tolerancia a fallos de zona.
- **Notificación:** Microsoft no le notifica automáticamente cuando una zona está inactiva. Sin embargo, puede usar Azure Service Health para comprender el estado general del servicio, incluidos los errores de zona, y puede configurar alertas de Service Health para notificarle problemas.
- **Solicitudes activas:** Si se produce un error en una zona, es probable que todos los contenedores que se ejecutan en esa zona se detengan, incluido cualquier trabajo activo que controlen.
- **Pérdida de datos esperada:** Dado que los contenedores y los grupos de contenedores no tienen estado, no se espera ninguna pérdida de datos debido a un fallo de zona. Sin

embargo, es responsable de garantizar que cada componente de la carga de trabajo sea resistente a la zona, incluidos los servicios de almacenamiento y las bases de datos.

- **Tiempo de inactividad esperado:** El tiempo de inactividad que puede esperar de un error de zona depende de cómo implemente los grupos de contenedores.
  - *Grupos de contenedores creados manualmente:* En el caso de los grupos de contenedores zonales, cuando una zona no está disponible, el grupo de contenedores y sus contenedores no están disponibles hasta que se recupere la zona de disponibilidad.
  - *NGroups:* En el caso de NGroups, si una zona deja de funcionar, la aplicación sigue estando disponible porque los grupos de contenedores restantes de NGroups siguen ejecutándose en otras zonas. No se espera ningún tiempo de inactividad.
  - *Grupos en espera:* Los grupos en espera no proporcionan resistencia de zona. Si todos los grupos de contenedores del grupo en espera se encuentran en una sola zona de disponibilidad, es posible que todos los grupos de contenedores y sus contenedores queden inaccesibles hasta que se recupere la zona de disponibilidad.
- **Reenrutamiento del tráfico:** Dado que es responsable del enrutamiento del tráfico a los contenedores, también es responsable de volver a enrutar el tráfico si se produce un error en un grupo de contenedores debido a una interrupción de zona de disponibilidad.

Recuperación de zona

Una vez recuperada la zona, la plataforma Azure reinicia automáticamente los grupos de contenedores que se habían detenido. No se requiere ninguna acción del cliente.

Prueba de fallos de zona

No hay ninguna manera de simular una interrupción de la zona de disponibilidad que contiene el grupo de contenedores. Sin embargo, puede configurar manualmente puertas de enlace ascendentes o equilibradores de carga para redirigir el tráfico a otro grupo de contenedores en una zona de disponibilidad diferente.

Resistencia a errores en toda la región

Container Instances es un servicio de una sola región. Si la región deja de estar disponible, los grupos de contenedores y sus contenedores tampoco están disponibles.

Soluciones de varias regiones personalizadas para la resistencia

Opcionalmente, puede implementar grupos de contenedores independientes en varias regiones. Es responsable de implementar y configurar los grupos de contenedores en cada región. También debe configurar el equilibrio de carga mediante un servicio como Azure Traffic Manager o Azure Front Door. Usted es responsable de cualquier sincronización, conmutación por error y recuperación después de falla de datos.

Acuerdo de nivel de servicio

El acuerdo de nivel de servicio (SLA) para Azure servicios describe la disponibilidad esperada de cada servicio y las condiciones que la solución debe cumplir para lograr esa expectativa de disponibilidad. Para obtener más información, vea SLAs for servicios en línea.

Contenido relacionado

- Confiabilidad en Azure

ⓘ**Nota:** El autor ha creado este artículo con ayuda de inteligencia artificial. Más información

---

Last updated on 22/01/2026

Confiabilidad en Azure Data Factory

Azure Data Factory le permite crear canalizaciones de datos flexibles y eficaces para la integración de datos sin servidor y la transformación de datos. Como servicio de Azure, Data Factory proporciona una variedad de funcionalidades para satisfacer los requisitos de fiabilidad.

Cuando se usa Azure, la confiabilidad es una responsabilidad compartida. Microsoft proporciona una variedad de funcionalidades para admitir resistencia y recuperación. Es responsable de comprender cómo funcionan esas funcionalidades dentro de todos los servicios que usa y de seleccionar las funcionalidades que necesita para cumplir los objetivos empresariales y los objetivos de tiempo de actividad.

En este artículo se describe cómo hacer que Data Factory sea resistente a una variedad de posibles interrupciones y problemas, incluidos errores transitorios, interrupciones de zona de disponibilidad y interrupciones de regiones. También se describe cómo puede usar copias de seguridad para recuperarse de otros tipos de problemas y se resalta cierta información clave sobre el acuerdo de nivel de servicio (SLA) de Data Factory.

> ⓘ **Nota:**
>
> Al considerar la confiabilidad de la factoría de datos, también debe tener en cuenta la confiabilidad de los almacenes de datos a los que se conecta. La mejora de la resistencia de la factoría de datos por sí sola podría tener un impacto limitado si los almacenes de datos no son igualmente resistentes. En función de los requisitos de resistencia, es posible que tenga que realizar cambios de configuración en varias áreas. Para ayudar a garantizar que los almacenes de datos cumplan los requisitos de continuidad empresarial, consulte la documentación y las instrucciones de confiabilidad del producto.

Introducción a la arquitectura de confiabilidad

Data Factory consta de varios componentes de infraestructura. Cada componente admite la confiabilidad de la infraestructura de maneras diferentes.

Los componentes de Data Factory incluyen:

- **El servicio principal de Data Factory**, que administra los desencadenadores de canalización y supervisa la coordinación de las actividades de canalización. El servicio principal también administra los metadatos de cada componente de la factoría de datos. Microsoft administra el servicio principal.

- **Entornos de ejecución de integración (IR)**, que se conectan a almacenes de datos y realizan actividades definidas en la canalización. Hay diferentes tipos de infrarrojos.
    - *IR administrados por Microsoft*, que incluyen Azure IR y Azure-SQL Server Integration Services (Azure-SSIS) IR. Microsoft administra los componentes que componen estos entornos de ejecución. En algunos escenarios, se configuran las opciones que afectan a la resiliencia del IR.
    - *Entornos de ejecución de integración autohospedados (SHIR)*. Microsoft proporciona software que puede ejecutar en su propia infraestructura informática para realizar algunas partes de sus propias canalizaciones de Data Factory. Es responsable de la implementación y administración de los recursos de computación y de la resiliencia de esos recursos de computación.

Resistencia a errores transitorios

Los errores transitorios son errores breves e intermitentes en los componentes. Se producen con frecuencia en un entorno distribuido como la nube y son una parte normal de las operaciones. Los errores transitorios se corrigen después de un breve período de tiempo. Es importante que las aplicaciones puedan controlar errores transitorios, normalmente mediante el reintento de solicitudes afectadas.

Todas las aplicaciones hospedadas en la nube deben seguir las instrucciones de control de errores transitorios de Azure cuando se comunican con cualquier API, bases de datos y otros componentes hospedados en la nube. Para obtener más información, consulte Recomendaciones para controlar errores transitorios.

Cuando se usa Data Factory, es importante prepararse para errores transitorios, especialmente al diseñar canalizaciones y actividades.

Idempotencia

Las actividades de canalización deben ser *idempotentes*, lo cual significa que se puedan volver a ejecutar varias veces sin causar efectos adversos. Si se produce un error transitorio como un error de red o una interrupción de zona de disponibilidad, Data Factory podría volver a ejecutar actividades de canalización. Esta repetición puede crear registros duplicados.

Para evitar la inserción de registros duplicados debido a un error transitorio, implemente los procedimientos recomendados siguientes:

- *Use identificadores únicos* para cada registro antes de escribir en la base de datos. Este enfoque puede ayudarle a encontrar y eliminar registros duplicados.

- *Use una estrategia upsert* para los conectores que admiten upsert. Antes de que se produzca la inserción de registros duplicados, use este enfoque para comprobar si ya existe un registro. Si existe, actualícela. Si no existe, insértelo. Por ejemplo, comandos SQL como `MERGE` o `ON DUPLICATE KEY UPDATE` usan este enfoque de upsert.

- *Use estrategias de acción de copia.* Para obtener más información, consulte Comprobación de coherencia de datos en la actividad de copia.

Directivas de reintentos

Puede usar directivas de reintento para configurar partes de la canalización para reintentar si hay un problema, como errores transitorios en los recursos conectados. En Data Factory, puede configurar directivas de reintento en los siguientes tipos de objetos de canalización:

- Desencadenadores periódicos
- Actividades de ejecución

Para obtener más información sobre cómo cambiar o deshabilitar las directivas de reintento para los desencadenadores y actividades de Data Factory, consulte Ejecuciones y desencadenadores de canalización.

Resistencia a errores de zona de disponibilidad

Las zonas de disponibilidad son grupos físicamente independientes de centros de datos dentro de una región de Azure. Cuando una zona falla, los servicios pueden transferirse a una de las zonas restantes.

Data Factory admite *redundancia de zona*, lo que proporciona resistencia a errores en las zonas de disponibilidad.

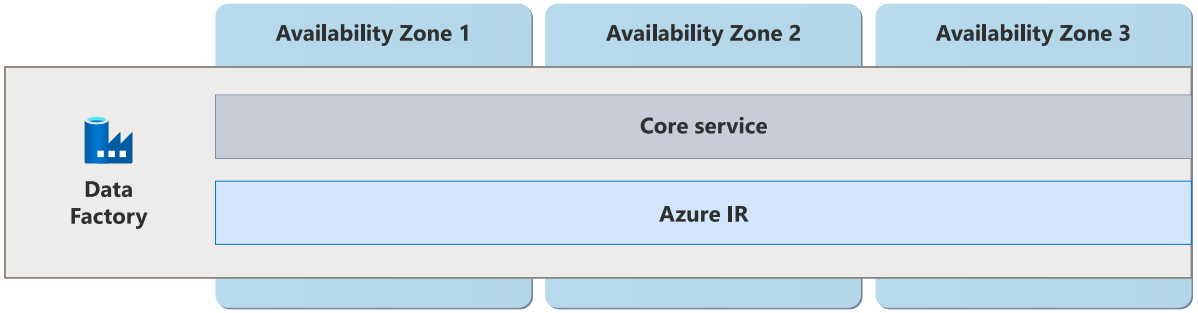
Cada parte de Data Factory admite redundancia de zona:

- **Servicio principal:** Microsoft administra los componentes en el servicio principal de Data Factory y los distribuye entre zonas de disponibilidad.

  Pero después de un error de zona, Microsoft no garantiza el estado de los desencadenadores de ventana de saltos de tamaño constante.

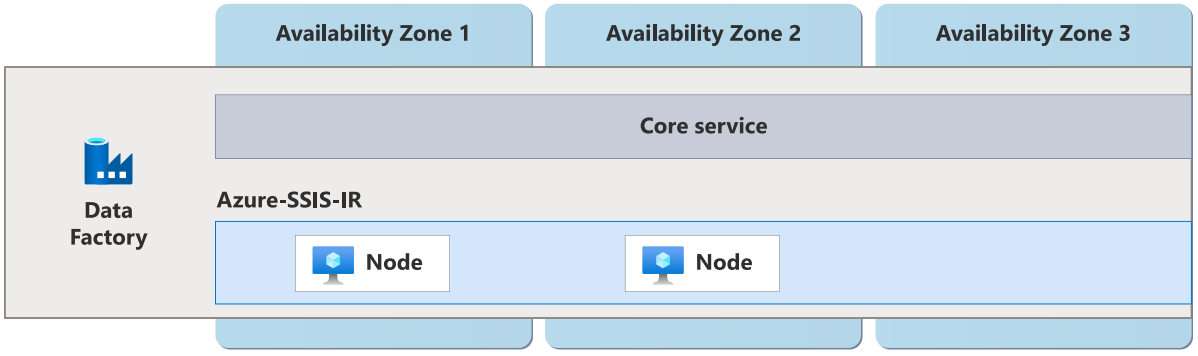
- **IRs:** La compatibilidad con la redundancia de zona depende del tipo de IR que se utilice.

  - *Azure IR* admite redundancia de zona y Microsoft administra esta funcionalidad.



- *Un Azure-SSIS IR* requiere la implementación de al menos dos nodos. Estos nodos se asignan automáticamente a diferentes zonas de disponibilidad.

  En el diagrama siguiente se muestra una canalización con redundancia de zona y un entorno de ejecución de integración de Azure-SSIS con dos nodos que se implementan en zonas diferentes:



- *Un SHIR* le ofrece la responsabilidad de implementar la infraestructura de proceso para hospedar el entorno de ejecución. Puede implementar varios nodos, como máquinas virtuales individuales y configurarlos para alta disponibilidad. Después, puede distribuir esos nodos en varias zonas de disponibilidad. Para más información, consulte Alta disponibilidad y escalabilidad.

Requisitos

Los recursos de Data Factory con redundancia de zona se pueden implementar en cualquier región que admita zonas de disponibilidad.

Cost

- **Servicio principal:** No se aplica ningún costo adicional para la redundancia de zona.
- **IR:** El costo de la redundancia de zona varía en función del tipo de IR que use.
  - *Azure IR* incluye redundancia de zona sin costo adicional.

- *Un Azure-SSIS IR* requiere que despliegue al menos dos nodos para lograr redundancia de zona. Para obtener más información sobre cómo se factura cada nodo, vea Ejemplo de precios: Ejecutar paquetes SSIS en un IR de Azure-SSIS.
- *Un SHIR* requiere que se implemente y administre la infraestructura de proceso. Para lograr la resiliencia de la zona, debe distribuir los recursos informáticos entre varias zonas. Dependiendo del número de nodos que implemente y cómo los configure, puede incurrir en costos adicionales de los servicios de proceso subyacentes y otros servicios auxiliares. No hay ningún cargo adicional para ejecutar el SHIR en varios nodos.

Configurar soporte de zonas de disponibilidad

- **Servicio principal:** No se requiere ninguna configuración. El servicio principal de Data Factory admite automáticamente la redundancia de zona.
- **IRs:**
  - *Una instancia de Azure IR:* No se requiere ninguna configuración. Azure IR habilita automáticamente la redundancia de zona.
  - *Un Azure-SSIS IR:* no se requiere configuración. Un IR de nivel Azure-SSIS habilita automáticamente la redundancia de zona cuando se implementa con dos o más nodos.
  - *Una SHIR* requiere que configure su propia resistencia, lo que incluye la propagación de los nodos entre varias zonas de disponibilidad.

Planeamiento y administración de capacidad

- **Servicio principal:** El servicio principal de Data Factory se escala automáticamente en función de la demanda y no es necesario planear ni administrar la capacidad.
- **IRs:**
  - *Una instancia de Azure IR* se escala automáticamente en función de la demanda y no es necesario planear ni administrar la capacidad.
  - *Un Azure-SSIS IR* requiere que configure específicamente el número de nodos que use. Para prepararse para errores de zona de disponibilidad, considere la posibilidad de sobreaprovisionar la capacidad del IR. El sobreaprovisionamiento permite a la solución tolerar cierto grado de pérdida de capacidad y seguir funcionando sin degradar el

rendimiento. Para obtener más información, consulte Administración de la capacidad mediante el aprovisionamiento excesivo.

- *Un SHIR* requiere que configure su propia capacidad y escalamiento. Considere la posibilidad de sobreaprovisionar al implementar una SHIR.

Comportamiento cuando todas las zonas están en buen estado

En esta sección se describe qué esperar cuando los recursos de Data Factory están configurados para la redundancia de zona y todas las zonas de disponibilidad están operativas.

- **Enrutamiento de tráfico entre zonas:** Durante las operaciones normales, Data Factory distribuye automáticamente las actividades de canalización, los desencadenadores y otros trabajos entre instancias correctas en cada zona de disponibilidad.

- **Replicación de datos entre zonas:** En general, Data Factory es un servicio sin estado, por lo que no es necesario replicar ningún estado entre zonas.

  Sin embargo, los desencadenadores de ventanas deslizantes contienen estado, que podría no estar completamente replicado entre zonas.

Comportamiento durante un fallo de zona

En esta sección se describe qué puede ocurrir cuando los recursos de Data Factory están configurados para la redundancia de zona y se produce una interrupción en la zona de disponibilidad.

- **Detección y respuesta:** La plataforma data Factory es responsable de detectar un error en una zona de disponibilidad y responder. No es necesario hacer nada para iniciar una conmutación por error de zona en las canalizaciones u otros componentes.

- **Notificación**: Microsoft no le notifica automáticamente cuando una zona está inactiva. Sin embargo, puede usar Azure Resource Health para supervisar el estado de un recurso individual y puede configurar alertas de Resource Health para notificarle problemas. También puede usar Azure Service Health para comprender el estado general del servicio, incluidos los errores de zona, y puede configurar alertas de Service Health para notificarle problemas.

- **Solicitudes activas:** cualquier canalización y desencadenador en proceso sigue ejecutándose, y no experimentará ninguna interrupción inmediata debido a un error en la zona. Sin embargo, las actividades en curso durante un error de zona pueden producir un error y reiniciarse. Es importante diseñar actividades para que sean idempotentes, lo que

les ayuda a recuperarse de errores de zona y de otros errores. Para obtener más información, consulte Resistencia a errores transitorios.

- **Tiempo de inactividad esperado:** No se espera ningún tiempo de inactividad durante un error de zona.
- **Pérdida de datos esperada:** En general, Data Factory es un servicio sin estado, por lo que no se espera ninguna pérdida de datos durante un error de zona.

  Pero si usa un desencadenador periódico, el estado del desencadenador podría perderse después de un error de zona. Debe reiniciar o volver a ejecutar los triggers que estaban en ejecución en el momento del fallo de zona.

Recuperación de zona

Cuando se recupera la zona de disponibilidad, Data Factory vuelve automáticamente a la zona original. No es necesario hacer nada para iniciar una conmutación por recuperación de zona en las canalizaciones u otros componentes.

Sin embargo, si usa una SHIR, es posible que tenga que reiniciar los recursos de cómputo si se han detenido.

Prueba de fallos de zona

Para el servicio principal, así como para los Azure y los Azure-SSIS IR, Data Factory administra el enrutamiento del tráfico, la conmutación por error y la conmutación por recuperación para los recursos con redundancia de zona. Dado que esta característica está totalmente administrada, no es necesario iniciar ni validar los procesos de error de zona de disponibilidad.

En el caso de los SHIR, puede usar Azure Chaos Studio para simular un error de zona de disponibilidad en una máquina virtual de Azure.

Resistencia a errores en toda la región

Los recursos de Data Factory se implementan en una sola región de Azure. Si la región deja de estar disponible, la factoría de datos tampoco está disponible. Sin embargo, hay enfoques que puede usar para ayudar a garantizar la resistencia a las interrupciones de la región. Estos enfoques dependen de si la factoría de datos está en una región emparejada o no emparejada y en sus requisitos y configuración específicos.

Conmutación por error administrada por Microsoft en una región emparejada

Data Factory admite la conmutación por error administrada por Microsoft para factorías de datos en regiones emparejadas, excepto Sur de Brasil y Sudeste de Asia. En el improbable caso de un error prolongado en la región, Microsoft podría iniciar una conmutación por error regional de la instancia de Data Factory.

Debido a los requisitos de residencia de datos en el Sur de Brasil y sudeste de Asia, los datos de Data Factory solo se almacenan en la región local mediante el almacenamiento con redundancia de zona de Azure Storage. Para sudeste asiático, todos los datos se almacenan en Singapur. Para el Sur de Brasil, todos los datos se almacenan en Brasil.

En el caso de las factorías de datos en regiones no emparejadas o en el Sur de Brasil o Sudeste de Asia, Microsoft no realiza la conmutación por error regional en su nombre.

> **ⓘ Importante**
>
> Microsoft inicia la conmutación por error administrada por Microsoft. Es probable que se produzca después de un retraso significativo y se realice con el mejor esfuerzo. También hay algunas excepciones a este proceso. Es posible que experimente alguna pérdida de los metadatos de la factoría de datos. La conmutación por error de los recursos de Data Factory puede producirse en un momento diferente del tiempo de conmutación por error de otros servicios de Azure.
>
> Si necesita ser resistente a interrupciones de regiones, considere la posibilidad de usar una de las **soluciones personalizadas de varias regiones para lograr resistencia**.

Conmutación por error de los IR

Para prepararse para una conmutación por error, puede haber algunas consideraciones adicionales, en función del entorno de ejecución de integración utilizado.

- Puede configurar *Azure IR* para resolver automáticamente la región que usa. Si la región está establecida para *resolverse automáticamente* y hay una interrupción en la región principal, el Azure IR cambia automáticamente a la región emparejada. Esta conmutación por error está sujeta a limitaciones. Para configurar la región de Azure IR para la implementación de la actividad o el envío en la configuración de IR, establezca la región para *resolverse automáticamente*.
- La conmutación por error de *Azure-SSIS IR* se administra por separado en la conmutación por error administrada por Microsoft de Data Factory. Para obtener más información,

consulte Soluciones personalizadas de varias regiones para lograr resistencia.

- *Una SHIR* se ejecuta en la infraestructura por la que es responsable, por lo que una conmutación por error administrada por Microsoft no se aplica a las SHIR. Para obtener más información, consulte Soluciones personalizadas de varias regiones para lograr resistencia.

Reconfiguración posterior a la conmutación por error

Una vez completada una conmutación por error administrada por Microsoft, puede acceder a la canalización de Data Factory en la región emparejada. Sin embargo, una vez completada la conmutación por error, es posible que tenga que realizar alguna reconfiguración para los IR u otros componentes. Este proceso incluye volver a establecer la configuración de red.

Soluciones personalizadas de varias regiones para la resistencia

Si necesita que las canalizaciones sean resistentes a las interrupciones regionales y necesita controlar el proceso de conmutación por error, considere la posibilidad de usar una canalización controlada por metadatos.

- **Configure el control de código fuente para que Data Factory** realice un seguimiento y audite los cambios realizados en los metadatos. Puede usar este enfoque para acceder a los archivos JSON de metadatos para canalizaciones, conjuntos de datos, servicios vinculados y desencadenadores. Data Factory admite diferentes tipos de repositorio de Git, como Azure DevOps y GitHub. Para más información, consulte Control de código fuente en Data Factory.

- **Utilice un sistema de integración continua y entrega continua (CI/CD),** como Azure DevOps, para gestionar los metadatos de la canalización y las implementaciones. Puede usar CI/CD para restaurar rápidamente las operaciones en una instancia de otra región. Si una región no está disponible, puede aprovisionar una nueva factoría de datos manualmente o a través de la automatización. Una vez creada la nueva factoría de datos, puede restaurar las canalizaciones, los conjuntos de datos y los servicios vinculados JSON desde el repositorio de Git existente. Para más información, consulte Continuidad empresarial y recuperación ante desastres (BCDR) para canalizaciones de Data Factory y Azure Synapse Analytics.

Dependiendo del IR que utilice, puede haber otras consideraciones.

- Un *Azure-SSIS IR* usa una base de datos almacenada en Azure SQL Database o Azure SQL Managed Instance. Puede configurar la replicación geográfica o un grupo de

conmutación por error para esta base de datos. La base de datos Azure-SSIS se encuentra en una región primaria de Azure que tiene acceso de lectura y escritura. La base de datos se replica continuamente en una región secundaria que tiene acceso de solo lectura. Si la región primaria no está disponible, se desencadena una conmutación por error, lo que hace que las bases de datos principales y secundarias intercambien roles.

También puede configurar un par de Azure-SSIS IR en espera dual que funcione en sincronización con el grupo de conmutación por error de Azure SQL Database o SQL Managed Instance.

Para obtener más información, consulte Configurar un IR de Azure-SSIS para BCDR.

- *Una SHIR* se ejecuta en la infraestructura que administra. Si el SHIR se implementa en una máquina virtual de Azure, puede usar Azure Site Recovery para desencadenar la conmutación por error de la máquina virtual en otra región.

Copias de seguridad y restauración

Data Factory admite CI/CD a través de la integración del control de código fuente, de modo que pueda realizar una copia de seguridad de los metadatos de una instancia de Data Factory. Las canalizaciones de CI/CD implementan estos metadatos sin problemas en un nuevo entorno. Para más información, consulte CI/CD en Data Factory.

Acuerdo de nivel de servicio

El contrato de nivel de servicio (SLA) para los servicios de Azure describe la disponibilidad esperada de cada servicio y las condiciones que la solución deberá cumplir para lograr esa expectativa de disponibilidad. Para obtener más información, consulte Acuerdos de Nivel de Servicio para servicios en línea.

Data Factory proporciona acuerdos de nivel de servicio de disponibilidad independientes para:

- La tasa de éxito de las llamadas API que hagas, como aquellas para gestionar tu fábrica de datos.
- Número de ejecuciones de actividad que comienzan a ejecutarse.

Las ejecuciones de actividad se pueden retrasar brevemente y requieren que se cumplan todas las dependencias para ejecutar el trabajo.

Contenido relacionado

- Confiabilidad en Azure

Last updated on 22/01/2026

Confiabilidad de Azure Functions

Azure Functions es un servicio de proceso controlado por eventos que ejecuta pequeños bloques de código o *funciones* sin aprovisionamiento o administración de infraestructuras explícitas. Las funciones pueden responder a eventos como solicitudes HTTP, temporizadores, mensajes de cola y cambios en otros servicios de Azure. Esta funcionalidad hace que Functions sea adecuado para el procesamiento de datos, la integración del sistema y las tareas en segundo plano.

Cuando se usa Azure, la confiabilidad es una responsabilidad compartida. Microsoft proporciona una variedad de funcionalidades para admitir resistencia y recuperación. Es responsable de comprender cómo funcionan esas funcionalidades dentro de todos los servicios que usa y de seleccionar las funcionalidades que necesita para cumplir los objetivos empresariales y los objetivos de tiempo de actividad.

En este artículo se describe cómo hacer que Functions sea resistente a varias posibles interrupciones y problemas, incluidos errores transitorios, errores de zona de disponibilidad y errores en toda la región. También resalta información clave sobre el acuerdo de nivel de servicio (SLA) de Functions.

Recomendaciones de implementación de producción

Azure Well-Architected Framework proporciona recomendaciones sobre confiabilidad, seguridad, costo, operaciones y rendimiento. Para obtener información sobre cómo estas áreas influyen entre sí y contribuyen a una solución confiable de Functions, consulte Procedimientos recomendados de arquitectura para Functions.

Introducción a la arquitectura de confiabilidad

Al implementar Functions, es importante familiarizarse con estos conceptos:

- **Planes de hospedaje:** Los planes representan el entorno de hospedaje de las aplicaciones de funciones. El plan determina los recursos de proceso disponibles, el modelo de precios y el comportamiento de escalado.
- **Cuentas de almacenamiento:** Al crear una aplicación de funciones, debe especificar una cuenta de almacenamiento de host. La cuenta de almacenamiento gestiona aspectos de las operaciones internas de la aplicación de funciones, incluidos el almacenamiento de código

de funciones, el registro y la gestión de concurrencia (como en el caso de las concesiones de blobs para tipos específicos de desencadenadores).

También puede usar una cuenta de almacenamiento para la implementación. Esta cuenta de almacenamiento puede ser la misma que la cuenta de almacenamiento de host o una cuenta de almacenamiento diferente.

> ⓘ **Importante**
>
> Las cuentas de almacenamiento son una parte fundamental de la arquitectura de confiabilidad de Functions. Configúrelos para satisfacer los requisitos de resistencia de la aplicación de funciones.

- **Desencadenadores y enlaces**: Los desencadenadores y enlaces permiten que la función responda a eventos, reciba datos de otros servicios y escriba datos en otros servicios.
- **Durable Functions**: Durable Functions es una característica de Functions. Proporciona funciones con estado, como orquestaciones de larga duración y entidades con estado.

  Cuando se usan funciones duraderas, se configura un proveedor de almacenamiento que almacena el estado. Evalúe las características de confiabilidad del almacén de estado que elija y configúrela para satisfacer los requisitos de resistencia.

Resistencia a errores transitorios

Los errores transitorios son errores breves e intermitentes en los componentes. Se producen con frecuencia en un entorno distribuido como la nube y son una parte normal de las operaciones. Los errores transitorios se corrigen después de un breve período de tiempo. Es importante que las aplicaciones puedan controlar errores transitorios, normalmente mediante el reintento de solicitudes afectadas.

Todas las aplicaciones hospedadas en la nube deben seguir las instrucciones de control de errores transitorios de Azure cuando se comunican con cualquier API, bases de datos y otros componentes hospedados en la nube. Para obtener más información, consulte Recomendaciones para controlar errores transitorios.

Tenga en cuenta las siguientes recomendaciones para controlar errores transitorios en las aplicaciones de funciones:

- **Desencadenadores y enlaces:** La plataforma Functions incluye el control de errores transitorios integrado para desencadenadores y enlaces. Cuando se produce un error transitorio mientras se desencadena un desencadenador compatible o se leen o escriben datos de enlace admitidos, la plataforma puede reintentar automáticamente la operación. Este comportamiento de reintento integrado garantiza que los problemas de conectividad temporal o las interrupciones del servicio no impidan que la función se ejecute. Para obtener más información, consulte Control de errores y reintentos de Functions.

  Esta protección solo cubre errores transitorios. La plataforma no reintenta los errores persistentes, como un cadena de conexión mal configurado o un recurso eliminado.

  La plataforma trata los errores persistentes y los errores transitorios repetidos como errores. Configure el registro para capturar información sobre los errores de ejecución de funciones. Para obtener más información, consulte Configuración de la supervisión de Funciones.

- **Tu código de función:** En tu código de función, eres responsable de controlar errores transitorios cuando tu función llame a servicios externos. Implemente la lógica de reintento, los tiempos de espera y los patrones de cortacircuito según corresponda para las llamadas de servicio externas realizadas en su código de función. Diseñe las funciones para que sean idempotentes siempre que sea posible, de modo que los reintentos no creen efectos secundarios repetidos.

- **Clientes:** Las aplicaciones cliente que se conectan a funciones de forma sincrónica, como a través de una conexión HTTP, deben ser resistentes a errores transitorios.

Resistencia a errores de zona de disponibilidad

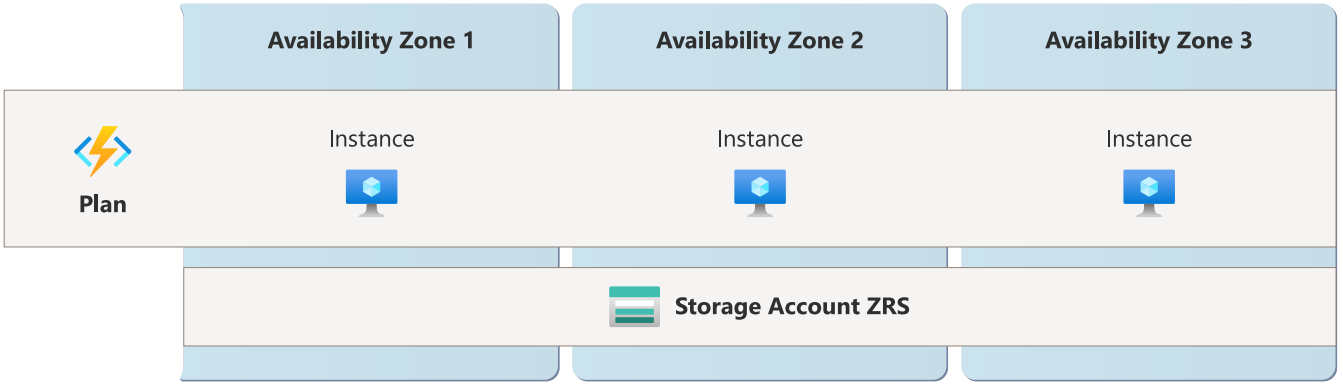
Las zonas de disponibilidad son grupos físicamente independientes de centros de datos dentro de una región de Azure. Cuando una zona falla, los servicios pueden transferirse a una de las zonas restantes.

Los planes Flex Consumption admiten implementaciones con redundancia de zona.

Cuando se habilita la redundancia de zona, la plataforma distribuye automáticamente las instancias del plan en todas las zonas de disponibilidad de la región seleccionada. Si alguna zona de disponibilidad de la región tiene un problema, las funciones continúan ejecutándose en instancias de zonas saludables.

Debe habilitar el almacenamiento con redundancia de zona (ZRS) en la cuenta de almacenamiento del host, para garantizar que también sea resistente a las interrupciones de

zona.



El plan es *no zonal* o *regional* cuando no se habilita la redundancia de zona. La plataforma puede ubicar instancias de plan en cualquier zona de disponibilidad de la región o en la misma zona. Las instancias del plan no son tolerantes a fallos en las zonas de disponibilidad. Es posible que su plan experimente tiempo de inactividad durante una interrupción en cualquier zona de la región.

Requisitos

- **Compatibilidad con regiones:** Puede implementar planes de consumo flexible con redundancia de zona en un conjunto específico de regiones. Puede recuperar la lista actual de regiones admitidas mediante la CLI de Azure. Para más información, consulte Visualización de regiones que admiten zonas de disponibilidad.

- **Cuenta de almacenamiento de host:** Debe configurar la cuenta de almacenamiento de host predeterminada de la aplicación de funciones para usar ZRS. Si usa una cuenta de almacenamiento de host que no está configurada para ZRS, es posible que la aplicación se comporte inesperadamente durante una interrupción de zona.

- **Cuenta de almacenamiento de contenedor de implementación:** Si usa una cuenta de almacenamiento independiente para el contenedor de implementación de la aplicación, actualícela también para que sea redundante de zona.

Consideraciones

Comportamientos no en tiempo de ejecución: La redundancia de zona solo garantiza un tiempo de actividad continuado para las aplicaciones implementadas. Una interrupción de zona de disponibilidad puede afectar a aspectos de Functions, aunque la aplicación siga sirviendo tráfico. Estos comportamientos incluyen el escalado del plan, la creación de aplicaciones, la configuración de la aplicación y la publicación de aplicaciones.

Distribución de instancias a través de zonas

Al configurar las aplicaciones del plan de consumo Flex como redundantes por zona, la plataforma distribuye automáticamente las instancias del plan en varias zonas de la región seleccionada, aplicando reglas diferentes para instancias siempre activas en comparación con las instancias bajo demanda.

- **Las instancias siempre disponibles** se distribuyen entre al menos dos zonas mediante la distribución circular.

  Para garantizar la resistencia de zona, la plataforma mantiene automáticamente al menos dos instancias listas para siempre para cada función o grupo de escalado por función, independientemente de la configuración siempre lista para la aplicación. Las instancias que crea la plataforma son administradas por la plataforma, facturadas como instancias listas para siempre y no cambian las opciones de configuración siempre preparadas.

- **Las instancias a petición** se generan a partir de volúmenes fuente de eventos cuando la aplicación se escala más allá del número de instancias siempre listas. Las instancias a petición se distribuyen entre zonas de disponibilidad de la mejor manera posible. La plataforma prioriza un escalado horizontal más rápido en lugar de una distribución uniforme entre zonas. La plataforma intenta equilibrar la distribución a lo largo del tiempo.

Costo

No se incurre en ningún costo adicional al habilitar la redundancia de zona. Los precios de un plan con redundancia de zona son los mismos que los de un plan de zona única. Sin embargo, habilitar la redundancia de zona afecta al número mínimo de instancias del plan.

Cuando se habilitan zonas de disponibilidad en una aplicación con una configuración de instancia del tipo siempre lista, y hay menos de dos instancias para cada función o grupo de escalado por función, la plataforma crea automáticamente dos instancias del tipo siempre lista para cada función o grupo de escalado por función. Estas nuevas instancias también se facturan como instancias siempre listas.

Para más información sobre los precios completos, consulte Precios de Functions.

Configurar soporte de zonas de disponibilidad

- **Cree un nuevo plan de Funciones con redundancia de zona.** Puede habilitar la redundancia de zona al crear un nuevo plan. Para obtener más información, consulte Creación de una

aplicación de funciones con redundancia de zona.

- **Habilite la redundancia de zona en un plan existente.** Puede activar o desactivar las zonas de disponibilidad para los planes de Elastic Premium existentes. Los planes elastic Premium tienen un comportamiento de capacidad específico que difiere de los planes dedicados (App Service) y requiere pasos de configuración adicionales. Para ver los pasos detallados, consulte Habilitación de la redundancia de zona en un plan existente.

Planeamiento y administración de capacidad

Las aplicaciones de funciones con redundancia de zona siguen ejecutándose incluso cuando las zonas de la región experimentan una interrupción.

Durante una interrupción en la zona, Functions detecta instancias perdidas e intenta automáticamente buscar o crear instancias de reemplazo en las zonas saludables. La plataforma realiza este proceso en base al mejor esfuerzo y no garantiza el éxito. Si la carga de trabajo requiere un número específico de instancias para mantener el nivel de servicio esperado, considere *la posibilidad de sobreaprovisionar* el número de instancias siempre listas. El sobreaprovisionamiento permite a la solución tolerar la pérdida de capacidad y seguir funcionando sin reducir el rendimiento. Para más información, consulte Administración de la capacidad mediante el sobreaprovisionamiento.

Comportamiento cuando todas las zonas están en buen estado

En esta sección se describe qué esperar cuando un plan es con redundancia de zona, la cuenta de almacenamiento de host usa ZRS y todas las zonas de disponibilidad están operativas.

- **Operación entre zonas:** Al configurar la redundancia de zona en Functions, las solicitudes se distribuyen automáticamente entre las instancias de cada zona de disponibilidad. Una solicitud puede ir a cualquier instancia de cualquier zona de disponibilidad.

- **Replicación de datos entre zonas:** Functions es un servicio de proceso sin estado, por lo que no hay datos para replicar entre zonas. La plataforma replica la configuración entre zonas automáticamente.

  Si la cuenta de almacenamiento de host usa ZRS, Azure Storage replica sincrónicamente sus datos en varias zonas de disponibilidad.

  Para funciones duraderas, revise el proveedor de almacenamiento para obtener información sobre cómo replica datos entre zonas.

Comportamiento durante un fallo de zona

En esta sección se describe qué esperar cuando un plan tiene redundancia de zona, la cuenta de almacenamiento del host usa ZRS y hay una interrupción en una zona de disponibilidad.

- **Detección y respuesta:** La plataforma Functions es responsable de detectar un error en una zona de disponibilidad. No es necesario iniciar una conmutación por error de zona.
- **Notificación**: Microsoft no le notifica automáticamente cuando una zona está inactiva. Sin embargo, puede usar Azure Resource Health para supervisar el estado de un recurso individual y puede configurar alertas de Resource Health para notificarle problemas. También puede usar Azure Service Health para comprender el estado general del servicio, incluidos los errores de zona, y puede configurar alertas de Service Health para notificarle problemas.
- **Solicitudes activas:** Cuando una zona de disponibilidad no está disponible, las solicitudes en curso que se conectan a una instancia de la zona de disponibilidad errónea finalizan y se deben reintentar. Asegúrese de que las aplicaciones están preparadas siguiendo las instrucciones de control de errores transitorios.
- **Pérdida de datos esperada:** No se espera que los errores de zona causen pérdida de datos porque Functions es un servicio sin estado.

  Si la cuenta de almacenamiento de host utiliza ZRS, el almacenamiento garantiza que no se pierdan datos debido a un fallo de zona.

  Para las funciones duraderas, compruebe su proveedor de almacenamiento para saber si la pérdida de datos sea posible durante un fallo de zona.
- **Tiempo de inactividad esperado:** Durante las interrupciones de zona, las conexiones pueden experimentar interrupciones breves que suelen durar unos segundos a medida que se redistribuye el tráfico. Asegúrese de que las aplicaciones están preparadas siguiendo las instrucciones de control de errores transitorios.
- **Reenrutamiento del tráfico:** Functions detecta las instancias perdidas de esa zona e intenta buscar nuevas instancias de reemplazo. Una vez que Functions encuentra reemplazos, distribuye el tráfico entre las nuevas instancias según sea necesario.

  > ⓘ **Importante**

> Azure no garantiza que las solicitudes de más instancias se realicen correctamente en un escenario de zona descendente. La plataforma intenta reponer instancias perdidas de forma óptima. Si necesita capacidad garantizada durante un error de zona de disponibilidad, cree y configure los planes para tener en cuenta la pérdida de zona mediante el sobreaprovisionamiento de la capacidad.

- **Comportamientos no en tiempo de ejecución:** Las aplicaciones de un plan de aplicación de funciones con redundancia de zona siguen ejecutándose y atendiendo el tráfico incluso si una zona de disponibilidad experimenta una interrupción. Sin embargo, los comportamientos que no son de ejecución pueden verse afectados durante una interrupción de la zona de disponibilidad. Estos comportamientos incluyen el escalado de aplicaciones de funciones, la creación de aplicaciones, la configuración de la aplicación y la publicación de aplicaciones.

Recuperación de zona

Cuando se recupera la zona de disponibilidad, Functions restaura automáticamente las instancias de la zona de disponibilidad, quita las instancias temporales creadas en las otras zonas de disponibilidad y vuelve a enrutar el tráfico entre las instancias como normal.

Prueba de fallos de zona

La plataforma Functions administra el enrutamiento del tráfico, la conmutación por error y la recuperación de zona para los recursos con redundancia de zona. No es necesario iniciar nada. Azure administra completamente esta característica, por lo que no es necesario validar los procesos de error de zona de disponibilidad.

Resistencia a errores en toda la región

Functions es un servicio de una sola región. Si la región deja de estar disponible, el recurso de Functions tampoco está disponible.

Soluciones personalizadas de varias regiones para la resistencia

Para evitar interrupciones en su servicio durante caídas regionales, puede implementar de forma redundante las mismas funciones en aplicaciones de funciones en varias regiones.

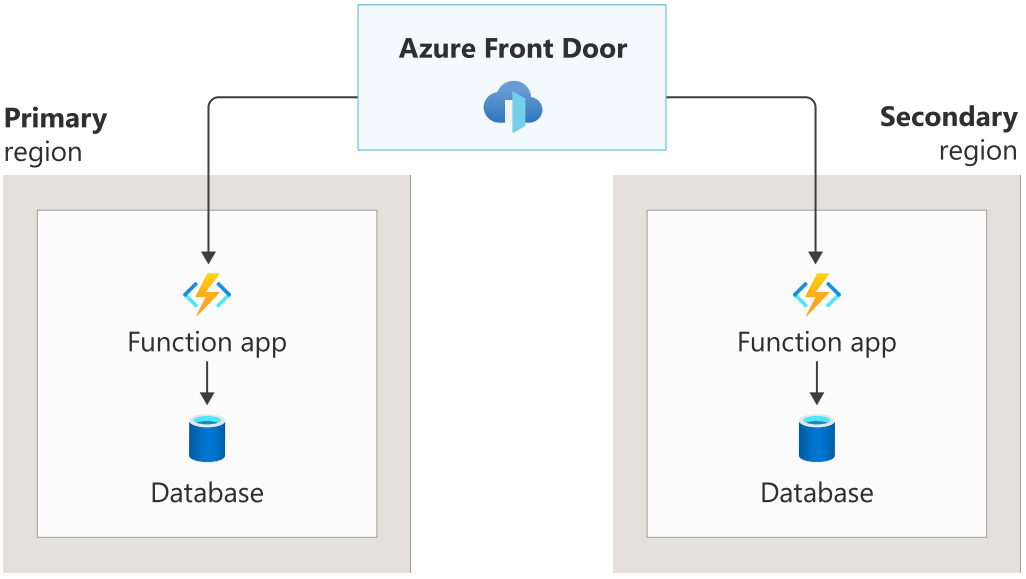
Usted es responsable de:

- Implementación de aplicaciones de funciones en varias regiones.
- Administración de la distribución del tráfico entre regiones.
- Implementación de mecanismos de conmutación por error.
- Garantizar la coherencia de los datos entre regiones (si procede).
- Supervisión y administración de implementaciones entre regiones.

Al ejecutar el mismo código de función en varias regiones, tenga en cuenta los patrones *activo-activo* y *activo-pasivo* . En las secciones siguientes se presentan estos patrones, pero no se proporcionan instrucciones detalladas ni pasos de configuración.

Patrón activo-activo para funciones de desencadenador HTTP

En un patrón activo-activo, las funciones en las dos regiones ejecutan activamente y procesan eventos, ya sea de forma duplicada o de manera alternada. Utilice un patrón activo-activo en combinación con Azure Front Door para sus funciones críticas desencadenadas por HTTP, que pueden enrutar y distribuir solicitudes HTTP entre funciones que se están ejecutando en varias regiones. Azure Front Door también puede comprobar periódicamente el estado de cada punto de conexión. Si un servicio de una región deja de responder a las comprobaciones de estado, Azure Front Door lo quita de la rotación y solo reenvía el tráfico a los servicios saludables restantes.



Patrón activo-pasivo para funciones de desencadenador que no son HTTP

En el caso de funciones dirigidas por eventos que no se activan mediante HTTP (como activadores de Azure Service Bus y Azure Event Hubs), use un patrón activo-pasivo. En un patrón

activo-pasivo, las instancias de función se ejecutan en la región que recibe eventos, mientras que las instancias de la región secundaria permanecen inactivas. Este patrón garantiza que solo una función procese cada mensaje, lo que ayuda a mantener la coherencia de los datos. Permite conmutar por error a la región secundaria durante un desastre, como una interrupción de la región.

Considere la conmutación por error de la app de funciones junto con los comportamientos de conmutación por error de otros servicios que use, como:

- Service Bus, replicación geográfica y recuperación ante desastres
- Replicación geográfica de Event Hubs y recuperación ante desastres geográfica

Considere una topología de ejemplo que usa un activador de Event Hubs, donde el espacio de nombres de Event Hubs está configurado para la recuperación ante desastres geográficos. En este caso, el patrón activo-pasivo requiere los siguientes componentes:

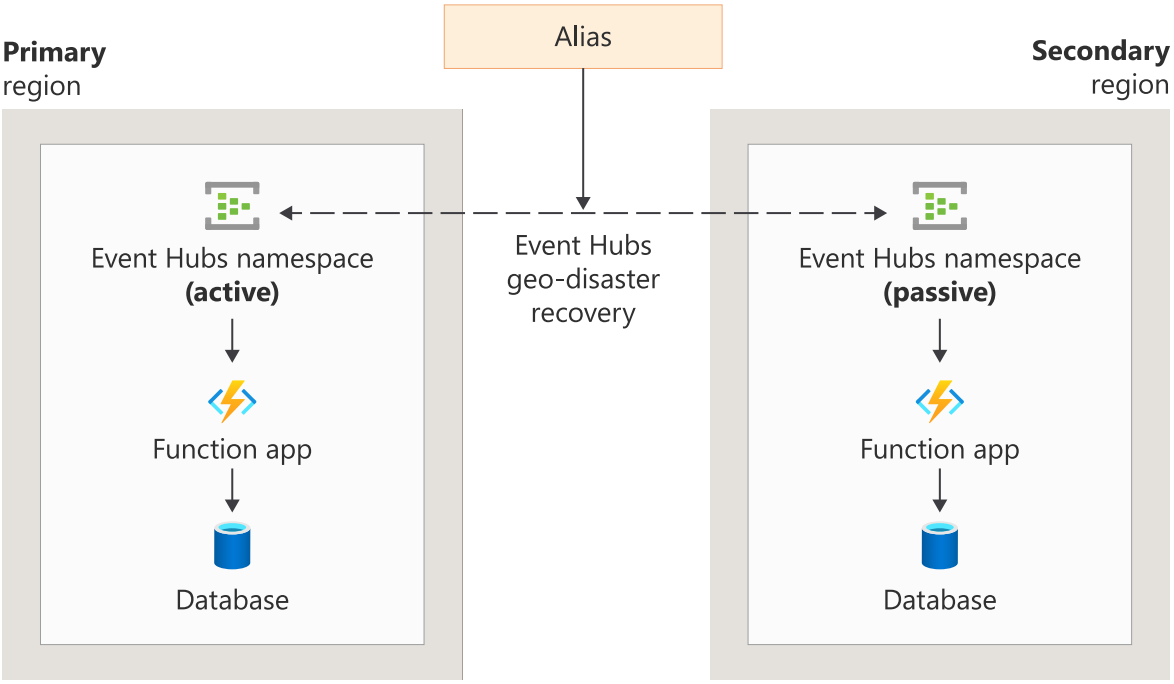
- Espacios de nombres de Event Hubs implementados tanto en una región primaria como en una secundaria.

- Se habilitó la recuperación ante desastres geográfica para vincular los centros de eventos principales y secundarios. Esta configuración también crea un *alias* que puede usar para conectarse al espacio de nombres de Event Hubs y cambiar el alias de la base de datos principal a la secundaria sin cambios en la información de conexión.

- Aplicaciones de funciones desplegadas tanto en la región primaria como en la secundaria. La aplicación de la región secundaria permanece inactiva porque no recibe mensajes.

- Los desencadenadores de cada aplicación de funciones utilizan la cadena de conexión *direct* (noalias) para su espacio de nombres correspondiente de Event Hubs.

- Los publicadores del espacio de nombres de Event Hubs publican al alias de la cadena de conexión.



Antes de la conmutación por error, los publicadores que envían eventos al alias compartido dirigen el tráfico hacia el centro de eventos principal. La aplicación de funciones principal escucha exclusivamente al hub de eventos principal. La aplicación de funciones secundaria permanece pasiva e inactiva.

Cuando se inicia el failover, los publicadores que envían eventos al alias compartido son dirigidos al centro de eventos secundario. La aplicación de funciones secundaria se activa y se desencadena automáticamente. El centro de eventos puede controlar todo el proceso de conmutación por error y cada aplicación de funciones solo se ejecuta cuando su centro de eventos correspondiente está activo.

Durable Functions

Para la recuperación ante desastres en varias regiones para funciones duraderas, consulte Disaster recovery and geo-distribution in Azure durable functions.

Resistencia al mantenimiento del servicio

Functions realiza actualizaciones de servicio periódicas y otras tareas de mantenimiento.

- **Resistencia de errores transitorios:** Durante el mantenimiento del servicio, las instancias que ejecutan la aplicación de funciones pueden reiniciarse o experimentar interrupciones temporales. Asegúrese de que las aplicaciones cliente que interactúan con la aplicación de funciones son resistentes a errores transitorios.

- **Habilite la redundancia de zona:** Al habilitar la redundancia de zona en el plan, también se mejora la resistencia durante las actualizaciones de la plataforma. Implementar varias instancias en tu plan y habilitar la redundancia de zonas proporciona una capa adicional de resiliencia si una instancia o una zona se vuelven defectuosas durante un proceso de actualización.

Resistencia a las implementaciones de aplicaciones

Las implementaciones de aplicaciones presentan el riesgo de problemas en un entorno de producción. Prepárese para revertir una actualización si provoca problemas. Controlar cómo se implementan las actualizaciones para reducir la interrupción de los reinicios de la aplicación.

Los planes de consumo flexible admiten estrategias de actualización del sitio, que proporcionan varias maneras de implementar las actualizaciones de la aplicación. Estas estrategias incluyen actualizaciones graduales para implementaciones sin tiempo de inactividad.

Acuerdo de nivel de servicio

El acuerdo de nivel de servicio (SLA) para Azure servicios describe la disponibilidad esperada de cada servicio y las condiciones que la solución debe cumplir para lograr esa expectativa de disponibilidad. Para obtener más información, consulte Acuerdos de Nivel de Servicio para servicios en línea.

Functions proporciona acuerdos de nivel de servicio de disponibilidad distintos para el plan de consumo y para otros tipos de plan.

Contenido relacionado

- Recuperación ante desastres y distribución geográfica en funciones duraderas de Azure
- Creación de una instancia de Azure Front Door
- Consideraciones sobre la conmutación por error de Event Hubs
- Estrategias de arquitectura para usar regiones y zonas de disponibilidad
- Confiabilidad en Azure

Last updated on 28/04/2026

Confiabilidad en Azure Kubernetes Service (AKS)

Azure Kubernetes Service (AKS) es un servicio de orquestación de contenedores administrado que simplifica la implementación, la administración y las operaciones de Kubernetes.

Cuando se usa Azure, la confiabilidad es una responsabilidad compartida. Microsoft proporciona una variedad de funcionalidades para admitir resistencia y recuperación. Es responsable de comprender cómo funcionan esas funcionalidades dentro de todos los servicios que usa y de seleccionar las funcionalidades que necesita para cumplir los objetivos empresariales y los objetivos de tiempo de actividad.

En este artículo se describe cómo hacer que Azure Kubernetes Service (AKS) sea resistente a una variedad de posibles interrupciones y problemas, incluidos errores transitorios, interrupciones de zona de disponibilidad y interrupciones en regiones. También se describe cómo puede usar copias de seguridad para recuperarse de otros tipos de problemas y se resalta cierta información clave sobre el acuerdo de nivel de servicio (SLA) de Azure Kubernetes Service (AKS).

Recomendaciones de implementación de producción

Para obtener recomendaciones sobre cómo implementar cargas de trabajo de producción confiables en AKS, consulte los siguientes artículos:

- Procedimientos recomendados de implementación y confiabilidad de clústeres para AKS
- Descripción general de la alta disponibilidad (HA) y la recuperación ante desastres (DR) para AKS
- Consideraciones de resistencia de zona para AKS

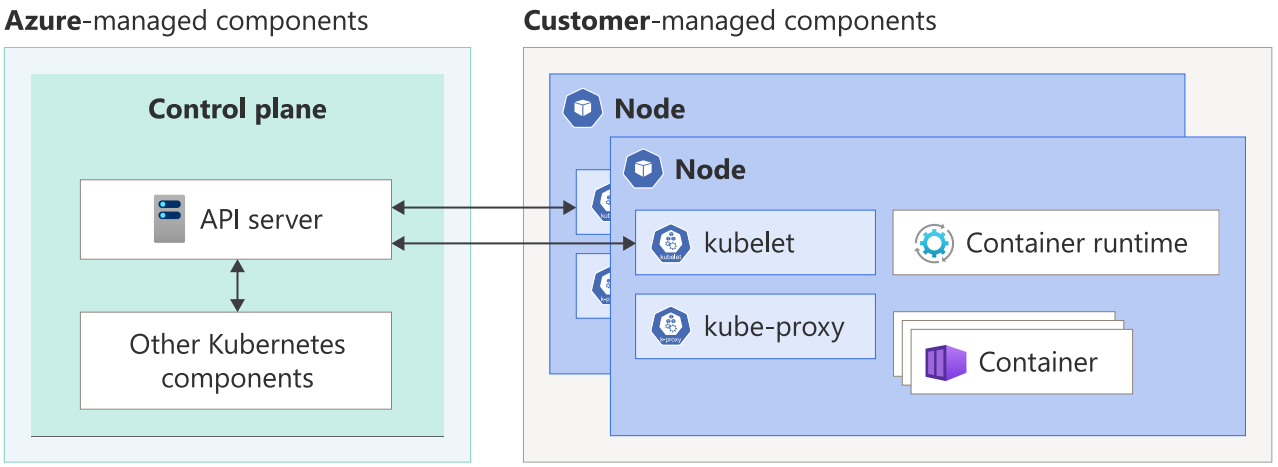
Introducción a la arquitectura de confiabilidad

Al crear un clúster de AKS, la plataforma Azure crea y configura automáticamente:

- Un plano de control que tiene el servidor de API, etcd, el programador y otros pods necesarios para administrar la carga de trabajo.

- Un grupo de nodos del sistema a tu suscripción que hospeda los complementos y pods adicionales que se ejecutan en el espacio de nombres *kube-system*.

Una vez completada la configuración del grupo de nodos inicial, puede agregar o eliminar grupos de nodos para sus propias cargas de trabajo de usuario. AKS no administra grupos de nodos para la confiabilidad y debe asegurarse de que las cargas de trabajo son resistentes a los errores de infraestructura.



La resistencia es una responsabilidad compartida entre usted y Microsoft. Como servicio de cálculo, AKS administra algunos aspectos de la fiabilidad del clúster, pero usted es responsable de gestionar otros aspectos.

- **Microsoft administra** el plano de control y otros componentes administrados de AKS.
- **Es su responsabilidad**:
  - *Defina cómo deben configurarse los componentes, incluidos los grupos de nodos y los equilibradores de carga que se conectan a los servicios, para satisfacer los requisitos de confiabilidad.* Después de definir los componentes, Microsoft los implementa y administra en su nombre.
  - *Administre los componentes fuera del clúster de AKS, incluido el almacenamiento y las bases de datos.* Compruebe que estos componentes cumplen los requisitos de confiabilidad. Al implementar las cargas de trabajo, asegúrese de que otros componentes de Azure también están configurados para la resistencia siguiendo los procedimientos recomendados para esos servicios.

Resistencia a errores transitorios

Los errores transitorios son errores breves e intermitentes en los componentes. Se producen con frecuencia en un entorno distribuido como la nube y son una parte normal de las operaciones. Los errores transitorios se corrigen después de un breve período de tiempo. Es importante que las aplicaciones puedan controlar errores transitorios, normalmente mediante el reintento de solicitudes afectadas.

Todas las aplicaciones hospedadas en la nube deben seguir las instrucciones de control de errores transitorios de Azure cuando se comunican con cualquier API, bases de datos y otros componentes hospedados en la nube. Para obtener más información, consulte Recomendaciones para controlar errores transitorios.

Cuando utilizas AKS, pueden ocurrir fallos transitorios por varias razones, como fallos de aplicaciones, operaciones de escalado y balanceo de pods, actualización de nodos y fallos temporales en la infraestructura, como problemas de hardware o de redes.

Es imposible eliminar todos los errores transitorios, por lo que los clientes que acceden a las aplicaciones hospedadas en AKS deben estar preparados para reintentar las solicitudes con errores y seguir otras recomendaciones de control de errores transitorios. Puede minimizar la probabilidad de errores transitorios y evitar o mitigar el tiempo de inactividad que podrían provocar siguiendo los procedimientos recomendados de Kubernetes y Azure en la implementación.

- ✓ Establece los presupuestos de interrupciones de pod (PDB) en el pod YAML para especificar cuántos pods necesitas tener en un estado `Ready` en un momento dado. Al establecer archivos PDB, AKS garantiza una disponibilidad mínima de réplicas cuando ejecuta operaciones para acordonar y purgar los nodos. Si no se puede satisfacer una PDB durante procesos como actualizaciones, el pod continúa funcionando y es posible que se produzca un error en la operación. Para obtener más información, consulte PDB.
- ✓ Use `maxUnavailable` para definir el número máximo de réplicas que pueden dejar de estar disponibles en un momento dado. Por ejemplo, cuando se realiza un reinicio gradual, AKS garantiza que no más del número `maxUnavailable` de pods se reinician en un momento dado. Para obtener más información, consulte maxUnavailable.
- ✓ Siga los procedimientos recomendados de implementación. Las réplicas de pod también pueden fallar debido a problemas de la aplicación. Para más información, consulte Procedimientos recomendados de nivel de implementación para la confiabilidad del clúster de AKS.

> ⓘ **Nota:**
>
> Si quiere que AKS valide las implementaciones para cumplir los procedimientos recomendados y proporcione notificaciones de bloqueo o advertencia, puede usar medidas de seguridad de implementación. Las medidas de seguridad de implementación son una oferta administrada que ayuda a aplicar los procedimientos recomendados del producto antes de que el código se implemente en el clúster.

Resistencia a errores de zona de disponibilidad

Las zonas de disponibilidad son grupos físicamente independientes de centros de datos dentro de una región de Azure. Cuando una zona falla, los servicios pueden transferirse a una de las zonas restantes.

Al implementar un clúster de AKS en una región que admita zonas de disponibilidad, los distintos componentes requieren distintos tipos de configuración.

El plano de control de AKS es resiliente a zonas por defecto. Si se produce un error en una zona, el plano de control no requiere ninguna configuración ni administración para lograr resistencia. Sin embargo, la resistencia del plano de control no es suficiente para que el clúster permanezca operativo cuando se produce un error en una zona. Para el grupo de nodos del sistema y los grupos de nodos de usuario que implemente, debe habilitar la compatibilidad con zonas de disponibilidad para ayudar a garantizar que las cargas de trabajo son resistentes a los errores de zona de disponibilidad.

Requisitos

Compatibilidad con regiones: Puede implementar clústeres de AKS resistentes a zonas en cualquier región que admita zonas de disponibilidad.

Consideraciones

Para mejorar la confiabilidad y la resistencia de las cargas de trabajo de producción de AKS en una región, debe configurar AKS para la redundancia de zona mediante la realización de las siguientes configuraciones:

- **Implemente varias réplicas.** Kubernetes distribuye los pods entre nodos en función de las etiquetas de nodo. Para distribuir la carga de trabajo entre zonas, debe implementar varias réplicas del pod. Por ejemplo, si configura el grupo de nodos con tres zonas pero solo implementa una única réplica de su pod, su implementación no es resiliente a las zonas.

- **Habilite el escalado automático.** Los grupos de nodos de Kubernetes proporcionan opciones de escalado manuales y automáticas. Con el escalado manual, puedes agregar o eliminar nodos según sea necesario y los pods pendientes esperan hasta que escales verticalmente un grupo de nodos. El escalado administrado por AKS usa el escalador automático del clúster o el aprovisionamiento automático de nodos (NAP). AKS aumenta o reduce el grupo de nodos según los requisitos del pod y teniendo en cuenta la cuota y capacidad de la SKU de tu suscripción. Este método ayuda a garantizar que los pods se programen en los nodos disponibles en las zonas de disponibilidad.

- **Establece las restricciones de la topología del pod.** Usa restricciones de propagación de topología de pods para controlar cómo se distribuyen los pods entre distintos nodos o zonas. Las restricciones le ayudan a lograr la alta disponibilidad, la resistencia y el uso eficaz de los recursos. Si prefieres propagar pods estrictamente entre zonas, puedes establecer restricciones para forzar un pod en un estado pendiente a fin de mantener el equilibrio de pods entre zonas. Para más información, consulta Restricciones de propagación de topología de pod.

- **Configurar redes con resiliencia zonal.** Si los pods sirven tráfico externo, configure la arquitectura de red del clúster mediante servicios como Azure Application Gateway, Azure Load Balancer o Azure Front Door.

- **Asegúrese de que las dependencias son resistentes a la zona.** La mayoría de las aplicaciones de AKS usan otros servicios para el almacenamiento, la seguridad o las redes. Asegúrese de revisar las recomendaciones de resistencia de zona para esos servicios.

Cost

No hay ningún cargo adicional para habilitar la compatibilidad con la zona de disponibilidad en AKS. Paga por las máquinas virtuales (VM) y otros recursos que implemente en las zonas de disponibilidad.

Configurar soporte de zonas de disponibilidad

- **Cree un nuevo clúster de AKS que tenga compatibilidad con la zona de disponibilidad:** Para configurar la compatibilidad con zonas de disponibilidad, consulte Creación de un clúster de Azure Kubernetes Service (AKS) que use zonas de disponibilidad.
- **Migración:** No se puede habilitar la compatibilidad con zonas de disponibilidad después de crear un clúster. En su lugar, debe crear un nuevo clúster que tenga habilitada la compatibilidad con la zona de disponibilidad y eliminar el clúster existente.
- **Deshabilite la compatibilidad con zonas de disponibilidad:** No se puede deshabilitar la compatibilidad con zonas de disponibilidad después de crear un clúster. En su lugar, debe crear un clúster nuevo que tenga compatibilidad con la zona de disponibilidad deshabilitada y eliminar el clúster existente.

Comportamiento cuando todas las zonas están en buen estado

En esta sección se describe qué esperar cuando los clústeres de AKS están configurados para la compatibilidad con zonas de disponibilidad y todas las zonas de disponibilidad están operativas.

- **Enrutamiento de tráfico entre zonas:** Al implementar un clúster de AKS que usa zonas de disponibilidad, es importante asegurarse de que los componentes de red también son resistentes a la zona. Según los equilibradores de carga y otros componentes de red que use, es posible que tenga que configurar explícitamente los componentes para enrutar el tráfico a los nodos correctos de las zonas correctas y responder a interrupciones de zona. Para más información, consulte Consideraciones de resistencia de zona para AKS.

- **Replicación de datos entre zonas:** Si ejecuta una carga de trabajo sin estado, debe usar servicios administrados de Azure, como bases de datos de Azure, Azure Managed Redis o Azure Storage para almacenar los datos de la aplicación. Puede usar estos servicios para ayudar a garantizar que el tráfico se pueda mover entre nodos y zonas sin riesgo de pérdida de datos ni que afecte a la experiencia del usuario. Puedes usar implementaciones, servicios y sondeos de estado de Kubernetes para administrar pods sin estado y garantizar la distribución uniforme entre zonas.

  Si necesita almacenar el estado dentro del clúster mediante discos de Azure, use el almacenamiento con redundancia de zona de Azure para asegurarse de que los datos se replican en varias zonas de disponibilidad. Para obtener más información, consulte Elección del tipo de disco adecuado en función de las necesidades de la aplicación.

Comportamiento durante un fallo de zona

En esta sección se describe qué esperar cuando se produce una interrupción de zona de disponibilidad mientras los clústeres de AKS están configurados con compatibilidad con zonas de disponibilidad.

- **Detección y respuesta**: cuando se produce una interrupción de zona, el plano de control conmuta automáticamente por error. Si los grupos de nodos usan zonas de disponibilidad y siguen losprocedimientos recomendados de resistencia de zona, puedes esperar que AKS abra nodos y réplicas en las zonas que están operativas. AKS realiza esta tarea automáticamente cuando se usan soluciones administradas como cluster autoscaler o NAP. Sin el escalado automático, los nodos y las réplicas permanecen en estado *Pendiente* y esperan a que la intervención manual escale verticalmente el grupo de nodos.

  AKS también intenta reequilibrar los pods en las zonas correctas. Si decides escalar manualmente el grupo de nodos en un escenario de reducción de zona, los pods pueden dejarse en el estado *Pendiente* cuando no haya nodos disponibles en las zonas correctas. El escalado horizontal en las zonas restantes también está sujeto a la disponibilidad de la cuota y la capacidad de la SKU de máquina virtual que uses.

- **Notificación**: Microsoft no le notifica automáticamente cuando una zona está inactiva. Sin embargo, puede usar Azure Resource Health para supervisar el estado de un recurso

individual y puede configurar alertas de Resource Health para notificarle problemas. También puede usar Azure Service Health para comprender el estado general del servicio, incluidos los errores de zona, y puede configurar alertas de Service Health para notificarle problemas.

También puede usar las métricas de estado del nodo o pod para supervisar el estado de los nodos y pods.

- **Solicitudes activas:** Las solicitudes activas pueden experimentar interrupciones. Algunas solicitudes pueden fallar, y la latencia podría aumentar mientras la carga de trabajo se transfiere a otra zona.
- **Pérdida de datos esperada:** Si almacena el estado dentro del clúster mediante discos de Azure y usa almacenamiento con redundancia de zona, no se espera que un error de zona cause ninguna pérdida de datos.
- **Tiempo de inactividad esperado:** Si configura correctamente la resistencia de zona para el clúster y los pods, no se espera que un error de zona cause tiempo de inactividad para la carga de trabajo de AKS.
- **Redireccionamiento del tráfico:** los equilibradores de carga redirigen las nuevas solicitudes entrantes a pods que se ejecutan en nodos saludables.

Para más información, consulte Consideraciones de resistencia de zona para AKS.

Recuperación de zona

Cuando se recupera la zona de disponibilidad, el comportamiento de la conmutación por recuperación depende del componente:

- **Plano de control:** AKS restaura automáticamente las operaciones del plano de control en todas las zonas de disponibilidad. No se requiere intervención manual.
- **Grupos de nodos y nodos**: inmediatamente después de la conmutación por recuperación, los nodos permanecen en las zonas previamente correctas y no se restauran en la zona recuperada. Pero la próxima vez que se realice una operación de escalado de nodos, como al escalar horizontalmente el grupo de nodos, el grupo de nodos puede crear nodos en la zona recuperada.
- **Pods**: inmediatamente después de la conmutación por recuperación, los pods siguen ejecutándose en los nodos en los que se están ejecutando. Cuando se crean nuevos pods o se vuelven a crear los pods existentes, son aptos para usar nodos en la zona recuperada.

- **Almacenamiento:** cualquier almacenamiento conectado a pods se recupera en función de cómo funciona el almacenamiento con redundancia de zona.

Prueba de fallos de zona

Puede probar la resistencia a los errores de zona de disponibilidad mediante los métodos siguientes:

- Aislar y drenar nodos en una única zona de disponibilidad
- Simulación de un error de zona de disponibilidad mediante Azure Chaos Studio

Resistencia a errores en toda la región

Los clústeres de AKS son recursos de una sola región. Si la región no está disponible, el clúster de AKS tampoco está disponible.

Soluciones personalizadas de varias regiones para la resistencia

Si necesita implementar la carga de trabajo de Kubernetes en varias regiones de Azure, tiene dos opciones para administrar la orquestación de estos clústeres.

- Use Fleet Manager de Azure Kubernetes si desea una experiencia administrada más sencilla. Con Fleet Manager de Azure Kubernetes, puede hacer lo siguiente:
  - Administre un conjunto de clústeres de AKS como una sola unidad y esos clústeres se pueden distribuir entre varias regiones de Azure.
  - Automatice aspectos específicos de la administración de clústeres, como las actualizaciones de imágenes de clúster y nodo.
  - Utilice las capacidades de distribución de tráfico para repartir el tráfico entre los clústeres y realizar una conmutación automática si una región no está disponible.
- Organice la conmutación por error mediante un modelo de implementación activo-activo o activo-pasivo manual si la carga de trabajo requiere un control más matizado sobre los distintos componentes de las conmutaciones por error interregionales. Para obtener más información, consulte Información general sobre alta disponibilidad y recuperación ante desastres para AKS.

Copias de seguridad y restauración

Azure Backup tiene una extensión que puede usar para realizar copias de seguridad de los recursos del clúster de AKS y volúmenes persistentes que se conectan al clúster. El almacén de Backup se comunica con el clúster de AKS a través de la extensión para realizar operaciones de copia de seguridad y restauración.

Si el clúster de AKS está en una región emparejada, puede configurar las copias de seguridad que se almacenarán en el almacenamiento con redundancia geográfica. Puede restaurar copias de seguridad georedundantes en la región emparejada.

Para obtener más información, consulte los artículos siguientes:

- ¿Qué es la copia de seguridad de Azure Kubernetes Service?
- Copia de seguridad de AKS mediante Azure Backup

Para la mayoría de las soluciones, no debe confiar exclusivamente en copias de seguridad. En su lugar, utilice las otras capacidades descritas en esta guía para apoyar los requisitos de resiliencia. Sin embargo, las copias de seguridad protegen contra algunos riesgos que otros enfoques no. Para más información, consulte ¿Qué son la redundancia, la replicación y la copia de seguridad?.

Intente usar clústeres sin estado que minimicen la necesidad de copia de seguridad. Almacene datos en bases de datos y sistemas de almacenamiento externos en lugar de en el clúster.

Resistencia al mantenimiento del servicio

AKS realiza el mantenimiento en el clúster, incluidas las actualizaciones del clúster y las imágenes de nodo. Para asegurarse de que Kubernetes mantiene el número mínimo de instancias de pod necesarias para atender el tráfico de producción incluso durante las actualizaciones, debe configurar los pods para que usen presupuestos de interrupciones del pod.

Para reducir las interrupciones del servicio durante períodos de tiempo críticos, AKS proporciona controles para que pueda especificar los tiempos de mantenimiento planeados. Para más información, consulte Uso del mantenimiento planeado para programar y controlar las actualizaciones del clúster de Azure Kubernetes Service.

Acuerdo de nivel de servicio

El contrato de nivel de servicio (SLA) para los servicios de Azure describe la disponibilidad esperada de cada servicio y las condiciones que la solución deberá cumplir para lograr esa expectativa de disponibilidad. Para obtener más información, consulte Acuerdos de Nivel de Servicio para servicios en línea.

AKS proporciona tres planes de tarifa para la administración de clústeres: **Gratis**, **Estándar** y **Premium**. El nivel Gratis permite usar AKS para probar las cargas de trabajo. Los niveles Estándar y Premium están diseñados para cargas de trabajo de producción. Al implementar un clúster de AKS con zonas de disponibilidad habilitadas, aumenta el porcentaje de tiempo de actividad definido en el Acuerdo de Nivel de Servicio. Sin embargo, el Acuerdo de Nivel de Servicio solo se aplica si implementa un clúster en el plan de tarifa Estándar o Premium.

Contenido relacionado

- Confiabilidad en Azure

Last updated on 22/01/2026

Implementación de un clúster de Azure Service Fabric en Availability Zones

Las zonas de disponibilidad de Azure son una función de alta disponibilidad que protege las aplicaciones y los datos frente a errores del centro de datos. Una zona de disponibilidad es una ubicación física única equipada con alimentación independiente, refrigeración y redes dentro de una región de Azure.

Para admitir clústeres que abarcan distintas instancias de Availability Zones, Azure Service Fabric proporciona los dos métodos de configuración que se describen en el artículo siguiente. Las zonas de disponibilidad solo se ofrecen en algunas regiones. Para más información, consulte la información general sobre las zonas de disponibilidad.

Hay plantillas de ejemplo disponibles en el repositorio de plantillas para uso entre zonas de disponibilidad de Service Fabric.

Topología para abarcar un tipo de nodo principal en diversas zonas de disponibilidad

> **Nota:**
>
> La ventaja de abarcar el tipo de nodo principal entre zonas de disponibilidad solo se ve para tres zonas y no solo dos.

- El nivel de confiabilidad del clúster ha de establecerse en `Platinum`.
- Un único recurso de IP pública que use la SKU Estándar.
- Un único recurso de Load Balancer que use la SKU Estándar.
- Un grupo de seguridad de red (NSG) referenciado por la subred en la que implementar los conjuntos de escalado de máquinas virtuales.

> **Nota:**
>
> La propiedad de grupo de selección única del conjunto de escalado de máquinas virtuales debe haberse establecido en `true`.

La siguiente lista de nodos de ejemplo muestra los formatos FD/UD en un conjunto de escalado de máquinas virtuales que expande zonas:

ALL NODES EVENTS

Search list Reset All

Name	Address	Node Type	Upgrade Domain	Fault Domain	Is Seed Node
_mgmt_0	10.0.0.16	mgmt	ud:/zone1/0	fd:/zone1/0	true
_mgmt_1	10.0.0.17	mgmt	ud:/zone2/0	fd:/zone2/0	false
_mgmt_10	10.0.0.18	mgmt	ud:/zone3/3	fd:/zone3/3	false
_mgmt_11	10.0.0.19	mgmt	ud:/zone1/3	fd:/zone1/3	false
_mgmt_12	10.0.0.20	mgmt	ud:/zone2/4	fd:/zone2/4	true
_mgmt_13	10.0.0.21	mgmt	ud:/zone3/4	fd:/zone3/4	false

Distribución de réplicas de servicio entre zonas

Al implementar un servicio en los tipos de nodo que expanden Availability Zones, las réplicas se incluyen para garantizar que se encuentren en zonas independientes. Los dominios de error en los nodos de cada uno de estos tipos de nodo se configuran con la información de zona (es decir, FD = fd:/zone1/1, etc.). Por ejemplo, para cinco réplicas o instancias de servicio, la distribución es 2-2-1 y el entorno de ejecución intentará garantizar una distribución uniforme entre zonas.

Configuración de la réplica de servicio de usuario

Los servicios de usuario con estado implementados en los tipos de nodo entre Availability Zones deben configurarse de esta forma: número de réplicas con destino = 9, min = 5. Esta configuración permite que el servicio funcione incluso cuando una zona deje de estar disponible, ya que seguirán habiendo seis réplicas activas en las otras dos zonas. Una actualización de la aplicación en este escenario también tendrá éxito.

Elemento ReliabilityLevel del clúster

Este valor define el número de nodos raíz del clúster y el tamaño de réplica de los servicios del sistema. Una configuración entre zonas de disponibilidad tiene un mayor número de nodos, que se distribuyen entre zonas para habilitar la resistencia de zona.

Un valor `ReliabilityLevel` más elevado garantiza que haya presentes más nodos raíz y réplicas de servicio del sistema y se distribuyan uniformemente entre zonas, de tal manera que si se produce un error en una zona, el clúster y los servicios del sistema no se verán afectados. El valor `ReliabilityLevel = Platinum` (recomendado) garantiza que nueve nodos raíz se extiendan entre zonas del clúster, con tres valores en cada zona.

Escenario de zona fuera de servicio

Cuando una zona está fuera de servicio, todos los nodos y las réplicas de servicio de esa zona aparecen fuera de servicio. Como hay réplicas en las demás zonas, el servicio seguirá respondiendo. Las réplicas principales conmutan por error en las zonas funcionales. Los servicios están en estado de advertencia aparentemente porque el número de réplicas de destino aún no se ha alcanzado y el número de máquinas virtuales sigue siendo mayor que el tamaño mínimo de réplica de destino.

El equilibrador de carga de Service Fabric abrirá las réplicas en las zonas funcionales para que coincidan con el número de réplicas de destino. En este momento, los servicios aparecerán correctamente. Cuando la zona que estaba fuera de servicio vuelve a ser funcional, el equilibrador de carga redistribuye todas las réplicas de servicio uniformemente entre las zonas.

Próximas optimizaciones

- Para proporcionar actualizaciones de infraestructura confiables, Service Fabric requiere que la durabilidad del conjunto de escalado de máquinas virtuales se establezca al menos en Silver. Esto permite que el conjunto de escalado de máquinas virtuales subyacente y el entorno de ejecución de Service Fabric proporcionen actualizaciones confiables. Esto también requiere que cada zona tenga como mínimo 5 VM. Estamos trabajando para bajar este requisito a 3 y 2 VM por zona para los tipos de nodo principales y no principales, respectivamente.
- Todas las configuraciones mencionadas a continuación y el próximo trabajo proporcionan migración local a los clientes donde se puede actualizar el mismo clúster para usar la nueva configuración mediante la adición de nuevos elementos nodeTypes y la retirada de los antiguos.

Requisitos de red

Recurso de IP pública y equilibrador de carga

Para habilitar la propiedad `zones` en un recurso de conjunto de escalado de máquinas virtuales, el recurso de equilibrador de carga e IP referenciado por ese conjunto de escalado de máquinas virtuales deben utilizar una SKU Estándar. La creación de un recurso de IP sin la propiedad SKU da lugar a un SKU Básico, que no admite Availability Zones. De forma predeterminada, un equilibrador de carga de SKU Estándar bloquea todo el tráfico desde fuera. Para permitir el tráfico externo, implemente un NSG en la subred.

JSON

```
{
  "apiVersion": "2018-11-01",
```

```
  "type": "Microsoft.Network/publicIPAddresses",
  "name": "[concat('LB','-', parameters('clusterName')]",
  "location": "[parameters('computeLocation')]",
  "sku": {
    "name": "Standard"
  }
}
```

JSON

```
{
  "apiVersion": "2018-11-01",
  "type": "Microsoft.Network/loadBalancers",
  "name": "[concat('LB','-', parameters('clusterName')]",
  "location": "[parameters('computeLocation')]",
  "dependsOn": [
    "[concat('Microsoft.Network/networkSecurityGroups/', concat('nsg',
parameters('subnet0Name')))]"
  ],
  "properties": {
    "addressSpace": {
      "addressPrefixes": [
        "[parameters('addressPrefix')]"
      ]
    },
    "subnets": [
      {
        "name": "[parameters('subnet0Name')]",
        "properties": {
          "addressPrefix": "[parameters('subnet0Prefix')]",
          "networkSecurityGroup": {
            "id": "[resourceId('Microsoft.Network/networkSecurityGroups',
concat('nsg', parameters('subnet0Name')))]"
          }
        }
      }
    ]
  },
  "sku": {
    "name": "Standard"
  }
}
```

> ⓘ **Nota:**
>
> No es posible realizar un cambio local de SKU en los recursos de IP pública. Si va a hacer una migración a partir de recursos existentes que tienen un SKU Básico, consulte la sección sobre migración en este artículo.

Reglas NAT para conjuntos de escalado de máquinas virtuales

Las reglas de traducción de direcciones de red (NAT) de entrada para el equilibrador de carga deben coincidir con los grupos NAT del conjunto de escalado de máquinas virtuales. Cada conjunto de escalado de máquinas virtuales debe tener un único grupo NAT de entrada.

JSON

```
{
  "inboundNatPools": [
    {
      "name": "LoadBalancerBEAddressNatPool0",
      "properties": {
        "backendPort": "3389",
        "frontendIPConfiguration": {
          "id": "[variables('lbIPConfig0')]"
        },
        "frontendPortRangeEnd": "50999",
        "frontendPortRangeStart": "50000",
        "protocol": "tcp"
      }
    },
    {
      "name": "LoadBalancerBEAddressNatPool1",
      "properties": {
        "backendPort": "3389",
        "frontendIPConfiguration": {
          "id": "[variables('lbIPConfig0')]"
        },
        "frontendPortRangeEnd": "51999",
        "frontendPortRangeStart": "51000",
        "protocol": "tcp"
      }
    },
    {
      "name": "LoadBalancerBEAddressNatPool2",
      "properties": {
        "backendPort": "3389",
        "frontendIPConfiguration": {
          "id": "[variables('lbIPConfig0')]"
        },
        "frontendPortRangeEnd": "52999",
        "frontendPortRangeStart": "52000",
        "protocol": "tcp"
      }
    }
  ]
}
```

Reglas de salida para un equilibrador de carga de SKU Estándar

La IP pública del SKU Estándar presenta nuevas funciones y comportamientos diferentes para la conectividad de salida en comparación con el uso de SKU básicos. Si desea obtener conectividad de salida cuando trabaje con las SKU estándar, debe definirlas explícitamente con una dirección IP pública o un equilibrador de carga de SKU Estándar. Para más información, consulte Conexiones de salida y ¿Qué es Azure Load Balancer?

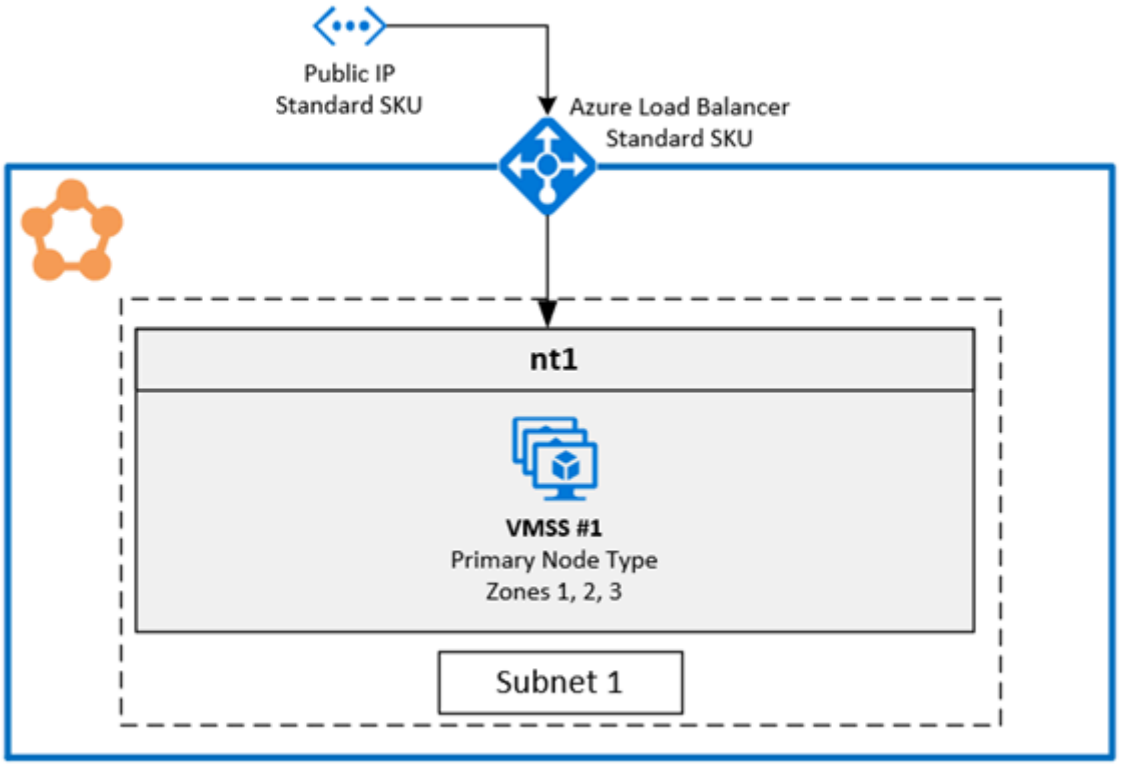
> **Nota:**
>
> La plantilla estándar hace referencia a un NSG que permite todo el tráfico saliente de forma predeterminada. El tráfico de entrada se limita a los puertos necesarios para las operaciones de administración de Service Fabric. Las reglas de NSG pueden modificarse para satisfacer sus requisitos.

> **Importante**
>
> Cada tipo de nodo de un clúster de Service Fabric que use un equilibrador de carga de SKU Estándar requiere una regla que permita el tráfico saliente en el puerto 443. Esto es necesario para completar la configuración del clúster. Cualquier implementación que no incluya esta regla generará un error.

1. Habilitación de varias zonas de disponibilidad en un único conjunto de escalado de máquinas virtuales

Esta solución permite a los usuarios abarcar tres instancias de Availability Zones en el mismo tipo de nodo. Se trata de la topología de implementación recomendada, ya que permite realizar la implementación entre zonas de disponibilidad mientras se mantiene un único conjunto de escalado de máquinas virtuales.

GitHub tiene una plantilla de ejemplo a su disposición.



Configuración de zonas en un conjunto de escalado de máquinas virtuales

Para habilitar zonas en un conjunto de escalado de máquinas virtuales, incluya los tres valores siguientes en el recurso del conjunto de escalado de máquinas virtuales:

- El primer valor es la propiedad `zones`, que especifica las zonas de disponibilidad que están presentes en el conjunto de escalado de máquinas virtuales.
- El segundo valor es la propiedad `singlePlacementGroup`, que se debe establecer en `true`. El conjunto de escalado que se extiende entre tres zona de disponibilidad puede admitir hasta 300 máquinas virtuales incluso con `singlePlacementGroup = true`.
- El tercer valor es `zoneBalance`, que garantiza el equilibrio de zona estricta. Este valor debe ser `true`. Esto garantiza que las distribuciones de máquinas virtuales entre zonas no sean desiguales, lo cual significa que cuando una de las zonas deja de funcionar, las otras dos zonas disponen de suficientes máquinas virtuales para que el clúster sigue ejecutándose.

  Es posible que un clúster con una distribución desigual de máquinas virtuales no sobreviva a un escenario de zona fuera de servicio, ya que esa zona podría acaparar la mayoría de las máquinas virtuales. La distribución desigual de máquinas virtuales entre zonas también provoca problemas relacionados con la prestación de servicios y el bloqueo de las actualizaciones de infraestructura. Más información sobre zoneBalancing.

No es necesario configurar las invalidaciones `FaultDomain` y `UpgradeDomain`.

JSON

```
{
  "apiVersion": "2018-10-01",
  "type": "Microsoft.Compute/virtualMachineScaleSets",
  "name": "[parameters('vmNodeType1Name')]",
  "location": "[parameters('computeLocation')]",
  "zones": [ "1", "2", "3" ],
  "properties": {
    "singlePlacementGroup": true,
    "zoneBalance": true
  }
}
```

Nota:

- Los clústeres de Service Fabric deben tener al menos un tipo principal de nodo. El nivel de durabilidad de los tipos de nodo principal debe ser Silver o superior.
- Una Azure Availability Zones de disponibilidad que extiende los conjuntos de escalado de máquinas virtuales debe configurarse con tres zona de disponibilidad como mínimo, independientemente del nivel de durabilidad.
- Una zona de disponibilidad que extienden conjuntos de escalado de máquinas virtuales con el nivel de durabilidad Silver o superior deben tener al menos 15 máquinas virtuales (5 por región).
- La zona de disponibilidad que abarca el conjunto de escalado de máquinas virtuales con durabilidad Bronze debe tener al menos 6 máquinas virtuales.

Habilitación de la compatibilidad con varias zonas en el tipo de nodo de Service Fabric

El tipo de nodo de Service Fabric debe estar habilitado para admitir varias zonas de disponibilidad.

- El primer valor es `multipleAvailabilityZones`, que se debe establecer en `true` para el tipo de nodo.
- El segundo valor es `sfZonalUpgradeMode` y es opcional. Esta propiedad no se puede modificar si ya existe un tipo de nodo con varias zonas de disponibilidad en el clúster. Esta propiedad determina la agrupación lógica de máquinas virtuales en los dominios de actualización (UD).

- Si este valor se establece en `Parallel`, las máquinas virtuales del tipo de nodo se agrupan en UD y omiten la información de zona en cinco UD. Esta configuración hace que los UD de todas las zonas se actualicen al mismo tiempo. Este modo de implementación es más rápido para las actualizaciones. No se recomienda su uso, ya que se trata de instrucciones SDP que indican que las actualizaciones deben aplicarse en una zona cada vez.
- Si el valor se omite o se establece en `Hierarchical`, las máquinas virtuales se agruparán para reflejar la distribución de zonas en un máximo de 15 UD. Cada una de las tres zonas tiene cinco UD. De este modo, se garantiza que las zonas se actualicen de una en una y que solo se pase a la zona siguiente después de completar cinco UD en la primera zona. Este proceso de actualización es más seguro para el clúster y la aplicación de usuario.

La propiedad solo define el comportamiento de actualización de la aplicación Service Fabric y las actualizaciones de código. Las actualizaciones del conjunto subyacente de escalado de máquinas virtuales seguirán produciéndose simultáneamente en todas las zonas de disponibilidad. Esta propiedad no afecta a la distribución de UD para los tipos de nodo que no tengan habilitadas varias zonas.

- El tercer valor es `vmssZonalUpgradeMode`, es opcional y se puede actualizar en cualquier momento. Esta propiedad define el esquema de actualización para que el conjunto de escalado de máquinas virtuales se pueda realizar en paralelo o secuencialmente en Availability Zones.
  - Si este valor se establece en `Parallel`: todas las actualizaciones del conjunto de escalado se suceden en paralelo en todas las zonas. Este modo de implementación es más rápido para las actualizaciones. No se recomienda su uso, ya que se trata de instrucciones SDP que indican que las actualizaciones deben aplicarse en una zona cada vez.
  - Si este valor se omite o se establece en `Hierarchical`: se garantiza que las zonas se actualicen de una en una y que solo se pase a la zona siguiente después de completar cinco UD en la primera zona. Este proceso de actualización es más seguro para el clúster y la aplicación de usuario.

> **ⓘ Importante**
>
> La versión de API del recurso de clúster de Service Fabric debe ser la 2020-12-01 (versión preliminar) o superior.
>
> La versión del código del clúster debe ser 8.1.321 o posterior.

JSON

```
{
  "apiVersion": "2020-12-01-preview",
  "type": "Microsoft.ServiceFabric/clusters",
  "name": "[parameters('clusterName')]",
  "location": "[parameters('clusterLocation')]",
  "dependsOn": [
    "[concat('Microsoft.Storage/storageAccounts/',
parameters('supportLogStorageAccountName'))]"
  ],
  "properties": {
    "reliabilityLevel": "Platinum",
    "sfZonalUpgradeMode": "Hierarchical",
    "vmssZonalUpgradeMode": "Parallel",
    "nodeTypes": [
      {
        "name": "[parameters('vmNodeType0Name')]",
        "multipleAvailabilityZones": true
      }
    ]
  }
}
```

Nota:

- Los recursos de IP pública y equilibrar la carga deben usar la SKU Estándar, tal como se describió anteriormente en el artículo.
- La propiedad `multipleAvailabilityZones` del tipo de nodo solo se puede definir cuando se haya creado el tipo de nodo y no se podrá modificar más adelante. Los tipos de nodo existentes no se pueden configurar con esta propiedad.
- Cuando `sfZonalUpgradeMode` se omite o se establece en `Hierarchical`, las implementaciones de clúster y aplicación serán más lentas porque habrá más dominios de actualización en el clúster. Es importante ajustar correctamente los tiempos de inactividad de la directiva de actualización para tener en cuenta el tiempo de actualización necesario para 15 dominios de actualización. La directiva de actualización para la aplicación y el clúster debe actualizarse para garantizar que la implementación no supere el límite de tiempo de implementación de 12 horas de Azure Resource Service. Esto significa que la implementación no debe tardar más de 12 horas para 15 UD (es decir, no debe tardar más de 40 minutos con cada UD).
- Establezca el nivel de confiabilidad del clúster en `Platinum` para garantizar que el clúster sobreviva al escenario con una zona fuera de servicio.
- No se admite la actualización de DurabilityLevel para un tipo de nodo con multipleAvailabilityZones. En su lugar, cree un tipo de nodo con mayor durabilidad.

- SF solo admite tres AvailabilityZones. Por el momento, no se admite ningún número mayor.

Sugerencia

Se recomienda establecer `sfZonalUpgradeMode` en `Hierarchical`, o bien omitirlo. La implementación seguirá la distribución de zonas de las máquinas virtuales y afecta a un número más reducido de réplicas o instancias, por lo estas son más seguras. Use `sfZonalUpgradeMode` establecido en `Parallel` si la velocidad de implementación es una prioridad o solo se ejecuta una carga de trabajo sin estado en el tipo de nodo con varias zonas de disponibilidad. Esto hará que el recorrido del dominio de actualización tenga lugar al mismo tiempo en todas las zonas de disponibilidad.

Migración al tipo de nodo con varias zonas de disponibilidad

En todos los escenarios de migración, es necesario agregar un nuevo tipo de nodo que admita varias zonas de disponibilidad. No se puede migrar un tipo de nodo existente para admitir varias zonas. El artículo Escalado vertical de un del tipo de nodo principal de un clúster de Service Fabric incluye pasos detallados para agregar un nuevo tipo de nodo y los demás recursos necesarios para el nuevo tipo de nodo, como los recursos de IP y equilibrador de carga. En el artículo se describe también cómo retirar el tipo de nodo existente después de agregar al clúster un nuevo tipo de nodo con varias zonas de disponibilidad.

- Migración desde un tipo de nodo que usa recursos de IP básicos: este proceso ya se describe en una subsección posterior para la solución con un tipo de nodo por zona de disponibilidad.

  Para el nuevo tipo de nodo, la única diferencia es que solo hay un conjunto de escalado de máquinas virtuales y un tipo de nodo para todas las zonas de disponibilidad, en lugar de uno por zona de disponibilidad.

- Migración desde un tipo de nodo que usa los recursos de IP de SKU Estándar con un NSG: siga el procedimiento descrito anteriormente. Sin embargo, no es necesario agregar nuevos recursos de IP y NSG. Se pueden reutilizar los mismos recursos en el nuevo tipo de nodo.

2. Implementación de zonas mediante el anclaje de un conjunto de escalado de máquinas virtuales a

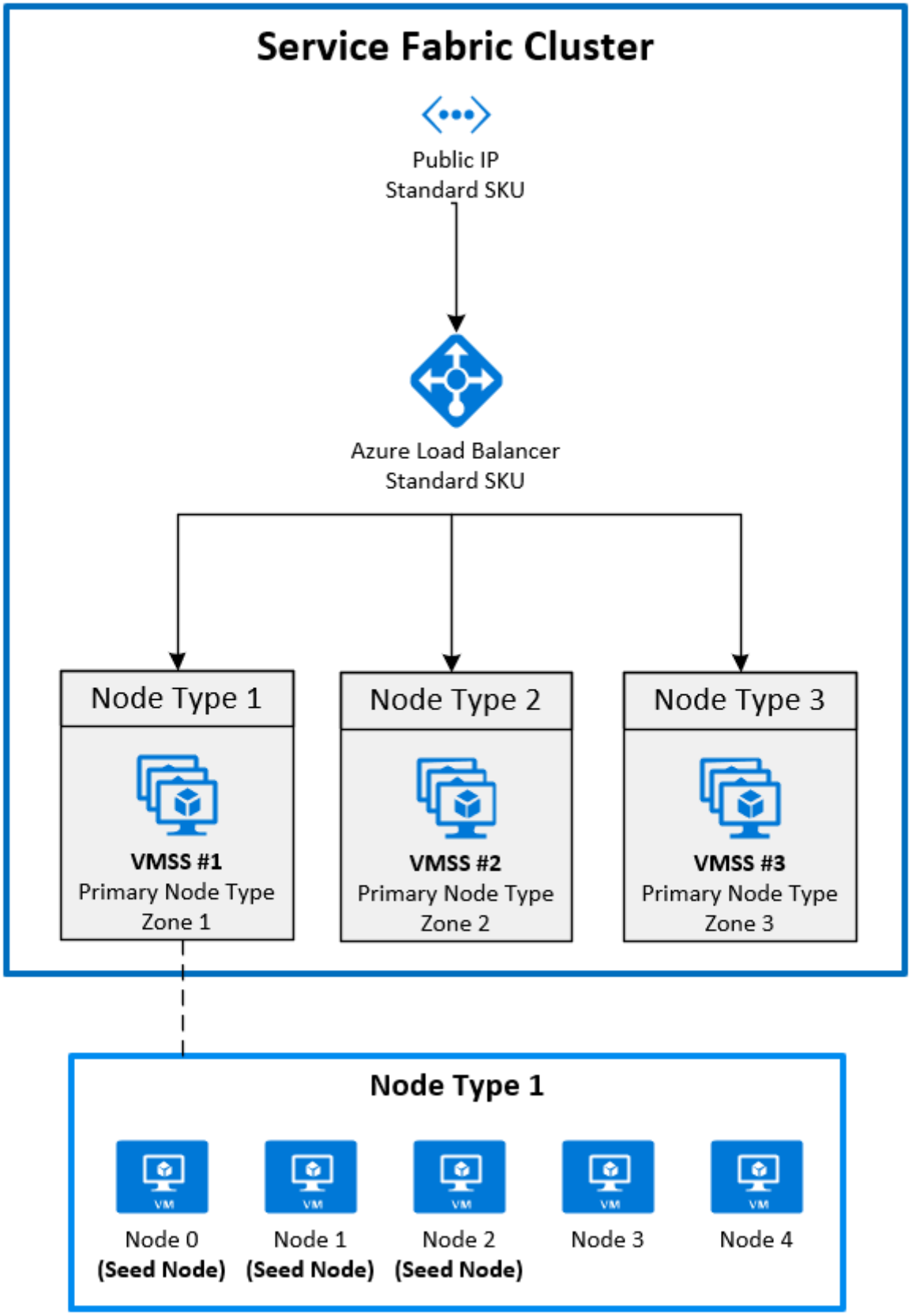
cada zona

Esta es la configuración disponible con carácter general en este momento. Para extender un clúster de Service Fabric entre Availability Zones, debe crear un tipo de nodo principal en cada zona de disponibilidad admitida por la región. De este modo, se distribuyen nodos raíz de manera uniforme entre todos los tipos de nodos principales.

La topología recomendada para el tipo de nodo principal requiere lo siguiente:

- Tres tipos de nodo deben marcarse como principales.
  - Cada tipo de nodo debe asignarse a su propio conjunto de escalado de máquinas virtuales ubicado en distintas zonas.
  - Cada conjunto de escalado de máquinas virtuales debe tener al menos cinco nodos (durabilidad Silver).

El diagrama siguiente muestra la arquitectura de la zona de disponibilidad de Azure Service Fabric.



Habilitación de zonas en un conjunto de escalado de máquinas virtuales

Para habilitar una zona en un conjuntos de escalado de máquinas virtuales, incluya estos tres valores en el recurso del conjunto de escalado de máquinas virtuales:

- El primer valor es la propiedad `zones`, que especifica en qué zona de disponibilidad se implementará el conjunto de escalado de máquinas virtuales.
- El segundo valor es la propiedad `singlePlacementGroup`, que se debe establecer en `true`.
- El tercer valor es la propiedad `faultDomainOverride` en la extensión de conjunto de escalado de máquinas virtuales de Service Fabric. Esta propiedad debe incluir solo la zona en la que se incluirá este conjunto de escalado de máquinas virtuales. Ejemplo: `"faultDomainOverride": "az1"`. Todos los recursos del conjunto de escalado de máquinas virtuales se deben incluir en la misma región porque los clústeres de Azure Service Fabric no ofrecen compatibilidad entre regiones.

JSON

```
{
  "apiVersion": "2018-10-01",
  "type": "Microsoft.Compute/virtualMachineScaleSets",
  "name": "[parameters('vmNodeType1Name')]",
  "location": "[parameters('computeLocation')]",
  "zones": [
    "1"
  ],
  "properties": {
    "singlePlacementGroup": true
  },
  "virtualMachineProfile": {
    "extensionProfile": {
      "extensions": [
        {
          "name": "[concat(parameters('vmNodeType1Name'),'_ServiceFabricNode')]",
          "properties": {
            "type": "ServiceFabricNode",
            "autoUpgradeMinorVersion": false,
            "publisher": "Microsoft.Azure.ServiceFabric",
            "settings": {
              "clusterEndpoint": "
[reference(parameters('clusterName')).clusterEndpoint]",
              "nodeTypeRef": "[parameters('vmNodeType1Name')]",
              "dataPath": "D:\\\\SvcFab",
              "durabilityLevel": "Silver",
              "certificate": {
                "thumbprint": "[parameters('certificateThumbprint')]",
                "x509StoreName": "[parameters('certificateStoreValue')]"
              },
              "systemLogUploadSettings": {
                "Enabled": true
              },
              "faultDomainOverride": "az1"
            },
```

```
            "typeHandlerVersion": "1.0"
          }
        }
      ]
    }
  }
}
```

Habilitación de varios tipos de nodos principales en el recurso de clúster de Service Fabric

Para establecer uno o más tipos de nodo como principales en un recurso de clúster, establezca la propiedad `isPrimary` en `true`. Al implementar un clúster de Service Fabric en Availability Zones, debe tener tres tipos de nodo en distintas zonas.

JSON

```
{
  "reliabilityLevel": "Platinum",
  "nodeTypes": [
    {
      "name": "[parameters('vmNodeType0Name')]",
      "applicationPorts": {
        "endPort": "[parameters('nt0applicationEndPort')]",
        "startPort": "[parameters('nt0applicationStartPort')]"
      },
      "clientConnectionEndpointPort": "[parameters('nt0fabricTcpGatewayPort')]",
      "durabilityLevel": "Silver",
      "ephemeralPorts": {
        "endPort": "[parameters('nt0ephemeralEndPort')]",
        "startPort": "[parameters('nt0ephemeralStartPort')]"
      },
      "httpGatewayEndpointPort": "[parameters('nt0fabricHttpGatewayPort')]",
      "isPrimary": true,
      "vmInstanceCount": "[parameters('nt0InstanceCount')]"
    },
    {
      "name": "[parameters('vmNodeType1Name')]",
      "applicationPorts": {
        "endPort": "[parameters('nt1applicationEndPort')]",
        "startPort": "[parameters('nt1applicationStartPort')]"
      },
      "clientConnectionEndpointPort": "[parameters('nt1fabricTcpGatewayPort')]",
      "durabilityLevel": "Silver",
      "ephemeralPorts": {
        "endPort": "[parameters('nt1ephemeralEndPort')]",
        "startPort": "[parameters('nt1ephemeralStartPort')]"
      },
      "httpGatewayEndpointPort": "[parameters('nt1fabricHttpGatewayPort')]",
      "isPrimary": true,
      "vmInstanceCount": "[parameters('nt1InstanceCount')]"
```

```
    },
    {
      "name": "[parameters('vmNodeType2Name')]",
      "applicationPorts": {
        "endPort": "[parameters('nt2applicationEndPort')]",
        "startPort": "[parameters('nt2applicationStartPort')]"
      },
      "clientConnectionEndpointPort": "[parameters('nt2fabricTcpGatewayPort')]",
      "durabilityLevel": "Silver",
      "ephemeralPorts": {
        "endPort": "[parameters('nt2ephemeralEndPort')]",
        "startPort": "[parameters('nt2ephemeralStartPort')]"
      },
      "httpGatewayEndpointPort": "[parameters('nt2fabricHttpGatewayPort')]",
      "isPrimary": true,
      "vmInstanceCount": "[parameters('nt2InstanceCount')]"
    }
  ]
}
```

Migración a Availability Zones desde un clúster mediante una dirección IP de SKU Básico

Para migrar un clúster que usa una dirección IP con un SKU Básico, primero debe crear un recurso de IP completamente nuevo mediante el SKU Estándar. No es posible actualizar estos recursos.

Haga referencia a la IP en los nuevos tipos de nodo entre zonas de disponibilidad que desea usar. En el ejemplo anterior, se agregaron tres nuevos recursos de conjunto de escalado de máquinas virtuales en las zonas 1, 2 y 3. Estos conjuntos de escalado de máquinas virtuales hacen referencia a la IP recién creada y se identifican como tipos de nodo principal en el recurso de clúster de Service Fabric.

1. Para empezar, agregue los nuevos recursos a la plantilla existente de Azure Resource Manager. Estos recursos incluyen:

    - Un recurso de IP pública que usa la SKU Estándar.
    - Un recurso de equilibrador de carga que usa la SKU Estándar.
    - Un NSG referenciado por la subred en la que se implementan los conjuntos de escalado de máquinas virtuales.
    - Tres tipos de nodo deben marcarse como principales.
        - Cada tipo de nodo debe asignarse a su propio conjunto de escalado de máquinas virtuales ubicado en distintas zonas.
        - Cada conjunto de escalado de máquinas virtuales debe tener al menos cinco nodos (durabilidad Silver).

Puede encontrar un ejemplo de estos recursos en la plantilla de ejemplo.

PowerShell

```
New-AzureRmResourceGroupDeployment `
    -ResourceGroupName $ResourceGroupName `
    -TemplateFile $Template `
    -TemplateParameterFile $Parameters
```

2. Una vez que ha terminado la implementación de los recursos, puede deshabilitar los nodos en el tipo de nodo principal del clúster original. Cuando los nodos estén deshabilitados, los servicios del sistema se migrarán al nuevo tipo de nodo principal que implementó anteriormente.

PowerShell

```
Connect-ServiceFabricCluster -ConnectionEndpoint $ClusterName `
    -KeepAliveIntervalInSec 10 `
    -X509Credential `
    -ServerCertThumbprint $thumb  `
    -FindType FindByThumbprint `
    -FindValue $thumb `
    -StoreLocation CurrentUser `
    -StoreName My

Write-Host "Connected to cluster"

$nodeNames = @("_nt0_0", "_nt0_1", "_nt0_2", "_nt0_3", "_nt0_4")

Write-Host "Disabling nodes..."
foreach($name in $nodeNames) {
    Disable-ServiceFabricNode -NodeName $name -Intent RemoveNode -Force
}
```

3. Una vez que todos los nodos están deshabilitados, los servicios del sistema se ejecutarán en el tipo de nodo principal, que se extiende entre zonas. A continuación, puede quitar los nodos deshabilitados del clúster. Después de quitar los nodos, puede quitar los recursos de IP original, de equilibrador de carga y de conjunto de escalado de máquinas virtuales.

PowerShell

```
foreach($name in $nodeNames){
    # Remove the node from the cluster
    Remove-ServiceFabricNodeState -NodeName $name -TimeoutSec 300 -Force
    Write-Host "Removed node state for node $name"
}

$scaleSetName="nt0"
```

```
Remove-AzureRmVmss -ResourceGroupName $groupname -VMScaleSetName $scaleSetName
-Force

$lbname="LB-cluster-nt0"
$oldPublicIpName="LBIP-cluster-0"
$newPublicIpName="LBIP-cluster-1"

Remove-AzureRmLoadBalancer -Name $lbname -ResourceGroupName $groupname -Force
Remove-AzureRmPublicIpAddress -Name $oldPublicIpName -ResourceGroupName
$groupname -Force
```

4. A continuación, quite las referencias a estos recursos de la plantilla de Resource Manager que se implementó.

5. Para finalizar, actualice el nombre DNS y la dirección IP pública.

PowerShell

```
$oldprimaryPublicIP = Get-AzureRmPublicIpAddress -Name $oldPublicIpName  -
ResourceGroupName $groupname
$primaryDNSName = $oldprimaryPublicIP.DnsSettings.DomainNameLabel
$primaryDNSFqdn = $oldprimaryPublicIP.DnsSettings.Fqdn

Remove-AzureRmLoadBalancer -Name $lbname -ResourceGroupName $groupname -Force
Remove-AzureRmPublicIpAddress -Name $oldPublicIpName -ResourceGroupName $groupname
-Force

$PublicIP = Get-AzureRmPublicIpAddress -Name $newPublicIpName  -ResourceGroupName
$groupname
$PublicIP.DnsSettings.DomainNameLabel = $primaryDNSName
$PublicIP.DnsSettings.Fqdn = $primaryDNSFqdn
Set-AzureRmPublicIpAddress -PublicIpAddress $PublicIP
```

Last updated on 22/03/2026

Recuperación ante desastres en Service Fabric de Azure

Una parte fundamental de la entrega de alta disponibilidad es garantizar que los servicios sobreviven a cualquier tipo de error. Esto es especialmente importante para los errores imprevistos y fuera de control.

En este artículo se describen algunos modos de error comunes que podrían ser desastres si no se hubieran modelado y administrado correctamente. También se explican las mitigaciones y las acciones que deben realizarse en caso de desastre. El objetivo es limitar o eliminar el riesgo de pérdida de datos o de tiempo de inactividad cuando se produzcan errores, tanto planeados como imprevistos.

Impedimento de desastres

El objetivo principal de Azure Service Fabric es ayudarle a modelar tanto el entorno como los servicios de tal manera que los tipos de error comunes no se conviertan en desastres.

En general, hay dos tipos de escenarios de desastre o error:

- Errores de hardware y software
- Errores operativos

Errores de hardware y software

Los errores de hardware y software son imprevisibles. La manera más fácil de sobrevivir a errores es ejecutar más copias del servicio más allá de los límites del hardware o software defectuoso.

Por ejemplo, si el servicio se ejecuta solamente en una máquina concreta, el error que se produce en esa máquina es un desastre para el servicio. La manera más sencilla de evitar este desastre es asegurarse de que el servicio se ejecuta en varias máquinas. Puede ser necesario realizar pruebas para garantizar que el error de una máquina no interrumpe el servicio en ejecución. El planeamiento de capacidad garantiza que se puede crear una instancia de reemplazo en otra parte y que la reducción de capacidad no sobrecarga el resto de servicios.

Este patrón funciona independientemente de dónde intente evitar el error. Por ejemplo, si le preocupa el error de una red de área de almacenamiento, debe ejecutarlo en varias redes de área de almacenamiento. Si le preocupa perder un conjunto de servidores, debe ejecutarlo en varios conjuntos. Si le preocupa la pérdida de centros de datos, el servicio debe ejecutarse en

varias regiones de Azure, varias instancias de Azure Availability Zones o entre sus propios centros de datos.

Cuando un servicio se distribuye entre varias instancias físicas (máquinas, bastidores, centros de datos o regiones), aún existe la posibilidad de experimentar algunos tipos de errores simultáneos. No obstante, los errores únicos e incluso los múltiples de un tipo determinado (por ejemplo, una sola máquina virtual o un error de vínculo de red) se controlan automáticamente, por lo que ya no son un "desastre".

Service Fabric proporciona mecanismos para expandir el clúster y se ocupa de recuperar nodos y servicios defectuosos. También permite ejecutar varias instancias de los servicios para evitar que los errores imprevistos se conviertan en verdaderos desastres.

Pueden existir razones por las que ejecutar una gran implementación para abarcar los errores no sea posible. Por ejemplo, es posible que sea necesario dedicar más recursos de hardware de los que esté dispuesto a pagar por el riesgo de error. Cuando se trabaja con aplicaciones distribuidas, los costos de los saltos de comunicación o de replicación de estados adicionales a distancia geográfica podrían provocar una latencia inaceptable. Esta línea se traza en función de cada aplicación.

En concreto, para errores de software, el error puede encontrarse en el servicio que intenta escalar. En este caso, el hecho de tener más copias no evita el desastre, ya que la condición de error se encuentra en todas las instancias.

Errores operativos

Aunque se distribuya el servicio por todo el mundo con redundancias, pueden seguir produciéndose eventos de desastre. Por ejemplo, alguien podría volver a configurar por error el nombre de DNS del servicio o directamente eliminarlo.

A modo de ejemplo, supongamos que tuviera un servicio de Service Fabric con estado y que alguien lo elimina por error. A menos que haya otra mitigación, ese servicio y su estado habrán desaparecido. Estos tipos de desastres operativos ("vaya") requieren más mitigaciones y pasos para la recuperación que los errores imprevistos normales.

Lo mejor para evitar estos tipos de errores operativos es:

- Restringir el acceso operativo al entorno.
- Inspeccionar minuciosamente las operaciones peligrosas.
- Imponer la automatización, evitar los cambios manuales o fuera de banda y validar los cambios concretos en el entorno antes de establecerlos.
- Asegurarse de que las operaciones destructivas "se suavizan". Las operaciones de software suavizadas no surten efecto inmediatamente o se pueden deshacer en un plazo

determinado.

Service Fabric proporciona mecanismos para evitar errores operativos, como un control de acceso basado en rol para las operaciones del clúster. Sin embargo, la mayoría de estos errores operativos requiere esfuerzos organizativos y otros sistemas. Service Fabric proporciona mecanismos para sobrevivir a errores operativos, en especial, copia de seguridad y restauración para los servicios con estado.

Administración de errores

El objetivo de Service Fabric es la administración automática de los errores. Sin embargo, para abordar algunos tipos de errores, los servicios deben tener código adicional. Otros tipos de errores *no* deben solucionarse automáticamente por motivos de seguridad y continuidad empresarial.

Abordaje de errores únicos

Una máquina puede fallar por todo tipo de razones. En ocasiones se debe al hardware, como los errores en las fuentes de alimentación o el hardware de las redes. Otros errores son de software. Se trata de errores en el sistema operativo y el propio servicio. Service Fabric detecta automáticamente estos tipos de errores, incluidos los casos en los que la máquina se aísla de otras por problemas de red.

Independientemente del tipo de servicio, ejecutar una sola instancia se traduce en tiempo de inactividad de ese servicio si se produce un error en la copia única del código por cualquier motivo.

Para abordar cualquier error único, lo más sencillo es garantizar que los servicios se ejecutan en más de un nodo de forma predeterminada. En el caso de los servicios sin estado, asegúrese de que `InstanceCount` sea superior a 1. Para los servicios con estado, la recomendación es que `TargetReplicaSetSize` y `MinReplicaSetSize` estén establecidos en 3 como mínimo. La ejecución de más copias del código de servicio garantiza que este puede abordar cualquier error único automáticamente.

Abordaje de errores coordinados

En un clúster se pueden producir errores coordinados por errores y cambios de infraestructura planeados o imprevistos, o por cambios de software planeados. Service Fabric modela las zonas de infraestructura que experimentan errores coordinados como *dominios de error*. Las áreas que experimentarán cambios coordinados de software se modelan como *dominios de actualización*. Para más información acerca de los dominios de error, los dominios de

actualización y la topología de clúster, consulte Descripción de un clúster de Service Fabric con Cluster Resource Manager.

De forma predeterminada, Service Fabric considera los dominios de error y de actualización al planear dónde deben ejecutarse los servicios. De forma predeterminada, Service Fabric intenta asegurarse de que los servicios se ejecutan en varios dominios de error y de actualización, de manera que si se realizan cambios planeados o imprevistos, los servicios permanecen disponibles.

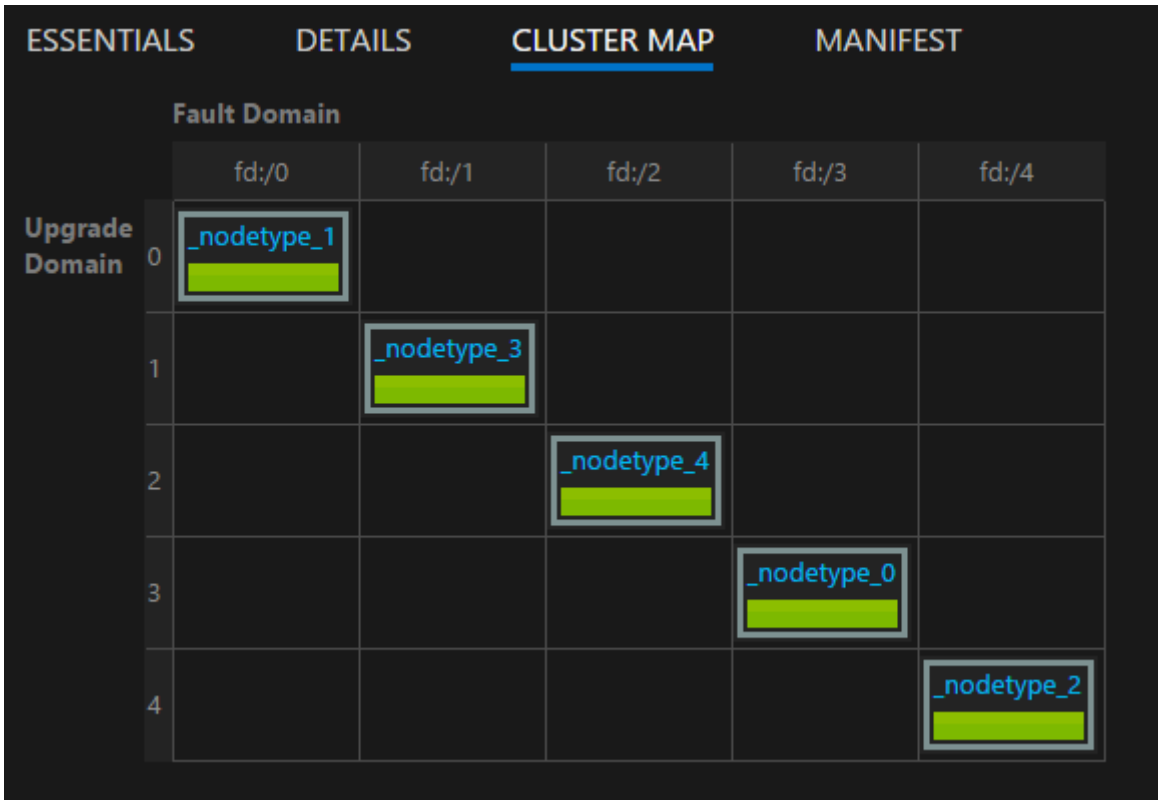
Por ejemplo, supongamos que un error en una fuente de alimentación provoca un error simultáneo en todas las máquinas de un bastidor. Con varias copias del servicio en ejecución, la pérdida de varias máquinas del dominio de error se convierte simplemente en otro ejemplo de error único para un servicio. Por este motivo la administración de dominios de error y actualización es crucial para garantizar la alta disponibilidad de los servicios.

Al ejecutar Service Fabric en Azure, los dominios de error y actualización se administran automáticamente. En otros entornos puede que esto no sea así. Si va a compilar sus propios clústeres locales, asegúrese de asignar y planear el diseño de los dominios de error correctamente.

Los dominios de actualización son útiles para modelar las áreas donde se va a actualizar software de manera simultánea. Por este motivo, los dominios de actualización a menudo también definen los límites de desconexión del software durante las actualizaciones planeadas. Las actualizaciones de Service Fabric y los servicios siguen el mismo modelo. Para más información acerca de las actualizaciones graduales, los dominios de actualización y el modelo de estado de Service Fabric que ayuda a evitar que los cambios por error afecten al clúster y al servicio, consulte:

- Actualización de aplicaciones
- Tutorial sobre la actualización de aplicaciones
- Modelo de estado de Service Fabric

Puede visualizar el diseño del clúster mediante el mapa de clúster que se proporciona en Service Fabric Explorer:



> ⓘ **Nota:**
>
> El modelado de áreas de error, las actualizaciones graduales, la ejecución de muchas instancias del código de servicio y el estado, las reglas de colocación para asegurarse de que los servicios se ejecutan entre dominios de error y de actualización, y la supervisión de estado integrada son solo *algunas* de las características que Service Fabric proporciona para evitar que los errores y problemas operativos normales se conviertan en desastres.

Abordaje de errores simultáneos en el hardware y en el software

Hasta ahora se han tratado errores únicos. Como puede ver, son fáciles de solucionar para los servicios con y sin estado con más copias del código (y del estado) en varios dominios de error y de actualización.

También se pueden producir varios errores simultáneos y aleatorios. Estos tienen mayor probabilidad de provocar un tiempo de inactividad o un verdadero desastre.

Servicios sin estado

El recuento de instancias de un servicio sin estado indica el número deseado de instancias que deben estar en ejecución. Si se produce un error en alguna de las instancias (o en todas), Service Fabric responderá creando instancias de reemplazo automáticamente en otros nodos.

Service Fabric seguirá creando reemplazos hasta que el servicio vuelva al recuento de instancias deseado.

Por ejemplo, supongamos que el servicio sin estado tiene un valor `InstanceCount` de -1. Este valor significa que se debe ejecutar una instancia en cada nodo del clúster. Si se produjera un error en alguna de esas instancias, Service Fabric detectará que el servicio no está en su estado deseado e intentará crear las instancias en los nodos donde faltan.

Servicios con estado

Existen dos tipos de servicios con estado:

- Con estado con estado persistente.
- Con estado con estado no persistente (el estado se almacena en la memoria).

La recuperación de errores de servicios con estado depende del tipo de servicio con estado, así como del número de réplicas que tenía el servicio y cuántas de ellas han generado un error.

En un servicio con estado, los datos entrantes se replican entre las réplicas (principal y secundarias activas). Si la mayoría de las réplicas reciben los datos, se considera que los datos están confirmados en el *cuórum* (para cinco réplicas, tres serán un cuórum). Esto significa que, en cualquier momento, habrá al menos un cuórum de réplicas con los datos más recientes. Si se produce un error en las réplicas (por ejemplo, en dos de cinco), podemos usar el valor de cuórum para calcular si se puede recuperar (dado que las tres réplicas restantes de las cinco siguen estando activas, se garantiza que al menos una réplica tendrá datos completos).

Cuando se produce un error en un cuórum de réplicas, se declara que la partición está en estado de *pérdida de cuórum*. Supongamos que una partición tiene cinco réplicas, lo que significa que se garantiza que al menos tres tienen datos completos. Si se produce un error en un cuórum (tres de cinco) de réplicas, Service Fabric no puede determinar si las réplicas restantes (dos de cinco) tienen suficientes datos para restaurar la partición. En los casos en los que Service Fabric detecta pérdida de cuórum, su comportamiento predeterminado es evitar escrituras adicionales en la partición, declarar pérdida de cuórum y esperar a que se restaure un cuórum de réplicas.

La determinación de si se ha producido un desastre para un servicio con estado y su administración se producen en tres etapas:

1. Determinación de si se ha producido pérdida de cuórum o no.

   La pérdida de cuórum se declara cuando la mayoría de las réplicas de un servicio con estado están inactivas al mismo tiempo.

2. Determinación de si la pérdida de cuórum es permanente o no.

La mayoría de las veces, los errores son temporales. Los procesos, los nodos y las máquinas virtuales se reinician y las particiones de la red se reparan. Sin embargo, a veces los errores son permanentes. El hecho de que los errores sean permanentes o no depende de si el servicio con estado conserva su estado o si lo mantiene solo en la memoria:

- Para servicios sin estado persistente, un error de un cuórum o varios en las réplicas deriva *inmediatamente* en una pérdida de cuórum permanente. Cuando Service Fabric detecta una pérdida de cuórum en un servicio con estado no persistente, declara una (posible) pérdida de datos y pasa inmediatamente al paso 3. Pasar a pérdida de datos tiene sentido, ya que Service Fabric sabe que no merece la pena esperar el retorno de las réplicas. Aunque se recuperaran, se perderán los datos debido a la naturaleza no persistente del servicio.
- En los servicios con estado persistente, un error de uno más cuórums de réplicas provoca que Service Fabric espere el retorno de las réplicas y la restauración del cuórum. Esto supone la interrupción del servicio de *escritura* en las particiones afectadas (o "conjunto de réplicas") del servicio. Sin embargo, la lectura seguirá siendo posible, con garantía de coherencia reducida. La cantidad de tiempo predeterminada que Service Fabric espera a que el cuórum se restaure es *infinita*, ya que continuar supone un (posible) evento de pérdida de datos y acarrea otros riesgos. Esto significa que Service Fabric no continuará con el paso siguiente a menos que un administrador actúe para declarar pérdida de datos.

3. Determinación de si se ha producido pérdida de datos y restauración desde copias de seguridad.

   Si se ha declarado pérdida de cuórum (ya sea de forma automática o mediante una acción administrativa), Service Fabric y los servicios pasan a determinar si realmente se han perdido los datos. En este punto, Service Fabric también sabe que las demás réplicas no se van a recuperar. Esa fue la decisión que se tomó al detener la espera a que la pérdida de cuórum se resolviera por sí sola. La mejor forma de proceder con el servicio es la inmovilización mientras se espera la intervención administrativa específica.

   Siempre que Service Fabric llama al método `OnDataLossAsync` se debe a una *sospecha* de pérdida de datos. Service Fabric garantiza que esta llamada se entrega a la *mejor* de las réplicas restantes. Esta es la réplica que haya realizado más progreso.

   El motivo por el que siempre decimos *sospecha* de pérdida de datos es que puede que el resto de réplicas tengan el mismo estado que la principal al perderse el cuórum. Sin embargo, sin ese estado con el que compararlo, no hay forma de que Service Fabric o los operadores lo sepan con seguridad.

   Entonces, ¿qué hace una implementación típica del método `OnDataLossAsync`?

a. La implementación registra que se ha desencadenado `OnDataLossAsync` y activa las alertas administrativas necesarias.

b. Normalmente, la implementación se detiene y espera a la toma de otras decisiones y a la realización de acciones manuales. Esto es así porque aunque haya copias de seguridad disponibles, puede ser necesario prepararlas.

   Por ejemplo, si dos servicios diferentes coordinan información, puede que sea necesario modificar las copias de seguridad para garantizar que, cuando se haya realizado la restauración, la información importante para los dos servicios sea coherente.

c. A menudo también hay otros datos de telemetría o salidas del servicio. Estos metadatos pueden estar contenidos en otros servicios o registros. Esta información puede usarse según sea necesario para determinar si hubo llamadas recibidas y procesadas en la réplica principal que no estaban presentes en la copia de seguridad o en esta réplica determinada. Puede que sea necesario reproducirlas o agregarlas a la copia de seguridad para poder realizar la restauración.

d. La implementación compara el estado de las réplicas restantes con el de aquellas contenidas en las copias de seguridad disponibles. Si se usan las colecciones de confianza de Service Fabric, hay herramientas y procesos disponibles para ello. El objetivo consiste en ver si el estado de la réplica es suficiente y si falta algo en la copia de seguridad.

e. Una vez realizada la comparación, y tras realizarse la restauración (si es necesaria), el código del servicio debe devolver **true** si se han realizado cambios de estado. Si la réplica determinó que era la mejor copia disponible del estado y no realizó ningún cambio, el código devolverá **false**.

   El valor **true** indica que *cualquier* réplica restante ahora puede ser incoherente con esta. Se descartarán y se volverán a crear a partir de esta réplica. El valor **false** indica que no se realizó ningún cambio de estado, por lo que las demás réplicas pueden conservar el contenido.

Es muy importante que los autores de los servicios practiquen con escenarios de posible pérdida de datos y error antes de implementarlos en producción. Como protección ante la posible pérdida de datos, es importante realizar periódicamente una copia de seguridad del estado de los servicios con estado en almacenes con redundancia geográfica.

También debe asegurarse de que puede restaurar el estado. Dado que se realizan copias de seguridad de muchos servicios diferentes en momentos distintos, deberá asegurarse de que una vez restaurados, los servicios se ven de manera coherente entre ellos.

Por ejemplo, considere una situación donde un servicio genera un número y lo almacena y, a continuación, lo envía a otro servicio que también lo almacena. Después de una restauración, puede que descubra que el segundo servicio tiene el número pero el primero no, porque su copia de seguridad no incluye esa operación.

Si descubre que las réplicas restantes son insuficientes para continuar en un escenario de pérdida de datos y no se puede reconstruir el estado del servicio a partir de los datos de telemetría o por agotamiento, la frecuencia de las copias de seguridad determina el mejor objetivo de punto de recuperación (RPO) posible. Service Fabric facilita muchas herramientas para probar diversos escenarios de error, como el cuórum permanente y la pérdida de datos, que requiere la restauración a partir de una copia de seguridad. Estos escenarios forman parte de las herramientas de capacidad de prueba de Service Fabric, administradas por el servicio de análisis de errores. Consulte Introducción al servicio de análisis de errores para más información sobre estas herramientas y patrones.

> ⓘ **Nota:**
>
> Los servicios del sistema también pueden sufrir pérdida de cuórum. El impacto es específico en el servicio en cuestión. Por ejemplo, la pérdida de cuórum en el servicio de nombres afecta a la resolución de nombres, mientras que la pérdida de cuórum en Administrador de conmutación por error bloquea la creación de nuevos servicios y las conmutaciones por error.
>
> Los servicios de sistema de Service Fabric siguen el mismo patrón que los servicios de administración de estado, si bien no se recomienda intentar sacarlos de la pérdida de cuórum para incluirlos en posible pérdida de datos. En su lugar, se recomienda **buscar asistencia** para encontrar una solución específica para su situación. Suele ser preferible esperar a que las réplicas inactivas vuelvan a estar disponibles.

Solución de problemas de pérdida de cuórum

Es posible que las réplicas estén inactivas intermitentemente debido a un error transitorio. Espere un tiempo mientras Service Fabric intenta activarlas. Si las réplicas llevan inactivas más tiempo del esperado, siga las acciones de solución de problemas que se indican a continuación:

- Es posible que las réplicas se bloqueen. Compruebe los informes de estado del nivel de réplica y los registros de la aplicación. Recopile volcados de memoria y tome las medidas necesarias para realizar la recuperación.
- Es posible que el proceso de réplica haya dejado de responder. Inspeccione los registros de la aplicación para comprobarlo. Recopile volcados del proceso y, a continuación,

detenga el proceso que no responde. Service Fabric creará un proceso de reemplazo e intentará devolver la réplica.
- Es posible que los nodos que hospedan réplicas estén inactivos. Reinicie la máquina virtual subyacente para poner en marcha los nodos.

En ocasiones, puede no ser posible recuperar réplicas. Por ejemplo, si el error se ha producido en las unidades o las máquinas no responden físicamente. En estos casos, es necesario indicar a Service Fabric que no espere la recuperación de la réplica.

No use estos métodos si la posible pérdida de datos es inaceptable para poner el servicio en línea. En ese caso, se deben realizar todos los esfuerzos para la recuperación de máquinas físicas.

Las siguientes acciones podrían provocar pérdida de datos. Compruébelo antes de realizarlas.

> ⓘ **Nota:**
>
> *Nunca* es seguro utilizar estos métodos de manera que no sea directamente con particiones específicas.

- Use la API de `Repair-ServiceFabricPartition -PartitionId` o `System.Fabric.FabricClient.ClusterManagementClient.RecoverPartitionAsync(Guid partitionId)`. Esta API permite especificar el identificador de la partición que va a cambiar de pérdida de cuórum a posible pérdida de datos.
- Si el clúster detecta errores frecuentes que hacen que los servicios entren en estado de pérdida de cuórum y la *pérdida de datos potencial es aceptable*, especificar un valor de QuorumLossWaitDuration adecuado puede ayudarle con la recuperación automática del servicio. Service Fabric esperará el valor de `QuorumLossWaitDuration` proporcionado (el valor predeterminado es infinito) antes de realizar la recuperación. *No* se recomienda este método porque puede provocar pérdidas de datos inesperadas.

Disponibilidad del clúster de Service Fabric

Por lo general, el propio clúster de Service Fabric es un entorno altamente distribuido sin puntos de error único. Un error en cualquiera de los nodos no provoca problemas de disponibilidad o confiabilidad en el clúster, principalmente porque los servicios de sistema de Service Fabric siguen las directrices anteriores. Es decir, siempre se ejecutan con tres o más réplicas de forma predeterminada y los servicios de sistema sin estado se ejecutan en todos los nodos.

Las capas subyacentes de detección de errores y redes de Service Fabric están totalmente distribuidas. La mayoría de los servicios de sistema se regeneran a partir de los metadatos del clúster o saben cómo volver a sincronizar su estado desde otros lugares. La disponibilidad del clúster podría deteriorarse si los servicios de sistema se encuentran en situaciones de pérdida de cuórum como las descritas anteriormente. En estos casos, quizá no pueda realizar algunas operaciones en el clúster (como comenzar una actualización o implementar servicios nuevos), pero el clúster continúa activo.

Los servicios en un clúster en ejecución seguirán ejecutándose en estas condiciones, a menos que requieran que las operaciones de escritura en los servicios de sistema continúen funcionando. Por ejemplo, si Administrador de conmutación por error está en pérdida de cuórum, todos los servicios seguirán ejecutándose. Sin embargo, los servicios con error no podrán reiniciarse automáticamente, ya que esto requiere la implicación de Administrador de conmutación por error.

Errores en un centro de datos o una región de Azure

En raras ocasiones, los centros de datos físicos pueden dejar de estar disponibles temporalmente debido a la pérdida de conectividad de red o de alimentación. En estos casos, los clústeres de Service Fabric y los servicios de ese centro de datos o la región de Azure dejarán de estar disponibles. Sin embargo, *los datos se conservan*.

Para los clústeres que se ejecutan en Azure, puede ver las actualizaciones en caso de interrupciones en la página de estado de Azure. En el caso muy poco probable de que se destruyera un centro de datos físico parcial o totalmente, los clústeres de Service Fabric que hospede o sus servicios se podrían perder. Esto incluye cualquier estado del que no se haya realizado copia de seguridad fuera de ese centro de datos o esa región.

Hay varias estrategias diferentes para sobrevivir al error permanente o duradero de un único centro de datos o región:

- Ejecutar clústeres de Service Fabric independientes en varias de estas regiones y utilizar algún mecanismo de conmutación por error y conmutación por recuperación entre estos entornos. Este tipo de modelo de varios clústeres activos/activos o activos/pasivos requiere código adicional de administración y operación. Este modelo también requiere la coordinación de las copias de seguridad de los servicios de un centro de datos o región para que estén disponibles en otros centros de datos o regiones cuando uno produzca un error.

- Ejecutar un único clúster de Service Fabric que abarque varios centros de datos. La configuración mínima admitida para esta estrategia es tres centros de datos. Para más

información, consulte Implementación de un clúster de Azure Service Fabric en Availability Zones.

Este modelo requiere configuración adicional. Sin embargo, la ventaja es que el error de un centro de datos se transforma de desastre a error normal. Estos errores se solucionan mediante los mecanismos que funcionan para los clústeres de una única región. Los dominios de error, los dominios de actualización y las reglas de colocación de Service Fabric aseguran que las cargas de trabajo se distribuyen para tolerar los errores normales.

Para más información sobre las directivas que pueden ayudar a que los servicios funcionen en este tipo de clúster, consulte Directivas de colocación de servicios de Service Fabric.

- Ejecutar un único clúster de Service Fabric que abarque varias regiones mediante el modelo independiente. El número recomendado de regiones es tres. Consulte Creación de un clúster independiente para más información sobre la configuración de Service Fabric independiente.

Errores aleatorios que dan lugar a errores en el clúster

Service Fabric tiene el concepto de *nodos raíz*. Se trata de nodos que conservan la disponibilidad del clúster subyacente.

Estos nodos raíz ayudan a garantizar que el clúster permanece activo, al establecer concesiones con otros nodos y servir de elemento diferenciador durante ciertos tipos de errores. Si los errores aleatorios eliminan la mayoría de los nodos raíz del clúster y no se recuperan rápidamente, el clúster se cierra automáticamente. A continuación, se produce un error en el clúster.

En Azure, el proveedor de recursos de Service Fabric administra configuraciones de clúster de Service Fabric. De forma predeterminada, el proveedor de recursos distribuye los nodos raíz entre los dominios de error y de actualización para el *tipo de nodo principal*. Si el tipo de nodo principal está marcado como durabilidad Silver o Gold, al quitar un nodo raíz (ya sea mediante el escalado en el tipo de nodo principal o mediante su eliminación manual), el clúster intentará promover otro nodo que no sea de raíz de la capacidad disponible del tipo de nodo principal. Este intento producirá un error si tiene menos capacidad disponible de la que requiere el nivel de confiabilidad del clúster para el tipo de nodo principal.

En los clústeres independientes de Service Fabric y Azure, el tipo de nodo principal es el que ejecuta los nodos raíz. Al definir un tipo de nodo principal, Service Fabric aprovechará automáticamente el número de nodos proporcionado mediante la creación de hasta nueve nodos raíz y siete réplicas de cada servicio del sistema. Si un conjunto de errores aleatorios deshabilita la mayoría de las réplicas al mismo tiempo, los servicios de sistema entran en

pérdida del cuórum. Si se pierde la mayoría de los nodos raíz, el clúster se apaga poco después.

Pasos siguientes

- Obtenga más información sobre cómo simular varios errores mediante la plataforma de comprobación.
- Lea otros recursos sobre alta disponibilidad y recuperación ante desastres. Microsoft ha publicado una gran cantidad de instrucciones sobre estos temas. Aunque algunos de estos recursos se refieren a técnicas específicas para su uso en otros productos, contienen muchos procedimientos recomendados generales que se pueden aplicar en el contexto de Service Fabric:
  - Lista de comprobación de disponibilidad
  - Exploración en profundidad de la recuperación ante desastres
  - Recuperación ante desastres y alta disponibilidad para aplicaciones de Azure
- Más información sobre las opciones de soporte técnico de Service Fabric.

Last updated on 22/03/2026

Confiabilidad en Azure Virtual Machines

Azure Virtual Machines proporciona recursos de proceso escalables a petición. Como servicio de infraestructura fundamental, está diseñado para ofrecer confiabilidad y disponibilidad de nivel empresarial para cargas de trabajo críticas.

Al usar Azure, relibilidad es una responsabilidad compartida. Microsoft proporciona una variedad de capacidades para admitir resiliencia y recuperación. Es responsable de comprender cómo funcionan esas funcionalidades dentro de todos los servicios que usa y de seleccionar las funcionalidades que necesita para cumplir los objetivos empresariales y los objetivos de tiempo de actividad.

En este artículo se describe cómo hacer que Virtual Machines sean resistentes a una variedad de posibles interrupciones y problemas, incluidos errores transitorios, interrupciones de zona de disponibilidad, interrupciones de regiones y mantenimiento del servicio. También se describe cómo puede usar copias de seguridad para recuperarse de otros tipos de problemas y se resalta cierta información clave sobre el acuerdo de nivel de servicio (SLA) de Virtual Machines.

> **ⓘ Importante**
>
> Al considerar la confiabilidad de una máquina virtual (VM), también debe tener en cuenta la confiabilidad de los discos, la infraestructura de red y las aplicaciones que se ejecutan en las máquinas virtuales. La mejora de la resistencia de la máquina virtual por sí sola podría tener un impacto limitado si los otros componentes no son igualmente resistentes. En función de los requisitos de resistencia, es posible que tenga que realizar cambios de configuración en varias áreas.

Recomendaciones de implementación de producción

Para obtener más información sobre cómo implementar máquinas virtuales para cumplir con los requisitos de confiabilidad de la solución y cómo la confiabilidad afecta a otros aspectos de la arquitectura, consulte Mejores prácticas de arquitectura para máquinas virtuales y conjuntos de escalado en el marco bien arquitectado de Azure.

Introducción a la arquitectura de confiabilidad

Las máquinas virtuales son la unidad de proceso fundamental en Azure, tanto si aprovisiona las máquinas virtuales usted mismo como si usa otros servicios de proceso Azure que aprovisionan y administran de forma transparente.

Una máquina virtual individual también se conoce como una *sola máquina virtual de instancia*. Se ejecuta en un host específico, que es un servidor físico. La mayoría de las máquinas virtuales comparten su host con otras máquinas virtuales.

Al crear las máquinas virtuales, puede influir en dónde se ejecutan dentro de la infraestructura subyacente. Normalmente, se toman decisiones de selección de ubicación en función de los requisitos de confiabilidad, latencia y aislamiento. Azure proporciona varias opciones de configuración que afectan a cómo se colocan las máquinas virtuales.

- **Region:** Puede seleccionar en qué Azure región que debe ejecutar la máquina virtual. Una región es un área geográfica que puede contener varios centros de datos, cada uno con un gran número de hosts.

- **Availability zone:**Availability zones son grupos de centros de datos físicamente independientes dentro de cada región de Azure. En regiones que admiten zonas de disponibilidad, puede seleccionar en qué zona se ejecuta la máquina virtual. Para obtener más información, consulte Resilience to availability zone failures (Resistencia a errores de zona de disponibilidad).

- **Availability sets:** Un conjunto de disponibilidad es una agrupación lógica de máquinas virtuales que permite a Azure comprender cómo se crea la aplicación para proporcionar redundancia y disponibilidad.

  Al usar conjuntos de disponibilidad, Azure distribuye un grupo de máquinas virtuales entre diferentes dominios de errores. Esta distribución minimiza el riesgo de errores de hardware localizados mediante la agrupación de máquinas virtuales que comparten una fuente de alimentación común y un conmutador de red.

  Los conjuntos de disponibilidad también pueden colocar diferentes máquinas virtuales en diferentes dominios de actualización, lo que controla cómo la plataforma de Azure gestiona la implementación de las actualizaciones de la plataforma. Mediante el uso de dominios de actualización, puede asegurarse de que solo se reinicia un subconjunto de las máquinas virtuales para las actualizaciones al mismo tiempo.

- **Grupos de colocación de proximidad:** Para las cargas de trabajo que necesitan lograr la menor latencia posible entre las máquinas virtuales, puede usar un grupo de proximidad

para asegurarse de que Azure coloque las máquinas virtuales físicamente cerca. Sin embargo, la ubicación de proximidad significa que una interrupción del centro de datos puede afectar a todas las máquinas virtuales del grupo. Para asegurar una fiabilidad elevada, es posible que tenga que aprovisionar varios grupos de colocación de proximidad en diferentes zonas de disponibilidad.

- **Dedicated hosts:** Puede usar Azure Dedicated Host para aprovisionar su propio servidor físico que ejecute una o varias máquinas virtuales, como para requisitos de cumplimiento estrictos. Sin embargo, al aprovisionar un host dedicado, una interrupción en su centro de datos puede afectar a todas las máquinas virtuales de ese host. Para lograr una alta confiabilidad, es posible que tenga que aprovisionar varios hosts dedicados en distintas zonas de disponibilidad.

Si crea un conjunto de máquinas virtuales que realizan funciones similares, considere la posibilidad de usar Azure Virtual Machine Scale Sets para crear y administrar las máquinas virtuales como un grupo. Los conjuntos de escalado también proporcionan más opciones de confiabilidad, como distribuir las máquinas virtuales entre varias zonas de disponibilidad.

Para obtener más información sobre la disponibilidad de las máquinas virtuales, consulte Opciones de disponibilidad para Virtual Machines.

Resistencia a errores transitorios

Los errores transitorios son errores breves e intermitentes en los componentes. Se producen con frecuencia en un entorno distribuido como la nube y son una parte normal de las operaciones. Los errores transitorios se corrigen después de un breve período de tiempo. Es importante que las aplicaciones puedan controlar errores transitorios, normalmente mediante el reintento de solicitudes afectadas.

Todas las aplicaciones hospedadas en la nube deben seguir las instrucciones de control de errores transitorios Azure cuando se comunican con cualquier API, bases de datos y otros componentes hospedados en la nube. Para obtener más información, consulte Recomendaciones para controlar errores transitorios.

Las aplicaciones que se ejecutan en las máquinas virtuales deben implementar estrategias adecuadas de control de errores para asegurarse de que las interrupciones temporales del servicio no afecten a la carga de trabajo.

Resistencia a errores de zona de disponibilidad

Availability zones son grupos de centros de datos físicamente independientes dentro de una región de Azure. Cuando una zona falla, los servicios pueden transferirse a una de las zonas restantes.

Una máquina virtual individual se puede implementar en una configuración *zonal* , lo que significa que está anclada a una sola zona de disponibilidad que seleccione. Por sí misma, una máquina virtual zonal no es resistente a interrupciones de zona. Sin embargo, puede crear varias máquinas virtuales y colocarlas en distintas zonas de disponibilidad y, a continuación, distribuir las aplicaciones y los datos en las instancias de máquina virtual. Como alternativa, puede usar conjuntos de escalado de máquinas virtuales para implementar un conjunto de máquinas virtuales en varias zonas de disponibilidad.

Si no configura una máquina virtual para que sea zonal, se considera *no zonal* o *regional*. Las máquinas virtuales no zonales se pueden colocar en cualquier zona de disponibilidad dentro de la región. Si alguna zona de disponibilidad de la región experimenta una interrupción, las máquinas virtuales nozonales podrían estar en la zona afectada y pueden experimentar tiempo de inactividad.

Requisitos

Compatibilidad con regiones: Las máquinas virtuales zonales se pueden implementar en cualquier región que admita zonas de disponibilidad.

Sin embargo, algunos tipos y tamaños de máquina virtual solo están disponibles en regiones específicas o zonas específicas dentro de una región. Para comprobar qué regiones y zonas admiten los tipos de máquina virtual que necesita, use los siguientes recursos:

- Para comprobar los tipos de máquina virtual disponibles en cada región, consulte Productos disponibles por región .
- Para comprobar los tamaños y tipos de máquina virtual admitidos dentro de cada zona de una región específica, consulte Comprobación de la disponibilidad de la SKU de máquina virtual.

Cost

No hay ninguna diferencia de costo entre una máquina virtual zonal y no zonal.

Configurar soporte de zonas de disponibilidad

En esta sección se explica cómo configurar la compatibilidad con la zona de disponibilidad para la instancia de máquina virtual.

> ⓘ **Nota:**
>
> Cuando selecciona qué zonas de disponibilidad usar, en realidad está seleccionando la *zona de disponibilidad lógica*. Si implementa otros componentes de carga de trabajo en una suscripción de Azure diferente, pueden usar un número de zona de disponibilidad lógica diferente para acceder a la misma zona de disponibilidad física. Para obtener más información, consulte **zonas de disponibilidad física y lógica**.

- **Cree una máquina virtual zonal.** Puede crear una máquina virtual zonal mediante las siguientes guías:
  - Crear máquinas virtuales en una zona de disponibilidad mediante el portal de Azure
  - Crear una máquina virtual en una zona de disponibilidad mediante el CLI de Azure
  - Crear una máquina virtual en una zona de disponibilidad mediante Azure PowerShell

- **Convierta las máquinas virtuales existentes en una configuración zonal.** Puede pasar de una máquina virtual no zonal a una máquina virtual zonal. Este proceso crea una nueva máquina virtual en la zona de disponibilidad de destino y requiere que la máquina virtual se detenga durante el proceso de traslado. Para obtener más información, consulte Mover máquinas virtuales de instancia única de Azure de zonas de disponibilidad regionales a zonales.

- **Cambie la zona de disponibilidad de una máquina virtual zonal existente.** Las máquinas virtuales zonales no se pueden mover a otra zona de disponibilidad. En su lugar, debe implementar una nueva máquina virtual zonal en la zona de disponibilidad deseada.

- **Convierta una máquina virtual zonal en una configuración no zonal.** Las máquinas virtuales zonales no se pueden convertir en una configuración no zonal. En su lugar, debe implementar una nueva máquina virtual no zonal.

Comportamiento cuando todas las zonas están en buen estado

En esta sección se describe qué esperar cuando las instancias de máquina virtual están configuradas con compatibilidad con la zona de disponibilidad y todas las zonas de disponibilidad están operativas.

- **Enrutamiento de tráfico entre zonas:** Es responsable del enrutamiento del tráfico entre máquinas virtuales, incluidas las máquinas virtuales que se encuentran en distintas zonas de disponibilidad. Entre los enfoques comunes se incluyen Azure Load Balancer y Azure Application Gateway. Para obtener más información, consulte Opciones de equilibrio de carga.

- **Replicación de datos entre zonas:** Es responsable de cualquier replicación de datos que necesite producirse entre máquinas virtuales, incluidas las máquinas virtuales de diferentes zonas de disponibilidad. Las bases de datos y otras aplicaciones con estado similares que se ejecutan en máquinas virtuales suelen proporcionar funcionalidades para replicar datos.

Comportamiento durante un fallo de zona

En esta sección se describe qué esperar cuando las instancias de máquina virtual están configuradas con compatibilidad con la zona de disponibilidad y hay una interrupción en sus zonas de disponibilidad.

- **Detección y respuesta:** Es responsable de detectar y responder a errores de zona que afectan a las máquinas virtuales.

- **Notification**: Microsoft no le notifica automáticamente cuando una zona está inactiva. Sin embargo, puede usar Azure Resource Health para supervisar el estado de un recurso individual y puede configurar Resource Health alertas para notificarle problemas. También puede usar Azure Service Health para comprender el estado general del servicio, incluidos los errores de zona, y puede configurar alertas de Service Health para notificarle problemas.

- **Solicitudes activas:** Es probable que se finalicen las solicitudes activas u otro trabajo que se produzca en la máquina virtual durante el error de zona.

- **Pérdida de datos esperada:** Es posible que los discos LRS zonales no estén disponibles durante un error de zona.

  Si usa discos de almacenamiento de zona redundante (ZRS) y una interrupción afecta a su máquina virtual, puede forzar la desasociación de los discos ZRS de la máquina virtual que ha fallado. Este enfoque permite conectar los discos ZRS a otra máquina virtual.

  Para obtener más información sobre cómo hacer que los discos sean resistentes a errores de zona de disponibilidad y otros problemas, consulte Reliability in Azure Disk Storage.

- **Tiempo de inactividad esperado:** Las máquinas virtuales permanecen inactivas hasta que se recupera la zona de disponibilidad.
- **Reenrutamiento del tráfico:** Es responsable de redirigir el tráfico a otras máquinas virtuales en zonas correctas.

  Si configura un equilibrador de carga resistente a zonas y realiza comprobaciones de estado, el equilibrador de carga normalmente detecta máquinas virtuales con errores y puede enrutar el tráfico a otras instancias de máquina virtual en zonas correctas.

Recuperación de zona

Cuando la zona está en buen estado, las máquinas virtuales de la zona se reinician. Es responsable de cualquier procedimiento de recuperación de zona y sincronización de datos que requieran las cargas de trabajo.

Prueba de fallos de zona

Puede usar Azure Chaos Studio para simular la pérdida de una máquina virtual como parte de un experimento. Chaos Studio proporciona errores integrados para máquinas virtuales, incluida la capacidad de apagar una máquina virtual. Puede usar estas capacidades para simular fallos a nivel de zona y probar los procesos de conmutación por error.

Soluciones personalizadas multizona para la resiliencia

Al implementar múltiples máquinas virtuales en distintas zonas, usted es responsable de configurar y administrar la replicación, el equilibrio de carga, la conmutación por error y los procesos de conmutación por recuperación.

Algunas aplicaciones proporcionan funcionalidades integradas que pueden ayudar al implementar en varias máquinas virtuales. Por ejemplo, SQL Server en máquinas virtuales Azure proporciona un conjunto de funcionalidades para simplificar los procesos de configuración y administración en todas las zonas de disponibilidad.

Puede considerar el uso de Azure Site Recovery recuperación ante desastres de zona a zona (DR) cuando su aplicación se ejecuta en una sola zona a la vez y no requiere un failover casi instantáneo entre zonas. La recuperación ante desastres de zona a zona tiene algunas limitaciones importantes, por lo que debe revisar los requisitos exhaustivamente.

Resistencia a errores en toda la región

Las máquinas virtuales son recursos de una sola región. Si la región deja de estar disponible, la máquina virtual tampoco está disponible.

Soluciones personalizadas de varias regiones para la resistencia

Puede implementar varias máquinas virtuales en diferentes regiones, pero debe implementar procesos de replicación, equilibrio de carga y conmutación por error.

Site Recovery es un servicio que permite la recuperación ante desastres mediante la replicación de máquinas virtuales y sus datos en una región secundaria. Puede seleccionar casi cualquier región de Azure como región secundaria, incluidas combinaciones de regiones no emparejadas. Para obtener más información, consulte la arquitectura de recuperación ante desastres de Azure a Azure.

Algunas aplicaciones crean clústeres u otras construcciones para replicar datos y distribuir el trabajo entre varias máquinas virtuales, incluidas en regiones diferentes. Estas aplicaciones pueden simplificar la configuración de una solución de varias regiones.

Para obtener una arquitectura de ejemplo que ilustra cómo usar máquinas virtuales en varias regiones, consulte Equilibrio de carga en varias regiones con Azure Traffic Manager, Azure Firewall y Application Gateway.

Resistencia al mantenimiento del servicio

Azure realiza un mantenimiento periódico normal en máquinas virtuales para garantizar la confiabilidad. Hay varias maneras de asegurarse de que las cargas de trabajo permanecen operativas durante las actividades de mantenimiento:

- Al usar conjuntos de disponibilidad o conjuntos de escalado de máquinas virtuales, puede configurar dominios de actualización. Los dominios de actualización ayudan a distribuir las actividades de mantenimiento en diferentes máquinas virtuales en momentos diferentes, por lo que las máquinas virtuales no se reinician simultáneamente.
- Puede personalizar el tiempo que se aplica el mantenimiento a las máquinas virtuales mediante *el control de mantenimiento*. Puede usar configuraciones de mantenimiento para programarla en un momento que se adapte a la carga de trabajo.
- Puede recibir notificaciones de las próximas actividades de mantenimiento.

Para obtener más información, consulte Información general sobre las actualizaciones de invitado y el mantenimiento del host.

Copias de seguridad y restauración

Azure Backup proporciona compatibilidad nativa de copia de seguridad con máquinas virtuales. Azure Backup crea y administra copias de seguridad y proporciona protección coherente con la aplicación para toda la máquina virtual, incluidos todos los discos conectados. Una solución de copia de seguridad de máquina virtual con Azure Backup es ideal cuando se necesita una copia de seguridad coordinada de varios discos o copias de seguridad compatibles con aplicaciones. Sin embargo, en el caso de las cargas de trabajo de base de datos, considere las soluciones de copia de seguridad específicas de la aplicación que proporcionan protección coherente con la transacción y opciones de recuperación más rápidas.

Con Azure Backup para máquinas virtuales, puede personalizar la frecuencia de copia de seguridad, la duración de retención y la configuración de almacenamiento para satisfacer sus necesidades. Para obtener más información, consulte Azure Backup for VMs.

La copia de seguridad también admite discos conectados a máquinas virtuales. Para obtener más información, vea Overview of Azure Disk Backup.

Para la mayoría de las soluciones, no debe confiar exclusivamente en copias de seguridad. En su lugar, utilice las otras capacidades descritas en esta guía para apoyar los requisitos de resiliencia. Sin embargo, las copias de seguridad protegen contra algunos riesgos que otros enfoques no. Para más información, consulte ¿Qué son la redundancia, la replicación y la copia de seguridad?.

Acuerdo de nivel de servicio

El acuerdo de nivel de servicio (SLA) para Azure servicios describe la disponibilidad esperada de cada servicio y las condiciones que la solución debe cumplir para lograr esa expectativa de disponibilidad. Para obtener más información, vea SLAs for servicios en línea.

Para Virtual Machines, el Acuerdo de Nivel de Servicio proporciona un nivel base de disponibilidad. El porcentaje de tiempo de actividad definido en el Acuerdo de Nivel de Servicio aumenta cuando tiene dos o más máquinas virtuales y realiza las siguientes acciones:

- Configure esas máquinas virtuales para que se implementen en dos o más zonas de disponibilidad.

- Configure esas máquinas virtuales para su implementación en un grupo de disponibilidad.

Pasos siguientes

- VMs en Azure
- Conjuntos de Escala de Máquinas Virtuales

Recurso relacionado

- Reliability in Azure Disk Storage
- Fiabilidad en Azure

Last updated on 08/04/2026

Confiabilidad en Azure Image Builder (AIB)

Este artículo contiene la recuperación ante desastres entre regiones y continuidad empresarial.

Azure Image Builder (AIB) es un servicio regional con un clúster que atiende a regiones únicas. La configuración regional de AIB mantiene los datos y los recursos dentro del límite regional. AIB como servicio no realiza la conmutación por error para clústeres y bases de datos SQL en escenarios de regiones inactivas.

> **Nota:**
>
> Azure Image Builder no admite zonas de disponibilidad.

Recuperación ante desastres entre regiones y continuidad empresarial

La recuperación ante desastres (DR) hace referencia a las prácticas que las organizaciones usan para recuperarse de eventos de alto impacto, como desastres naturales o implementaciones con errores que producen tiempo de inactividad y pérdida de datos. Independientemente de la causa, el mejor remedio para un desastre es un plan de recuperación ante desastres bien definido y probado y un diseño de aplicaciones que apoye activamente la recuperación ante desastres. Antes de empezar a crear el plan de recuperación ante desastres, consulte Recomendaciones para diseñar una estrategia de recuperación ante desastres.

Para la recuperación ante desastres, Microsoft usa el modelo de responsabilidad compartida. En este modelo, Microsoft garantiza que la infraestructura de línea base y los servicios de plataforma estén disponibles. Sin embargo, muchos servicios de Azure no replican automáticamente datos ni se revierten de una región con errores para realizar la replicación cruzada en otra región habilitada. En esos servicios, es responsable de configurar un plan de recuperación ante desastres válido para su carga de trabajo. La mayoría de los servicios que se ejecutan en ofertas de plataforma como servicio (PaaS) de Azure proporcionan características e instrucciones para admitir la recuperación ante desastres. Puede usar características específicas del servicio para admitir la recuperación rápida con el fin de contribuir al desarrollo del plan de recuperación ante desastres.

Para garantizar una recuperación rápida y sencilla para Azure Image Builder (AIB), se recomienda ejecutar una plantilla de imagen en pares de regiones o en varias regiones al diseñar la solución de AIB. También debe replicar los recursos desde el principio cuando configure las plantillas de imagen.

Recuperación ante desastres en una geografía multirregional

Cuando se produce un desastre regional, Microsoft es responsable de la detección de interrupciones, las notificaciones y la compatibilidad con AIB. Sin embargo, usted es responsable de configurar la recuperación ante desastres para los planes de control (lado del servicio) y de datos.

Detección, notificación y administración de interrupciones

Microsoft envía una notificación si hay una interrupción del servicio Azure Image Builder (AIB). Un síntoma de interrupción común incluye plantillas de imagen que obtienen errores 500 al intentar ejecutarse. Puede revisar las notificaciones de interrupción y las actualizaciones de estado de Azure Image Builder mediante la administración de solicitudes de soporte técnico.

Configuración de la recuperación ante desastres y la detección de interrupciones

Usted es responsable de configurar la recuperación ante desastres para su entorno de Azure Image Builder (AIB), ya que no hay una conmutación por error de región en el lado del servicio AIB. Debe configurar el plano de control (lado de servicio) y el plano de datos.

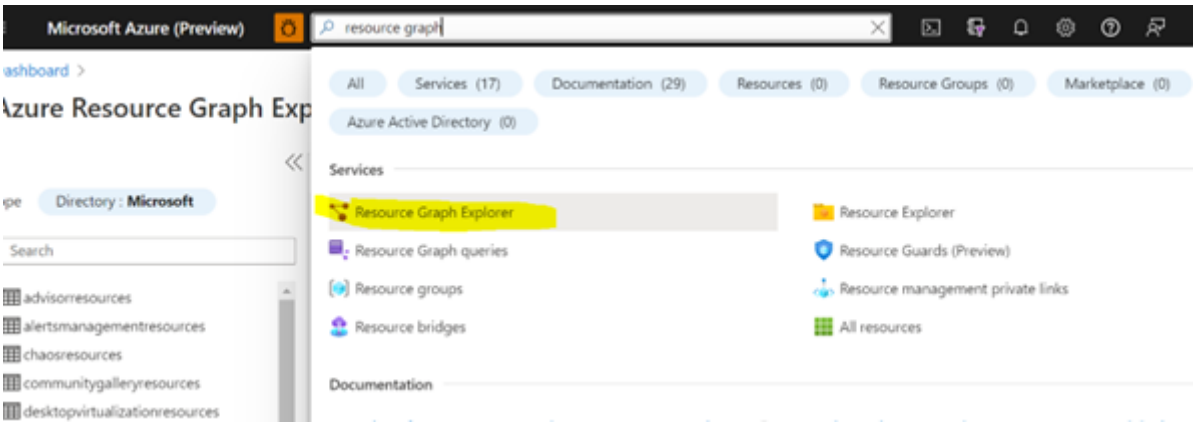
Se recomienda crear un recurso de AIB en otra región cercana, en la que puede replicar los recursos. Para obtener más información, consulte las regiones admitidas y qué recursos se incluyen en una creación de AIB.

Recuperación ante desastres de geografía de una sola región

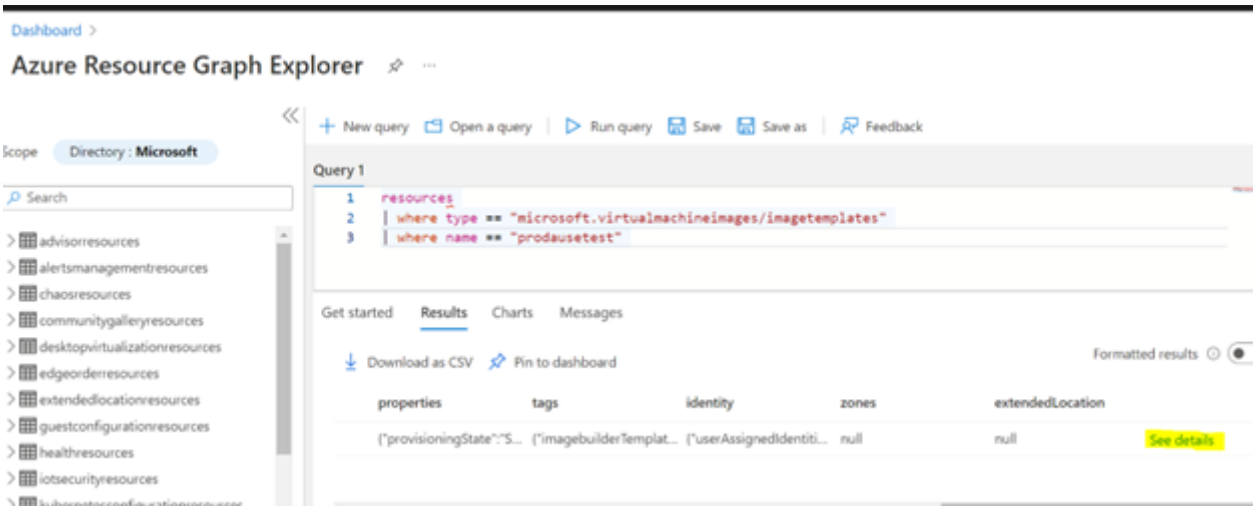
En el caso de un desastre para una sola región, todavía debe obtener un recurso de plantilla de imagen de esa región incluso cuando esa región no esté disponible. Puede mantener una copia de una plantilla de imagen localmente o pueden usar Azure Resource Graph desde Azure Portal para obtener un recurso de plantilla de imagen.

Para obtener un recurso de plantilla de imagen usando Resource Graph desde Azure Portal:

1. Vaya a la barra de búsqueda de Azure Portal y busque *explorador de Resource Graph*.



2. Use la barra de búsqueda del extremo izquierdo para buscar recursos por tipo y nombre, y así podrá consultar los detalles que le proporcionan las propiedades de la plantilla de imagen. La opción *Ver detalles* de la parte inferior derecha muestra el atributo de propiedades y las etiquetas de la plantilla de imagen por separado. El nombre de plantilla, la ubicación, el identificador y el identificador de inquilino se pueden usar para obtener el recurso de plantilla de imagen correcto.



Capacidad y resistencia proactiva de la recuperación ante desastres

Microsoft y sus clientes operan bajo el modelo de responsabilidad compartida. En la recuperación ante desastres habilitada para el cliente (servicios responsables del cliente), es responsable de abordar la recuperación ante desastres para cualquier servicio que implemente y controle. Para asegurarse de que la recuperación es proactiva, siempre debe implementar secundarias previamente. Sin secundarias implementadas previamente, no hay ninguna garantía de capacidad en el momento del impacto.

Al planear dónde replicar una plantilla, tenga en cuenta lo siguiente:

- Disponibilidad de la región de AIB:
  - Elija regiones compatibles con AIB cerca de sus usuarios.

    - AIB se expande continuamente en nuevas regiones.
- Regiones emparejadas de Azure:
    - Para su área geográfica, elija dos regiones emparejadas entre sí.
    - Esfuerzos de recuperación para regiones emparejadas en las que es necesario establecer prioridades.

Instrucciones adicionales

En lo que respecta a la información sobre el procesamiento de sus datos, consulte los detalles sobre la residencia de datos de Azure Image Builder.

Pasos siguientes

- Confiabilidad en Azure
- Habilitación de la recuperación ante desastres de máquinas virtuales de Azure entre zonas de disponibilidad
- Introducción a Azure Image Builder

Last updated on 22/01/2026

Fiabilidad en Azure Virtual Machine Scale Sets

Azure Virtual Machine Scale Sets es un recurso de proceso de Azure que puede usar para crear y administrar un grupo de instancias de máquina virtual (VM). El número de instancias de máquina virtual puede aumentar o disminuir automáticamente según la demanda, o de acuerdo a una programación definida. Los conjuntos de escalado de máquinas virtuales ayudan a que las aplicaciones sean de alta disponibilidad y resistentes mediante la distribución de máquinas virtuales entre varias zonas de disponibilidad y dominios de error.

Cuando se usa Azure, la confiabilidad es una responsabilidad compartida. Microsoft proporciona una variedad de funcionalidades para admitir resistencia y recuperación. Es responsable de comprender cómo funcionan esas funcionalidades dentro de todos los servicios que usa y de seleccionar las funcionalidades que necesita para cumplir los objetivos empresariales y los objetivos de tiempo de actividad.

En este artículo se describe cómo hacer que los conjuntos de escalado de máquinas virtuales sean resistentes a varias posibles interrupciones y problemas, incluidos errores transitorios, interrupciones de zona de disponibilidad, interrupciones de regiones, reconfiguración de máquinas virtuales y mantenimiento del servicio. También describe cómo puede usar copias de seguridad para recuperarse de otros tipos de problemas y resalta información clave sobre el acuerdo de nivel de servicio (SLA) de virtual Machine Scale Sets.

> **ⓘ Importante**
>
> Cuando considere la confiabilidad de un conjunto de escalado y sus máquinas virtuales, también debe tener en cuenta la confiabilidad de los discos, la infraestructura de red y las aplicaciones que se ejecutan en las máquinas virtuales. Mejorar la resistencia de las máquinas virtuales por sí solas podría tener un efecto limitado si los otros componentes no son igualmente resistentes. En función de los requisitos de resistencia, es posible que tenga que realizar cambios de configuración en varias áreas.

Recomendaciones de implementación de producción

Azure Well-Architected Framework proporciona recomendaciones para la confiabilidad, el rendimiento, la seguridad, el costo y las operaciones. Para obtener información sobre cómo estas áreas influyen entre sí y contribuyen a una solución de conjuntos de escalado de

máquinas virtuales confiables, consulte Procedimientos recomendados de arquitectura para Azure Virtual Machines y conjuntos de escalado en Well-Architected Framework.

Introducción a la arquitectura de confiabilidad

Un conjunto de escalado agrupa varias instancias de máquina virtual y aplica la configuración centralizada, las reglas de escalado automático y las actualizaciones graduales.

Los conjuntos de escalas admiten dos modos de orquestación distintos:

- *Los conjuntos de escalado flexibles* (recomendados) proporcionan más flexibilidad para implementar y administrar instancias de máquina virtual individuales.
- *Los conjuntos de escalado uniforme* implementan máquinas virtuales que tienen una configuración idéntica y las administra como un grupo.

Propagación de dominios de error

Los dominios de error son grupos de aislamiento de errores dentro de un centro de datos. Cada dominio de error es como un bastidor de servidor, que es una colección de nodos de hardware que comparten la misma alimentación, redes, enfriamiento y cronograma de mantenimiento de la plataforma. Dado que las instancias de máquina virtual de cada conjunto de escalado se distribuyen entre varios dominios de error, es probable que una interrupción planeada o no planeada que se produzca en un dominio de error no afecte a las instancias de máquina virtual en otros dominios de error.

Al implementar un conjunto de escalado, puede controlar el número de dominios de error entre los que se distribuyen las máquinas virtuales. Para la mayoría de los escenarios, use el comportamiento de *propagación máxima* , que usa tantos dominios de error como sea posible. Para obtener más información, consulte Elegir el Número Correcto de Dominios de Error para Virtual Machine Scale Sets.

En las regiones que tienen zonas de disponibilidad, cada zona tiene un conjunto distinto de dominios de error. Al crear un conjunto de escalado que abarca regiones, las instancias se distribuyen entre dominios de fallo en cada región que usa el conjunto de escalado.

Equilibrio de carga

Los conjuntos de escalado se pueden integrar con los servicios de equilibrio de carga de Azure, incluido Azure Load Balancer y Azure Application Gateway. Cuando el conjunto de escalado agrega o quita instancias, la integración integrada del equilibrador de carga actualiza automáticamente la configuración del equilibrador de carga. Para obtener más información, consulte Redes para conjuntos de escalado de máquinas virtuales.

Los conjuntos de escalado incluyen muchos otros controles y funcionalidades que afectan a la implementación, escalado, distribución y actualización de instancias. Para obtener más información, consulte Introducción a los conjuntos de escalado de máquinas virtuales.

Resistencia a errores transitorios

Los errores transitorios son errores breves e intermitentes en los componentes. Se producen con frecuencia en un entorno distribuido como la nube y son una parte normal de las operaciones. Los errores transitorios se corrigen después de un breve período de tiempo. Es importante que las aplicaciones puedan controlar errores transitorios, normalmente mediante el reintento de solicitudes afectadas.

Todas las aplicaciones hospedadas en la nube deben seguir las instrucciones de control de errores transitorios de Azure cuando se comunican con cualquier API, bases de datos y otros componentes hospedados en la nube. Para obtener más información, consulte Recomendaciones para controlar errores transitorios.

Las aplicaciones que se ejecutan en las máquinas virtuales deben implementar estrategias adecuadas de control de errores para asegurarse de que las interrupciones temporales del servicio no afecten a la carga de trabajo.

Resiliencia ante problemas de instancia

Cuando un conjunto de escalado inicia una tarea de creación o eliminación de una instancia de máquina virtual, es posible que se produzca un error en la operación. Para reintentar automáticamente las tareas de creación o eliminación de instancias de máquina virtual con errores, considere la posibilidad de usar la característica resistente de creación y eliminación de conjuntos de escalado de máquinas virtuales (versión preliminar).

Es posible que se produzcan problemas durante la ejecución de instancias. Por ejemplo, una instancia podría dejar de responder debido a bloqueos de aplicación o agotamiento de recursos. Use reparaciones automáticas de instancias para supervisar el estado de la aplicación y reiniciar, restablecer la imagen o reemplazar automáticamente una instancia de máquina virtual cuando sea necesario.

Resistencia a errores de zona de disponibilidad

Las zonas de disponibilidad son grupos físicamente independientes de centros de datos dentro de una región de Azure. Cuando una zona falla, los servicios pueden transferirse a una de las zonas restantes.

Virtual Machine Scale Sets admite zonas de disponibilidad en configuraciones zonales y de expansión de zona.

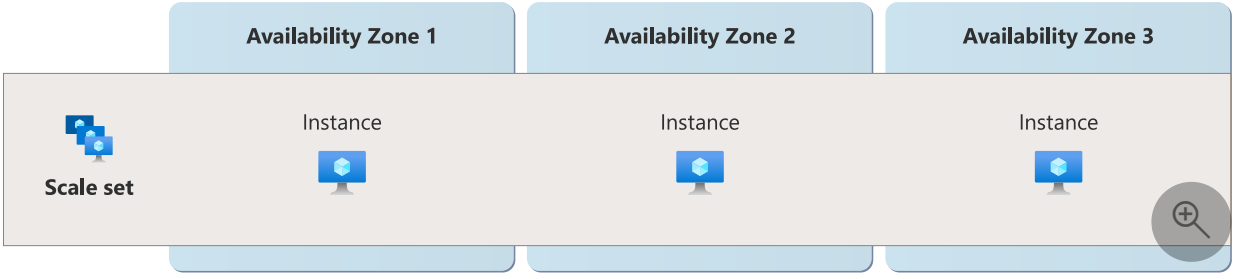
- Un conjunto de escalado que abarca zonas distribuye instancias en varias zonas de disponibilidad que seleccionas.

  Distribuir instancias de máquina virtual entre zonas de disponibilidad te ofrece el nivel más alto de SLA. Recomendamos utilizar conjuntos de escalado que abarcan zonas para la mayoría de las cargas de trabajo basadas en máquinas virtuales en Azure.

  En un conjunto de escala que abarca zonas, cada instancia de máquina virtual y sus discos están vinculados a una zona de disponibilidad específica. Cuando todas las zonas están en buen estado, las instancias pueden comunicarse entre zonas mediante una red de alto rendimiento y baja latencia. Si una zona experimenta un problema de interrupción o conectividad, las instancias de las otras zonas siguen sin verse afectadas.

  De forma predeterminada, el conjunto de escalado usa un enfoque de esfuerzo máximo para repartir equitativamente las instancias entre las zonas seleccionadas. Pero si necesita un equilibrio estricto, puede cambiar la configuración de equilibrio de zona.

  En el diagrama siguiente se muestra un conjunto de instancias escalables que abarca tres zonas, incluyendo una instancia en cada zona.



  La expansión de zona es similar a la redundancia de zona en otros servicios de Azure, pero los conjuntos de escalado no realizan automáticamente la replicación de datos entre zonas ni la conmutación por error cuando las zonas están inactivas. En algunas situaciones inusuales, un conjunto de escalado de expansión de zona también podría tener sus instancias implementadas en una sola zona, como cuando se conectan máquinas virtuales individuales de una zona a un conjunto de escalado flexible de expansión de zona que tiene una sola instancia en la misma zona.

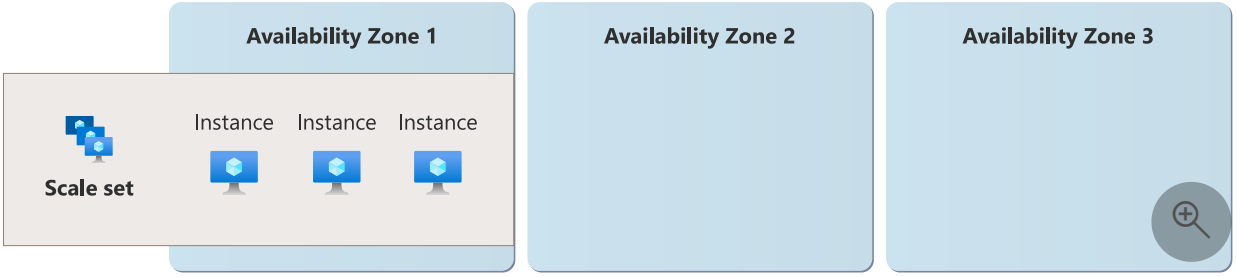
  > ⓘ **Nota:**
  >
  > Si usa el modo de orquestación flexible y adjunta, desasocia o quita máquinas virtuales individuales, asegúrese de que las máquinas virtuales se reparten entre

varias zonas. Si todas las máquinas virtuales están en una sola zona, es posible que el conjunto de escalado no sea resistente a una interrupción en dicha zona.

- Un conjunto de escala *zonal*, también llamado *alineado con la zona*, coloca todas sus instancias en una sola zona de disponibilidad que usted especifique. Cada máquina virtual y sus discos son zonales, por lo que se anclan a esa zona específica.

  > **ⓘ Importante**
  >
  > La asignación a una sola zona de disponibilidad solo se recomienda cuando la latencia entre zonas es demasiado alta para sus necesidades y después de comprobar que la latencia no cumple sus requisitos. Por sí mismo, un recurso zonal no proporciona resistencia a una interrupción de zona de disponibilidad. Para mejorar la resiliencia de un recurso zonal, debe desplegar explícitamente recursos independientes en múltiples zonas de disponibilidad y configurar el enrutamiento del tráfico y la conmutación por error. Para obtener más información, consulte Recursos zonales y resistencia de zona.

  En el diagrama siguiente se muestra un conjunto de escalado zonal en una sola zona, con tres instancias en la zona.



Si no especifica zonas de disponibilidad para el conjunto de escalado, es *no zonal* o *regional*. En este escenario, es posible que las instancias se coloquen en cualquier zona dentro de la región y no se distribuyan uniformemente ni se encuentren en la misma zona. Cuando se usa un conjunto de escalado no zonal, se garantiza la colocación de disco en la misma zona para discos Ultra y Premium v2. La coubicación se proporciona de la mejor manera posible para los discos Premium v1 y no se garantiza para los discos de SKU estándar, incluidos los discos de unidad de estado sólido (SSD) o de disco duro (HDD). Si se produce un error en alguna zona de la región, el conjunto de escalado podría experimentar tiempo de inactividad.

Requisitos

- **Soporte para regiones:** Puede implementar conjuntos de escalado que abarquen zonas y conjuntos zonales en cualquier región que soporte zonas de disponibilidad.

  Sin embargo, algunos tipos y tamaños de máquina virtual solo están disponibles en regiones específicas o zonas específicas dentro de una región. Para comprobar qué regiones y zonas admiten los tipos de máquina virtual que necesita, use los siguientes recursos:

  - Para comprobar los tipos de máquina virtual disponibles en cada región, consulte Productos disponibles por región.
  - Para comprobar los tamaños y tipos de máquina virtual admitidos dentro de cada zona de una región específica, consulte Comprobación de la disponibilidad de la SKU de máquina virtual.

  Si una SKU de máquina virtual específica no está disponible en ninguna de las zonas que seleccione para el conjunto de escalamiento, es posible que el conjunto de escalamiento no pueda escalar para satisfacer los requisitos de capacidad.

- **Hosts dedicados:** Las implementaciones de hosts dedicados de Azure no admiten conjuntos de escalado que abarcan zonas ni conjuntos de escalado zonal.

- **Tipos:** La compatibilidad con zonas de disponibilidad está disponible para todos los tipos de conjuntos de escalado, incluidos los conjuntos de escalado flexibles y uniformes.

Consideraciones

- **Propagación de dominios de error:** Cuando el conjunto de escalado usa zonas de disponibilidad, debe seleccionar entre enfoques específicos de propagación de dominios de error. Se recomienda usar la propagación máxima, que usa tantos dominios de error como sea posible, para la mayoría de las cargas de trabajo. Para obtener más información, consulte Elegir el Número Correcto de Dominios de Error para Virtual Machine Scale Sets.

- **Equilibrio de zona:** el equilibrio de zona determina si las instancias de máquina virtual de un conjunto de escalado se distribuyen uniformemente entre las zonas que seleccione. Un conjunto de escalado se considera equilibrado si cada zona tiene el mismo número de máquinas virtuales, más o menos una máquina virtual. Puede establecer el modo de balanceo de zona en mejor esfuerzo o estricto. Esta configuración controla si el conjunto de escalado puede escalar horizontalmente de forma desigual, incluidos los escenarios de interrupción de zona.

- **Grupos de selección de ubicación:** Para conjuntos de escalado uniformes, si configura varios grupos de selección de ubicación, Azure implementa varios grupos de selección de ubicación en cada zona que usa el conjunto de escalado.

Cost

No hay ninguna diferencia de costo entre un conjunto de escalado que abarca zonas, un conjunto zonal y uno no zonal que tengan el mismo número y tipo de instancias de máquinas virtuales.

Configurar soporte de zonas de disponibilidad

En esta sección se explica cómo configurar la compatibilidad con zonas de disponibilidad para el conjunto de escalas.

- **Cree un conjunto de escalado zonal o que abarque múltiples zonas.** Puede configurar zonas de disponibilidad al crear un nuevo conjunto de escalado. Para obtener más información, consulte Creación de un conjunto de escalado de máquinas virtuales que usa zonas de disponibilidad.

  > ⓘ **Nota:**
  >
  > Cuando selecciona qué zonas de disponibilidad usar, en realidad está seleccionando la *zona de disponibilidad lógica*. Si implementa otros componentes de la carga de trabajo en una suscripción de Azure diferente, podrían usar un número de zona de disponibilidad lógica distinto para acceder a la misma zona de disponibilidad física. Para obtener más información, consulte **zonas de disponibilidad física y lógica**.

- **Convierta los conjuntos de escalado existentes para usar zonas de disponibilidad.** Puede convertir un conjunto de escalado no zonal existente para usar zonas de disponibilidad. Para más información, consulte Actualización de conjuntos de escalado para agregar zonas de disponibilidad.

- **Cambie la configuración de zona de disponibilidad de un conjunto de escalado existente.** Puede agregar zonas a un conjunto de escalado existente, pero no puede quitar zonas. Para más información, consulte Actualización de conjuntos de escalado para agregar zonas de disponibilidad.

  > ⓘ **Importante**

> Al expandir un conjunto de escalado a más zonas, las instancias de máquina virtual originales no migran ni cambian inmediatamente. Al escalar, se crean nuevas instancias y se distribuyen uniformemente en las zonas de disponibilidad seleccionadas. Si necesita obtener datos de las instancias originales, será responsable de migrar esos datos a instancias en las nuevas zonas. Al escalar en el conjunto de escalado, las instancias regionales se priorizan primero para la eliminación. A continuación, las instancias se quitan en función de la directiva de escalado del conjunto de escalado. Para obtener más información, consulte **Cómo equilibrar manualmente el conjunto de escalado**.

Planeamiento y administración de capacidad

Para prepararse para errores de zona de disponibilidad, considere la *posibilidad de sobreaprovisionar* el número de instancias de máquina virtual del conjunto de escalado. Este enfoque permite a la solución tolerar cierta pérdida de capacidad y seguir funcionando sin degradar el rendimiento y garantiza que las zonas restantes tengan capacidad suficiente para controlar la carga de producción completa. Para obtener más información, consulte Administración de la capacidad mediante el aprovisionamiento excesivo.

Comportamiento cuando todas las zonas están en buen estado

En esta sección se describe qué esperar cuando los conjuntos de escalado están configurados con compatibilidad con la zona de disponibilidad y todas las zonas de disponibilidad están operativas.

- **Enrutamiento de tráfico entre zonas:** Es responsable del enrutamiento del tráfico entre máquinas virtuales del conjunto de escalado, incluidas las máquinas virtuales que se encuentran en distintas zonas de disponibilidad. Entre los enfoques comunes se incluyen Load Balancer y Application Gateway, que proporcionan integración integrada con conjuntos de escalado. Para obtener más información, consulte Redes para conjuntos de escalado de máquinas virtuales.

- **Replicación de datos entre zonas:** Es responsable de cualquier replicación de datos que necesite producirse entre máquinas virtuales, incluidas las máquinas virtuales de diferentes zonas de disponibilidad. Las bases de datos y otras aplicaciones con estado similares que se ejecutan en máquinas virtuales suelen proporcionar funcionalidades para replicar datos.

Comportamiento durante un fallo de zona

En esta sección se describe qué esperar cuando los conjuntos de escalado están configurados con soporte para zonas de disponibilidad y se produce una interrupción en estas.

- **Detección y respuesta:** Es responsable de detectar la pérdida de una zona de disponibilidad y decidir cómo responder.

  En el caso de los conjuntos de escalado que abarcan zonas, las instancias de máquina virtual en la zona afectada podrían no estar disponibles. Las instancias de las zonas saludables permanecen operativas.

  En el caso de los conjuntos de escalado zonales implementados en la zona afectada, es posible que todas las instancias de máquina virtual no estén disponibles. Debe planificar cómo responder a una falla de zona. Por ejemplo, puede redirigir el tráfico a otro conjunto de escalado en una zona o región diferente.

- **Notificación**: Microsoft no le notifica automáticamente cuando una zona está inactiva. Sin embargo, puede usar Azure Resource Health para supervisar el estado de un recurso individual y puede configurar alertas de Resource Health para notificarle problemas. También puede usar Azure Service Health para comprender el estado general del servicio, incluidos los errores de zona, y puede configurar alertas de Service Health para notificarle problemas.

- **Solicitudes activas:** Es probable que se finalicen las solicitudes activas u otro trabajo que se produzca en las máquinas virtuales de la zona de disponibilidad afectada.

- **Pérdida de datos esperada:** Es posible que los discos de máquina virtual zonales no estén disponibles durante un error de zona.

  Si usa discos de almacenamiento de zona redundante (ZRS) y una interrupción afecta a su máquina virtual, puede forzar la desasociación de los discos ZRS de la máquina virtual que ha fallado. Este enfoque permite conectar los discos ZRS a otra máquina virtual.

- **Tiempo de inactividad esperado:** Las máquinas virtuales de la zona afectada permanecen inactivas hasta que se recupere la zona de disponibilidad. Al usar conjuntos de escalado que abarcan múltiples zonas, las máquinas virtuales ubicadas en zonas saludables siguen funcionando.

- **Reenrutamiento del tráfico:** Es responsable de redirigir el tráfico a otras máquinas virtuales en zonas correctas.

  Si configura un equilibrador de carga resistente a zonas que realiza comprobaciones de estado, el equilibrador de carga suele detectar máquinas virtuales con errores y puede enrutar el tráfico a otras instancias de máquina virtual en zonas correctas.

- **Reemplazo de instancia:** No se garantiza que los conjuntos de escalado de máquinas virtuales agreguen automáticamente nuevas instancias a zonas saludables.

  Si tiene un conjunto de escalado que abarca zonas, puede escalar horizontalmente para agregar más instancias. Si el error de zona está restringido a conjuntos específicos de servidores dentro de la zona, la operación de escalabilidad horizontal podría agregar instancias correctas a la misma zona o podría agregar instancias a otras zonas. Sin embargo, si el conjunto de escalado usa un equilibrio de zona estricto, el conjunto de escalado bloquea las operaciones de escalado horizontal que provocan un desequilibrio.

  > **Sugerencia**
  >
  > Se recomienda configurar reglas de **escalado automático** en función del uso de CPU o memoria. Las reglas de escalado automático pueden permitir que el conjunto de escalado responda a una pérdida de las instancias de máquina virtual en una zona aumentando el escalado para agregar nuevas instancias en las zonas operativas restantes.

Recuperación de zona

Cuando la zona está en buen estado, las máquinas virtuales de la zona se reinician. Es responsable de cualquier procedimiento de recuperación de zona y sincronización de datos que requieran las cargas de trabajo.

Si agrega instancias temporales al grupo de escalado durante un fallo de zona, cuando se restaura la zona, es posible que tenga que reducir el tamaño del grupo de escalado a la capacidad original.

Prueba de fallos de zona

Puede usar Azure Chaos Studio para simular la pérdida de máquinas virtuales en una o varias zonas de disponibilidad como parte de un experimento. Chaos Studio proporciona errores integrados para conjuntos de escalado, incluida la capacidad de apagar máquinas virtuales en zonas específicas. Puede usar estas capacidades para simular fallos a nivel de zona y probar los procesos de conmutación por error.

Resistencia a errores en toda la región

Los conjuntos de escalado son recursos de una sola región. Si la región no está disponible, los conjuntos de escalado de la región tampoco están disponibles.

Soluciones personalizadas de varias regiones para la resistencia

Puede implementar varios conjuntos de escalado en diferentes regiones, pero debe implementar procesos de replicación, equilibrio de carga y conmutación por error. Por ejemplo, puede implementar conjuntos de escalado idénticos en varias regiones y usar Azure Front Door o Azure Traffic Manager, junto con sondeos de estado, para enrutar el tráfico. Es responsable de replicar el estado mediante mecanismos de aplicación o servicios de datos administrados.

Copias de seguridad y restauración

Azure Backup proporciona compatibilidad nativa con copias de seguridad para máquinas virtuales. Azure Backup crea y administra copias de seguridad y proporciona protección coherente con la aplicación para toda la máquina virtual, incluidos todos los discos conectados. Una solución de copia de seguridad de máquina virtual con Azure Backup es ideal cuando se necesita una copia de seguridad coordinada de varios discos o copias de seguridad compatibles con aplicaciones. Sin embargo, en el caso de las cargas de trabajo de base de datos, considere las soluciones de copia de seguridad específicas de la aplicación que proporcionan protección coherente con la transacción y opciones de recuperación más rápidas.

Con Azure Backup para vm, puede personalizar la frecuencia de copia de seguridad, la duración de retención y la configuración de almacenamiento para satisfacer sus necesidades. Para más información, consulte Azure Backup para VM.

La copia de seguridad también admite discos conectados a máquinas virtuales. Para obtener más información, consulte Información general sobre Azure Disk Backup.

Para la mayoría de las soluciones, no debe confiar exclusivamente en copias de seguridad. En su lugar, utilice las otras capacidades descritas en esta guía para apoyar los requisitos de resiliencia. Sin embargo, las copias de seguridad protegen contra algunos riesgos que otros enfoques no. Para más información, consulte ¿Qué son la redundancia, la replicación y la copia de seguridad?.

Resistencia a la reconfiguración de máquinas virtuales

Los conjuntos de escalado permiten controlar cómo se aplican los cambios de configuración a las máquinas virtuales, como cambiar la SKU de la máquina virtual, cambiar la imagen que usa

cada máquina virtual y agregar o quitar extensiones de máquina virtual. Puede controlar el *modo de directiva de actualización*, que determina cómo se aplican las actualizaciones. Para obtener más información, consulte modos de actualización de la política para conjuntos de máquinas virtuales a escala.

Algunos tipos de actualización requieren la reestructuración o la reimplementación de una instancia. Para excluir instancias específicas de las actualizaciones automáticas, considere la posibilidad de usar la protección de instancias. Puede excluir instancias que contengan estado que necesite conservar o configuración que no pueda replicar en otras instancias.

Resistencia al mantenimiento del servicio

Azure realiza periódicamente actualizaciones para mejorar la confiabilidad, el rendimiento y la seguridad de la infraestructura de host para las máquinas virtuales. Los conjuntos de escalado proporcionan varias maneras de comprender y controlar el mantenimiento planeado:

- Las notificaciones de mantenimiento planeado le indican cuándo se debe realizar el mantenimiento y le permiten controlar cuándo se produce el mantenimiento.
- Las configuraciones de mantenimiento permiten programar una ventana de mantenimiento a la vez que se adapte a sus necesidades empresariales.
- Scheduled Events for Linux VMs and for Windows VMs (Eventos programados para máquinas virtuales Linux y para máquinas virtuales Windows) proporciona tiempo a la aplicación para prepararse para el mantenimiento de máquinas virtuales. Proporciona información sobre los próximos eventos de mantenimiento, como un reinicio, para que la aplicación pueda prepararse para ellos y limitar la interrupción.

Acuerdo de nivel de servicio

El contrato de nivel de servicio (SLA) para los servicios de Azure describe la disponibilidad esperada de cada servicio y las condiciones que la solución deberá cumplir para lograr esa expectativa de disponibilidad. Para obtener más información, consulte Acuerdos de Nivel de Servicio para servicios en línea.

Los conjuntos de escalado de máquinas virtuales comparten el SLA de disponibilidad para las máquinas virtuales. Puede lograr un porcentaje de tiempo de actividad mayor para las máquinas virtuales mediante un conjunto de escalado que cumpla ambos criterios:

- El conjunto de escalado contiene dos o más instancias.
- El conjunto de escalado distribuye esas instancias en dos o más zonas de disponibilidad.

Contenido relacionado

- ¿Qué son las zonas de disponibilidad?
- ¿Qué son los conjuntos de escalado de máquinas virtuales?

Last updated on 14/02/2026

Confiabilidad en Azure VMware Solution

Azure VMware Solution proporciona nubes privadas que contienen clústeres de VMware vSphere creados a partir de una infraestructura dedicada de Azure sin sistema operativo. Puede migrar cargas de trabajo desde los entornos locales, implementar nuevas máquinas virtuales (VM) y consumir servicios Azure desde las nubes privadas. Puede usar una combinación de funcionalidades nativas de VMware y Azure para habilitar la alta disponibilidad y la resistencia de las cargas de trabajo.

Al usar Azure, relibilidad es una responsabilidad compartida. Microsoft proporciona una variedad de capacidades para admitir resiliencia y recuperación. Es responsable de comprender cómo funcionan esas funcionalidades dentro de todos los servicios que usa y de seleccionar las funcionalidades que necesita para cumplir los objetivos empresariales y los objetivos de tiempo de actividad.

En este artículo se describe cómo hacer de Azure VMware Solution una solución resiliente a posibles interrupciones y problemas, incluidos errores transitorios, interrupciones de zona de disponibilidad e interrupciones de región. También se describe cómo puede usar copias de seguridad para recuperarse de otros tipos de problemas y se resalta cierta información clave sobre el contrato de nivel de servicio (SLA) de Azure VMware Solution.

Recomendaciones de implementación de producción

Azure VMware Solution implementaciones requieren una planeación cuidadosa en una variedad de áreas y a menudo requieren varios servicios de Azure. Para obtener más información, consulte Azure VMware Solution workloads en Azure Well-Architected Framework.

Introducción a la arquitectura de confiabilidad

Azure VMware Solution usa una infraestructura hiperconvergida (HCI) con clústeres de VMware vSphere.

Al implementar Azure VMware Solution, se implementa un *private cloud*, que tiene uno o varios clústeres. Cada clúster contiene hosts ESXi que proporcionan proceso, almacenamiento a través de SAN virtual (vSAN) y redes a través de VMware NSX. Hay dos generaciones de Azure VMware Solution:

- Gen 1 usa hardware bare-metal especializado para nodos y utiliza enfoques de redes dedicados. Para obtener más información sobre los conceptos clave, consulte Azure VMware Solution conceptos de nube privada y clúster.

- Gen 2 usa tipos de máquina virtual estándar Azure y redes virtuales Azure. Esta arquitectura simplifica la arquitectura de red, mejora las velocidades de transferencia de datos, reduce la latencia de las cargas de trabajo y mejora el rendimiento cuando accede a otros servicios de Azure.

Tolerancia a errores

Azure VMware Solution proporciona varios mecanismos para controlar los errores en el nivel de infraestructura y aplicación:

- *vSphere High Availability (HA):* vSphere HA supervisa los anfitriones ESXi y las máquinas virtuales. Si se produce un error en un host, reinicia automáticamente las máquinas virtuales afectadas en hosts correctos. vSphere HA está activado de forma predeterminada y reserva la capacidad de proceso y memoria para un único fallo de nodo.

- *Tolerancia a errores de vSAN:* las directivas de almacenamiento de vSAN protegen contra errores transitorios de nivel de almacenamiento manteniendo varias copias de datos entre hosts. Si una ruta de acceso de almacenamiento o un disco experimenta problemas transitorios, vSAN gestiona automáticamente el failover a rutas de acceso de almacenamiento disponibles.

- *Network redundancy:* Azure VMware Solution proporciona rutas de acceso de red redundantes y varios adaptadores de red VMkernel para controlar errores transitorios de nivel de red.

Resistencia a errores transitorios

Los errores transitorios son errores breves e intermitentes en los componentes. Se producen con frecuencia en un entorno distribuido como la nube y son una parte normal de las operaciones. Los errores transitorios se corrigen después de un breve período de tiempo. Es importante que las aplicaciones puedan controlar errores transitorios, normalmente mediante el reintento de solicitudes afectadas.

Todas las aplicaciones hospedadas en la nube deben seguir las instrucciones de control de errores transitorios Azure cuando se comunican con cualquier API, bases de datos y otros

componentes hospedados en la nube. Para obtener más información, consulte Recomendaciones para controlar errores transitorios.

En el caso de las aplicaciones que se ejecutan en máquinas virtuales de Azure VMware Solution, implemente procedimientos estándar para controlar errores transitorios:

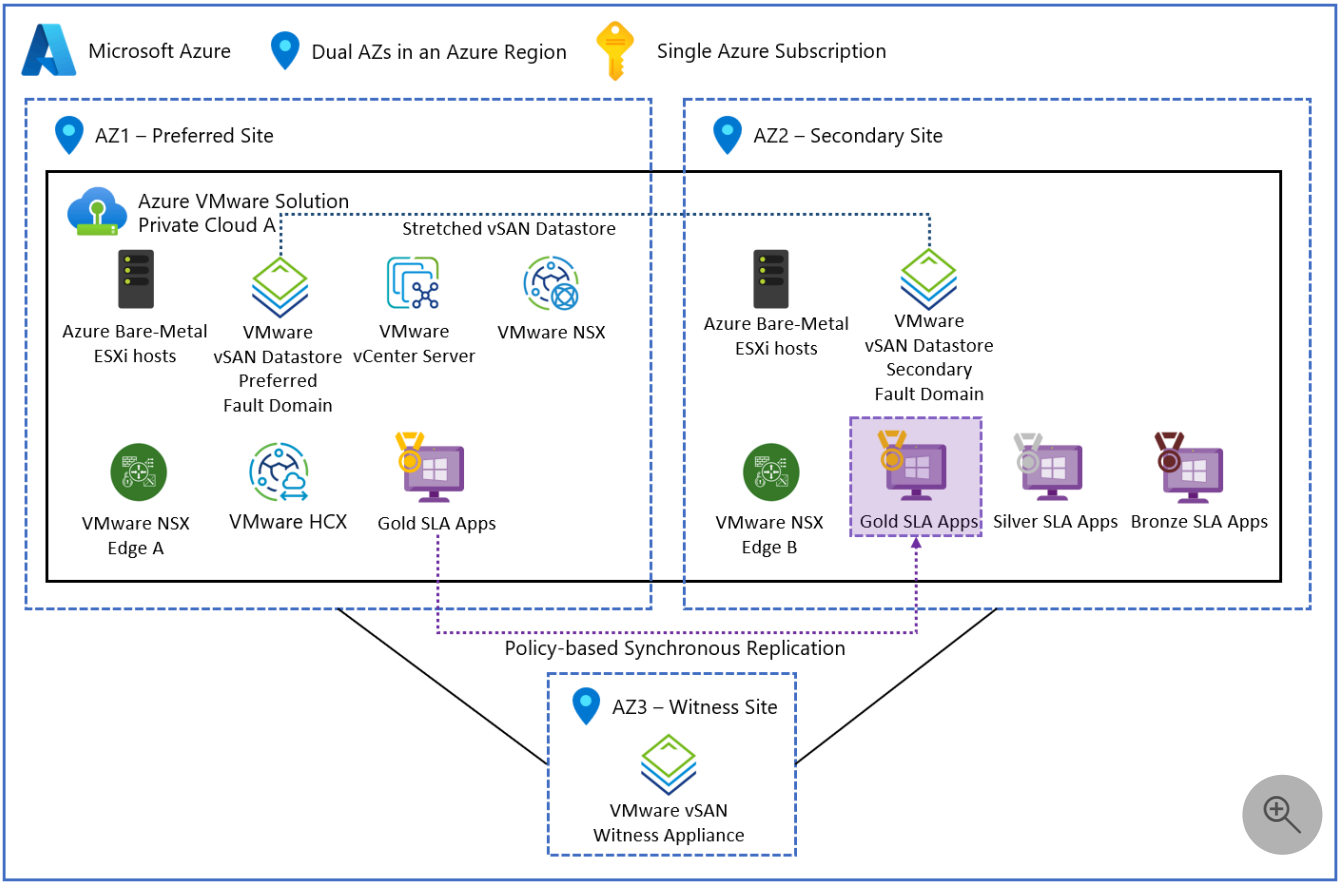
- Configura las directrices de reintento adecuadas con aplazamiento exponencial.
- Utilice patrones de cortocircuito para llamadas de servicio externas.
- Monitoree el estado de la aplicación e implemente una degradación progresiva.
- Diseñe aplicaciones sin estado siempre que sea posible para reducir el impacto de los reinicios de la máquina virtual.

Resistencia a errores de zona de disponibilidad

Availability zones son grupos de centros de datos físicamente independientes dentro de una región de Azure. Cuando una zona falla, los servicios pueden transferirse a una de las zonas restantes.

Azure VMware Solution Gen 1 admite zonas de disponibilidad a través de clústeres de *stretched*, que distribuyen hosts ESXi entre dos zonas de disponibilidad dentro de una región. Microsoft selecciona las zonas que se van a usar. El clúster se ejecuta en una configuración activa-activa en las dos zonas y vSAN también abarca varias zonas. Puede designar si cada carga de trabajo se implementa en una o dos zonas.

Un nodo testigo se implementa automáticamente en una tercera zona de disponibilidad para proporcionar cuórum para escenarios de cerebro dividido. Microsoft administra automáticamente el nodo testigo.



Un *clúster estándar* es un clúster que no se extiende entre zonas. En un clúster estándar, el clúster y todos sus hosts ESXi se consideran *no zonales* o *regionales*. Los clústeres no zonales se pueden colocar en cualquier zona de disponibilidad dentro de la región y Microsoft selecciona la zona. Si una zona de disponibilidad de la región experimenta una interrupción, los clústeres y hosts no zonales podrían estar en la zona afectada y podrían experimentar tiempo de inactividad.

Para obtener más información sobre la compatibilidad con zonas de disponibilidad para otras generaciones, seleccione la generación adecuada al principio de este artículo.

Requisitos

- **Soporte de regiones:** Los clústeres extendidos están disponibles solo en las regiones de Azure que admiten la configuración de clústeres extendidos. Consulte la tabla de asignación de tipos de host en la zona de disponibilidad de Azure para ver la compatibilidad de la región actual.

- **Hosts mínimos:** Implemente un mínimo de seis hosts en dos zonas de disponibilidad (tres hosts para cada zona) para habilitar la configuración de clúster extendido. Al escalar horizontalmente, hay que escalar en pares para que cada zona tenga un número igual de hosts.

- **SKU de host:** Los tipos de host AV36, AV36P, y AV52 admiten clústeres extendidos. La SKU de AV64 no admite clústeres extendidos.

Consideraciones

Cada zona de disponibilidad de una región puede admitir tipos de host específicos. Para obtener una lista detallada de los tipos de host disponibles en cada zona, consulte la tabla de asignación de tipo de host por zona de disponibilidad en la Región de Azure.

Cost

Se incurre en costos para cada nodo del clúster, independientemente de la configuración de la zona de disponibilidad del clúster. Para obtener información detallada sobre los precios, consulte Azure VMware Solution precios.

Configurar soporte de zonas de disponibilidad

- **Desplegar un nuevo clúster:** Cuando creas una nueva nube privada de Azure VMware Solution en una región compatible, puedes configurarla como un clúster extendido durante la implementación. Esta configuración distribuye los hosts entre dos zonas de disponibilidad automáticamente. Para más información, consulte Implementación de clústeres extendidos de vSAN.

- **Clústeres existentes:** No se puede convertir un clúster estándar en un clúster extendido y no se puede convertir un clúster extendido en un clúster estándar. En su lugar, debe implementar un nuevo clúster y migrar las cargas de trabajo.

Comportamiento cuando todas las zonas están en buen estado

En esta sección se describe qué esperar cuando el clúster está extendido y todas las zonas de disponibilidad están operativas.

- **Operación entre zonas:** Las máquinas virtuales se pueden ejecutar en hosts en cualquiera de las zonas de disponibilidad. Puede controlar la selección de ubicación de las máquinas virtuales mediante la afinidad de vSphere Distributed Resource Scheduler (DRS) y las reglas de antiafinidad para optimizar los requisitos de rendimiento o disponibilidad.

- **Replicación de datos entre zonas:** vSAN replica los datos de forma sincrónica entre zonas de disponibilidad. Ambas zonas confirman cada operación de escritura antes de que se

complete para garantizar la integridad de los datos de manera consistente.

Comportamiento durante un fallo de zona

En esta sección se describe qué esperar cuando se amplía el clúster y se produce una interrupción de zona de disponibilidad.

- **Detección y respuesta:** Azure VMware Solution gestiona la respuesta a nivel de infraestructura ante los fallos de zona. vSphere HA detecta automáticamente errores de zona e inicia los procedimientos de reinicio de la máquina virtual si es necesario.

- **Notification**: Microsoft no le notifica automáticamente cuando una zona está inactiva. Sin embargo, puede usar Azure Resource Health para supervisar el estado de un recurso individual y puede configurar Resource Health alertas para notificarle problemas. También puede usar Azure Service Health para comprender el estado general del servicio, incluidos los errores de zona, y puede configurar alertas de Service Health para notificarle problemas.

- **Solicitudes activas:** Cualquier máquina virtual que se ejecute en la zona de disponibilidad fallida se reinicia en hosts de la zona de disponibilidad sana. Las solicitudes activas y las conexiones a las máquinas virtuales afectadas finalizan y los clientes son responsables de reintentarlas.

- **Tiempo de inactividad esperado:** El tiempo de reinicio de las máquinas virtuales con errores en la zona correcta suele ser de unos minutos, en función de los procedimientos de inicio y configuración de la máquina virtual. El clúster extendido permanece operativo con una capacidad reducida.

  Si la zona de disponibilidad con errores contiene el nodo testigo, el testigo deja de ser accesible. Siempre que haya suficientes réplicas de datos disponibles, los hosts de datos y las cargas de trabajo en ejecución seguirán funcionando sin pérdida de datos inmediata. Sin embargo, vSAN pierde la conciencia del cuórum en este estado. La pérdida de cuórum impide que tome decisiones de colocación y recuperación de forma segura. También bloquea ciertas operaciones, como el encendido de la máquina virtual después de errores, el reequilibrio y las reparaciones.

- **Pérdida de datos esperada:** Dado que vSAN usa la replicación sincrónica entre zonas, no se espera ninguna pérdida de datos durante un error de zona.

- **Redistribución:** vSphere DRS redistribuye automáticamente las cargas de trabajo de máquina virtual a la zona de disponibilidad correcta. El enrutamiento del tráfico de red a través de VMware NSX se adapta automáticamente a la nueva ubicación de la máquina virtual.

Recuperación de zona

Cuando se recupera la zona de disponibilidad, vSphere DRS puede redistribuir las máquinas virtuales de nuevo a la zona recuperada en función de las reglas de afinidad y configuración de DRS. También puede controlar manualmente la ubicación de la máquina virtual mediante operaciones de vMotion.

Prueba de fallos de zona

Para prepararse para errores de zona, pruebe la resistencia de la aplicación a los reinicios de la máquina virtual y los cambios de ruta de acceso de red, especialmente cuando haya extendido clústeres o implemente aplicaciones en clústeres independientes en diferentes zonas.

Dado que Azure VMware Solution administra la respuesta de la infraestructura a los errores de zona, principalmente debe probar la respuesta de la aplicación a los reinicios de la máquina virtual.

Resistencia a errores en toda la región

Cada clúster de Azure VMware Solution se implementa dentro de una sola región de Azure. Si la región deja de estar disponible, la nube privada y todos los recursos dentro de ella no estarán disponibles.

Sin embargo, también puede diseñar soluciones personalizadas de varias regiones que combinen diferentes enfoques o se integren con la infraestructura existente para satisfacer sus requisitos empresariales y objetivos de recuperación específicos.

Soluciones personalizadas de varias regiones para la resistencia

Para lograr resistencia en varias regiones con Azure VMware Solution, debe implementar nubes privadas independientes en varias regiones e implementar la conmutación por error y otras soluciones de recuperación ante desastres (DR).

Una variedad de opciones admiten diferentes requisitos de resistencia. Para obtener más información, consulte Disaster recovery solutions for Azure VMware Solution virtual machines.

Copias de seguridad y restauración

Azure VMware Solution realiza automáticamente copias de seguridad de los componentes de administración, como vCenter Server, NSX Manager y HCX Manager si está habilitado. Para restaurar componentes de estas copias de seguridad de administración, cree una solicitud de Soporte técnico de Azure.

Para las cargas de trabajo de máquina virtual, Azure VMware Solution admite varios enfoques de copia de seguridad. Para obtener más información, consulte Backup solutions for Azure VMware Solution VMs.

Resistencia al mantenimiento del servicio

Azure realiza el mantenimiento automático de la plataforma para aplicar actualizaciones de seguridad, implementar nuevas características y mejorar la confiabilidad del servicio.

Para obtener información sobre cómo afecta el mantenimiento a los componentes de Azure VMware Solution y comprender los componentes que es responsable de mantener frente a los componentes que Microsoft mantiene, consulte Azure VMware Solution mantenimiento de la nube privada.

Puede configurar las ventanas de mantenimiento del clúster para reducir la probabilidad de que el mantenimiento afecte a las cargas de trabajo de producción. Para obtener más información, consulte Planificación del mantenimiento de autoservicio para Azure VMware Solution.

Acuerdo de nivel de servicio

El acuerdo de nivel de servicio (SLA) para Azure servicios describe la disponibilidad esperada de cada servicio y las condiciones que la solución debe cumplir para lograr esa expectativa de disponibilidad. Para obtener más información, vea SLAs for servicios en línea.

Azure VMware Solution proporciona diferentes acuerdos de nivel de servicio de disponibilidad para la infraestructura de cargas de trabajo y para las operaciones de administración.

Los clústeres configurados como clústeres distribuidos tienen un nivel de disponibilidad de infraestructura de carga de trabajo más alto en su Acuerdo de Nivel de Servicio.

Sin embargo, para calificar para los Acuerdos de Nivel de Servicio de disponibilidad, debe configurar el clúster de maneras específicas. Para obtener más información, consulte el texto del Acuerdo de Nivel de Servicio.

Contenido relacionado

- Fiabilidad en Azure
- Información general de Azure VMware Solution
- Implementación de clústeres extendidos de vSAN
- Implementación de recuperación ante desastres mediante VMware HCX
- Continuidad del negocio y recuperación ante desastres (BCDR) para Azure VMware Solution
- Cargas de trabajo de Azure VMware Solution

ⓘ **Nota:** El autor ha creado este artículo con ayuda de inteligencia artificial. Más información

---

Last updated on 09/04/2026

Confiabilidad en Azure App Configuration

Azure App Configuration almacena y administra centralmente las opciones de configuración de la aplicación y las marcas de características, que reemplazan los archivos de configuración incrustados directamente dentro de las aplicaciones. Con este enfoque, puede actualizar dinámicamente los valores de configuración, realizar un seguimiento del historial de versiones y mantener un registro de los cambios de configuración a lo largo del tiempo. La disponibilidad y confiabilidad de App Configuration son consideraciones importantes porque el comportamiento de la aplicación puede depender directamente del acceso a los datos de configuración en tiempo de ejecución.

Al usar Azure, relibilidad es una responsabilidad compartida. Microsoft proporciona una variedad de capacidades para admitir resiliencia y recuperación. Es responsable de comprender cómo funcionan esas funcionalidades dentro de todos los servicios que usa y de seleccionar las funcionalidades que necesita para cumplir los objetivos empresariales y los objetivos de tiempo de actividad.

En este artículo se describe la arquitectura de confiabilidad de App Configuration y cómo el servicio está diseñado para permanecer disponible durante errores transitorios, errores de zona de disponibilidad y interrupciones de región.

Recomendaciones de implementación de producción para la confiabilidad

Para la mayoría de las implementaciones de producción de App Configuration, tenga en cuenta las siguientes recomendaciones:

- ✓ **SKU:** Use la SKU Estándar o Premium.
- ✓ **Eliminación temporal y protección de purga:** Active la eliminación temporal y la protección de purga para ayudar a evitar la eliminación de datos.
- ✓ **Para escenarios críticos:** Use la SKU Premium y configure la réplica incluida para replicar en varias regiones. Este enfoque mejora la alta disponibilidad y la resistencia a las interrupciones de la región.

Para obtener una lista de los procedimientos recomendados y la configuración de las cargas de trabajo de producción, consulte Compilación de aplicaciones que tienen alta resistencia.

Introducción a la arquitectura de confiabilidad

Al implementar App Configuration, se implementa una *tienda*. Tu tienda contiene varios tipos de configuración que la aplicación puede usar, incluidas las claves y los valores, y las banderas de características. El servicio también incluye funcionalidades integradas para organizar, proteger, control de versiones y implementar de forma segura los cambios de configuración en todos los entornos. Para obtener más información, consulte ¿Qué es App Configuration?

App Configuration es un servicio totalmente administrado. Microsoft es responsable de realizar el mantenimiento en el servicio y almacenar y administrar la configuración.

Al crear aplicaciones cliente que se conectan a App Configuration, opcionalmente puede usar App Configuration con Azure Front Door (versión preliminar) para habilitar el caché y la aceleración global del tráfico. Esta configuración presenta otras consideraciones para la replicación geográfica, que se resaltan en este artículo cuando corresponda.

Resistencia a errores transitorios

Los errores transitorios son errores breves e intermitentes en los componentes. Se producen con frecuencia en un entorno distribuido como la nube y son una parte normal de las operaciones. Los errores transitorios se corrigen después de un breve período de tiempo. Es importante que las aplicaciones puedan controlar errores transitorios, normalmente mediante el reintento de solicitudes afectadas.

Todas las aplicaciones hospedadas en la nube deben seguir las instrucciones de control de errores transitorios Azure cuando se comunican con cualquier API, bases de datos y otros componentes hospedados en la nube. Para obtener más información, consulte Recomendaciones para controlar errores transitorios.

Al usar App Configuration, tenga en cuenta los siguientes procedimientos recomendados para minimizar el efecto de los errores transitorios en el acceso a la configuración, especialmente dentro de las rutas de acceso de código críticas.

- **Proveedores de configuración:** Use proveedores de App Configuration, que tienen integradas funcionalidades de reintento, almacenamiento en caché y otras capacidades de resiliencia.

- **SDK de Azure:** Utilice los SDK de App Configuration si su aplicación necesita enviar solicitudes de escritura. Los SDKs reintentan automáticamente en las respuestas de código de estado HTTP 429 y otros errores transitorios.

- **Lógica de reintento:** Incluya lógica de reintento en clientes personalizados si no puede utilizar proveedores o SDK de App Configuration. El `retry-after-ms` encabezado de la respuesta proporciona un tiempo de espera sugerido en milisegundos antes de que el cliente vuelva a intentar la solicitud.

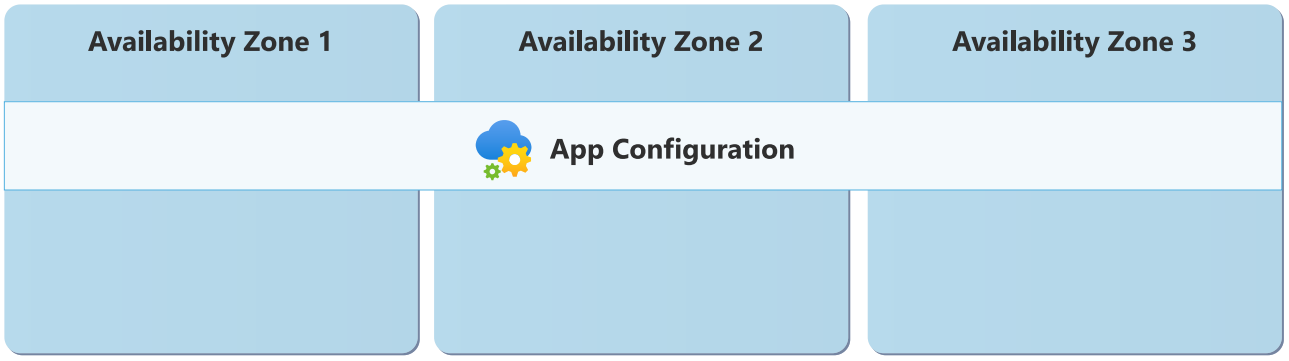
- **Almacenamiento en caché:** Configure los ajustes de caché en la memoria cuando sea posible para reducir las solicitudes directas a su tienda.

Para obtener otras instrucciones de configuración de aplicaciones, consulte Preguntas más frecuentes sobre App Configuration.

Resistencia a errores de zona de disponibilidad

Availability zones son grupos de centros de datos físicamente independientes dentro de una región de Azure. Cuando una zona falla, los servicios pueden transferirse a una de las zonas restantes.

App Configuration proporciona automáticamente redundancia de zona en regiones que admiten zonas de disponibilidad. Esta redundancia proporciona alta disponibilidad dentro de una región sin necesidad de ninguna configuración específica.



Cuando una zona de disponibilidad deja de estar disponible, App Configuration redirige automáticamente las solicitudes a otras zonas de disponibilidad correctas para garantizar una alta disponibilidad.

Requisitos

Compatibilidad con regiones: Los almacenes implementados en las siguientes regiones son automáticamente redundantes de zona.

Expandir tabla

Americas	Europa	Oriente Medio	África	Asia Pacífico
Sur de Brasil	Centro de Francia	Israel Central		Australia East
Canada Central	Centro-oeste de Alemania	Qatar Central		Centro de la India
Central US	Norte de Italia	Norte de Emiratos Árabes Unidos		Norte de China 3
East US	Norte de Europa			Este de Asia
Este de EE. UU. 2	Norway East			Japón Oriental
Centro de México	Centro de Polonia			Centro de Corea del Sur
Centro-sur de EE. UU.	Centro de España			Sudeste de Asia
Gobierno de EE. UU. - Virginia	Centro de Suecia			
Oeste de EE. UU. 2	Norte de Suiza			
Oeste de EE. UU. 3	Sur de Reino Unido			
	Oeste de Europa			

Cost

No hay ningún costo adicional para la redundancia de zona para App Configuration.

Configurar soporte de zonas de disponibilidad

Microsoft configura automáticamente la redundancia de zona para un almacén cuando se encuentra en a región que admite zonas de disponibilidad.

Si App Configuration agrega compatibilidad con zonas de disponibilidad a una región existente, no es necesario realizar ninguna acción para beneficiarse de la compatibilidad con la zona de disponibilidad. La tienda se beneficia de la compatibilidad con la zona de disponibilidad que está disponible para los almacenes de App Configuration en la región.

Comportamiento cuando todas las zonas están en buen estado

En esta sección describimos lo que puede esperar cuando tiene un almacén de App Configuration con redundancia de zona y todas las zonas están operativas.

- **Operación entre zonas:** App Configuration administra automáticamente el enrutamiento de tráfico entre zonas de disponibilidad. Durante las operaciones normales, distribuye de forma transparente las solicitudes entre zonas.
- **Replicación de datos entre zonas:** En las regiones que admiten zonas, App Configuration replica de forma sincrónica los datos entre zonas de disponibilidad. Esta replicación garantiza que la configuración siga siendo coherente y disponible incluso si una zona deja de estar disponible.

Comportamiento durante un fallo de zona

En esta sección se describe qué cabe esperar al usar un almacén de App Configuration con redundancia de zona y se produce una interrupción en una de las zonas.

- **Detección y respuesta:** El servicio App Configuration detecta errores de zona y responde automáticamente a ellos. No es necesario realizar ninguna acción ante un fallo de zona.
- **Notification:** Microsoft no le notifica automáticamente cuando una zona está inactiva. Sin embargo, puede usar Azure Service Health para comprender el estado general del servicio, incluidos los errores de zona, y puede configurar alertas de Service Health para notificarle problemas.
- **Solicitudes activas:** durante un error de zona, es posible que la zona afectada no controle las solicitudes en curso, lo que requiere que las aplicaciones cliente vuelvan a intentarlas. Las aplicaciones cliente deben seguir los procedimientos transitorios de control de errores para asegurarse de que pueden reintentar solicitudes si se produce un error de zona.
- **Pérdida de datos esperada:** No se espera ninguna pérdida de datos durante un error de zona debido a la replicación sincrónica entre zonas.
- **Tiempo de inactividad esperado:** No se espera ningún tiempo de inactividad.
- **Redistribución:** App Configuration vuelve a enrutar automáticamente el tráfico de la zona afectada a zonas correctas sin necesidad de intervención del cliente.

Recuperación de zona

Cuando se recupera una zona no disponible anteriormente, App Configuration restaura automáticamente las operaciones normales en todas las zonas de disponibilidad. No es necesario realizar ninguna acción para recuperar de una falla de zona.

Prueba de fallos de zona

La plataforma App Configuration gestiona el enrutamiento del tráfico, la conmutación ante fallos y la recuperación de zona para almacenes con redundancia de zona. Microsoft administra completamente este proceso, por lo que no es necesario validar los procesos de error de zona de disponibilidad.

Resistencia a errores en toda la región

App Configuration proporciona capacidades nativas de geo-replicación para admitir la resiliencia durante interrupciones regionales. La replicación geográfica permite que los datos de configuración se repliquen entre regiones como una característica de servicio administrado.

Replicación geográfica

Con la replicación geográfica, puede replicar un almacén en varias regiones de Azure. Cada almacén puede tener varias *réplicas* en regiones diferentes. El almacén original también es una réplica. Esta funcionalidad ayuda a proteger las aplicaciones frente a interrupciones en toda la región.

Requisitos

- **Region support:** Puede crear réplicas en cualquier región de Azure compatible con App Configuration, incluso si las regiones no son regiones emparejadas de Azure.
- **Nivel:** El almacén de configuración debe usar un nivel admitido para habilitar la replicación geográfica. Para obtener más información, consulte Habilitación de la replicación geográfica.

Consideraciones

Al habilitar la replicación geográfica, tenga en cuenta los siguientes factores:

- **Réplicas redundantes por zona:** Cualquier réplica creada en una región donde App Configuration admita zonas de disponibilidad será automáticamente redundante por zona.
- **Azure Front Door:** Para habilitar la entrega de configuración con redundancia geográfica con Azure Front Door, configure réplicas de App Configuration como orígenes dentro de un grupo de origen. Los orígenes configurados correctamente son necesarios para que Azure Front Door realice el enrutamiento basado en la salud, el equilibrio de carga y la

conmutación automática por error entre regiones. Para obtener más información, consulte Métodos de enrutamiento del tráfico al origen.

Cost

Cada región con replicación geográfica se factura por separado según los precios del plan y la región correspondientes. No se aplican cargos de salida de datos para la replicación entre regiones. Para obtener más información sobre los precios, consulte Precios de App Configuration.

Configuración de la compatibilidad con varias regiones

Para configurar la replicación para un almacén de configuración recién creado, consulte Habilitación de la replicación geográfica.

Comportamiento cuando todas las regiones están en buen estado

En esta sección se describe qué esperar al configurar un almacén de App Configuration para la replicación geográfica y todas las regiones están operativas.

- **Operación entre regiones:** Cada réplica es direccionable individualmente y tiene su propio nombre del sistema de nombres de dominio (DNS). Todas las réplicas pueden aceptar operaciones de lectura y escritura.

  App Configuration no enruta automáticamente el tráfico entre regiones. Cuando se usan proveedores de configuración de App Configuration, la aplicación puede usar opcionalmente la detección automática de réplicas. Alternativamente, puede especificar una lista priorizada de réplicas, y App Configuration selecciona la primera réplica saludable. Este enfoque permite a la aplicación controlar qué réplica usa.

  > **Nota:**
  >
  > Si usa Azure Front Door, el comportamiento del enrutamiento del tráfico es diferente. Para obtener más información, consulte Conmutación por error y equilibrio de carga.

- **Replicación de datos entre regiones:** Los datos se replican de forma asincrónica y finalmente son coherentes. Puede usar la métrica de latencia de replicación en Azure Monitor para supervisar la latencia de replicación actual entre réplicas.

Comportamiento durante una falla de región

En esta sección se describe qué se puede esperar al configurar una tienda de configuración de aplicaciones para la replicación geográfica, y cuando se produce una interrupción en una de las regiones de réplica.

- **Detección y respuesta:** Microsoft es responsable de detectar errores de región o réplica e iniciar procesos de recuperación.

  Cuando se usan proveedores de configuración de App Configuration con detección automática de réplicas o una lista de varias réplicas, la aplicación detecta automáticamente errores y realiza una conmutación en caso de error a una réplica sana.

  Si no usa proveedores de App Configuration, es responsable de cambiar su aplicación a una réplica saludable.

- **Notification:** Microsoft no le notifica automáticamente cuando una región está inactiva. Sin embargo, puede usar Azure Service Health para comprender el estado general del servicio, incluidos los errores de región, y puede configurar alertas de Service Health para notificarle problemas.

- **Solicitudes activas:** Las solicitudes activas contra una réplica en la región pueden fallar. Las aplicaciones cliente deben reintentar las solicitudes en una réplica diferente.

- **Pérdida de datos esperada:** Si se produce un error en una réplica, es posible que los cambios recientes realizados en esa réplica aún no se repliquen en otras réplicas. Esos cambios pueden permanecer no disponibles hasta que se recupere la réplica. Para calcular la posible pérdida de datos, supervise la métrica de latencia de replicación en Azure Monitor.

- **Tiempo de inactividad esperado:** Cuando una réplica deja de estar disponible, permanece sin conexión hasta que se recupera su región. Otras réplicas siguen gestionando las solicitudes. Las aplicaciones pueden experimentar un breve tiempo de inactividad mientras detectan el error y cambian a una réplica correcta. La duración depende de la rapidez con la que cada aplicación realiza esta detección y conmutación por error.

- **Redistribución:** Las aplicaciones deben enrutar el tráfico a una réplica saludable cuando se produce un error.

  Si utiliza proveedores de configuración de App Configuration, estos gestionan automáticamente la selección de réplica y la conmutación por error.

Si coloca Azure Front Door delante del almacenamiento de datos y configura el grupo de origen con varias réplicas como orígenes para la conmutación por error, Azure Front Door vuelve a enrutar automáticamente las solicitudes a una réplica en buen estado.

Recuperación de regiones

Una vez que la región se recupera, App Configuration sincroniza la réplica con las demás réplicas sin intervención del usuario.

Usted es responsable de volver a configurar su aplicación para enrutar el tráfico de regreso a la instancia de la región recuperada. Las aplicaciones que utilizan App Configuration reanudan automáticamente el uso de la réplica.

Prueba de fallos de región

No se puede simular una conmutación por error de réplica directamente en App Configuration. Sin embargo, dado que las aplicaciones controlan la selección de réplica, puede probar el comportamiento de la conmutación por error forzando la aplicación a un estado en el que debe cambiar las réplicas.

Para validar el comportamiento de failover de réplica de su aplicación, puede introducir un error de conectividad controlado en un entorno de no producción y observar la respuesta de la aplicación.

Un enfoque consiste en usar la máquina local u otro entorno en el que tenga acceso administrativo. Siga estos pasos:

1. Active el registro exhaustivo para el SDK de Azure. En .NET, use la clase `AzureEventSourceListener` para configurar un registrador. Para obtener más información, vea el documento sobre supervisión y registro.

2. Configure manualmente el `hosts` archivo para que las solicitudes al almacén de App Configuration se enruten a una dirección IP que no pueda recibirlos, como `127.0.0.1` (localhost).

   > ⚠ **Advertencia**
   >
   > Este paso bloquea eficazmente el acceso desde tu equipo a tu almacén de Configuración de la Aplicación. Siga estos pasos solo en un entorno que no sea de producción.

3. Supervise los registros de un mensaje similar al ejemplo siguiente:

   Consola

   ```
   [Warning] Microsoft-Extensions-Configuration-AzureAppConfiguration-Refresh:
   Failed to get configuration settings from endpoint
   'https://myappconfigstore.azconfig.io'. Failing over to endpoint
   https://myappconfigstore-eus.azconfig.io'.
   ```

   Este mensaje indica que la aplicación conmutó con éxito para usar otra réplica de su almacén.

4. Después de completar la prueba, deshaga los cambios en el archivo `hosts`.

Copias de seguridad y restauración

Puede usar App Configuration para exportar datos de configuración desde un almacén y usarlos como parte de una estrategia de copia de seguridad más amplia.

Para la mayoría de las soluciones, no debe confiar exclusivamente en copias de seguridad. En su lugar, utilice las otras capacidades descritas en esta guía para apoyar los requisitos de resiliencia. Sin embargo, las copias de seguridad protegen contra algunos riesgos que otros enfoques no. Para más información, consulte ¿Qué son la redundancia, la replicación y la copia de seguridad?.

Resistencia a la eliminación accidental

App Configuration proporciona dos características clave de recuperación para evitar la eliminación accidental o malintencionada:

- **Eliminación temporal:** Al activar la eliminación temporal, puede recuperar almacenes y configuraciones eliminados durante un período de retención configurable. Las funciones de borrado suave actúan como una papelera de reciclaje para los recursos de App Configuration.

- **Protección de purga:** Al activar la protección de purga, el servicio evita la eliminación permanente del almacén y su configuración hasta que transcurre el período de retención. Esta protección evita que los actores malintencionados destruyan permanentemente la configuración.

Use ambas características para entornos de producción. Para más información, consulte Eliminación temporal y protección contra purgas.

Resistencia al mantenimiento del servicio

Microsoft realiza periódicamente actualizaciones de servicio y otro mantenimiento. El servicio controla estas actividades automáticamente, lo que hace que el mantenimiento sea transparente y sin problemas. No se espera ningún tiempo de inactividad durante los eventos de mantenimiento a menos que Azure Service Health proporcione un aviso de mantenimiento planeado.

Resistencia a problemas de configuración

Los cambios de configuración incorrectos o accidentales pueden provocar tiempo de inactividad de la aplicación. Utiliza instantáneas de configuración para desplegar los cambios en la configuración de forma segura. Supervise el estado de la aplicación después de cualquier cambio de configuración y vuelva a la última instantánea de configuración correcta conocida si los cambios presentan un problema.

Acuerdo de nivel de servicio

El acuerdo de nivel de servicio (SLA) para Azure servicios describe la disponibilidad esperada de cada servicio y las condiciones que la solución debe cumplir para lograr esa expectativa de disponibilidad. Para obtener más información, vea SLAs for servicios en línea.

Contenido relacionado

- Confiabilidad en Azure
- Documentación de Configuración de la aplicación

ⓘ**Nota:** El autor ha creado este artículo con ayuda de inteligencia artificial. Más información

---

Last updated on 15/04/2026

Confiabilidad en Azure Container Apps

Azure Container Apps es un servicio de hospedaje de contenedores sin servidor totalmente administrado para implementar microservicios y aplicaciones en contenedor.

Cuando se usa Azure, la confiabilidad es una responsabilidad compartida. Microsoft proporciona una variedad de funcionalidades para admitir resistencia y recuperación. Es responsable de comprender cómo funcionan esas funcionalidades dentro de todos los servicios que usa y de seleccionar las funcionalidades que necesita para cumplir los objetivos empresariales y los objetivos de tiempo de actividad.

En este artículo se describe cómo hacer que Las aplicaciones de contenedor sean resistentes a varias posibles interrupciones y problemas, incluidos errores transitorios, interrupciones de zona de disponibilidad, interrupciones de regiones y mantenimiento del servicio. También se describe cómo usar copias de seguridad para recuperarse de otros tipos de problemas y se resalta información clave sobre el acuerdo de nivel de servicio (SLA) de Container Apps.

Recomendaciones de implementación de producción

Para obtener información sobre cómo implementar Container Apps para admitir los requisitos de confiabilidad de la solución y cómo afecta la confiabilidad a otros aspectos de la arquitectura, consulte Procedimientos recomendados de arquitectura para Aplicaciones de contenedor en Azure Well-Architected Framework.

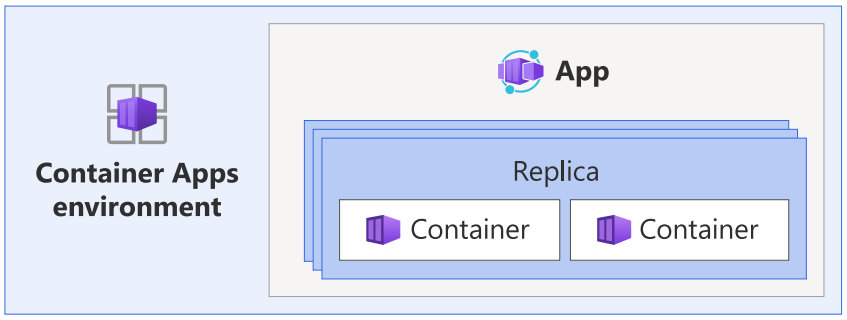
Introducción a la arquitectura de confiabilidad

Al usar Container Apps, implementará un *entorno* que actúa como unidad de implementación básica y define un límite seguro alrededor de un grupo de aplicaciones contenedoras. El entorno es donde se configuran los valores principales, incluida la compatibilidad de la zona de disponibilidad y la configuración de red. Los dos tipos de entornos son entornos de perfiles de carga de trabajo y entornos de solo consumo. Para obtener más información, consulte Estructuras de proceso y facturación en Container Apps.

Puede implementar varias *aplicaciones* en un único entorno. Cada aplicación ejecuta uno o varios *contenedores*. Un entorno también puede ejecutar uno o varios *trabajos*, que representan tareas no interactivas. Para obtener más información, consulte Contenedores en Aplicaciones de contenedor y trabajos en Aplicaciones de contenedor.

Cada aplicación tiene una o varias *réplicas*, que representan las instancias en ejecución de la aplicación. Puedes controlar cómo se escala la aplicación, incluido el número mínimo y máximo

de réplicas y cómo la aplicación agrega y quita réplicas dinámicamente. El planificador de la plataforma garantiza una distribución óptima entre hosts físicos al mismo tiempo que cumple con los requisitos mínimos de cantidad de réplicas. Para obtener más información, consulte Establecimiento de reglas de escalado en Container Apps.



Container Apps apoya la fiabilidad de las aplicaciones mediante diferentes capacidades.

- **Supervisión automática del estado:** El controlador de ingreso integrado equilibra automáticamente la carga del tráfico entre réplicas saludables. Si se produce un error en las comprobaciones de estado de una réplica o su infraestructura subyacente deja de estar disponible durante un tiempo prolongado, el servicio reinicia automáticamente los contenedores con errores o crea réplicas de reemplazo. También redistribuye el tráfico fuera de las réplicas no saludables y administra los reintentos de red en el clúster. Este proceso de recuperación automática no requiere ninguna intervención del cliente y mantiene el recuento de réplicas especificado. Para más información, consulte Sondeos de estado.

- **Resistencia de la aplicación a través de Dapr:** Container Apps proporciona una estrecha integración con Dapr, que es un marco que admite microservicios de nivel de producción y aplicaciones en contenedor. Dapr incluye características que ayudan a mejorar la resistencia, incluido el control de errores en otros servicios. Para obtener más información, consulte Microservicios con Aplicaciones de Contenedores.

- **Resistencia de la infraestructura para los componentes del sistema:** Esta resistencia incluye el plano de control, los controladores de entrada y el entorno de ejecución del contenedor. En las regiones que tienen zonas de disponibilidad, Container Apps proporciona redundancia de zona. Para obtener más información, consulte Resilience to availability zone failures (Resistencia a errores de zona de disponibilidad).

Resistencia a errores transitorios

Los errores transitorios son errores breves e intermitentes en los componentes. Se producen con frecuencia en un entorno distribuido como la nube y son una parte normal de las operaciones. Los errores transitorios se corrigen después de un breve período de tiempo. Es

importante que las aplicaciones puedan controlar errores transitorios, normalmente mediante el reintento de solicitudes afectadas.

Todas las aplicaciones hospedadas en la nube deben seguir las instrucciones de control de errores transitorios de Azure cuando se comunican con cualquier API, bases de datos y otros componentes hospedados en la nube. Para obtener más información, consulte Recomendaciones para controlar errores transitorios.

Container Apps gestiona automáticamente muchos errores transitorios a través de sus mecanismos de reintento a nivel de plataforma y supervisión de salud. Para asegurarse de que las aplicaciones son resistentes a errores transitorios, realice las siguientes acciones:

- **Configure sondeos de estado** que permitan a la plataforma detectar y responder a condiciones de error específicas de la aplicación. Establezca los umbrales de error y los valores de tiempo de espera adecuados en función de las características de inicio de la aplicación. Por ejemplo, para evitar reinicios prematuros del contenedor durante problemas temporales, use un umbral de fallos de 3 con un período de 10 segundos para los probes de vivacidad. Para más información, consulte Sondeos de estado.

- **Utilice directivas de resiliencia para el descubrimiento de servicios (versión preliminar)** para prevenir, detectar y recuperar proactivamente de los fallos en las solicitudes de servicio. Por ejemplo, cuando se usa una directiva de resistencia, cada solicitud entrante a la aplicación se puede reintentar automáticamente si hay un error transitorio que impide que la aplicación responda. Para más información, consulte Resistencia de detección de servicios (versión preliminar).

- **Implemente la lógica de reintento** en las aplicaciones para llamadas de servicio externas, conexiones de base de datos y solicitudes de API.

  Si la aplicación usa Dapr para integrarse con los servicios en la nube, use la resistencia de componentes de Dapr (versión preliminar) para configurar reintentos, tiempos de espera y disyuntores.

  Para otras dependencias, la aplicación debe controlar los errores transitorios. Use estrategias de retroceso exponencial y patrones de cortacircuitos al llamar a servicios externos para evitar fallos en cascada durante interrupciones en servicios posteriores. Las características integradas de detección y equilibrio de carga de Container Apps enrutan automáticamente el tráfico de las instancias con errores, pero las políticas de reintento a nivel de aplicación garantizan una gestión eficaz de manera elegante de los errores transitorios antes de que las comprobaciones de estado a nivel de plataforma desencadenen el reinicio del contenedor.

- **Diseñe trabajos para ser resistentes a errores transitorios**, incluidos los errores durante la ejecución del trabajo o en sus dependencias. Diseñe los trabajos para reanudarlos si se

reinician, o diseñe para la idempotencia de modo que se puedan ejecutar nuevamente de forma segura.

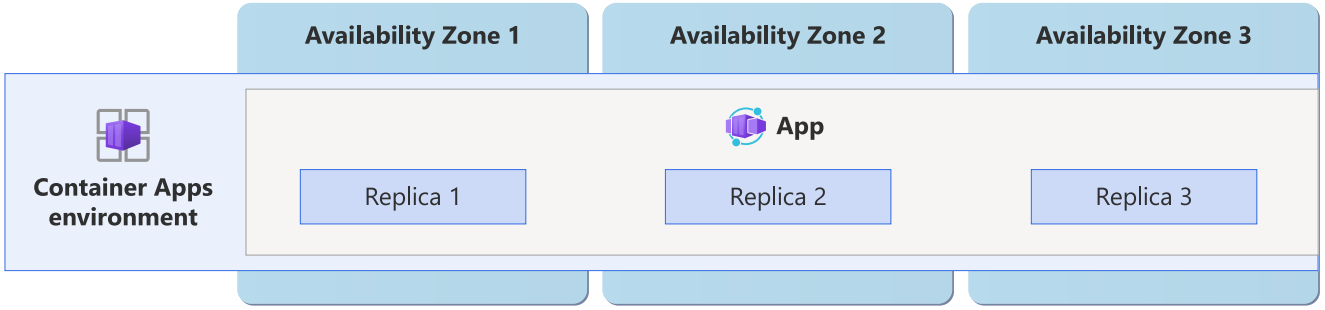
Resistencia a errores de zona de disponibilidad

Las zonas de disponibilidad son grupos físicamente independientes de centros de datos dentro de una región de Azure. Cuando una zona falla, los servicios pueden transferirse a una de las zonas restantes.

Al crear un entorno de Container Apps, puede habilitar la *redundancia de zona* para distribuir la infraestructura subyacente en varias zonas de disponibilidad en la región de Azure elegida. Container Apps programa automáticamente las réplicas de las aplicaciones entre zonas. Esta distribución se produce de forma transparente, lo que significa que no es necesario especificar la ubicación de zona para las réplicas individuales.

La redundancia de zona mejora la resistencia de la aplicación a errores de nivel de zona al garantizar que las réplicas de la aplicación contenedor se reparten entre varias zonas.

En el diagrama siguiente se muestra una aplicación contenedora con redundancia de zona con tres réplicas. Cada réplica se ejecuta en una zona de disponibilidad independiente.



Requisitos

- **Compruebe la compatibilidad con regiones.** La redundancia de zona está disponible en todas las regiones que admiten aplicaciones de contenedores y zonas de disponibilidad.

  Para ver qué regiones admiten zonas de disponibilidad, consulte Regiones de Azure con compatibilidad con zonas de disponibilidad.

  Para ver qué regiones admiten Container Apps, consulte Disponibilidad del producto por región.

- **Use perfiles de carga de trabajo.** La redundancia de zona está disponible para todos los planes de Aplicaciones de Contenedor, incluidos los perfiles de carga de trabajo de consumo y dedicado.

- **Habilite la redundancia de zona durante la creación del entorno.** Esta configuración no se puede cambiar después de crear el entorno.
- **Implemente un entorno de Container Apps en una red virtual.** La red virtual debe estar en una región que admita zonas de disponibilidad. Asegúrese de que la red virtual tiene una subred de tamaño adecuado. Los entornos de solo consumo necesitan una subred con un `/23` intervalo de enrutamiento de Inter-Domain sin clases (CIDR) o mayor, mientras que los entornos de perfiles de carga de trabajo requieren un `/27` intervalo CIDR o superior.
- **Establezca el número mínimo de réplicas en al menos dos para garantizar la distribución entre varias zonas de disponibilidad.** Considere la posibilidad de establecer un número mínimo de réplicas mayor si se aplica alguna de las condiciones siguientes:
  - La carga máxima esperada necesita más de dos réplicas.
  - Debe ser resistente a varias interrupciones de zona simultáneas.
  - Quieres minimizar el tiempo que esperas a que se creen nuevas réplicas en otras zonas durante una interrupción de zona.

Cost

No incurre en cargos adicionales más allá de los precios estándar de Container Apps al habilitar la redundancia de zona. Se pagan las mismas tarifas por los recursos de computación, las solicitudes y los segundos de vCore, independientemente de si la redundancia de zona está habilitada o no. Para más información, consulte Precios de Container Apps y facturación de Container Apps.

Configurar soporte de zonas de disponibilidad

- **Cree un entorno de Container Apps con zonas redundantes.** Para obtener instrucciones de implementación que abarcan Azure Portal, la CLI de Azure y Azure PowerShell, consulte Creación de una aplicación de contenedor con redundancia de zona.
- **Migre a una implementación redundante por zonas.** No se puede habilitar la redundancia de zona en un entorno de Container Apps existente. Para actualizar los entornos existentes que no tienen redundancia de zona, cree un nuevo entorno con redundancia de zona habilitada en una región admitida. A continuación, vuelva a implementar las aplicaciones de contenedor.
- **Desactivar la redundancia de zona.** La redundancia de zona no se puede deshabilitar después de habilitarla durante la creación del entorno. Si necesita una implementación

sin redundancia de zona, debe crear un nuevo entorno sin habilitar la opción de redundancia de zona ni implementarla en una región que no admita zonas de disponibilidad.

- **Compruebe la redundancia de zona.** Puede usar Azure Portal, la CLI de Azure y Azure PowerShell para comprobar el estado de redundancia de zona del entorno.

Planeamiento y administración de capacidad

Si una zona de disponibilidad deja de estar disponible, la plataforma Container Apps usa las reglas de escalado para decidir cuándo reemplazar las réplicas perdidas de esa zona. Es importante configurar correctamente las reglas de escalado para que el programador pueda tomar las decisiones de programación adecuadas.

Para configurar correctamente las reglas de escalado, siga estos principios:

- **Establezca un número mínimo de réplicas** que la aplicación pueda tolerar. Puede tardar un breve período de tiempo en reemplazar las réplicas perdidas porque la plataforma debe detectar que las réplicas antiguas han desaparecido. A continuación, las nuevas réplicas deben iniciarse y devolver un estado de sondeo de preparación correcto para poder recibir solicitudes entrantes. Si no puede tolerar ningún período con menos del número mínimo de réplicas que ha especificado, considere la posibilidad de sobreaprovisionar para mantener el rendimiento de la aplicación incluso si una zona deja de estar disponible.

- **Establezca las solicitudes de recursos y los límites** para ayudar al programador de Container Apps a tomar decisiones de ubicación óptimas entre zonas. Los requisitos de recursos infraespecificados pueden provocar errores de distribución o ubicación desiguales durante una carga alta.

Para obtener más información sobre las opciones de configuración, consulte Establecimiento de reglas de escalado.

Comportamiento cuando todas las zonas están en buen estado

En esta sección se describe qué esperar cuando los recursos de Container Apps están configurados para la redundancia de zona y todas las zonas de disponibilidad están operativas.

- **Enrutamiento de tráfico entre zonas:** Con Container Apps con redundancia de zonas, la plataforma funciona en un modelo activo-activo donde varias réplicas sirven tráfico simultáneamente. El controlador de entrada distribuye las solicitudes entrantes entre

todas las réplicas funcionales, independientemente de su zona, y utiliza el equilibrio de carga 'round robin' de forma predeterminada. Cada zona procesa solicitudes de forma independiente y la plataforma no da prioridad a ninguna zona específica para la distribución del tráfico. Los sondeos de salud se originan en todas las zonas para garantizar una evaluación precisa de la salud de cada réplica desde varias perspectivas.

- **Replicación de datos entre zonas:** Container Apps no replica los datos de la aplicación entre zonas porque está diseñado para cargas de trabajo sin estado. Los datos que la aplicación almacena en almacenamiento efímero, incluido el almacenamiento con ámbito de contenedor y el almacenamiento con ámbito de réplica, se eliminan cuando se cierra el contenedor o la réplica.

  Para los requisitos de datos con estado, monte un recurso compartido de archivos de Azure Files configurado para el almacenamiento con redundancia de zona o use otros servicios de Azure, como Azure Cosmos DB o Azure SQL Database, que proporcionan sus propias funcionalidades de replicación entre zonas.

  La plataforma solo replica los metadatos del plano de control, incluidas las configuraciones de la aplicación, las reglas de escalado y los secretos entre zonas para lograr una alta disponibilidad. Las imágenes de contenedor se extraen del registro de contenedor en cada zona según sea necesario cuando se crean réplicas.

Comportamiento durante un fallo de zona

En esta sección se describe qué esperar cuando los recursos de Container Apps están configurados para redundancia de zona y hay una interrupción en una zona de disponibilidad.

- **Detección y respuesta:** Azure detecta automáticamente errores de zona. Container Apps detiene inmediatamente la programación de nuevas réplicas en la zona fallida y comienza a redistribuir el tráfico a réplicas saludables en las zonas restantes. La plataforma gestiona todas las operaciones de conmutación por error automáticamente, sin necesidad de tu intervención.

- **Notificación:** Microsoft no le notifica automáticamente cuando una zona está inactiva. Sin embargo, puede usar Azure Service Health para comprender el estado general del servicio, incluidos los errores de zona, y puede configurar alertas de Service Health para notificarle problemas.

  También puede supervisar el estado de las aplicaciones a través de las métricas de Container Apps en Azure Monitor. Configure alertas en las caídas del recuento de réplicas y las tasas de error de solicitudes para recibir notificaciones inmediatas cuando se produzcan problemas relacionados con la zona.

- **Solicitudes activas:** Es posible que las solicitudes en vuelo a las réplicas de la zona fallida puedan eliminarse o experimenten tiempos de espera o errores de conexión. Las ejecuciones de trabajos en la zona afectada se anulan y se marcan como fallidas.

- **Pérdida de datos esperada:** No se produce ninguna pérdida de datos en el nivel de plataforma container Apps porque el servicio está diseñado para cargas de trabajo sin estado. Los datos almacenados en el almacenamiento efímero dentro de la zona de disponibilidad se pierden cuando finaliza la réplica y el almacenamiento efímero solo se debe usar para los datos temporales.

- **Tiempo de inactividad esperado:** Las aplicaciones experimentan un tiempo de inactividad mínimo o sin tiempo de inactividad durante los errores de zona. El impacto real depende de la configuración del sondeo de estado de la aplicación y del número de réplicas en zonas correctas. Asegúrese de que los clientes sigan las instrucciones de control de errores transitorios para minimizar los efectos.

  Los trabajos que se ejecutan en la zona afectada son abortados y se marcan como fallidos. Si necesita que un trabajo sea resistente a un error de zona, configure reintentos o configure paralelismo para que el trabajo ejecute varias copias de la misma ejecución. Para obtener más información, consulte Configuración avanzada del trabajo.

- **Reenrutamiento del tráfico:** Las sondas de estado del controlador de ingreso detectan rápidamente las réplicas inaccesibles y las eliminan del pool de equilibrio de carga. En función de la configuración del sondeo de estado de la aplicación, este proceso de conmutación por error suele producirse en unos 30 segundos. El tráfico entrante subsiguiente se distribuye entre las réplicas saludables restantes. Esta redirección de tráfico se produce de forma transparente a los clientes, que siguen usando la misma dirección URL de la aplicación.

  Si la afinidad de sesión está habilitada y una zona deja de funcionar, los clientes que se enrutaron previamente a réplicas de esa zona se enrutan a nuevas réplicas porque las réplicas anteriores ya no están disponibles. Se pierde cualquier estado asociado a las réplicas anteriores.

  No se iniciarán nuevas instancias de trabajos en la zona defectuosa.

- **Administración de instancias:** Es posible que se creen nuevas instancias de réplica en zonas saludables si las reglas de escalado automático se activan debido a un aumento en la carga.

Recuperación de zona

Cuando una zona de disponibilidad se recupera de un error, Container Apps vuelve a reinsertar automáticamente la zona en un servicio activo sin necesidad de intervención. Los sondeos de estado de la plataforma detectan cuándo la infraestructura de la zona recuperada está disponible y Container Apps comienza a programar nuevas réplicas en esa zona en función de la configuración de escalado. Las réplicas existentes en zonas saludables siguen atendiendo el tráfico durante el proceso de reintegración, lo que ayuda a evitar la interrupción del servicio.

Container Apps equilibra gradualmente la distribución de réplicas entre todas las zonas disponibles como parte de las operaciones de escalado estandarizados. Este reajuste automático ocurre cuando se crean réplicas en respuesta a eventos de escalado o cuando se reemplazan las réplicas defectuosas. La plataforma no fuerza la redistribución inmediata de réplicas correctas existentes, lo que evita reinicios innecesarios del contenedor y mantiene la estabilidad de la aplicación durante la recuperación.

Prueba de fallos de zona

La plataforma Container Apps administra el enrutamiento del tráfico, la conmutación por error y la conmutación por recuperación para aplicaciones de contenedor con redundancia de zona. Esta característica está totalmente administrada, por lo que no es necesario iniciar ni validar los procesos de error de zona de disponibilidad.

Para validar la resistencia de la aplicación a errores de zona, simule interrupciones de nivel de zona en el nivel de aplicación mediante enfoques de prueba controlados. Detenga o quite réplicas de zonas específicas mediante el escalado descendente de su aplicación y supervise cómo las réplicas restantes manejan la carga aumentada. Supervise las métricas clave durante las pruebas de resistencia, incluido el recuento de réplicas, las tasas de éxito de las solicitudes, los tiempos de respuesta y el comportamiento de escalado automático. Asegúrese de que el número mínimo de réplicas mantiene la disponibilidad del servicio cuando se eliminan las réplicas y compruebe que las reglas de escalado pueden manejar el aumento de carga en las réplicas restantes. Pruebe las configuraciones de sondeo de estado de salud fallando deliberadamente los puntos de conexión de estado para confirmar que la plataforma quita las instancias no saludables del ciclo de rotación durante los períodos de tiempo esperados.

Resistencia a errores en toda la región

Container Apps es un servicio de una sola región. Si la región deja de estar disponible, el entorno y las aplicaciones tampoco están disponibles.

Soluciones personalizadas de varias regiones para la resistencia

Para reducir el riesgo de un error de una sola región que afecte a la aplicación, puede implementar entornos en varias regiones. Los pasos siguientes ayudan a reforzar la resistencia:

- Implemente las aplicaciones en entornos de cada región. Cada entorno requiere su propia configuración de red virtual y los requisitos de subred se aplican de forma independiente a cada implementación regional. Las imágenes de contenedor deben estar disponibles en todas las regiones, lo que puede lograr mediante Azure Container Registry con la replicación geográfica habilitada.
- Configure las directivas de equilibrio de carga y conmutación por error usando servicios como Azure Front Door o Azure Traffic Manager.
- Replique los datos entre regiones para que pueda recuperar el último estado de la aplicación.

Copias de seguridad y restauración

Container Apps no proporciona funcionalidades de copia de seguridad integradas para las aplicaciones o los datos. Como plataforma de hospedaje de contenedores sin estado, Container Apps espera que las aplicaciones administren sus propias estrategias de persistencia y recuperación de datos a través de servicios externos. Los contenedores de aplicaciones y sus sistemas de archivos locales son efímeros y los datos almacenados localmente se pierden cuando las réplicas se reinician o se mueven.

Resistencia durante las actualizaciones de la aplicación

Use la administración de revisiones para implementar actualizaciones en la aplicación sin tiempo de inactividad. Puede crear nuevas revisiones con imágenes de contenedor actualizadas y realizar una transición mediante una estrategia de implementación azul-verde o cambiar gradualmente el tráfico mediante reglas de división de tráfico. Durante las actualizaciones de aplicaciones, la plataforma mantiene los recuentos mínimos de réplicas mediante la creación de nuevos contenedores antes de retirar los antiguos, lo que ayuda a evitar interrupciones del servicio.

Para obtener más información, consulte Actualización e implementación de cambios en Container Apps.

Resistencia al mantenimiento del servicio

Container Apps realiza el mantenimiento automático de la plataforma para aplicar actualizaciones de seguridad, implementar nuevas características y mejorar la confiabilidad del servicio. La plataforma usa actualizaciones graduales en dominios de error y zonas de disponibilidad para reducir la interrupción de las aplicaciones en ejecución. Durante las ventanas de mantenimiento, los contenedores siguen ejecutándose sin interrupción porque las actualizaciones se aplican a la infraestructura subyacente en fases.

Puede especificar sus propias ventanas de mantenimiento, que son períodos de tiempo que desea realizar el mantenimiento en las aplicaciones. Tenga en cuenta que las actualizaciones críticas pueden producirse fuera de las ventanas de mantenimiento. Para obtener más información, consulte Mantenimiento planeado de Container Apps.

Acuerdo de nivel de servicio

El contrato de nivel de servicio (SLA) para los servicios de Azure describe la disponibilidad esperada de cada servicio y las condiciones que la solución deberá cumplir para lograr esa expectativa de disponibilidad. Para obtener más información, consulte Acuerdos de Nivel de Servicio para servicios en línea.

El Acuerdo de Nivel de Servicio de disponibilidad para las Aplicaciones de Contenedor se basa en las reglas de escalado que establezca en las aplicaciones.

Contenido relacionado

- Introducción a Container Apps
- Sondas de estado en Container Apps
- Escalado automático en aplicaciones de contenedores
- Observabilidad en Aplicaciones de Contenedores
- Documentación sobre confiabilidad de Azure
- Guía de confiabilidad deWell-Architected Framework

ⓘ **Nota:** El autor ha creado este artículo con ayuda de inteligencia artificial. Más información

---

Last updated on 22/01/2026

Confiabilidad en Azure Container Instances

En este artículo se describe la compatibilidad con la confiabilidad en Azure Container Instances, que proporciona una manera sencilla de ejecutar contenedores de Linux o Windows en Azure, sin necesidad de administrar máquinas virtuales (VM) ni adoptar un servicio de nivel superior más complejo.

Al usar Azure, relibilidad es una responsabilidad compartida. Microsoft proporciona una variedad de funcionalidades para admitir resistencia y recuperación. Es responsable de comprender cómo funcionan esas funcionalidades dentro de todos los servicios que usa y de seleccionar las funcionalidades que necesita para cumplir los objetivos empresariales y los objetivos de tiempo de actividad.

En este artículo se describe cómo hacer que Azure Container Instances sean resistentes a una variedad de posibles interrupciones y problemas, incluidos errores transitorios, interrupciones de zona de disponibilidad y interrupciones de regiones. Resalta cierta información clave sobre el acuerdo de nivel de servicio (SLA) de Azure Container Instances.

Recomendaciones de implementación de producción para la confiabilidad

Para aumentar la confiabilidad de las aplicaciones de producción basadas en Container Instances, se recomienda realizar las siguientes acciones:

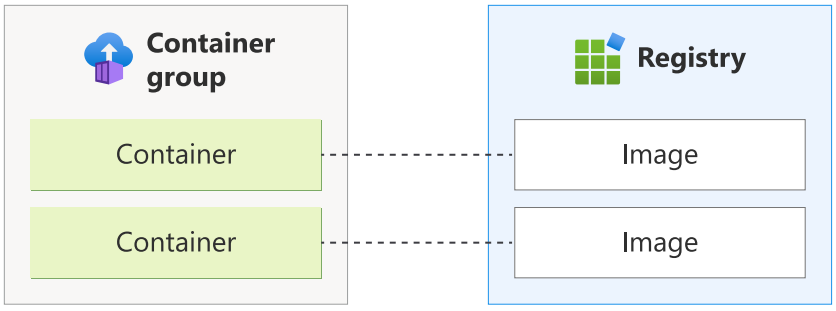
- ✓ Ejecute las aplicaciones en varias zonas de disponibilidad.
- ✓ Considere si también se deben ejecutar grupos de contenedores independientes en varias regiones.
- ✓ Use sondeos de actividad para detectar y automáticamente reiniciar contenedores no saludables.
- ✓ Utilice sondeos de preparación para esperar hasta que los contenedores estén listos antes de recibir tráfico.
- ✓ Use actualizaciones graduales para aplicar cambios progresivamente si usa NGroups. Este enfoque reduce la probabilidad de tiempo de inactividad debido a las actualizaciones.
- ✓ Revise las prácticas recomendadas y consideraciones para Container Instances.

Introducción a la arquitectura de confiabilidad

Para usar Container Instances, implemente un grupo *container*. Un grupo de contenedores contiene uno o varios *contenedores*. Cada contenedor se crea a partir de una imagen *container*, que se almacena en un registro como Azure Container Registry.

Todos los contenedores de un grupo de contenedores se implementan juntos como una sola unidad lógica y comparten la misma infraestructura física.

En el diagrama siguiente se muestra la relación entre grupos de contenedores, contenedores e imágenes.



Container Instances proporciona las siguientes características para administrar grupos de contenedores:

- NGroups (versión preliminar) proporciona un conjunto de funcionalidades para administrar varios grupos de contenedores relacionados. Al crear un NGroup, se define el número de grupos de contenedores para crear. Container Instances proporciona funcionalidades como lanzamientos de actualización automatizados y distribución de grupos de contenedores entre zonas de disponibilidad.
- Los grupos en espera crean un grupo de grupos de contenedores aprovisionados previamente que se pueden usar en respuesta al tráfico entrante. Los grupos en espera están concebidos para optimizar la creación de grupos de contenedores y no para aumentar la resiliencia.

Resistencia a errores transitorios

Los errores transitorios son errores breves e intermitentes en los componentes. Se producen con frecuencia en un entorno distribuido como la nube y son una parte normal de las operaciones. Los errores transitorios se corrigen después de un breve período de tiempo. Es importante que las aplicaciones puedan controlar errores transitorios, normalmente mediante el reintento de solicitudes afectadas.

Todas las aplicaciones hospedadas en la nube deben seguir las instrucciones de control de errores transitorios Azure cuando se comunican con cualquier API, bases de datos y otros componentes hospedados en la nube. Para obtener más información, consulte Recomendaciones para controlar errores transitorios.

Los SDK proporcionados por Microsoft suelen controlar errores transitorios. Dado que hospeda sus propias aplicaciones en Container Instances, siga los pasos necesarios para reducir la posibilidad de errores transitorios:

- Ejecute varios grupos de contenedores para cargas de trabajo importantes para asegurarse de que un error en un contenedor o grupo de contenedores no afecte a toda la aplicación.
- Construye el código de tu aplicación para que sea resistente a errores transitorios en los servicios a los que te conectas, por ejemplo, utilizando políticas de reintentos con estrategias de espera incremental.

Para obtener más información sobre otros errores que pueden producirse en tiempo de ejecución y cómo responder a ellos, consulte Problemas durante el tiempo de ejecución del grupo de contenedores.

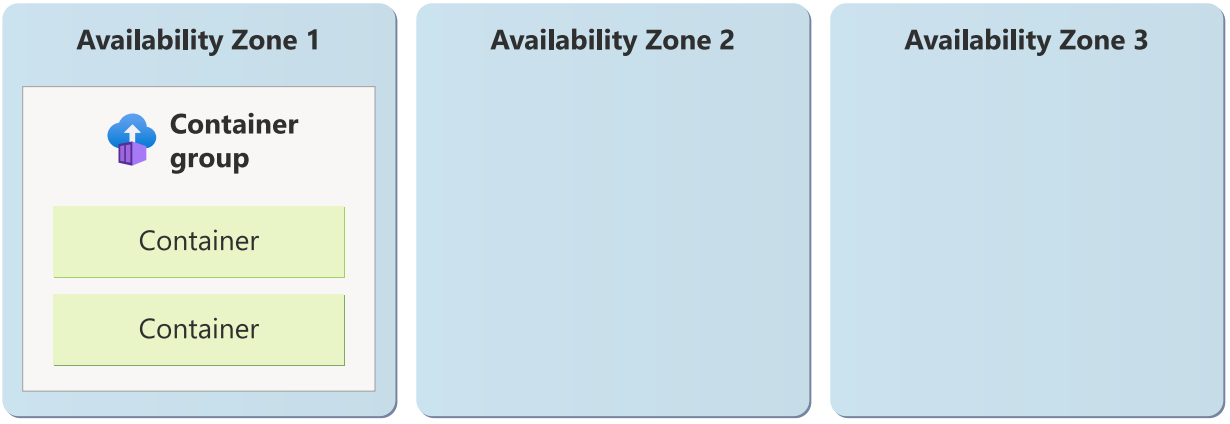
Resistencia a errores de zona de disponibilidad

Availability zones son grupos de centros de datos físicamente independientes dentro de una región de Azure. Cuando una zona falla, los servicios pueden transferirse a una de las zonas restantes.

Container Instances admite zonas de disponibilidad de diferentes maneras, en función de cómo implemente los grupos de contenedores:

- **Grupos de contenedores creados manualmente:** Un grupo de contenedores individual es un recurso *zonal* , lo que significa que se puede implementar en una sola zona de disponibilidad que seleccione. Todos los contenedores del grupo se despliegan en la misma zona de disponibilidad. Si esa zona de disponibilidad tiene una interrupción, el grupo de contenedores y todos sus contenedores podrían experimentar tiempo de inactividad.

  En el diagrama siguiente se muestra un grupo de contenedores que se ha implementado manualmente en la zona de disponibilidad 1:

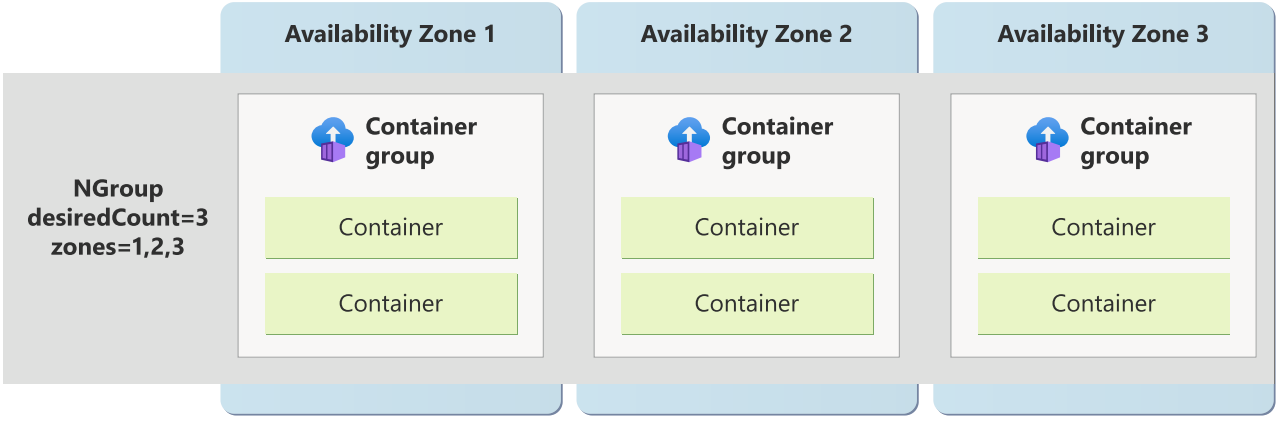


> ⓘ **Nota:**
>
> Para asegurarse de que la aplicación sigue ejecutándose cuando cualquier zona única de la región experimenta una interrupción, se recomienda crear un mínimo de dos grupos de contenedores en dos zonas de disponibilidad diferentes.

Si no especifica zonas de disponibilidad que se van a usar para el grupo de contenedores, es *no zonal* o *regional*, lo que significa que podría colocarse en cualquier zona de disponibilidad dentro de la región o dentro de la misma zona. Si alguna zona de disponibilidad de la región tiene un problema, el grupo de contenedores podría experimentar tiempo de inactividad.

- **NGroups:** Al implementar un NGroup, puede especificar una o varias zonas en las que implementarla. Si implementa un NGroup en dos o más zonas, se trata de un NGroup *con redundancia de zona* y una interrupción de una zona de disponibilidad solo provoca problemas para los grupos de contenedores dentro de la zona afectada.

  En el diagrama siguiente se muestra un NGroup que se implementa en tres zonas de disponibilidad:



Si no especifica zonas de disponibilidad para su NGroup, será no zonal y podría experimentar tiempo de inactividad si alguna zona de disponibilidad en la región presenta un problema.

- **Grupos en espera:** Al implementar un grupo en espera, puede especificar opcionalmente una o varias zonas. La plataforma puede solicitar contenedores en las zonas que seleccione.

  Sin embargo, los grupos en espera no tienen redundancia de zona ni resistencia de zona porque no hay garantía de que los contenedores se creen en varias zonas. Si se produce una interrupción de zona, es posible que todos los contenedores del grupo se coloquen en la zona afectada.

  Debido a que los grupos de respaldo no están diseñados para soportar la resiliencia ante fallos de zona, esta guía no describe el comportamiento detallado de los grupos de respaldo con zonas de disponibilidad.

  > ⓘ **Importante**
  >
  > Los grupos en espera no se diseñaron para tener resistencia de zona. No deben usarse para cargas de trabajo que requieran resistencia a errores de zona.

Requisitos

- **Compatibilidad con regiones:** Las implementaciones de grupos de contenedores zonales se admiten en todas las regiones con zonas de disponibilidad.
- Las implementaciones zonales están disponibles para Linux y Windows Server 2019 grupos de contenedores.
- Para seleccionar una zona de disponibilidad, debe usar la SKU estándar. Los grupos de contenedores zonales no están disponibles con la SKU confidencial.

Consideraciones

Los contenedores de acceso puntual no admiten zonas de disponibilidad y siempre son no zonales.

Cost

No hay ningún costo adicional para configurar zonas de disponibilidad para un grupo de contenedores.

Configurar soporte de zonas de disponibilidad

- **Cree grupos de contenedores con soporte para zonas de disponibilidad.** El enfoque que se usa para configurar zonas de disponibilidad depende de cómo cree grupos de contenedores.
  - *Grupos de contenedores creados manualmente:* Para crear un grupo de contenedores zonales en una zona específica, puede usar uno de los métodos siguientes:
    - El portal Azure
    - El CLI de Azure
    - Azure PowerShell
    - Bicep
    - Una plantilla de Azure Resource Manager (plantilla de ARM)
    - Terraform
  - *NGroups:* Puede implementar un NGroup con redundancia de zona mediante un archivo de Bicep o una plantilla de ARM y especificando varias zonas. Para obtener más información, consulte ejemplo de NGroups con zonas.
  - *Grupos en espera:* Puede implementar un grupo en espera que use zonas de disponibilidad especificando una o varias zonas al crear o actualizar el grupo. Sin embargo, es posible que los contenedores no se creen en varias zonas. Los pools de reserva no deben usarse para cargas de trabajo que requieran tolerancia a fallos de zona. Para obtener más información, consulte Crear un grupo en espera para Container Instances.
- **Habilitar soporte para zonas de disponibilidad en los recursos existentes.** El enfoque que se usa para configurar zonas de disponibilidad depende de cómo cree grupos de contenedores.
  - *Grupos de contenedores creados manualmente:* No se pueden habilitar zonas de disponibilidad en un grupo de contenedores nozonal existente. Debe eliminar el grupo de contenedores y crear un grupo de contenedores zonal.
  - *NGroups:* No se pueden habilitar zonas de disponibilidad en un NGroup no zonal existente. Debe eliminar el NGroup y crear un NGroup con redundancia zonal.

- *Grupos en espera:* No se pueden habilitar zonas de disponibilidad en un grupo de espera no zonal existente. Debe eliminar el grupo de contenedores y crear un nuevo grupo de espera que use varias zonas de disponibilidad.

- **Mueva grupos de contenedores entre zonas o deshabilite el soporte para zonas de disponibilidad.** El enfoque que se usa para modificar zonas de disponibilidad depende de cómo cree grupos de contenedores.

  - *Grupos de contenedores creados manualmente:* Para cambiar la zona de disponibilidad de un grupo de contenedores, debe eliminar el grupo de contenedores y crear otro grupo de contenedores con la nueva zona de disponibilidad. Para obtener información sobre cómo eliminar el grupo de contenedores, consulte los siguientes recursos:
    - El portal Azure
    - El CLI de Azure
    - Azure PowerShell
    - Bicep
    - Una plantilla de ARM
    - Terraform
    - La CLI de Docker

  - *NGroups:* Puede agregar zonas a un NGroup, pero no puede quitar zonas.

  - *Grupos en espera:* Puede agregar zonas a un grupo en espera, pero no puede quitar zonas.

Distribución de grupos de contenedores

La forma en que los grupos de contenedores se distribuyen entre zonas de disponibilidad depende de cómo implemente los grupos de contenedores.

- **Grupos de contenedores creados manualmente:** Es responsable de distribuir los grupos de contenedores creados manualmente en varias zonas de disponibilidad.

- **NGroups:** Durante las operaciones de escalado horizontal, NGroups elimina aleatoriamente instancias, lo que podría no mantener una distribución entre zonas de disponibilidad. Las operaciones de escalabilidad horizontal intentan reequilibrar la distribución entre zonas.

- **Grupos en espera:** Un grupo en espera puede crear contenedores en cualquiera de las zonas de disponibilidad que configure en el grupo. Sin embargo, es posible que los

contenedores no se creen en varias zonas. Los pools de reserva no deben usarse para cargas de trabajo que requieran tolerancia a fallos de zona.

Planeamiento y administración de capacidad

** Para prepararse para un fallo de zona de disponibilidad, considere el sobreaprovisionamiento del número de grupos de contenedores que despliegue. Este enfoque permite que la solución tolere cierta pérdida de capacidad y siga funcionando sin que se vea afectado su rendimiento. Para obtener más información, consulte Administración de la capacidad mediante el sobreaprovisionamiento.

El enfoque que se usa para sobreaprovisionar grupos de contenedores depende de cómo implemente los grupos de contenedores.

- **Grupos de contenedores creados manualmente:** Es responsable de planear la capacidad de los grupos de contenedores creados manualmente, incluido el planeamiento del número de grupos de contenedores que se van a implementar en cada zona.
- **NGroup:** Considere *la posibilidad de sobreaprovisionar* la capacidad de NGroup para tolerar la pérdida de una zona.
- **Grupos en espera:** Los grupos en espera no están diseñados para ser resistentes a los errores de zona. Considere la posibilidad de usar varios grupos en espera en diferentes zonas o usar NGroups.

Comportamiento cuando todas las zonas están en buen estado

En esta sección se describe qué esperar cuando los recursos de Container Instances están configurados para ser compatibles con zonas de disponibilidad y todas las zonas de disponibilidad están operativas.

- **Enrutamiento de tráfico entre zonas:** Usted es responsable de dirigir el tráfico a sus contenedores. Por ejemplo, puede usar Azure Application Gateway como puerta de enlace y equilibrador de carga para los grupos de contenedores.

  Si usa grupos de NGroup o grupos en espera, es responsable del equilibrio de carga en cada contenedor. También debe configurar el sistema de enrutamiento de tráfico para detectar el estado de cada grupo de contenedores y redirigir el tráfico a un grupo alternativo cuando sea necesario.

- **Replicación de datos entre zonas:** Los contenedores y los grupos de contenedores no tienen estado. Puede adjuntar su propio recurso compartido de archivos o conectarse a bases de datos u otros servicios de almacenamiento desde sus aplicaciones. Usted es responsable de asegurar que los recursos compartidos de archivos y los servicios de almacenamiento sean resilientes por zonas. Revise las guías de confiabilidad de cada servicio para comprender cómo hacer que cada zona de componentes sea resistente.

Comportamiento durante un fallo de zona

En esta sección se describe qué esperar cuando los recursos de Container Instances están configurados para admitir la zona de disponibilidad y se produce una interrupción en esta.

- **Detección y respuesta:** La responsabilidad de la detección de errores de zona y la respuesta asociada dependen de cómo implemente los grupos de contenedores.
  - *Grupos de contenedores creados manualmente:* debe detectar la pérdida de una zona de disponibilidad e iniciar una conmutación por error a un grupo de contenedores alternativo que ha creado en otra zona de disponibilidad.
  - *NGroups:* La plataforma Container Instances es responsable de detectar un error en una zona de disponibilidad y responder.

    Sin embargo, es responsable de garantizar que el tráfico se enrute a los contenedores en una zona saludable.
  - *Standby pools:* No se garantiza que la plataforma de Container Instances responda a errores de zona para pools en espera. Los pools de reserva no deben usarse para cargas de trabajo que requieran tolerancia a fallos de zona.
- **Notificación:** Microsoft no le notifica automáticamente cuando una zona está inactiva. Sin embargo, puede usar Azure Service Health para comprender el estado general del servicio, incluidos los errores de zona, y puede configurar alertas de Service Health para notificarle problemas.
- **Solicitudes activas:** Si se produce un error en una zona, es probable que todos los contenedores que se ejecutan en esa zona se detengan, incluido cualquier trabajo activo que controlen.
- **Pérdida de datos esperada:** Dado que los contenedores y los grupos de contenedores no tienen estado, no se espera ninguna pérdida de datos debido a un fallo de zona. Sin

embargo, es responsable de garantizar que cada componente de la carga de trabajo sea resistente a la zona, incluidos los servicios de almacenamiento y las bases de datos.

- **Tiempo de inactividad esperado:** El tiempo de inactividad que puede esperar de un error de zona depende de cómo implemente los grupos de contenedores.
  - *Grupos de contenedores creados manualmente:* En el caso de los grupos de contenedores zonales, cuando una zona no está disponible, el grupo de contenedores y sus contenedores no están disponibles hasta que se recupere la zona de disponibilidad.
  - *NGroups:* En el caso de NGroups, si una zona deja de funcionar, la aplicación sigue estando disponible porque los grupos de contenedores restantes de NGroups siguen ejecutándose en otras zonas. No se espera ningún tiempo de inactividad.
  - *Grupos en espera:* Los grupos en espera no proporcionan resistencia de zona. Si todos los grupos de contenedores del grupo en espera se encuentran en una sola zona de disponibilidad, es posible que todos los grupos de contenedores y sus contenedores queden inaccesibles hasta que se recupere la zona de disponibilidad.
- **Reenrutamiento del tráfico:** Dado que es responsable del enrutamiento del tráfico a los contenedores, también es responsable de volver a enrutar el tráfico si se produce un error en un grupo de contenedores debido a una interrupción de zona de disponibilidad.

Recuperación de zona

Una vez recuperada la zona, la plataforma Azure reinicia automáticamente los grupos de contenedores que se habían detenido. No se requiere ninguna acción del cliente.

Prueba de fallos de zona

No hay ninguna manera de simular una interrupción de la zona de disponibilidad que contiene el grupo de contenedores. Sin embargo, puede configurar manualmente puertas de enlace ascendentes o equilibradores de carga para redirigir el tráfico a otro grupo de contenedores en una zona de disponibilidad diferente.

Resistencia a errores en toda la región

Container Instances es un servicio de una sola región. Si la región deja de estar disponible, los grupos de contenedores y sus contenedores tampoco están disponibles.

Soluciones de varias regiones personalizadas para la resistencia

Opcionalmente, puede implementar grupos de contenedores independientes en varias regiones. Es responsable de implementar y configurar los grupos de contenedores en cada región. También debe configurar el equilibrio de carga mediante un servicio como Azure Traffic Manager o Azure Front Door. Usted es responsable de cualquier sincronización, conmutación por error y recuperación después de falla de datos.

Acuerdo de nivel de servicio

El acuerdo de nivel de servicio (SLA) para Azure servicios describe la disponibilidad esperada de cada servicio y las condiciones que la solución debe cumplir para lograr esa expectativa de disponibilidad. Para obtener más información, vea SLAs for servicios en línea.

Contenido relacionado

- Confiabilidad en Azure

ⓘ**Nota:** El autor ha creado este artículo con ayuda de inteligencia artificial. Más información

---

Last updated on 22/01/2026

Confiabilidad en Azure Container Registry

Azure Container Registry es un servicio de registro de contenedor administrado que se usa para almacenar y administrar las imágenes de contenedor privadas de Docker y los artefactos relacionados para las implementaciones de contenedor.

Cuando se usa Azure, la confiabilidad es una responsabilidad compartida. Microsoft proporciona una variedad de funcionalidades para admitir resistencia y recuperación. Es responsable de comprender cómo funcionan esas funcionalidades dentro de todos los servicios que usa y de seleccionar las funcionalidades que necesita para cumplir los objetivos empresariales y los objetivos de tiempo de actividad.

En este artículo se describe cómo hacer que Container Registry sea resistente a una variedad de posibles interrupciones y problemas, incluidos errores transitorios, interrupciones de zona de disponibilidad y interrupciones de regiones. También se describe cómo puede usar copias de seguridad para recuperarse de otros tipos de problemas y se resalta cierta información clave sobre el acuerdo de nivel de servicio (SLA) de Container Registry.

Recomendaciones de implementación de producción para la confiabilidad

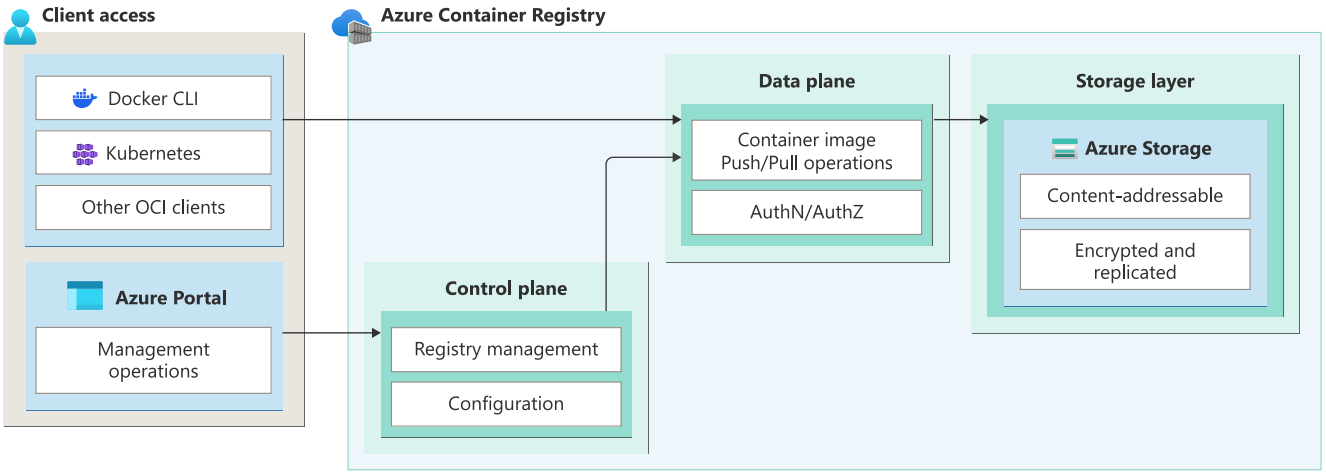
Para las cargas de trabajo de producción, le recomendamos que realice las siguientes acciones:

- Use el nivel Premium de Container Registry, que proporciona las características de confiabilidad más completas. El nivel Premium también proporciona límites de rendimiento más altos, características de seguridad mejoradas y funcionalidades avanzadas que son esenciales para las cargas de trabajo de contenedor de producción. Para más información sobre los niveles de servicio y las características, consulte Niveles de servicio de Container Registry.
- Aprovisione su registro de contenedores en una región que admita zonas de disponibilidad.
- En escenarios de varias regiones, configure la replicación geográfica para distribuir el registro en varias regiones en función de sus requisitos geográficos y de cumplimiento específicos.

Introducción a la arquitectura de confiabilidad

Container Registry se basa en la infraestructura distribuida de Azure para proporcionar alta disponibilidad y durabilidad de los datos. El servicio consta de varios componentes clave que

funcionan conjuntamente para garantizar la confiabilidad. En el siguiente diagrama se muestra la arquitectura de servicio principal.



- **El plano de control** es un componente de administración centralizado en la región principal para la configuración del registro, la configuración de autenticación y las directivas de replicación.
- **El plano de datos** es un servicio distribuido que controla las operaciones de inserción y extracción de imágenes de contenedor entre regiones y zonas de disponibilidad.
- **La capa de almacenamiento** es una solución de Azure Storage direccionable con contenido que conserva imágenes y artefactos de contenedor. Incluye desduplicación automática, cifrado en reposo y replicación integrada.

Microsoft es responsable de administrar la infraestructura subyacente de Container Registry, que incluye los siguientes tipos de mantenimiento:

- **Mantenimiento del plano de control** para la administración del registro
- **Mantenimiento del plano de datos** para las operaciones de imagen de contenedor en regiones y zonas de disponibilidad

Como cliente, es responsable de las siguientes acciones:

- **Resistencia de nivel de aplicación:** Implemente la lógica de reintento adecuada y el control de conmutación por error en las aplicaciones de contenedor y las plataformas de orquestación.
- **Configuración de resistencia de zona:** Asegúrese de que el registro de contenedor está implementado en una región que admita zonas de disponibilidad.
- *Configuración de replicación geográfica:* Elija las regiones adecuadas para la replicación geográfica en función de los requisitos de distribución geográfica, cumplimiento y rendimiento.

Container Registry también admite *tareas*, lo que puede ayudarle a automatizar las compilaciones de contenedor y las operaciones de mantenimiento. Las tareas se ejecutan en una infraestructura informática administrada por Microsoft y admiten desencadenadores manuales, basados en eventos o programados. Para más información, consulte Automatizar la creación y el mantenimiento de imágenes de contenedores mediante tareas de Container Registry.

> ⓘ **Nota:**
>
> Container Registry admite registros conectados, que son réplicas locales o remotas que se sincronizan con su Container Registry basado en la nube. Al usar registros conectados, es responsable de configurarlos para cumplir los requisitos de confiabilidad. Los registros conectados quedan fuera del ámbito de este artículo.

Resistencia a errores transitorios

Los errores transitorios son errores breves e intermitentes en los componentes. Se producen con frecuencia en un entorno distribuido como la nube y son una parte normal de las operaciones. Los errores transitorios se corrigen después de un breve período de tiempo. Es importante que las aplicaciones puedan controlar errores transitorios, normalmente mediante el reintento de solicitudes afectadas.

Todas las aplicaciones hospedadas en la nube deben seguir las instrucciones de control de errores transitorios de Azure cuando se comunican con cualquier API, bases de datos y otros componentes hospedados en la nube. Para obtener más información, consulte Recomendaciones para controlar errores transitorios.

Container Registry controla los errores transitorios internamente a través de varios mecanismos. El servicio implementa una lógica de reintento automático para las operaciones del registro y mantiene un grupo de conexiones para un uso eficiente de los recursos. Las operaciones de Container Registry están diseñadas para ser idempotentes, lo que permite reintentos seguros de las operaciones de envío y extracción. Las tareas administran automáticamente los fallos transitorios cuando realizan muchos tipos de operaciones.

En el caso de las aplicaciones cliente que usan Container Registry, implemente las directivas de reintento adecuadas con retroceso exponencial al realizar operaciones del Registro. Use el cliente oficial de Docker o los SDK de Container Registry, que incluyen mecanismos de reintento integrados para los errores transitorios más comunes.

Resistencia a errores de zona de disponibilidad

Las zonas de disponibilidad son grupos físicamente independientes de centros de datos dentro de una región de Azure. Cuando una zona falla, los servicios pueden transferirse a una de las zonas restantes.

La redundancia de zona protege su registro de contenedores contra errores de una sola zona distribuyendo los datos y las operaciones del registro entre varias zonas de disponibilidad dentro de la región. Las operaciones de extracción e inserción de imágenes de contenedor continúan funcionando durante las interrupciones de zona, con conmutación automática por error a zonas correctas.

La redundancia de zona está habilitada de forma predeterminada para todos los registros de regiones que admiten zonas de disponibilidad, lo que hace que los recursos sean más resistentes automáticamente y sin costo adicional. Esta mejora se aplica a todos los niveles de servicio, incluidos Básico y Estándar, y se ha aplicado a registros nuevos y existentes.

> ⓘ **Importante**
>
> Es posible que Azure Portal y otras herramientas aún no reflejen la actualización de redundancia de zona con precisión.
>
> La `zoneRedundancy` propiedad de la configuración del registro podría seguir apareciendo como `false`, pero la redundancia de zona está activa para todos los registros de las regiones admitidas.
>
> Estamos actualizando activamente las superficies del portal y de la API para reflejar este comportamiento predeterminado de forma más transparente. Todas las características habilitadas anteriormente siguen funcionando según lo previsto.

Consideraciones

- **Compatibilidad con regiones:** Los registros con redundancia de zona se pueden implementar en cualquier región que admita zonas de disponibilidad. Si su registro está en una región que no admite zonas de disponibilidad, para que sea redundante en zonas, debe crear un nuevo registro en una región que admita zonas de disponibilidad. A continuación, debe migrar las imágenes de contenedor mediante la creación de una canalización de transferencia o mediante la importación de imágenes de contenedor.

Consideraciones

- **Tareas:** Las tareas de Container Registry no admiten actualmente zonas de disponibilidad. La redundancia de zona se aplica al propio servicio del registro, pero no a las tareas ni a

sus operaciones.

- **Replicación geográfica:** Si el registro usa replicación geográfica, las réplicas creadas en regiones con zonas de disponibilidad son automáticamente redundantes a nivel de zona.

Cost

La redundancia de zona se incluye con registros de contenedor sin coste adicional.

Configurar soporte de zonas de disponibilidad

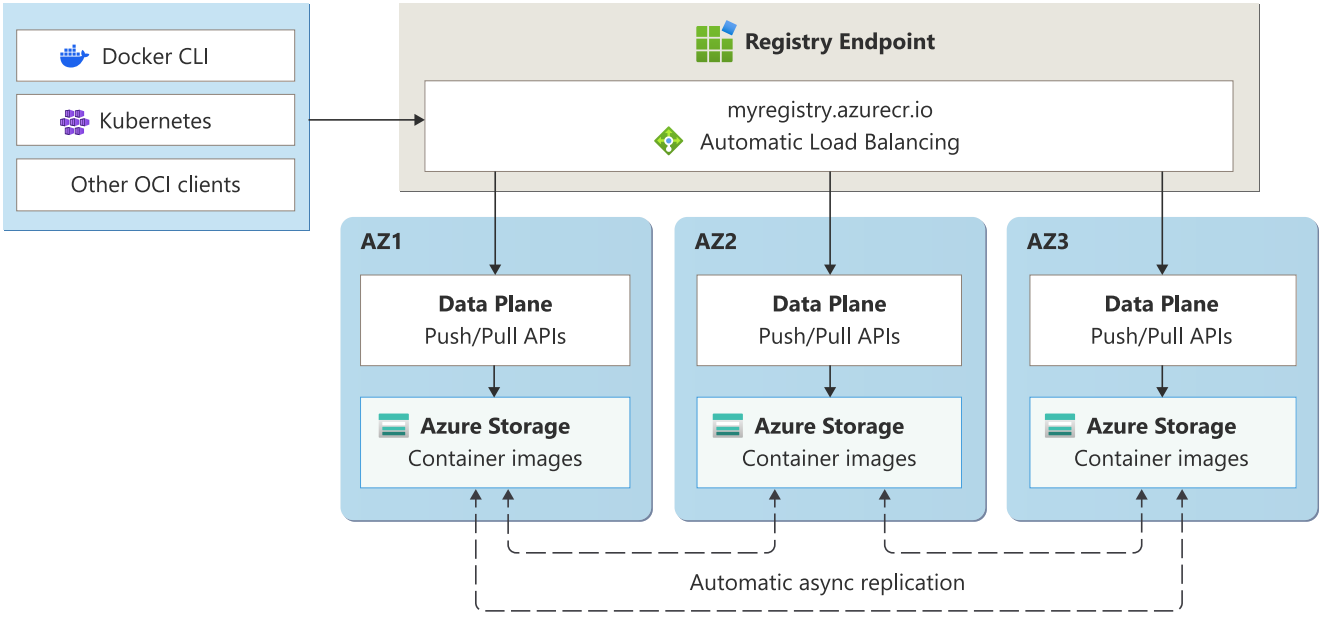
- **Cree un registro con redundancia de zona.** Cuando creas un registro en una región admitida, es automáticamente redundante a nivel de zona. Para más información sobre cómo crear un registro, consulte Creación de un registro con redundancia de zona en Container Registry.

- **Habilite la redundancia de zona en un registro existente.** Los registros que se encuentran en regiones con zonas de disponibilidad automáticamente tienen redundancia de zona. No es necesario habilitar la redundancia de zona.

- **Desactivar la redundancia de zona.** No se puede deshabilitar la redundancia de zona.

Comportamiento cuando todas las zonas están en buen estado

En esta sección se describe qué esperar cuando los recursos de Container Registry están configurados para la redundancia de zona y todas las zonas de disponibilidad están operativas.



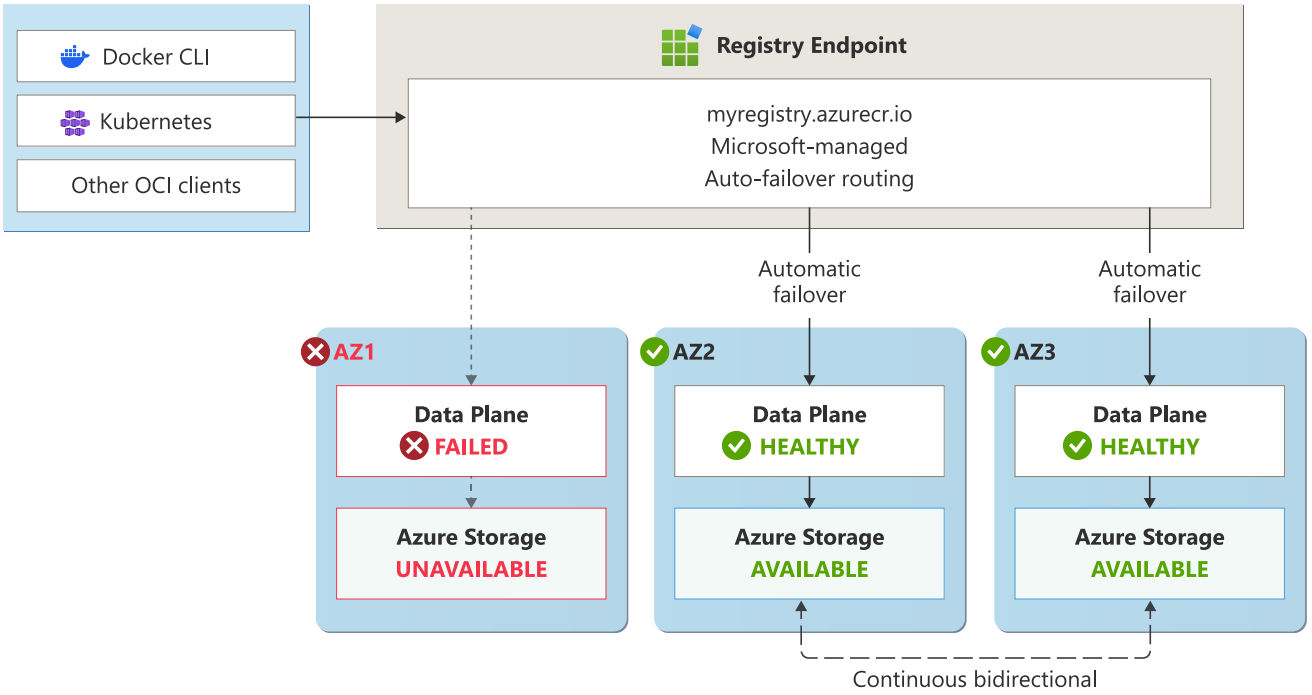
- **Enrutamiento de tráfico entre zonas:** Container Registry usa la funcionalidad de enrutamiento interno para distribuir automáticamente las operaciones del plano de datos en todas las zonas de disponibilidad dentro de una región. El servicio del Registro enruta automáticamente las solicitudes a zonas correctas sin necesidad de equilibradores de carga externos.

- **Replicación de datos entre zonas:** Los datos del Registro, incluidas las imágenes de contenedor, los manifiestos y los metadatos, se replican de forma asincrónica en varias zonas de disponibilidad. Los cambios se replican rápidamente entre zonas para mantener una alta disponibilidad y durabilidad de los datos. La replicación es asíncrona, pero normalmente se completa en cuestión de minutos, y todas las zonas permanecen disponibles para operaciones de lectura y escritura durante la replicación.

Comportamiento durante un fallo de zona

En esta sección se describe qué esperar cuando los recursos de Container Registry están configurados para la redundancia de zona y hay una interrupción de zona de disponibilidad.

Cuando una zona deja de estar disponible, Container Registry administra automáticamente el proceso de conmutación por error con un impacto mínimo en las operaciones del registro.



- **Detección y respuesta:** La plataforma Container Registry detecta automáticamente errores en una zona de disponibilidad e inicia una respuesta. El servicio enruta automáticamente el tráfico a las zonas correctas restantes. No se requiere ninguna intervención manual para iniciar una conmutación por error de zona.

- **Notificaciones:** Microsoft no le notifica automáticamente cuando una zona está inactiva. Sin embargo, puede usar Azure Service Health para comprender el estado general del servicio, incluidos los errores de zona, y puede configurar alertas de Service Health para notificarle problemas.

  También puede supervisar las métricas de disponibilidad del Registro en Azure Monitor.

- **Solicitudes activas:** Cuando una zona de disponibilidad no está disponible, se finalizan las solicitudes en curso conectadas a los recursos de la zona de disponibilidad errónea. Deben reintentarse.

- **Pérdida de datos esperada:** Las escrituras recientes realizadas en la zona defectuosa podrían no replicarse en otras regiones, lo que significa que podrían perderse hasta que la zona se recupere. Normalmente, se espera que la pérdida de datos sea inferior a 15 minutos, pero eso no está garantizado.

- **Tiempo de inactividad esperado:** Es posible que se produzca un breve periodo de inactividad durante la conmutación automática por error, ya que el tráfico se redirige a zonas correctas. Este tiempo de inactividad suele ser de unos segundos para la mayoría de las operaciones del registro. Se recomienda seguir los procedimientos recomendados de control de errores transitorios para minimizar el efecto de la conmutación por error de zona en las aplicaciones.

- **Reenrutamiento del tráfico:** La plataforma vuelve a enrutar automáticamente el tráfico a zonas correctas sin necesidad de realizar ningún cambio de configuración.

Recuperación de zona

Cuando se recupera la zona de disponibilidad afectada, Container Registry distribuye automáticamente las operaciones en todas las zonas disponibles, incluida la zona recuperada. El servicio reequilibra el tráfico y la distribución de datos sin necesidad de intervención manual ni provoca interrupciones del servicio.

Prueba de fallos de zona

La plataforma Container Registry administra el enrutamiento del tráfico, la conmutación por error y la conmutación por recuperación para registros redundantes por zona. Dado que esta característica está totalmente administrada, no es necesario iniciar ni validar los procesos de error de zona de disponibilidad.

Resistencia a errores en toda la región

Container Registry proporciona compatibilidad nativa con varias regiones a través de la replicación geográfica cuando el registro usa el nivel Premium. La replicación geográfica crea réplicas del Registro en varias regiones de su elección. La región en la que se implementa el recurso de registro se conoce como la *región principal*.

La replicación geográfica permite la resistencia a interrupciones regionales. Si su registro está replicado geográficamente y se produce una interrupción regional, los datos del registro seguirán estando disponibles desde las otras regiones que haya seleccionado. Si no habilita la replicación geográfica, es posible que sus datos no estén disponibles durante una interrupción del servicio en la región.

También puede utilizar la replicación geográfica para obtener acceso local a imágenes de contenedores dentro de esas regiones y reducir la latencia de las aplicaciones distribuidas a nivel mundial.

Al implementar Container Registry y habilitar la replicación geográfica, Microsoft usa Azure Traffic Manager para distribuir las solicitudes del plano de datos entre las réplicas y conmutar automáticamente por error entre réplicas si una réplica regional no está disponible.

La replicación geográfica de Container Registry no depende de las regiones emparejadas de Azure. Puede seleccionar cualquier combinación de regiones de Azure para la replicación en función de sus requisitos específicos de ubicación geográfica, rendimiento y cumplimiento normativo. Cada registro replicado geográficamente funciona como un punto de conexión de registro. Admite la mayoría de las operaciones del registro, incluyendo tareas de envío, extracción y administración de imágenes.

En esta sección se resume la información sobre la replicación geográfica en relación con la confiabilidad. Para más información, consulte Replicación geográfica en Container Registry.

Soporte para regiones

La replicación geográfica está disponible en todas las regiones de Azure en las que se admite el nivel Premium. Puede replicar en cualquier combinación de regiones, independientemente de si Azure empareja esas regiones.

Requisitos

Debe usar el nivel Premium para habilitar la replicación geográfica.

Consideraciones

- **Réplicas con redundancia de zona:** cualquier réplica que cree en una región con zonas de disponibilidad tiene redundancia de zona de manera automática.
- **Plano de control:** El plano de control se ejecuta en la región principal. Si la región principal no está disponible, las operaciones del plano de control no están disponibles y es posible que no pueda modificar la configuración del Registro.
- **Tareas:** Las tareas de Container Registry no admiten actualmente réplicas geográficas. Las tareas siempre se ejecutan en la región principal. Si la región principal no está disponible, la tarea no se ejecuta.

Cost

Cada región con replicación geográfica se factura por separado según los precios del plan Premium de la región correspondiente. Los cargos de salida también se aplican a la transferencia de datos entre regiones durante la replicación inicial y la sincronización continua.

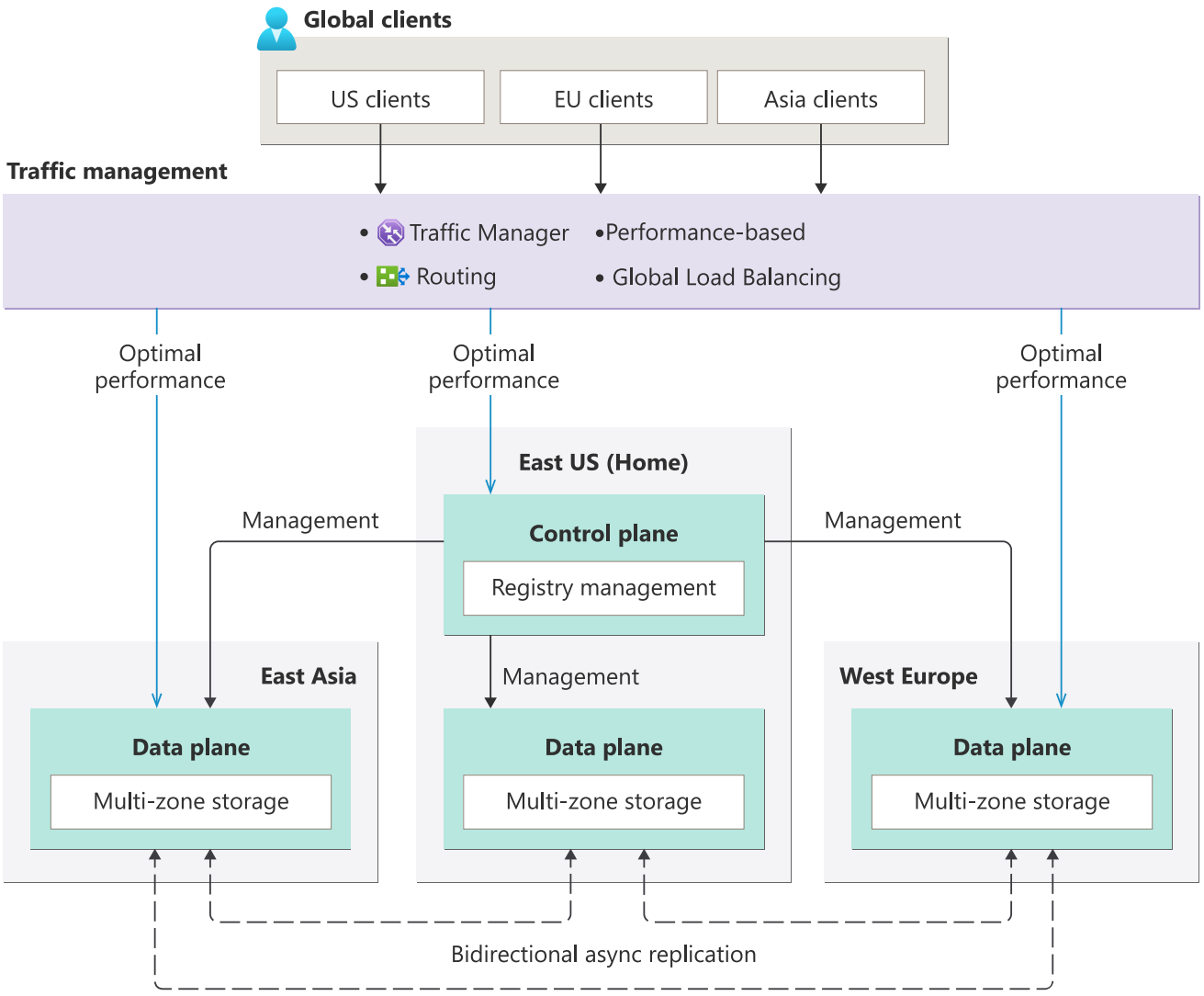
Configuración de la compatibilidad con varias regiones

La replicación geográfica se puede configurar durante la creación del registro o agregarse a los registros Premium existentes. La replicación geográfica se puede configurar a través de Azure Portal, la CLI de Azure, Azure PowerShell o las plantillas de Azure Resource Manager.

- **Crear un registro con replicación geográfica** Configure la replicación geográfica después de crear el registro especificando regiones adicionales.
- **Habilitar la replicación geográfica en un registro existente.** Para habilitar las funciones de replicación geográfica, actualice los registros existentes de nivel Básico o Estándar al nivel Premium. Puede cambiar las regiones de replicación en cualquier momento. Para obtener más información, consulte Configuración de replicación geográfica.
- **Deshabilitar la replicación geográfica.** Quite las réplicas regionales individuales a través de Azure Portal o las herramientas de línea de comandos. No se puede quitar el registro de la región principal.

Comportamiento cuando todas las regiones están en buen estado

En esta sección se describe qué esperar cuando un registro está configurado para la replicación geográfica y todas las regiones están operativas.



- **Enrutamiento del tráfico entre regiones:** Container Registry funciona en una configuración activa-activa donde cada punto de conexión regional puede atender todas las operaciones del plano de datos de forma independiente, incluidas las lecturas y escrituras. Las operaciones del plano de datos, como las operaciones de envío y extracción de contenedores, se enrutan automáticamente mediante Traffic Manager con criterios basados en el rendimiento para determinar el punto de conexión regional óptimo para el rendimiento.

- **Replicación de datos entre regiones:** La replicación geográfica sincroniza automáticamente las imágenes de contenedor y los artefactos en todas las regiones configuradas mediante la replicación asincrónica con coherencia final. El servicio utiliza almacenamiento direccionable por contenido para replicar de manera eficiente solo las capas de imagen únicas. Este enfoque minimiza el uso del ancho de banda y el tiempo de replicación. Las operaciones de lectura y escritura funcionan en todas las regiones con replicación geográfica. Los cambios realizados en cualquier región se replican en todas las demás regiones.

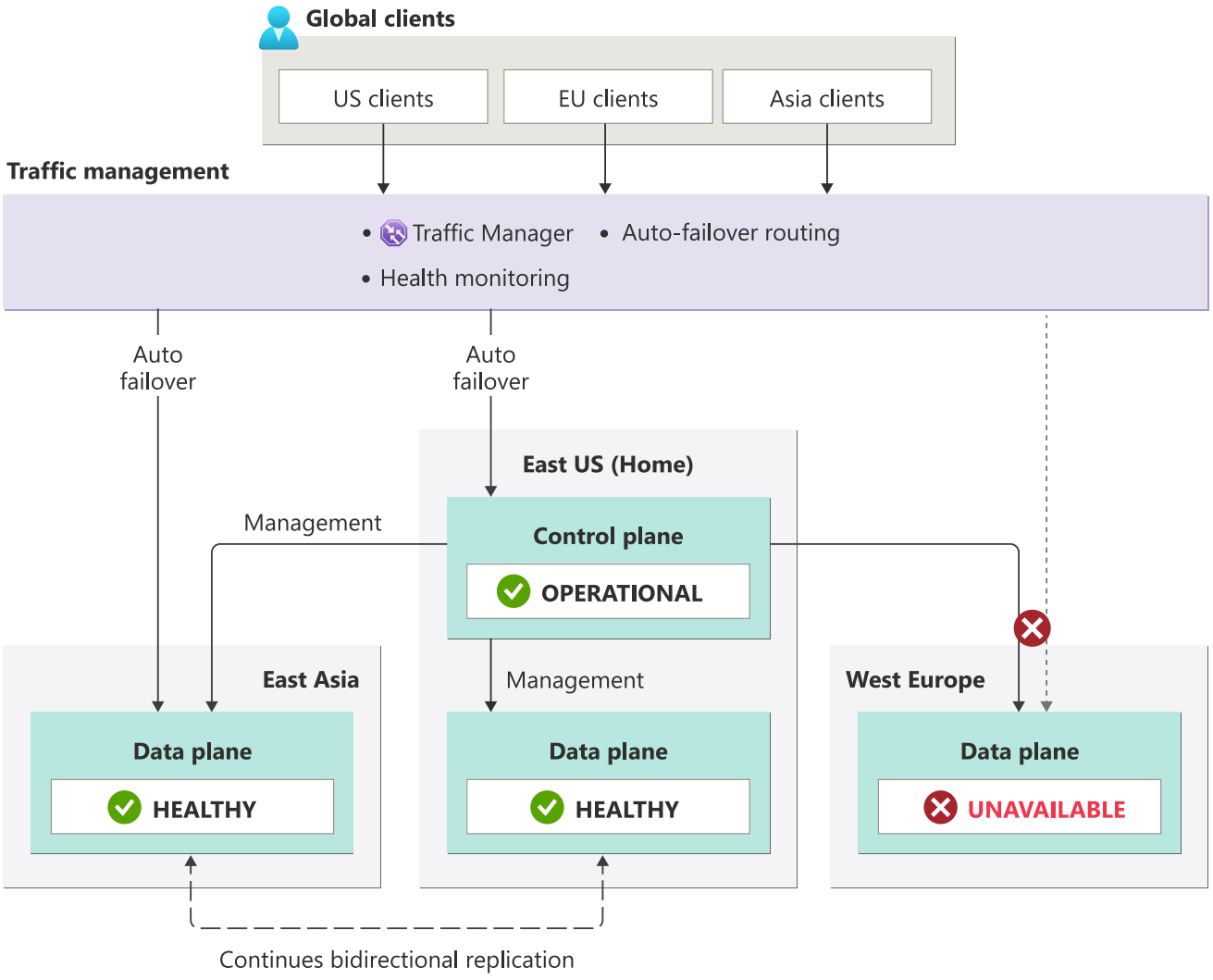
  La replicación se completa normalmente en cuestión de minutos después de los cambios. Sin embargo, no hay garantía sobre el tiempo de replicación de los datos. Las imágenes

de contenedores grandes o las actualizaciones de alta frecuencia pueden tardar más tiempo en replicarse en todas las regiones.

Comportamiento durante una falla de región

En esta sección se describe qué esperar cuando se configura un registro para la replicación geográfica y hay una interrupción en la región primaria.

Cuando una región deja de estar disponible, las operaciones de contenedor pueden seguir usando puntos de conexión regionales alternativos.



- **Detección y respuesta:** Container Registry supervisa el estado de cada réplica regional y es responsable de redirigir el tráfico a otra región.
- **Notificaciones:** Microsoft no le notifica automáticamente cuando una región está inactiva. Sin embargo, puede usar Azure Service Health para comprender el estado general del servicio, incluidos los errores de región, y puede configurar alertas de Service Health para notificarle problemas.

También puede supervisar las métricas de disponibilidad del Registro para cada punto de conexión regional en Azure Monitor.

- **Solicitudes activas:** Cualquier solicitud activa actualmente en tránsito hacia una región no disponible tendrá un error y deberá reintentarse para que pueda dirigirse a una región correcta.
- **Pérdida de datos esperada:** Es posible que las escrituras recientes realizadas en la región defectuosa no se repliquen en otras regiones. Este error significa que podrían perderse hasta que se recupere la región. Normalmente, se espera que la pérdida de datos sea inferior a 15 minutos, pero no se garantiza.
- **Tiempo de inactividad esperado:** En cuanto a las operaciones del plano de datos, se espera un breve periodo de inactividad mientras se completa la conmutación por error, lo que suele tardar entre 1 y 2 minutos.

  Si la región principal no está disponible, las operaciones del plano de control no estarán disponibles hasta que se recupere la región.
- **Reenrutamiento del tráfico:** Cuando una región deja de estar disponible, las operaciones de contenedor se enrutan automáticamente a otra réplica dentro de una región saludable. Los clientes no necesitan cambiar el punto de conexión en el que interactúan con el registro. Microsoft controla automáticamente el enrutamiento, la conmutación por error y la conmutación por recuperación.

Recuperación de regiones

Cuando se recupera una región, las operaciones del plano de datos se reanudan automáticamente para ese punto de conexión regional a través del enrutamiento de Traffic Manager. El servicio sincroniza cualquier cambio que se produzca durante la interrupción mediante la replicación asíncrona con coherencia eventual.

Prueba de fallos de región

No se puede simular el error de una de las regiones asociadas con el registro, pero se puede probar la capacidad de la aplicación para realizar una conmutación por error entre regiones. Puede simular una conmutación por error regional deshabilitando temporalmente las réplicas geográficas, lo que las quita del enrutamiento de Traffic Manager. A continuación, puede verificar que las operaciones de contenedores se transfieren correctamente a regiones alternativas en caso de error sin experimentar realmente una interrupción regional. Para obtener más información, consulte Deshabilitar temporalmente el enrutamiento a la replicación.

Al volver a activar la réplica, Traffic Manager reanuda el enrutamiento del tráfico a la réplica que se ha vuelto a habilitar. Además, los metadatos y las imágenes se sincronizan con coherencia eventual con la réplica reactivada para garantizar la coherencia de los datos en todas las regiones.

Copias de seguridad y restauración

Container Registry admite la exportación de imágenes de contenedor y artefactos desde el registro al almacenamiento externo o a registros alternativos. Use las funciones de importación y exportación de Container Registry o los comandos estándar de Docker para crear copias de imágenes de contenedores críticas para escenarios de recuperación ante desastres.

Para la mayoría de las soluciones, no debe confiar exclusivamente en copias de seguridad. En su lugar, utilice las otras capacidades descritas en esta guía para apoyar los requisitos de resiliencia. Sin embargo, las copias de seguridad protegen contra algunos riesgos que otros enfoques no. Para más información, consulte ¿Qué son la redundancia, la replicación y la copia de seguridad?.

Acuerdo de nivel de servicio

El contrato de nivel de servicio (SLA) para los servicios de Azure describe la disponibilidad esperada de cada servicio y las condiciones que la solución deberá cumplir para lograr esa expectativa de disponibilidad. Para obtener más información, consulte Acuerdos de Nivel de Servicio para servicios en línea.

Contenido relacionado

- Fiabilidad de Azure
- Niveles de servicio de Container Registry
- Procedimientos recomendados de Container Registry
- Supervisión de Container Registry
- Precios de Container Registry
- Importación de imágenes de contenedor en un registro de contenedor

---

Last updated on 22/01/2026

Confiabilidad en Azure Kubernetes Service (AKS)

Azure Kubernetes Service (AKS) es un servicio de orquestación de contenedores administrado que simplifica la implementación, la administración y las operaciones de Kubernetes.

Cuando se usa Azure, la confiabilidad es una responsabilidad compartida. Microsoft proporciona una variedad de funcionalidades para admitir resistencia y recuperación. Es responsable de comprender cómo funcionan esas funcionalidades dentro de todos los servicios que usa y de seleccionar las funcionalidades que necesita para cumplir los objetivos empresariales y los objetivos de tiempo de actividad.

En este artículo se describe cómo hacer que Azure Kubernetes Service (AKS) sea resistente a una variedad de posibles interrupciones y problemas, incluidos errores transitorios, interrupciones de zona de disponibilidad y interrupciones en regiones. También se describe cómo puede usar copias de seguridad para recuperarse de otros tipos de problemas y se resalta cierta información clave sobre el acuerdo de nivel de servicio (SLA) de Azure Kubernetes Service (AKS).

Recomendaciones de implementación de producción

Para obtener recomendaciones sobre cómo implementar cargas de trabajo de producción confiables en AKS, consulte los siguientes artículos:

- Procedimientos recomendados de implementación y confiabilidad de clústeres para AKS
- Descripción general de la alta disponibilidad (HA) y la recuperación ante desastres (DR) para AKS
- Consideraciones de resistencia de zona para AKS

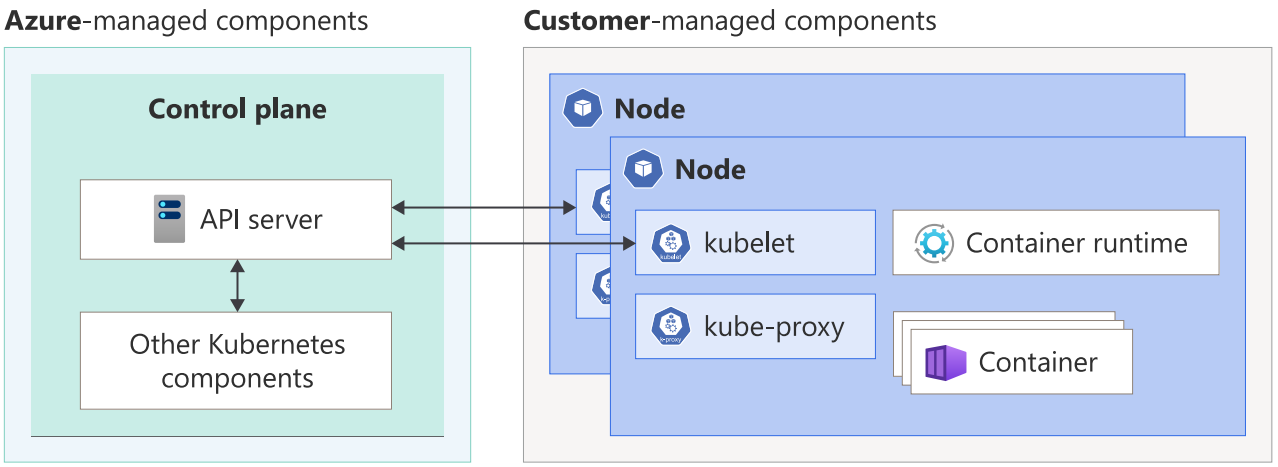
Introducción a la arquitectura de confiabilidad

Al crear un clúster de AKS, la plataforma Azure crea y configura automáticamente:

- Un plano de control que tiene el servidor de API, etcd, el programador y otros pods necesarios para administrar la carga de trabajo.

- Un grupo de nodos del sistema a tu suscripción que hospeda los complementos y pods adicionales que se ejecutan en el espacio de nombres *kube-system*.

Una vez completada la configuración del grupo de nodos inicial, puede agregar o eliminar grupos de nodos para sus propias cargas de trabajo de usuario. AKS no administra grupos de nodos para la confiabilidad y debe asegurarse de que las cargas de trabajo son resistentes a los errores de infraestructura.



La resistencia es una responsabilidad compartida entre usted y Microsoft. Como servicio de cálculo, AKS administra algunos aspectos de la fiabilidad del clúster, pero usted es responsable de gestionar otros aspectos.

- **Microsoft administra** el plano de control y otros componentes administrados de AKS.
- **Es su responsabilidad**:
  - *Defina cómo deben configurarse los componentes, incluidos los grupos de nodos y los equilibradores de carga que se conectan a los servicios, para satisfacer los requisitos de confiabilidad.* Después de definir los componentes, Microsoft los implementa y administra en su nombre.
  - *Administre los componentes fuera del clúster de AKS, incluido el almacenamiento y las bases de datos.* Compruebe que estos componentes cumplen los requisitos de confiabilidad. Al implementar las cargas de trabajo, asegúrese de que otros componentes de Azure también están configurados para la resistencia siguiendo los procedimientos recomendados para esos servicios.

Resistencia a errores transitorios

Los errores transitorios son errores breves e intermitentes en los componentes. Se producen con frecuencia en un entorno distribuido como la nube y son una parte normal de las operaciones. Los errores transitorios se corrigen después de un breve período de tiempo. Es importante que las aplicaciones puedan controlar errores transitorios, normalmente mediante el reintento de solicitudes afectadas.

Todas las aplicaciones hospedadas en la nube deben seguir las instrucciones de control de errores transitorios de Azure cuando se comunican con cualquier API, bases de datos y otros componentes hospedados en la nube. Para obtener más información, consulte Recomendaciones para controlar errores transitorios.

Cuando utilizas AKS, pueden ocurrir fallos transitorios por varias razones, como fallos de aplicaciones, operaciones de escalado y balanceo de pods, actualización de nodos y fallos temporales en la infraestructura, como problemas de hardware o de redes.

Es imposible eliminar todos los errores transitorios, por lo que los clientes que acceden a las aplicaciones hospedadas en AKS deben estar preparados para reintentar las solicitudes con errores y seguir otras recomendaciones de control de errores transitorios. Puede minimizar la probabilidad de errores transitorios y evitar o mitigar el tiempo de inactividad que podrían provocar siguiendo los procedimientos recomendados de Kubernetes y Azure en la implementación.

- ✓ Establece los presupuestos de interrupciones de pod (PDB) en el pod YAML para especificar cuántos pods necesitas tener en un estado `Ready` en un momento dado. Al establecer archivos PDB, AKS garantiza una disponibilidad mínima de réplicas cuando ejecuta operaciones para acordonar y purgar los nodos. Si no se puede satisfacer una PDB durante procesos como actualizaciones, el pod continúa funcionando y es posible que se produzca un error en la operación. Para obtener más información, consulte PDB.
- ✓ Use `maxUnavailable` para definir el número máximo de réplicas que pueden dejar de estar disponibles en un momento dado. Por ejemplo, cuando se realiza un reinicio gradual, AKS garantiza que no más del número `maxUnavailable` de pods se reinician en un momento dado. Para obtener más información, consulte maxUnavailable.
- ✓ Siga los procedimientos recomendados de implementación. Las réplicas de pod también pueden fallar debido a problemas de la aplicación. Para más información, consulte Procedimientos recomendados de nivel de implementación para la confiabilidad del clúster de AKS.

> ⓘ **Nota:**
>
> Si quiere que AKS valide las implementaciones para cumplir los procedimientos recomendados y proporcione notificaciones de bloqueo o advertencia, puede usar medidas de seguridad de implementación. Las medidas de seguridad de implementación son una oferta administrada que ayuda a aplicar los procedimientos recomendados del producto antes de que el código se implemente en el clúster.

Resistencia a errores de zona de disponibilidad

Las zonas de disponibilidad son grupos físicamente independientes de centros de datos dentro de una región de Azure. Cuando una zona falla, los servicios pueden transferirse a una de las zonas restantes.

Al implementar un clúster de AKS en una región que admita zonas de disponibilidad, los distintos componentes requieren distintos tipos de configuración.

El plano de control de AKS es resiliente a zonas por defecto. Si se produce un error en una zona, el plano de control no requiere ninguna configuración ni administración para lograr resistencia. Sin embargo, la resistencia del plano de control no es suficiente para que el clúster permanezca operativo cuando se produce un error en una zona. Para el grupo de nodos del sistema y los grupos de nodos de usuario que implemente, debe habilitar la compatibilidad con zonas de disponibilidad para ayudar a garantizar que las cargas de trabajo son resistentes a los errores de zona de disponibilidad.

Requisitos

Compatibilidad con regiones: Puede implementar clústeres de AKS resistentes a zonas en cualquier región que admita zonas de disponibilidad.

Consideraciones

Para mejorar la confiabilidad y la resistencia de las cargas de trabajo de producción de AKS en una región, debe configurar AKS para la redundancia de zona mediante la realización de las siguientes configuraciones:

- **Implemente varias réplicas.** Kubernetes distribuye los pods entre nodos en función de las etiquetas de nodo. Para distribuir la carga de trabajo entre zonas, debe implementar varias réplicas del pod. Por ejemplo, si configura el grupo de nodos con tres zonas pero solo implementa una única réplica de su pod, su implementación no es resiliente a las zonas.

- **Habilite el escalado automático.** Los grupos de nodos de Kubernetes proporcionan opciones de escalado manuales y automáticas. Con el escalado manual, puedes agregar o eliminar nodos según sea necesario y los pods pendientes esperan hasta que escales verticalmente un grupo de nodos. El escalado administrado por AKS usa el escalador automático del clúster o el aprovisionamiento automático de nodos (NAP). AKS aumenta o reduce el grupo de nodos según los requisitos del pod y teniendo en cuenta la cuota y capacidad de la SKU de tu suscripción. Este método ayuda a garantizar que los pods se programen en los nodos disponibles en las zonas de disponibilidad.

- **Establece las restricciones de la topología del pod.** Usa restricciones de propagación de topología de pods para controlar cómo se distribuyen los pods entre distintos nodos o zonas. Las restricciones le ayudan a lograr la alta disponibilidad, la resistencia y el uso eficaz de los recursos. Si prefieres propagar pods estrictamente entre zonas, puedes establecer restricciones para forzar un pod en un estado pendiente a fin de mantener el equilibrio de pods entre zonas. Para más información, consulta Restricciones de propagación de topología de pod.

- **Configurar redes con resiliencia zonal.** Si los pods sirven tráfico externo, configure la arquitectura de red del clúster mediante servicios como Azure Application Gateway, Azure Load Balancer o Azure Front Door.

- **Asegúrese de que las dependencias son resistentes a la zona.** La mayoría de las aplicaciones de AKS usan otros servicios para el almacenamiento, la seguridad o las redes. Asegúrese de revisar las recomendaciones de resistencia de zona para esos servicios.

Cost

No hay ningún cargo adicional para habilitar la compatibilidad con la zona de disponibilidad en AKS. Paga por las máquinas virtuales (VM) y otros recursos que implemente en las zonas de disponibilidad.

Configurar soporte de zonas de disponibilidad

- **Cree un nuevo clúster de AKS que tenga compatibilidad con la zona de disponibilidad:** Para configurar la compatibilidad con zonas de disponibilidad, consulte Creación de un clúster de Azure Kubernetes Service (AKS) que use zonas de disponibilidad.
- **Migración:** No se puede habilitar la compatibilidad con zonas de disponibilidad después de crear un clúster. En su lugar, debe crear un nuevo clúster que tenga habilitada la compatibilidad con la zona de disponibilidad y eliminar el clúster existente.
- **Deshabilite la compatibilidad con zonas de disponibilidad:** No se puede deshabilitar la compatibilidad con zonas de disponibilidad después de crear un clúster. En su lugar, debe crear un clúster nuevo que tenga compatibilidad con la zona de disponibilidad deshabilitada y eliminar el clúster existente.

Comportamiento cuando todas las zonas están en buen estado

En esta sección se describe qué esperar cuando los clústeres de AKS están configurados para la compatibilidad con zonas de disponibilidad y todas las zonas de disponibilidad están operativas.

- **Enrutamiento de tráfico entre zonas:** Al implementar un clúster de AKS que usa zonas de disponibilidad, es importante asegurarse de que los componentes de red también son resistentes a la zona. Según los equilibradores de carga y otros componentes de red que use, es posible que tenga que configurar explícitamente los componentes para enrutar el tráfico a los nodos correctos de las zonas correctas y responder a interrupciones de zona. Para más información, consulte Consideraciones de resistencia de zona para AKS.

- **Replicación de datos entre zonas:** Si ejecuta una carga de trabajo sin estado, debe usar servicios administrados de Azure, como bases de datos de Azure, Azure Managed Redis o Azure Storage para almacenar los datos de la aplicación. Puede usar estos servicios para ayudar a garantizar que el tráfico se pueda mover entre nodos y zonas sin riesgo de pérdida de datos ni que afecte a la experiencia del usuario. Puedes usar implementaciones, servicios y sondeos de estado de Kubernetes para administrar pods sin estado y garantizar la distribución uniforme entre zonas.

  Si necesita almacenar el estado dentro del clúster mediante discos de Azure, use el almacenamiento con redundancia de zona de Azure para asegurarse de que los datos se replican en varias zonas de disponibilidad. Para obtener más información, consulte Elección del tipo de disco adecuado en función de las necesidades de la aplicación.

Comportamiento durante un fallo de zona

En esta sección se describe qué esperar cuando se produce una interrupción de zona de disponibilidad mientras los clústeres de AKS están configurados con compatibilidad con zonas de disponibilidad.

- **Detección y respuesta**: cuando se produce una interrupción de zona, el plano de control conmuta automáticamente por error. Si los grupos de nodos usan zonas de disponibilidad y siguen losprocedimientos recomendados de resistencia de zona, puedes esperar que AKS abra nodos y réplicas en las zonas que están operativas. AKS realiza esta tarea automáticamente cuando se usan soluciones administradas como cluster autoscaler o NAP. Sin el escalado automático, los nodos y las réplicas permanecen en estado *Pendiente* y esperan a que la intervención manual escale verticalmente el grupo de nodos.

  AKS también intenta reequilibrar los pods en las zonas correctas. Si decides escalar manualmente el grupo de nodos en un escenario de reducción de zona, los pods pueden dejarse en el estado *Pendiente* cuando no haya nodos disponibles en las zonas correctas. El escalado horizontal en las zonas restantes también está sujeto a la disponibilidad de la cuota y la capacidad de la SKU de máquina virtual que uses.

- **Notificación**: Microsoft no le notifica automáticamente cuando una zona está inactiva. Sin embargo, puede usar Azure Resource Health para supervisar el estado de un recurso

individual y puede configurar alertas de Resource Health para notificarle problemas. También puede usar Azure Service Health para comprender el estado general del servicio, incluidos los errores de zona, y puede configurar alertas de Service Health para notificarle problemas.

También puede usar las métricas de estado del nodo o pod para supervisar el estado de los nodos y pods.

- **Solicitudes activas:** Las solicitudes activas pueden experimentar interrupciones. Algunas solicitudes pueden fallar, y la latencia podría aumentar mientras la carga de trabajo se transfiere a otra zona.
- **Pérdida de datos esperada:** Si almacena el estado dentro del clúster mediante discos de Azure y usa almacenamiento con redundancia de zona, no se espera que un error de zona cause ninguna pérdida de datos.
- **Tiempo de inactividad esperado:** Si configura correctamente la resistencia de zona para el clúster y los pods, no se espera que un error de zona cause tiempo de inactividad para la carga de trabajo de AKS.
- **Redireccionamiento del tráfico:** los equilibradores de carga redirigen las nuevas solicitudes entrantes a pods que se ejecutan en nodos saludables.

Para más información, consulte Consideraciones de resistencia de zona para AKS.

Recuperación de zona

Cuando se recupera la zona de disponibilidad, el comportamiento de la conmutación por recuperación depende del componente:

- **Plano de control:** AKS restaura automáticamente las operaciones del plano de control en todas las zonas de disponibilidad. No se requiere intervención manual.
- **Grupos de nodos y nodos**: inmediatamente después de la conmutación por recuperación, los nodos permanecen en las zonas previamente correctas y no se restauran en la zona recuperada. Pero la próxima vez que se realice una operación de escalado de nodos, como al escalar horizontalmente el grupo de nodos, el grupo de nodos puede crear nodos en la zona recuperada.
- **Pods**: inmediatamente después de la conmutación por recuperación, los pods siguen ejecutándose en los nodos en los que se están ejecutando. Cuando se crean nuevos pods o se vuelven a crear los pods existentes, son aptos para usar nodos en la zona recuperada.

- **Almacenamiento:** cualquier almacenamiento conectado a pods se recupera en función de cómo funciona el almacenamiento con redundancia de zona.

Prueba de fallos de zona

Puede probar la resistencia a los errores de zona de disponibilidad mediante los métodos siguientes:

- Aislar y drenar nodos en una única zona de disponibilidad
- Simulación de un error de zona de disponibilidad mediante Azure Chaos Studio

Resistencia a errores en toda la región

Los clústeres de AKS son recursos de una sola región. Si la región no está disponible, el clúster de AKS tampoco está disponible.

Soluciones personalizadas de varias regiones para la resistencia

Si necesita implementar la carga de trabajo de Kubernetes en varias regiones de Azure, tiene dos opciones para administrar la orquestación de estos clústeres.

- Use Fleet Manager de Azure Kubernetes si desea una experiencia administrada más sencilla. Con Fleet Manager de Azure Kubernetes, puede hacer lo siguiente:
  - Administre un conjunto de clústeres de AKS como una sola unidad y esos clústeres se pueden distribuir entre varias regiones de Azure.
  - Automatice aspectos específicos de la administración de clústeres, como las actualizaciones de imágenes de clúster y nodo.
  - Utilice las capacidades de distribución de tráfico para repartir el tráfico entre los clústeres y realizar una conmutación automática si una región no está disponible.
- Organice la conmutación por error mediante un modelo de implementación activo-activo o activo-pasivo manual si la carga de trabajo requiere un control más matizado sobre los distintos componentes de las conmutaciones por error interregionales. Para obtener más información, consulte Información general sobre alta disponibilidad y recuperación ante desastres para AKS.

Copias de seguridad y restauración

Azure Backup tiene una extensión que puede usar para realizar copias de seguridad de los recursos del clúster de AKS y volúmenes persistentes que se conectan al clúster. El almacén de Backup se comunica con el clúster de AKS a través de la extensión para realizar operaciones de copia de seguridad y restauración.

Si el clúster de AKS está en una región emparejada, puede configurar las copias de seguridad que se almacenarán en el almacenamiento con redundancia geográfica. Puede restaurar copias de seguridad georedundantes en la región emparejada.

Para obtener más información, consulte los artículos siguientes:

- ¿Qué es la copia de seguridad de Azure Kubernetes Service?
- Copia de seguridad de AKS mediante Azure Backup

Para la mayoría de las soluciones, no debe confiar exclusivamente en copias de seguridad. En su lugar, utilice las otras capacidades descritas en esta guía para apoyar los requisitos de resiliencia. Sin embargo, las copias de seguridad protegen contra algunos riesgos que otros enfoques no. Para más información, consulte ¿Qué son la redundancia, la replicación y la copia de seguridad?.

Intente usar clústeres sin estado que minimicen la necesidad de copia de seguridad. Almacene datos en bases de datos y sistemas de almacenamiento externos en lugar de en el clúster.

Resistencia al mantenimiento del servicio

AKS realiza el mantenimiento en el clúster, incluidas las actualizaciones del clúster y las imágenes de nodo. Para asegurarse de que Kubernetes mantiene el número mínimo de instancias de pod necesarias para atender el tráfico de producción incluso durante las actualizaciones, debe configurar los pods para que usen presupuestos de interrupciones del pod.

Para reducir las interrupciones del servicio durante períodos de tiempo críticos, AKS proporciona controles para que pueda especificar los tiempos de mantenimiento planeados. Para más información, consulte Uso del mantenimiento planeado para programar y controlar las actualizaciones del clúster de Azure Kubernetes Service.

Acuerdo de nivel de servicio

El contrato de nivel de servicio (SLA) para los servicios de Azure describe la disponibilidad esperada de cada servicio y las condiciones que la solución deberá cumplir para lograr esa expectativa de disponibilidad. Para obtener más información, consulte Acuerdos de Nivel de Servicio para servicios en línea.

AKS proporciona tres planes de tarifa para la administración de clústeres: **Gratis**, **Estándar** y **Premium**. El nivel Gratis permite usar AKS para probar las cargas de trabajo. Los niveles Estándar y Premium están diseñados para cargas de trabajo de producción. Al implementar un clúster de AKS con zonas de disponibilidad habilitadas, aumenta el porcentaje de tiempo de actividad definido en el Acuerdo de Nivel de Servicio. Sin embargo, el Acuerdo de Nivel de Servicio solo se aplica si implementa un clúster en el plan de tarifa Estándar o Premium.

Contenido relacionado

- Confiabilidad en Azure

Last updated on 22/01/2026

Implementación de un clúster de Azure Service Fabric en Availability Zones

Las zonas de disponibilidad de Azure son una función de alta disponibilidad que protege las aplicaciones y los datos frente a errores del centro de datos. Una zona de disponibilidad es una ubicación física única equipada con alimentación independiente, refrigeración y redes dentro de una región de Azure.

Para admitir clústeres que abarcan distintas instancias de Availability Zones, Azure Service Fabric proporciona los dos métodos de configuración que se describen en el artículo siguiente. Las zonas de disponibilidad solo se ofrecen en algunas regiones. Para más información, consulte la información general sobre las zonas de disponibilidad.

Hay plantillas de ejemplo disponibles en el repositorio de plantillas para uso entre zonas de disponibilidad de Service Fabric.

Topología para abarcar un tipo de nodo principal en diversas zonas de disponibilidad

> **Nota:**
>
> La ventaja de abarcar el tipo de nodo principal entre zonas de disponibilidad solo se ve para tres zonas y no solo dos.

- El nivel de confiabilidad del clúster ha de establecerse en `Platinum`.
- Un único recurso de IP pública que use la SKU Estándar.
- Un único recurso de Load Balancer que use la SKU Estándar.
- Un grupo de seguridad de red (NSG) referenciado por la subred en la que implementar los conjuntos de escalado de máquinas virtuales.

> **Nota:**
>
> La propiedad de grupo de selección única del conjunto de escalado de máquinas virtuales debe haberse establecido en `true`.

La siguiente lista de nodos de ejemplo muestra los formatos FD/UD en un conjunto de escalado de máquinas virtuales que expande zonas:

ALL NODES EVENTS

Search list Reset All

Name	Address	Node Type	Upgrade Domain	Fault Domain	Is Seed Node
_mgmt_0	10.0.0.16	mgmt	ud:/zone1/0	fd:/zone1/0	true
_mgmt_1	10.0.0.17	mgmt	ud:/zone2/0	fd:/zone2/0	false
_mgmt_10	10.0.0.18	mgmt	ud:/zone3/3	fd:/zone3/3	false
_mgmt_11	10.0.0.19	mgmt	ud:/zone1/3	fd:/zone1/3	false
_mgmt_12	10.0.0.20	mgmt	ud:/zone2/4	fd:/zone2/4	true
_mgmt_13	10.0.0.21	mgmt	ud:/zone3/4	fd:/zone3/4	false

Distribución de réplicas de servicio entre zonas

Al implementar un servicio en los tipos de nodo que expanden Availability Zones, las réplicas se incluyen para garantizar que se encuentren en zonas independientes. Los dominios de error en los nodos de cada uno de estos tipos de nodo se configuran con la información de zona (es decir, FD = fd:/zone1/1, etc.). Por ejemplo, para cinco réplicas o instancias de servicio, la distribución es 2-2-1 y el entorno de ejecución intentará garantizar una distribución uniforme entre zonas.

Configuración de la réplica de servicio de usuario

Los servicios de usuario con estado implementados en los tipos de nodo entre Availability Zones deben configurarse de esta forma: número de réplicas con destino = 9, min = 5. Esta configuración permite que el servicio funcione incluso cuando una zona deje de estar disponible, ya que seguirán habiendo seis réplicas activas en las otras dos zonas. Una actualización de la aplicación en este escenario también tendrá éxito.

Elemento ReliabilityLevel del clúster

Este valor define el número de nodos raíz del clúster y el tamaño de réplica de los servicios del sistema. Una configuración entre zonas de disponibilidad tiene un mayor número de nodos, que se distribuyen entre zonas para habilitar la resistencia de zona.

Un valor `ReliabilityLevel` más elevado garantiza que haya presentes más nodos raíz y réplicas de servicio del sistema y se distribuyan uniformemente entre zonas, de tal manera que si se produce un error en una zona, el clúster y los servicios del sistema no se verán afectados. El valor `ReliabilityLevel = Platinum` (recomendado) garantiza que nueve nodos raíz se extiendan entre zonas del clúster, con tres valores en cada zona.

Escenario de zona fuera de servicio

Cuando una zona está fuera de servicio, todos los nodos y las réplicas de servicio de esa zona aparecen fuera de servicio. Como hay réplicas en las demás zonas, el servicio seguirá respondiendo. Las réplicas principales conmutan por error en las zonas funcionales. Los servicios están en estado de advertencia aparentemente porque el número de réplicas de destino aún no se ha alcanzado y el número de máquinas virtuales sigue siendo mayor que el tamaño mínimo de réplica de destino.

El equilibrador de carga de Service Fabric abrirá las réplicas en las zonas funcionales para que coincidan con el número de réplicas de destino. En este momento, los servicios aparecerán correctamente. Cuando la zona que estaba fuera de servicio vuelve a ser funcional, el equilibrador de carga redistribuye todas las réplicas de servicio uniformemente entre las zonas.

Próximas optimizaciones

- Para proporcionar actualizaciones de infraestructura confiables, Service Fabric requiere que la durabilidad del conjunto de escalado de máquinas virtuales se establezca al menos en Silver. Esto permite que el conjunto de escalado de máquinas virtuales subyacente y el entorno de ejecución de Service Fabric proporcionen actualizaciones confiables. Esto también requiere que cada zona tenga como mínimo 5 VM. Estamos trabajando para bajar este requisito a 3 y 2 VM por zona para los tipos de nodo principales y no principales, respectivamente.
- Todas las configuraciones mencionadas a continuación y el próximo trabajo proporcionan migración local a los clientes donde se puede actualizar el mismo clúster para usar la nueva configuración mediante la adición de nuevos elementos nodeTypes y la retirada de los antiguos.

Requisitos de red

Recurso de IP pública y equilibrador de carga

Para habilitar la propiedad `zones` en un recurso de conjunto de escalado de máquinas virtuales, el recurso de equilibrador de carga e IP referenciado por ese conjunto de escalado de máquinas virtuales deben utilizar una SKU Estándar. La creación de un recurso de IP sin la propiedad SKU da lugar a un SKU Básico, que no admite Availability Zones. De forma predeterminada, un equilibrador de carga de SKU Estándar bloquea todo el tráfico desde fuera. Para permitir el tráfico externo, implemente un NSG en la subred.

JSON

```
{
  "apiVersion": "2018-11-01",
```

```
  "type": "Microsoft.Network/publicIPAddresses",
  "name": "[concat('LB','-', parameters('clusterName')]",
  "location": "[parameters('computeLocation')]",
  "sku": {
    "name": "Standard"
  }
}
```

JSON

```
{
  "apiVersion": "2018-11-01",
  "type": "Microsoft.Network/loadBalancers",
  "name": "[concat('LB','-', parameters('clusterName')]",
  "location": "[parameters('computeLocation')]",
  "dependsOn": [
    "[concat('Microsoft.Network/networkSecurityGroups/', concat('nsg',
parameters('subnet0Name')))]"
  ],
  "properties": {
    "addressSpace": {
      "addressPrefixes": [
        "[parameters('addressPrefix')]"
      ]
    },
    "subnets": [
      {
        "name": "[parameters('subnet0Name')]",
        "properties": {
          "addressPrefix": "[parameters('subnet0Prefix')]",
          "networkSecurityGroup": {
            "id": "[resourceId('Microsoft.Network/networkSecurityGroups',
concat('nsg', parameters('subnet0Name')))]"
          }
        }
      }
    ]
  },
  "sku": {
    "name": "Standard"
  }
}
```

> ⓘ **Nota:**
>
> No es posible realizar un cambio local de SKU en los recursos de IP pública. Si va a hacer una migración a partir de recursos existentes que tienen un SKU Básico, consulte la sección sobre migración en este artículo.

Reglas NAT para conjuntos de escalado de máquinas virtuales

Las reglas de traducción de direcciones de red (NAT) de entrada para el equilibrador de carga deben coincidir con los grupos NAT del conjunto de escalado de máquinas virtuales. Cada conjunto de escalado de máquinas virtuales debe tener un único grupo NAT de entrada.

JSON

```
{
  "inboundNatPools": [
    {
      "name": "LoadBalancerBEAddressNatPool0",
      "properties": {
        "backendPort": "3389",
        "frontendIPConfiguration": {
          "id": "[variables('lbIPConfig0')]"
        },
        "frontendPortRangeEnd": "50999",
        "frontendPortRangeStart": "50000",
        "protocol": "tcp"
      }
    },
    {
      "name": "LoadBalancerBEAddressNatPool1",
      "properties": {
        "backendPort": "3389",
        "frontendIPConfiguration": {
          "id": "[variables('lbIPConfig0')]"
        },
        "frontendPortRangeEnd": "51999",
        "frontendPortRangeStart": "51000",
        "protocol": "tcp"
      }
    },
    {
      "name": "LoadBalancerBEAddressNatPool2",
      "properties": {
        "backendPort": "3389",
        "frontendIPConfiguration": {
          "id": "[variables('lbIPConfig0')]"
        },
        "frontendPortRangeEnd": "52999",
        "frontendPortRangeStart": "52000",
        "protocol": "tcp"
      }
    }
  ]
}
```

Reglas de salida para un equilibrador de carga de SKU Estándar

La IP pública del SKU Estándar presenta nuevas funciones y comportamientos diferentes para la conectividad de salida en comparación con el uso de SKU básicos. Si desea obtener conectividad de salida cuando trabaje con las SKU estándar, debe definirlas explícitamente con una dirección IP pública o un equilibrador de carga de SKU Estándar. Para más información, consulte Conexiones de salida y ¿Qué es Azure Load Balancer?

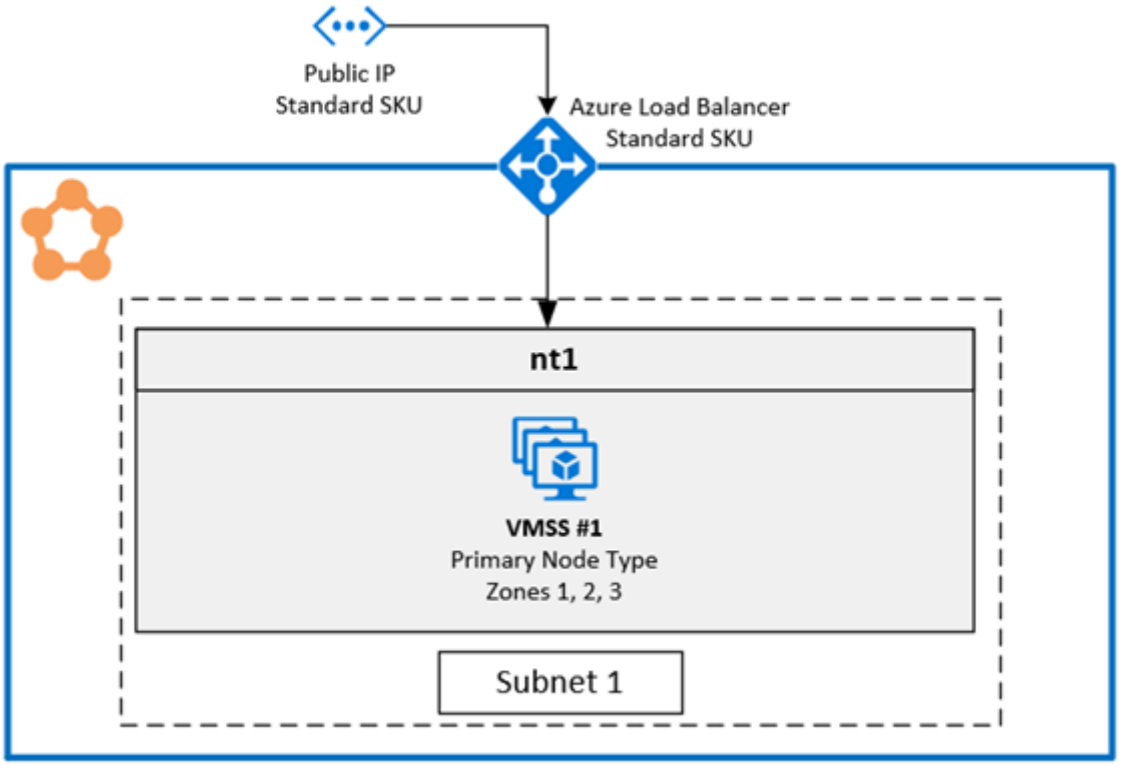
> **Nota:**
>
> La plantilla estándar hace referencia a un NSG que permite todo el tráfico saliente de forma predeterminada. El tráfico de entrada se limita a los puertos necesarios para las operaciones de administración de Service Fabric. Las reglas de NSG pueden modificarse para satisfacer sus requisitos.

> **Importante**
>
> Cada tipo de nodo de un clúster de Service Fabric que use un equilibrador de carga de SKU Estándar requiere una regla que permita el tráfico saliente en el puerto 443. Esto es necesario para completar la configuración del clúster. Cualquier implementación que no incluya esta regla generará un error.

1. Habilitación de varias zonas de disponibilidad en un único conjunto de escalado de máquinas virtuales

Esta solución permite a los usuarios abarcar tres instancias de Availability Zones en el mismo tipo de nodo. Se trata de la topología de implementación recomendada, ya que permite realizar la implementación entre zonas de disponibilidad mientras se mantiene un único conjunto de escalado de máquinas virtuales.

GitHub tiene una plantilla de ejemplo a su disposición.



Configuración de zonas en un conjunto de escalado de máquinas virtuales

Para habilitar zonas en un conjunto de escalado de máquinas virtuales, incluya los tres valores siguientes en el recurso del conjunto de escalado de máquinas virtuales:

- El primer valor es la propiedad `zones`, que especifica las zonas de disponibilidad que están presentes en el conjunto de escalado de máquinas virtuales.
- El segundo valor es la propiedad `singlePlacementGroup`, que se debe establecer en `true`. El conjunto de escalado que se extiende entre tres zona de disponibilidad puede admitir hasta 300 máquinas virtuales incluso con `singlePlacementGroup = true`.
- El tercer valor es `zoneBalance`, que garantiza el equilibrio de zona estricta. Este valor debe ser `true`. Esto garantiza que las distribuciones de máquinas virtuales entre zonas no sean desiguales, lo cual significa que cuando una de las zonas deja de funcionar, las otras dos zonas disponen de suficientes máquinas virtuales para que el clúster sigue ejecutándose.

  Es posible que un clúster con una distribución desigual de máquinas virtuales no sobreviva a un escenario de zona fuera de servicio, ya que esa zona podría acaparar la mayoría de las máquinas virtuales. La distribución desigual de máquinas virtuales entre zonas también provoca problemas relacionados con la prestación de servicios y el bloqueo de las actualizaciones de infraestructura. Más información sobre zoneBalancing.

No es necesario configurar las invalidaciones `FaultDomain` y `UpgradeDomain`.

JSON

```
{
  "apiVersion": "2018-10-01",
  "type": "Microsoft.Compute/virtualMachineScaleSets",
  "name": "[parameters('vmNodeType1Name')]",
  "location": "[parameters('computeLocation')]",
  "zones": [ "1", "2", "3" ],
  "properties": {
    "singlePlacementGroup": true,
    "zoneBalance": true
  }
}
```

Nota:

- Los clústeres de Service Fabric deben tener al menos un tipo principal de nodo. El nivel de durabilidad de los tipos de nodo principal debe ser Silver o superior.
- Una Azure Availability Zones de disponibilidad que extiende los conjuntos de escalado de máquinas virtuales debe configurarse con tres zona de disponibilidad como mínimo, independientemente del nivel de durabilidad.
- Una zona de disponibilidad que extienden conjuntos de escalado de máquinas virtuales con el nivel de durabilidad Silver o superior deben tener al menos 15 máquinas virtuales (5 por región).
- La zona de disponibilidad que abarca el conjunto de escalado de máquinas virtuales con durabilidad Bronze debe tener al menos 6 máquinas virtuales.

Habilitación de la compatibilidad con varias zonas en el tipo de nodo de Service Fabric

El tipo de nodo de Service Fabric debe estar habilitado para admitir varias zonas de disponibilidad.

- El primer valor es `multipleAvailabilityZones`, que se debe establecer en `true` para el tipo de nodo.
- El segundo valor es `sfZonalUpgradeMode` y es opcional. Esta propiedad no se puede modificar si ya existe un tipo de nodo con varias zonas de disponibilidad en el clúster. Esta propiedad determina la agrupación lógica de máquinas virtuales en los dominios de actualización (UD).

- Si este valor se establece en `Parallel`, las máquinas virtuales del tipo de nodo se agrupan en UD y omiten la información de zona en cinco UD. Esta configuración hace que los UD de todas las zonas se actualicen al mismo tiempo. Este modo de implementación es más rápido para las actualizaciones. No se recomienda su uso, ya que se trata de instrucciones SDP que indican que las actualizaciones deben aplicarse en una zona cada vez.
- Si el valor se omite o se establece en `Hierarchical`, las máquinas virtuales se agruparán para reflejar la distribución de zonas en un máximo de 15 UD. Cada una de las tres zonas tiene cinco UD. De este modo, se garantiza que las zonas se actualicen de una en una y que solo se pase a la zona siguiente después de completar cinco UD en la primera zona. Este proceso de actualización es más seguro para el clúster y la aplicación de usuario.

La propiedad solo define el comportamiento de actualización de la aplicación Service Fabric y las actualizaciones de código. Las actualizaciones del conjunto subyacente de escalado de máquinas virtuales seguirán produciéndose simultáneamente en todas las zonas de disponibilidad. Esta propiedad no afecta a la distribución de UD para los tipos de nodo que no tengan habilitadas varias zonas.

- El tercer valor es `vmssZonalUpgradeMode`, es opcional y se puede actualizar en cualquier momento. Esta propiedad define el esquema de actualización para que el conjunto de escalado de máquinas virtuales se pueda realizar en paralelo o secuencialmente en Availability Zones.
  - Si este valor se establece en `Parallel`: todas las actualizaciones del conjunto de escalado se suceden en paralelo en todas las zonas. Este modo de implementación es más rápido para las actualizaciones. No se recomienda su uso, ya que se trata de instrucciones SDP que indican que las actualizaciones deben aplicarse en una zona cada vez.
  - Si este valor se omite o se establece en `Hierarchical`: se garantiza que las zonas se actualicen de una en una y que solo se pase a la zona siguiente después de completar cinco UD en la primera zona. Este proceso de actualización es más seguro para el clúster y la aplicación de usuario.

> **ⓘ Importante**
>
> La versión de API del recurso de clúster de Service Fabric debe ser la 2020-12-01 (versión preliminar) o superior.
>
> La versión del código del clúster debe ser 8.1.321 o posterior.

JSON

```
{
  "apiVersion": "2020-12-01-preview",
  "type": "Microsoft.ServiceFabric/clusters",
  "name": "[parameters('clusterName')]",
  "location": "[parameters('clusterLocation')]",
  "dependsOn": [
    "[concat('Microsoft.Storage/storageAccounts/',
parameters('supportLogStorageAccountName'))]"
  ],
  "properties": {
    "reliabilityLevel": "Platinum",
    "sfZonalUpgradeMode": "Hierarchical",
    "vmssZonalUpgradeMode": "Parallel",
    "nodeTypes": [
      {
        "name": "[parameters('vmNodeType0Name')]",
        "multipleAvailabilityZones": true
      }
    ]
  }
}
```

Nota:

- Los recursos de IP pública y equilibrar la carga deben usar la SKU Estándar, tal como se describió anteriormente en el artículo.
- La propiedad `multipleAvailabilityZones` del tipo de nodo solo se puede definir cuando se haya creado el tipo de nodo y no se podrá modificar más adelante. Los tipos de nodo existentes no se pueden configurar con esta propiedad.
- Cuando `sfZonalUpgradeMode` se omite o se establece en `Hierarchical`, las implementaciones de clúster y aplicación serán más lentas porque habrá más dominios de actualización en el clúster. Es importante ajustar correctamente los tiempos de inactividad de la directiva de actualización para tener en cuenta el tiempo de actualización necesario para 15 dominios de actualización. La directiva de actualización para la aplicación y el clúster debe actualizarse para garantizar que la implementación no supere el límite de tiempo de implementación de 12 horas de Azure Resource Service. Esto significa que la implementación no debe tardar más de 12 horas para 15 UD (es decir, no debe tardar más de 40 minutos con cada UD).
- Establezca el nivel de confiabilidad del clúster en `Platinum` para garantizar que el clúster sobreviva al escenario con una zona fuera de servicio.
- No se admite la actualización de DurabilityLevel para un tipo de nodo con multipleAvailabilityZones. En su lugar, cree un tipo de nodo con mayor durabilidad.

- SF solo admite tres AvailabilityZones. Por el momento, no se admite ningún número mayor.

Sugerencia

Se recomienda establecer `sfZonalUpgradeMode` en `Hierarchical`, o bien omitirlo. La implementación seguirá la distribución de zonas de las máquinas virtuales y afecta a un número más reducido de réplicas o instancias, por lo estas son más seguras. Use `sfZonalUpgradeMode` establecido en `Parallel` si la velocidad de implementación es una prioridad o solo se ejecuta una carga de trabajo sin estado en el tipo de nodo con varias zonas de disponibilidad. Esto hará que el recorrido del dominio de actualización tenga lugar al mismo tiempo en todas las zonas de disponibilidad.

Migración al tipo de nodo con varias zonas de disponibilidad

En todos los escenarios de migración, es necesario agregar un nuevo tipo de nodo que admita varias zonas de disponibilidad. No se puede migrar un tipo de nodo existente para admitir varias zonas. El artículo Escalado vertical de un del tipo de nodo principal de un clúster de Service Fabric incluye pasos detallados para agregar un nuevo tipo de nodo y los demás recursos necesarios para el nuevo tipo de nodo, como los recursos de IP y equilibrador de carga. En el artículo se describe también cómo retirar el tipo de nodo existente después de agregar al clúster un nuevo tipo de nodo con varias zonas de disponibilidad.

- Migración desde un tipo de nodo que usa recursos de IP básicos: este proceso ya se describe en una subsección posterior para la solución con un tipo de nodo por zona de disponibilidad.

  Para el nuevo tipo de nodo, la única diferencia es que solo hay un conjunto de escalado de máquinas virtuales y un tipo de nodo para todas las zonas de disponibilidad, en lugar de uno por zona de disponibilidad.

- Migración desde un tipo de nodo que usa los recursos de IP de SKU Estándar con un NSG: siga el procedimiento descrito anteriormente. Sin embargo, no es necesario agregar nuevos recursos de IP y NSG. Se pueden reutilizar los mismos recursos en el nuevo tipo de nodo.

2. Implementación de zonas mediante el anclaje de un conjunto de escalado de máquinas virtuales a

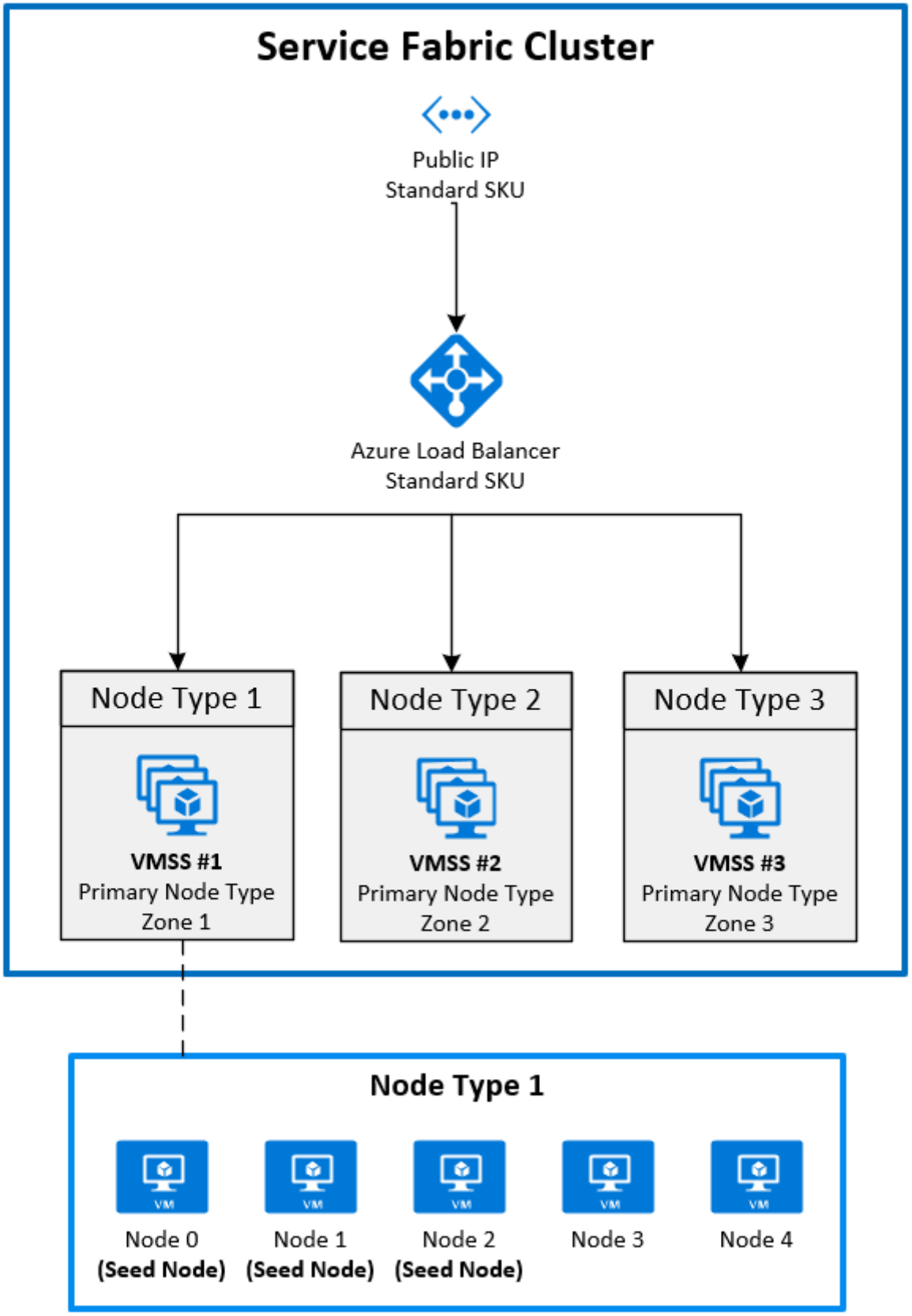
cada zona

Esta es la configuración disponible con carácter general en este momento. Para extender un clúster de Service Fabric entre Availability Zones, debe crear un tipo de nodo principal en cada zona de disponibilidad admitida por la región. De este modo, se distribuyen nodos raíz de manera uniforme entre todos los tipos de nodos principales.

La topología recomendada para el tipo de nodo principal requiere lo siguiente:

- Tres tipos de nodo deben marcarse como principales.
  - Cada tipo de nodo debe asignarse a su propio conjunto de escalado de máquinas virtuales ubicado en distintas zonas.
  - Cada conjunto de escalado de máquinas virtuales debe tener al menos cinco nodos (durabilidad Silver).

El diagrama siguiente muestra la arquitectura de la zona de disponibilidad de Azure Service Fabric.



Habilitación de zonas en un conjunto de escalado de máquinas virtuales

Para habilitar una zona en un conjuntos de escalado de máquinas virtuales, incluya estos tres valores en el recurso del conjunto de escalado de máquinas virtuales:

- El primer valor es la propiedad `zones`, que especifica en qué zona de disponibilidad se implementará el conjunto de escalado de máquinas virtuales.
- El segundo valor es la propiedad `singlePlacementGroup`, que se debe establecer en `true`.
- El tercer valor es la propiedad `faultDomainOverride` en la extensión de conjunto de escalado de máquinas virtuales de Service Fabric. Esta propiedad debe incluir solo la zona en la que se incluirá este conjunto de escalado de máquinas virtuales. Ejemplo: `"faultDomainOverride": "az1"`. Todos los recursos del conjunto de escalado de máquinas virtuales se deben incluir en la misma región porque los clústeres de Azure Service Fabric no ofrecen compatibilidad entre regiones.

JSON

```
{
  "apiVersion": "2018-10-01",
  "type": "Microsoft.Compute/virtualMachineScaleSets",
  "name": "[parameters('vmNodeType1Name')]",
  "location": "[parameters('computeLocation')]",
  "zones": [
    "1"
  ],
  "properties": {
    "singlePlacementGroup": true
  },
  "virtualMachineProfile": {
    "extensionProfile": {
      "extensions": [
        {
          "name": "[concat(parameters('vmNodeType1Name'),'_ServiceFabricNode')]",
          "properties": {
            "type": "ServiceFabricNode",
            "autoUpgradeMinorVersion": false,
            "publisher": "Microsoft.Azure.ServiceFabric",
            "settings": {
              "clusterEndpoint": "
[reference(parameters('clusterName')).clusterEndpoint]",
              "nodeTypeRef": "[parameters('vmNodeType1Name')]",
              "dataPath": "D:\\\\SvcFab",
              "durabilityLevel": "Silver",
              "certificate": {
                "thumbprint": "[parameters('certificateThumbprint')]",
                "x509StoreName": "[parameters('certificateStoreValue')]"
              },
              "systemLogUploadSettings": {
                "Enabled": true
              },
              "faultDomainOverride": "az1"
            },
```

```
            "typeHandlerVersion": "1.0"
          }
        }
      ]
    }
  }
}
```

Habilitación de varios tipos de nodos principales en el recurso de clúster de Service Fabric

Para establecer uno o más tipos de nodo como principales en un recurso de clúster, establezca la propiedad `isPrimary` en `true`. Al implementar un clúster de Service Fabric en Availability Zones, debe tener tres tipos de nodo en distintas zonas.

JSON

```
{
  "reliabilityLevel": "Platinum",
  "nodeTypes": [
    {
      "name": "[parameters('vmNodeType0Name')]",
      "applicationPorts": {
        "endPort": "[parameters('nt0applicationEndPort')]",
        "startPort": "[parameters('nt0applicationStartPort')]"
      },
      "clientConnectionEndpointPort": "[parameters('nt0fabricTcpGatewayPort')]",
      "durabilityLevel": "Silver",
      "ephemeralPorts": {
        "endPort": "[parameters('nt0ephemeralEndPort')]",
        "startPort": "[parameters('nt0ephemeralStartPort')]"
      },
      "httpGatewayEndpointPort": "[parameters('nt0fabricHttpGatewayPort')]",
      "isPrimary": true,
      "vmInstanceCount": "[parameters('nt0InstanceCount')]"
    },
    {
      "name": "[parameters('vmNodeType1Name')]",
      "applicationPorts": {
        "endPort": "[parameters('nt1applicationEndPort')]",
        "startPort": "[parameters('nt1applicationStartPort')]"
      },
      "clientConnectionEndpointPort": "[parameters('nt1fabricTcpGatewayPort')]",
      "durabilityLevel": "Silver",
      "ephemeralPorts": {
        "endPort": "[parameters('nt1ephemeralEndPort')]",
        "startPort": "[parameters('nt1ephemeralStartPort')]"
      },
      "httpGatewayEndpointPort": "[parameters('nt1fabricHttpGatewayPort')]",
      "isPrimary": true,
      "vmInstanceCount": "[parameters('nt1InstanceCount')]"
```

```
    },
    {
      "name": "[parameters('vmNodeType2Name')]",
      "applicationPorts": {
        "endPort": "[parameters('nt2applicationEndPort')]",
        "startPort": "[parameters('nt2applicationStartPort')]"
      },
      "clientConnectionEndpointPort": "[parameters('nt2fabricTcpGatewayPort')]",
      "durabilityLevel": "Silver",
      "ephemeralPorts": {
        "endPort": "[parameters('nt2ephemeralEndPort')]",
        "startPort": "[parameters('nt2ephemeralStartPort')]"
      },
      "httpGatewayEndpointPort": "[parameters('nt2fabricHttpGatewayPort')]",
      "isPrimary": true,
      "vmInstanceCount": "[parameters('nt2InstanceCount')]"
    }
  ]
}
```

Migración a Availability Zones desde un clúster mediante una dirección IP de SKU Básico

Para migrar un clúster que usa una dirección IP con un SKU Básico, primero debe crear un recurso de IP completamente nuevo mediante el SKU Estándar. No es posible actualizar estos recursos.

Haga referencia a la IP en los nuevos tipos de nodo entre zonas de disponibilidad que desea usar. En el ejemplo anterior, se agregaron tres nuevos recursos de conjunto de escalado de máquinas virtuales en las zonas 1, 2 y 3. Estos conjuntos de escalado de máquinas virtuales hacen referencia a la IP recién creada y se identifican como tipos de nodo principal en el recurso de clúster de Service Fabric.

1. Para empezar, agregue los nuevos recursos a la plantilla existente de Azure Resource Manager. Estos recursos incluyen:

    - Un recurso de IP pública que usa la SKU Estándar.
    - Un recurso de equilibrador de carga que usa la SKU Estándar.
    - Un NSG referenciado por la subred en la que se implementan los conjuntos de escalado de máquinas virtuales.
    - Tres tipos de nodo deben marcarse como principales.
        - Cada tipo de nodo debe asignarse a su propio conjunto de escalado de máquinas virtuales ubicado en distintas zonas.
        - Cada conjunto de escalado de máquinas virtuales debe tener al menos cinco nodos (durabilidad Silver).

Puede encontrar un ejemplo de estos recursos en la plantilla de ejemplo.

PowerShell

```
New-AzureRmResourceGroupDeployment `
    -ResourceGroupName $ResourceGroupName `
    -TemplateFile $Template `
    -TemplateParameterFile $Parameters
```

2. Una vez que ha terminado la implementación de los recursos, puede deshabilitar los nodos en el tipo de nodo principal del clúster original. Cuando los nodos estén deshabilitados, los servicios del sistema se migrarán al nuevo tipo de nodo principal que implementó anteriormente.

PowerShell

```
Connect-ServiceFabricCluster -ConnectionEndpoint $ClusterName `
    -KeepAliveIntervalInSec 10 `
    -X509Credential `
    -ServerCertThumbprint $thumb  `
    -FindType FindByThumbprint `
    -FindValue $thumb `
    -StoreLocation CurrentUser `
    -StoreName My

Write-Host "Connected to cluster"

$nodeNames = @("_nt0_0", "_nt0_1", "_nt0_2", "_nt0_3", "_nt0_4")

Write-Host "Disabling nodes..."
foreach($name in $nodeNames) {
    Disable-ServiceFabricNode -NodeName $name -Intent RemoveNode -Force
}
```

3. Una vez que todos los nodos están deshabilitados, los servicios del sistema se ejecutarán en el tipo de nodo principal, que se extiende entre zonas. A continuación, puede quitar los nodos deshabilitados del clúster. Después de quitar los nodos, puede quitar los recursos de IP original, de equilibrador de carga y de conjunto de escalado de máquinas virtuales.

PowerShell

```
foreach($name in $nodeNames){
    # Remove the node from the cluster
    Remove-ServiceFabricNodeState -NodeName $name -TimeoutSec 300 -Force
    Write-Host "Removed node state for node $name"
}

$scaleSetName="nt0"
```

```
Remove-AzureRmVmss -ResourceGroupName $groupname -VMScaleSetName $scaleSetName
-Force

$lbname="LB-cluster-nt0"
$oldPublicIpName="LBIP-cluster-0"
$newPublicIpName="LBIP-cluster-1"

Remove-AzureRmLoadBalancer -Name $lbname -ResourceGroupName $groupname -Force
Remove-AzureRmPublicIpAddress -Name $oldPublicIpName -ResourceGroupName
$groupname -Force
```

4. A continuación, quite las referencias a estos recursos de la plantilla de Resource Manager que se implementó.

5. Para finalizar, actualice el nombre DNS y la dirección IP pública.

PowerShell

```
$oldprimaryPublicIP = Get-AzureRmPublicIpAddress -Name $oldPublicIpName  -
ResourceGroupName $groupname
$primaryDNSName = $oldprimaryPublicIP.DnsSettings.DomainNameLabel
$primaryDNSFqdn = $oldprimaryPublicIP.DnsSettings.Fqdn

Remove-AzureRmLoadBalancer -Name $lbname -ResourceGroupName $groupname -Force
Remove-AzureRmPublicIpAddress -Name $oldPublicIpName -ResourceGroupName $groupname
-Force

$PublicIP = Get-AzureRmPublicIpAddress -Name $newPublicIpName  -ResourceGroupName
$groupname
$PublicIP.DnsSettings.DomainNameLabel = $primaryDNSName
$PublicIP.DnsSettings.Fqdn = $primaryDNSFqdn
Set-AzureRmPublicIpAddress -PublicIpAddress $PublicIP
```

Last updated on 22/03/2026

Recuperación ante desastres en Service Fabric de Azure

Una parte fundamental de la entrega de alta disponibilidad es garantizar que los servicios sobreviven a cualquier tipo de error. Esto es especialmente importante para los errores imprevistos y fuera de control.

En este artículo se describen algunos modos de error comunes que podrían ser desastres si no se hubieran modelado y administrado correctamente. También se explican las mitigaciones y las acciones que deben realizarse en caso de desastre. El objetivo es limitar o eliminar el riesgo de pérdida de datos o de tiempo de inactividad cuando se produzcan errores, tanto planeados como imprevistos.

Impedimento de desastres

El objetivo principal de Azure Service Fabric es ayudarle a modelar tanto el entorno como los servicios de tal manera que los tipos de error comunes no se conviertan en desastres.

En general, hay dos tipos de escenarios de desastre o error:

- Errores de hardware y software
- Errores operativos

Errores de hardware y software

Los errores de hardware y software son imprevisibles. La manera más fácil de sobrevivir a errores es ejecutar más copias del servicio más allá de los límites del hardware o software defectuoso.

Por ejemplo, si el servicio se ejecuta solamente en una máquina concreta, el error que se produce en esa máquina es un desastre para el servicio. La manera más sencilla de evitar este desastre es asegurarse de que el servicio se ejecuta en varias máquinas. Puede ser necesario realizar pruebas para garantizar que el error de una máquina no interrumpe el servicio en ejecución. El planeamiento de capacidad garantiza que se puede crear una instancia de reemplazo en otra parte y que la reducción de capacidad no sobrecarga el resto de servicios.

Este patrón funciona independientemente de dónde intente evitar el error. Por ejemplo, si le preocupa el error de una red de área de almacenamiento, debe ejecutarlo en varias redes de área de almacenamiento. Si le preocupa perder un conjunto de servidores, debe ejecutarlo en varios conjuntos. Si le preocupa la pérdida de centros de datos, el servicio debe ejecutarse en

varias regiones de Azure, varias instancias de Azure Availability Zones o entre sus propios centros de datos.

Cuando un servicio se distribuye entre varias instancias físicas (máquinas, bastidores, centros de datos o regiones), aún existe la posibilidad de experimentar algunos tipos de errores simultáneos. No obstante, los errores únicos e incluso los múltiples de un tipo determinado (por ejemplo, una sola máquina virtual o un error de vínculo de red) se controlan automáticamente, por lo que ya no son un "desastre".

Service Fabric proporciona mecanismos para expandir el clúster y se ocupa de recuperar nodos y servicios defectuosos. También permite ejecutar varias instancias de los servicios para evitar que los errores imprevistos se conviertan en verdaderos desastres.

Pueden existir razones por las que ejecutar una gran implementación para abarcar los errores no sea posible. Por ejemplo, es posible que sea necesario dedicar más recursos de hardware de los que esté dispuesto a pagar por el riesgo de error. Cuando se trabaja con aplicaciones distribuidas, los costos de los saltos de comunicación o de replicación de estados adicionales a distancia geográfica podrían provocar una latencia inaceptable. Esta línea se traza en función de cada aplicación.

En concreto, para errores de software, el error puede encontrarse en el servicio que intenta escalar. En este caso, el hecho de tener más copias no evita el desastre, ya que la condición de error se encuentra en todas las instancias.

Errores operativos

Aunque se distribuya el servicio por todo el mundo con redundancias, pueden seguir produciéndose eventos de desastre. Por ejemplo, alguien podría volver a configurar por error el nombre de DNS del servicio o directamente eliminarlo.

A modo de ejemplo, supongamos que tuviera un servicio de Service Fabric con estado y que alguien lo elimina por error. A menos que haya otra mitigación, ese servicio y su estado habrán desaparecido. Estos tipos de desastres operativos ("vaya") requieren más mitigaciones y pasos para la recuperación que los errores imprevistos normales.

Lo mejor para evitar estos tipos de errores operativos es:

- Restringir el acceso operativo al entorno.
- Inspeccionar minuciosamente las operaciones peligrosas.
- Imponer la automatización, evitar los cambios manuales o fuera de banda y validar los cambios concretos en el entorno antes de establecerlos.
- Asegurarse de que las operaciones destructivas "se suavizan". Las operaciones de software suavizadas no surten efecto inmediatamente o se pueden deshacer en un plazo

determinado.

Service Fabric proporciona mecanismos para evitar errores operativos, como un control de acceso basado en rol para las operaciones del clúster. Sin embargo, la mayoría de estos errores operativos requiere esfuerzos organizativos y otros sistemas. Service Fabric proporciona mecanismos para sobrevivir a errores operativos, en especial, copia de seguridad y restauración para los servicios con estado.

Administración de errores

El objetivo de Service Fabric es la administración automática de los errores. Sin embargo, para abordar algunos tipos de errores, los servicios deben tener código adicional. Otros tipos de errores *no* deben solucionarse automáticamente por motivos de seguridad y continuidad empresarial.

Abordaje de errores únicos

Una máquina puede fallar por todo tipo de razones. En ocasiones se debe al hardware, como los errores en las fuentes de alimentación o el hardware de las redes. Otros errores son de software. Se trata de errores en el sistema operativo y el propio servicio. Service Fabric detecta automáticamente estos tipos de errores, incluidos los casos en los que la máquina se aísla de otras por problemas de red.

Independientemente del tipo de servicio, ejecutar una sola instancia se traduce en tiempo de inactividad de ese servicio si se produce un error en la copia única del código por cualquier motivo.

Para abordar cualquier error único, lo más sencillo es garantizar que los servicios se ejecutan en más de un nodo de forma predeterminada. En el caso de los servicios sin estado, asegúrese de que `InstanceCount` sea superior a 1. Para los servicios con estado, la recomendación es que `TargetReplicaSetSize` y `MinReplicaSetSize` estén establecidos en 3 como mínimo. La ejecución de más copias del código de servicio garantiza que este puede abordar cualquier error único automáticamente.

Abordaje de errores coordinados

En un clúster se pueden producir errores coordinados por errores y cambios de infraestructura planeados o imprevistos, o por cambios de software planeados. Service Fabric modela las zonas de infraestructura que experimentan errores coordinados como *dominios de error*. Las áreas que experimentarán cambios coordinados de software se modelan como *dominios de actualización*. Para más información acerca de los dominios de error, los dominios de

actualización y la topología de clúster, consulte Descripción de un clúster de Service Fabric con Cluster Resource Manager.

De forma predeterminada, Service Fabric considera los dominios de error y de actualización al planear dónde deben ejecutarse los servicios. De forma predeterminada, Service Fabric intenta asegurarse de que los servicios se ejecutan en varios dominios de error y de actualización, de manera que si se realizan cambios planeados o imprevistos, los servicios permanecen disponibles.

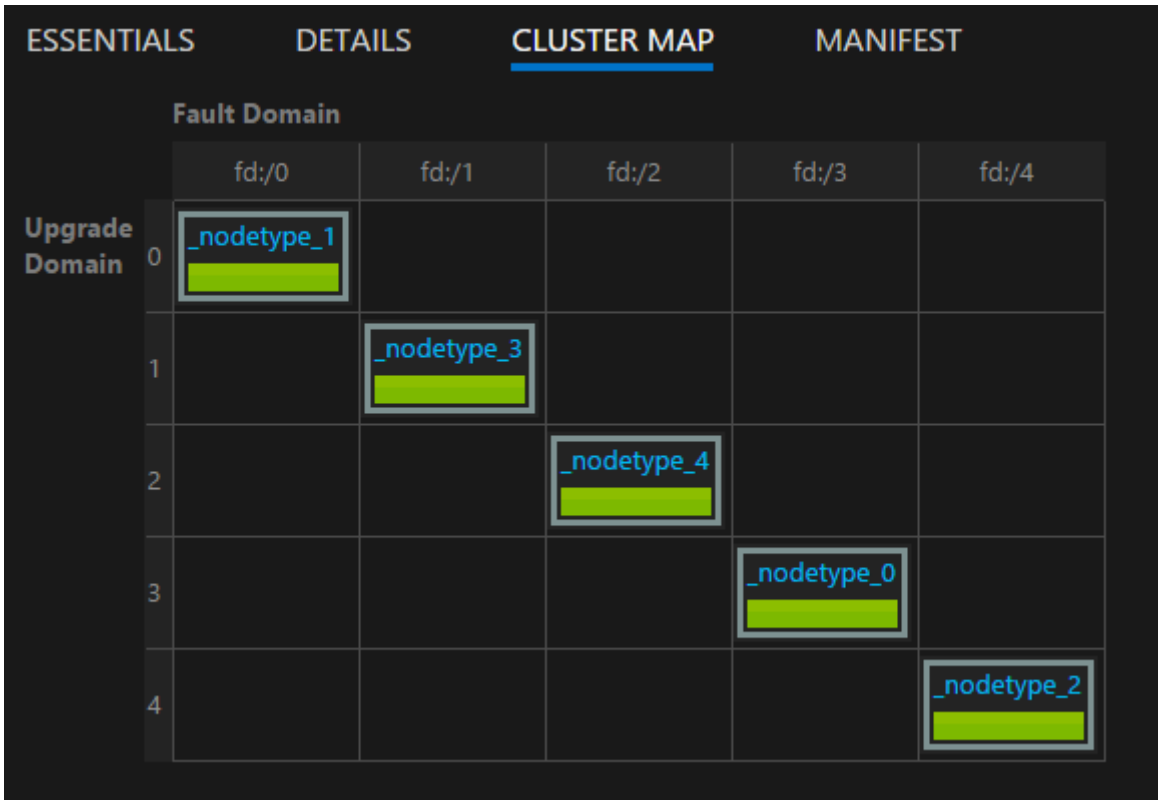
Por ejemplo, supongamos que un error en una fuente de alimentación provoca un error simultáneo en todas las máquinas de un bastidor. Con varias copias del servicio en ejecución, la pérdida de varias máquinas del dominio de error se convierte simplemente en otro ejemplo de error único para un servicio. Por este motivo la administración de dominios de error y actualización es crucial para garantizar la alta disponibilidad de los servicios.

Al ejecutar Service Fabric en Azure, los dominios de error y actualización se administran automáticamente. En otros entornos puede que esto no sea así. Si va a compilar sus propios clústeres locales, asegúrese de asignar y planear el diseño de los dominios de error correctamente.

Los dominios de actualización son útiles para modelar las áreas donde se va a actualizar software de manera simultánea. Por este motivo, los dominios de actualización a menudo también definen los límites de desconexión del software durante las actualizaciones planeadas. Las actualizaciones de Service Fabric y los servicios siguen el mismo modelo. Para más información acerca de las actualizaciones graduales, los dominios de actualización y el modelo de estado de Service Fabric que ayuda a evitar que los cambios por error afecten al clúster y al servicio, consulte:

- Actualización de aplicaciones
- Tutorial sobre la actualización de aplicaciones
- Modelo de estado de Service Fabric

Puede visualizar el diseño del clúster mediante el mapa de clúster que se proporciona en Service Fabric Explorer:



> ⓘ **Nota:**
>
> El modelado de áreas de error, las actualizaciones graduales, la ejecución de muchas instancias del código de servicio y el estado, las reglas de colocación para asegurarse de que los servicios se ejecutan entre dominios de error y de actualización, y la supervisión de estado integrada son solo *algunas* de las características que Service Fabric proporciona para evitar que los errores y problemas operativos normales se conviertan en desastres.

Abordaje de errores simultáneos en el hardware y en el software

Hasta ahora se han tratado errores únicos. Como puede ver, son fáciles de solucionar para los servicios con y sin estado con más copias del código (y del estado) en varios dominios de error y de actualización.

También se pueden producir varios errores simultáneos y aleatorios. Estos tienen mayor probabilidad de provocar un tiempo de inactividad o un verdadero desastre.

Servicios sin estado

El recuento de instancias de un servicio sin estado indica el número deseado de instancias que deben estar en ejecución. Si se produce un error en alguna de las instancias (o en todas), Service Fabric responderá creando instancias de reemplazo automáticamente en otros nodos.

Service Fabric seguirá creando reemplazos hasta que el servicio vuelva al recuento de instancias deseado.

Por ejemplo, supongamos que el servicio sin estado tiene un valor `InstanceCount` de -1. Este valor significa que se debe ejecutar una instancia en cada nodo del clúster. Si se produjera un error en alguna de esas instancias, Service Fabric detectará que el servicio no está en su estado deseado e intentará crear las instancias en los nodos donde faltan.

Servicios con estado

Existen dos tipos de servicios con estado:

- Con estado con estado persistente.
- Con estado con estado no persistente (el estado se almacena en la memoria).

La recuperación de errores de servicios con estado depende del tipo de servicio con estado, así como del número de réplicas que tenía el servicio y cuántas de ellas han generado un error.

En un servicio con estado, los datos entrantes se replican entre las réplicas (principal y secundarias activas). Si la mayoría de las réplicas reciben los datos, se considera que los datos están confirmados en el *cuórum* (para cinco réplicas, tres serán un cuórum). Esto significa que, en cualquier momento, habrá al menos un cuórum de réplicas con los datos más recientes. Si se produce un error en las réplicas (por ejemplo, en dos de cinco), podemos usar el valor de cuórum para calcular si se puede recuperar (dado que las tres réplicas restantes de las cinco siguen estando activas, se garantiza que al menos una réplica tendrá datos completos).

Cuando se produce un error en un cuórum de réplicas, se declara que la partición está en estado de *pérdida de cuórum*. Supongamos que una partición tiene cinco réplicas, lo que significa que se garantiza que al menos tres tienen datos completos. Si se produce un error en un cuórum (tres de cinco) de réplicas, Service Fabric no puede determinar si las réplicas restantes (dos de cinco) tienen suficientes datos para restaurar la partición. En los casos en los que Service Fabric detecta pérdida de cuórum, su comportamiento predeterminado es evitar escrituras adicionales en la partición, declarar pérdida de cuórum y esperar a que se restaure un cuórum de réplicas.

La determinación de si se ha producido un desastre para un servicio con estado y su administración se producen en tres etapas:

1. Determinación de si se ha producido pérdida de cuórum o no.

   La pérdida de cuórum se declara cuando la mayoría de las réplicas de un servicio con estado están inactivas al mismo tiempo.

2. Determinación de si la pérdida de cuórum es permanente o no.

La mayoría de las veces, los errores son temporales. Los procesos, los nodos y las máquinas virtuales se reinician y las particiones de la red se reparan. Sin embargo, a veces los errores son permanentes. El hecho de que los errores sean permanentes o no depende de si el servicio con estado conserva su estado o si lo mantiene solo en la memoria:

- Para servicios sin estado persistente, un error de un cuórum o varios en las réplicas deriva *inmediatamente* en una pérdida de cuórum permanente. Cuando Service Fabric detecta una pérdida de cuórum en un servicio con estado no persistente, declara una (posible) pérdida de datos y pasa inmediatamente al paso 3. Pasar a pérdida de datos tiene sentido, ya que Service Fabric sabe que no merece la pena esperar el retorno de las réplicas. Aunque se recuperaran, se perderán los datos debido a la naturaleza no persistente del servicio.
- En los servicios con estado persistente, un error de uno más cuórums de réplicas provoca que Service Fabric espere el retorno de las réplicas y la restauración del cuórum. Esto supone la interrupción del servicio de *escritura* en las particiones afectadas (o "conjunto de réplicas") del servicio. Sin embargo, la lectura seguirá siendo posible, con garantía de coherencia reducida. La cantidad de tiempo predeterminada que Service Fabric espera a que el cuórum se restaure es *infinita*, ya que continuar supone un (posible) evento de pérdida de datos y acarrea otros riesgos. Esto significa que Service Fabric no continuará con el paso siguiente a menos que un administrador actúe para declarar pérdida de datos.

3. Determinación de si se ha producido pérdida de datos y restauración desde copias de seguridad.

   Si se ha declarado pérdida de cuórum (ya sea de forma automática o mediante una acción administrativa), Service Fabric y los servicios pasan a determinar si realmente se han perdido los datos. En este punto, Service Fabric también sabe que las demás réplicas no se van a recuperar. Esa fue la decisión que se tomó al detener la espera a que la pérdida de cuórum se resolviera por sí sola. La mejor forma de proceder con el servicio es la inmovilización mientras se espera la intervención administrativa específica.

   Siempre que Service Fabric llama al método `OnDataLossAsync` se debe a una *sospecha* de pérdida de datos. Service Fabric garantiza que esta llamada se entrega a la *mejor* de las réplicas restantes. Esta es la réplica que haya realizado más progreso.

   El motivo por el que siempre decimos *sospecha* de pérdida de datos es que puede que el resto de réplicas tengan el mismo estado que la principal al perderse el cuórum. Sin embargo, sin ese estado con el que compararlo, no hay forma de que Service Fabric o los operadores lo sepan con seguridad.

   Entonces, ¿qué hace una implementación típica del método `OnDataLossAsync`?

a. La implementación registra que se ha desencadenado `OnDataLossAsync` y activa las alertas administrativas necesarias.

b. Normalmente, la implementación se detiene y espera a la toma de otras decisiones y a la realización de acciones manuales. Esto es así porque aunque haya copias de seguridad disponibles, puede ser necesario prepararlas.

   Por ejemplo, si dos servicios diferentes coordinan información, puede que sea necesario modificar las copias de seguridad para garantizar que, cuando se haya realizado la restauración, la información importante para los dos servicios sea coherente.

c. A menudo también hay otros datos de telemetría o salidas del servicio. Estos metadatos pueden estar contenidos en otros servicios o registros. Esta información puede usarse según sea necesario para determinar si hubo llamadas recibidas y procesadas en la réplica principal que no estaban presentes en la copia de seguridad o en esta réplica determinada. Puede que sea necesario reproducirlas o agregarlas a la copia de seguridad para poder realizar la restauración.

d. La implementación compara el estado de las réplicas restantes con el de aquellas contenidas en las copias de seguridad disponibles. Si se usan las colecciones de confianza de Service Fabric, hay herramientas y procesos disponibles para ello. El objetivo consiste en ver si el estado de la réplica es suficiente y si falta algo en la copia de seguridad.

e. Una vez realizada la comparación, y tras realizarse la restauración (si es necesaria), el código del servicio debe devolver **true** si se han realizado cambios de estado. Si la réplica determinó que era la mejor copia disponible del estado y no realizó ningún cambio, el código devolverá **false**.

   El valor **true** indica que *cualquier* réplica restante ahora puede ser incoherente con esta. Se descartarán y se volverán a crear a partir de esta réplica. El valor **false** indica que no se realizó ningún cambio de estado, por lo que las demás réplicas pueden conservar el contenido.

Es muy importante que los autores de los servicios practiquen con escenarios de posible pérdida de datos y error antes de implementarlos en producción. Como protección ante la posible pérdida de datos, es importante realizar periódicamente una copia de seguridad del estado de los servicios con estado en almacenes con redundancia geográfica.

También debe asegurarse de que puede restaurar el estado. Dado que se realizan copias de seguridad de muchos servicios diferentes en momentos distintos, deberá asegurarse de que una vez restaurados, los servicios se ven de manera coherente entre ellos.

Por ejemplo, considere una situación donde un servicio genera un número y lo almacena y, a continuación, lo envía a otro servicio que también lo almacena. Después de una restauración, puede que descubra que el segundo servicio tiene el número pero el primero no, porque su copia de seguridad no incluye esa operación.

Si descubre que las réplicas restantes son insuficientes para continuar en un escenario de pérdida de datos y no se puede reconstruir el estado del servicio a partir de los datos de telemetría o por agotamiento, la frecuencia de las copias de seguridad determina el mejor objetivo de punto de recuperación (RPO) posible. Service Fabric facilita muchas herramientas para probar diversos escenarios de error, como el cuórum permanente y la pérdida de datos, que requiere la restauración a partir de una copia de seguridad. Estos escenarios forman parte de las herramientas de capacidad de prueba de Service Fabric, administradas por el servicio de análisis de errores. Consulte Introducción al servicio de análisis de errores para más información sobre estas herramientas y patrones.

> ⓘ **Nota:**
>
> Los servicios del sistema también pueden sufrir pérdida de cuórum. El impacto es específico en el servicio en cuestión. Por ejemplo, la pérdida de cuórum en el servicio de nombres afecta a la resolución de nombres, mientras que la pérdida de cuórum en Administrador de conmutación por error bloquea la creación de nuevos servicios y las conmutaciones por error.
>
> Los servicios de sistema de Service Fabric siguen el mismo patrón que los servicios de administración de estado, si bien no se recomienda intentar sacarlos de la pérdida de cuórum para incluirlos en posible pérdida de datos. En su lugar, se recomienda **buscar asistencia** para encontrar una solución específica para su situación. Suele ser preferible esperar a que las réplicas inactivas vuelvan a estar disponibles.

Solución de problemas de pérdida de cuórum

Es posible que las réplicas estén inactivas intermitentemente debido a un error transitorio. Espere un tiempo mientras Service Fabric intenta activarlas. Si las réplicas llevan inactivas más tiempo del esperado, siga las acciones de solución de problemas que se indican a continuación:

- Es posible que las réplicas se bloqueen. Compruebe los informes de estado del nivel de réplica y los registros de la aplicación. Recopile volcados de memoria y tome las medidas necesarias para realizar la recuperación.
- Es posible que el proceso de réplica haya dejado de responder. Inspeccione los registros de la aplicación para comprobarlo. Recopile volcados del proceso y, a continuación,

detenga el proceso que no responde. Service Fabric creará un proceso de reemplazo e intentará devolver la réplica.
- Es posible que los nodos que hospedan réplicas estén inactivos. Reinicie la máquina virtual subyacente para poner en marcha los nodos.

En ocasiones, puede no ser posible recuperar réplicas. Por ejemplo, si el error se ha producido en las unidades o las máquinas no responden físicamente. En estos casos, es necesario indicar a Service Fabric que no espere la recuperación de la réplica.

No use estos métodos si la posible pérdida de datos es inaceptable para poner el servicio en línea. En ese caso, se deben realizar todos los esfuerzos para la recuperación de máquinas físicas.

Las siguientes acciones podrían provocar pérdida de datos. Compruébelo antes de realizarlas.

> ⓘ **Nota:**
>
> *Nunca* es seguro utilizar estos métodos de manera que no sea directamente con particiones específicas.

- Use la API de `Repair-ServiceFabricPartition -PartitionId` o `System.Fabric.FabricClient.ClusterManagementClient.RecoverPartitionAsync(Guid partitionId)`. Esta API permite especificar el identificador de la partición que va a cambiar de pérdida de cuórum a posible pérdida de datos.
- Si el clúster detecta errores frecuentes que hacen que los servicios entren en estado de pérdida de cuórum y la *pérdida de datos potencial es aceptable*, especificar un valor de QuorumLossWaitDuration adecuado puede ayudarle con la recuperación automática del servicio. Service Fabric esperará el valor de `QuorumLossWaitDuration` proporcionado (el valor predeterminado es infinito) antes de realizar la recuperación. *No* se recomienda este método porque puede provocar pérdidas de datos inesperadas.

Disponibilidad del clúster de Service Fabric

Por lo general, el propio clúster de Service Fabric es un entorno altamente distribuido sin puntos de error único. Un error en cualquiera de los nodos no provoca problemas de disponibilidad o confiabilidad en el clúster, principalmente porque los servicios de sistema de Service Fabric siguen las directrices anteriores. Es decir, siempre se ejecutan con tres o más réplicas de forma predeterminada y los servicios de sistema sin estado se ejecutan en todos los nodos.

Las capas subyacentes de detección de errores y redes de Service Fabric están totalmente distribuidas. La mayoría de los servicios de sistema se regeneran a partir de los metadatos del clúster o saben cómo volver a sincronizar su estado desde otros lugares. La disponibilidad del clúster podría deteriorarse si los servicios de sistema se encuentran en situaciones de pérdida de cuórum como las descritas anteriormente. En estos casos, quizá no pueda realizar algunas operaciones en el clúster (como comenzar una actualización o implementar servicios nuevos), pero el clúster continúa activo.

Los servicios en un clúster en ejecución seguirán ejecutándose en estas condiciones, a menos que requieran que las operaciones de escritura en los servicios de sistema continúen funcionando. Por ejemplo, si Administrador de conmutación por error está en pérdida de cuórum, todos los servicios seguirán ejecutándose. Sin embargo, los servicios con error no podrán reiniciarse automáticamente, ya que esto requiere la implicación de Administrador de conmutación por error.

Errores en un centro de datos o una región de Azure

En raras ocasiones, los centros de datos físicos pueden dejar de estar disponibles temporalmente debido a la pérdida de conectividad de red o de alimentación. En estos casos, los clústeres de Service Fabric y los servicios de ese centro de datos o la región de Azure dejarán de estar disponibles. Sin embargo, *los datos se conservan*.

Para los clústeres que se ejecutan en Azure, puede ver las actualizaciones en caso de interrupciones en la página de estado de Azure. En el caso muy poco probable de que se destruyera un centro de datos físico parcial o totalmente, los clústeres de Service Fabric que hospede o sus servicios se podrían perder. Esto incluye cualquier estado del que no se haya realizado copia de seguridad fuera de ese centro de datos o esa región.

Hay varias estrategias diferentes para sobrevivir al error permanente o duradero de un único centro de datos o región:

- Ejecutar clústeres de Service Fabric independientes en varias de estas regiones y utilizar algún mecanismo de conmutación por error y conmutación por recuperación entre estos entornos. Este tipo de modelo de varios clústeres activos/activos o activos/pasivos requiere código adicional de administración y operación. Este modelo también requiere la coordinación de las copias de seguridad de los servicios de un centro de datos o región para que estén disponibles en otros centros de datos o regiones cuando uno produzca un error.

- Ejecutar un único clúster de Service Fabric que abarque varios centros de datos. La configuración mínima admitida para esta estrategia es tres centros de datos. Para más

información, consulte Implementación de un clúster de Azure Service Fabric en Availability Zones.

Este modelo requiere configuración adicional. Sin embargo, la ventaja es que el error de un centro de datos se transforma de desastre a error normal. Estos errores se solucionan mediante los mecanismos que funcionan para los clústeres de una única región. Los dominios de error, los dominios de actualización y las reglas de colocación de Service Fabric aseguran que las cargas de trabajo se distribuyen para tolerar los errores normales.

Para más información sobre las directivas que pueden ayudar a que los servicios funcionen en este tipo de clúster, consulte Directivas de colocación de servicios de Service Fabric.

- Ejecutar un único clúster de Service Fabric que abarque varias regiones mediante el modelo independiente. El número recomendado de regiones es tres. Consulte Creación de un clúster independiente para más información sobre la configuración de Service Fabric independiente.

Errores aleatorios que dan lugar a errores en el clúster

Service Fabric tiene el concepto de *nodos raíz*. Se trata de nodos que conservan la disponibilidad del clúster subyacente.

Estos nodos raíz ayudan a garantizar que el clúster permanece activo, al establecer concesiones con otros nodos y servir de elemento diferenciador durante ciertos tipos de errores. Si los errores aleatorios eliminan la mayoría de los nodos raíz del clúster y no se recuperan rápidamente, el clúster se cierra automáticamente. A continuación, se produce un error en el clúster.

En Azure, el proveedor de recursos de Service Fabric administra configuraciones de clúster de Service Fabric. De forma predeterminada, el proveedor de recursos distribuye los nodos raíz entre los dominios de error y de actualización para el *tipo de nodo principal*. Si el tipo de nodo principal está marcado como durabilidad Silver o Gold, al quitar un nodo raíz (ya sea mediante el escalado en el tipo de nodo principal o mediante su eliminación manual), el clúster intentará promover otro nodo que no sea de raíz de la capacidad disponible del tipo de nodo principal. Este intento producirá un error si tiene menos capacidad disponible de la que requiere el nivel de confiabilidad del clúster para el tipo de nodo principal.

En los clústeres independientes de Service Fabric y Azure, el tipo de nodo principal es el que ejecuta los nodos raíz. Al definir un tipo de nodo principal, Service Fabric aprovechará automáticamente el número de nodos proporcionado mediante la creación de hasta nueve nodos raíz y siete réplicas de cada servicio del sistema. Si un conjunto de errores aleatorios deshabilita la mayoría de las réplicas al mismo tiempo, los servicios de sistema entran en

pérdida del cuórum. Si se pierde la mayoría de los nodos raíz, el clúster se apaga poco después.

Pasos siguientes

- Obtenga más información sobre cómo simular varios errores mediante la plataforma de comprobación.
- Lea otros recursos sobre alta disponibilidad y recuperación ante desastres. Microsoft ha publicado una gran cantidad de instrucciones sobre estos temas. Aunque algunos de estos recursos se refieren a técnicas específicas para su uso en otros productos, contienen muchos procedimientos recomendados generales que se pueden aplicar en el contexto de Service Fabric:
  - Lista de comprobación de disponibilidad
  - Exploración en profundidad de la recuperación ante desastres
  - Recuperación ante desastres y alta disponibilidad para aplicaciones de Azure
- Más información sobre las opciones de soporte técnico de Service Fabric.

Last updated on 22/03/2026

Confiabilidad en Azure Cosmos DB

Azure Cosmos DB para NoSQL es un servicio de base de datos multimodelo distribuido globalmente que admite modelos de datos de documentos con esquemas flexibles. Azure Cosmos DB ofrece características de confiabilidad completas, incluidos varios niveles de coherencia que permiten equilibrar el rendimiento y la disponibilidad, las implementaciones con redundancia de zona de disponibilidad que protegen frente a errores de zona de disponibilidad, replicación multiregión con conmutación por error administrada por el servicio o administrada por el cliente, y opciones de copia de seguridad continuas y periódicas para la protección de datos.

Cuando se usa Azure, la confiabilidad es una responsabilidad compartida. Microsoft proporciona una variedad de capacidades para apoyar la resiliencia y la recuperación. Es responsable de comprender cómo funcionan esas funcionalidades dentro de todos los servicios que usa y de seleccionar las funcionalidades que necesita para cumplir los objetivos empresariales y los objetivos de tiempo de actividad.

En este artículo se describe cómo hacer que Azure Cosmos DB sea resistente a varias posibles interrupciones y problemas, incluidos errores transitorios, interrupciones de zona de disponibilidad, interrupciones de regiones y mantenimiento del servicio. También se describe cómo usar copias de seguridad para recuperarse de otros tipos de problemas y se resalta información clave sobre el acuerdo de nivel de servicio (SLA) de Azure Cosmos DB.

Recomendaciones de implementación de producción

Azure Well-Architected Framework proporciona recomendaciones sobre confiabilidad, seguridad, costo, operaciones y rendimiento. Para comprender cómo estas áreas influyen entre sí y contribuir a una solución de Azure Cosmos DB confiable, consulte Procedimientos recomendados para Azure Cosmos DB.

Introducción a la arquitectura de confiabilidad

En esta sección se describen algunos de los aspectos importantes de cómo funciona el servicio que es más relevante desde una perspectiva de confiabilidad. En la sección se presenta la arquitectura lógica, que incluye algunos de los recursos y características que se implementan y usan. También se describe la arquitectura física, que proporciona detalles sobre cómo funciona el servicio en segundo plano.

Arquitectura lógica

El recurso principal que implemente es un Azure Cosmos DB *account*. Cada cuenta puede tener varias *bases de datos* con varios *contenedores*. Los contenedores sirven como unidades lógicas de distribución y escalabilidad. Puede crear contenedores como colecciones, tablas y gráficos, en función de la API que use para interactuar con Azure Cosmos DB. Para obtener más información sobre el modelo de recursos, consulte Databases, contenedores y elementos en Azure Cosmos DB. Cada contenedor usa particionamiento, que admite alto rendimiento y alta escalabilidad.

Configure el *rendimiento*, que representa la cantidad de recursos del sistema que puede utilizar para consultar y trabajar con sus datos. Puede aprovisionar manualmente el rendimiento, usar la escalabilidad automática para ajustar dinámicamente la capacidad en función de los requisitos de la carga de trabajo o usar el tipo de cuenta sin servidor para que se cobre por el uso real.

Una sola cuenta puede abarcar varias regiones de Azure, lo que aumenta su resiliencia ante interrupciones regionales. Puede configurar varias regiones para leer y, si usa el nivel Crítico para la empresa, puede usar varias regiones para escribir. Azure Cosmos DB replica automáticamente sus datos geográficamente. El comportamiento de la replicación geográfica se ve afectado por la configuración que se usa, como el nivel de coherencia, que indica cómo desea compensar el equilibrio entre la coherencia de los datos, la disponibilidad, la latencia y el rendimiento. Diferentes niveles de coherencia optimizan para diferentes preocupaciones, admiten diferentes garantías y proporcionan diferentes tipos de replicación entre regiones.

Arquitectura física

Azure Cosmos DB almacena varios *replicas* de los datos para la redundancia. El servicio mitiga automáticamente las interrupciones de las réplicas manteniendo el quórum entre las réplicas de cada región. Este enfoque garantiza la alta disponibilidad y protege contra la pérdida de datos durante errores de nodo individuales, sin necesidad de cambios o configuración de la aplicación.

Internamente, Azure Cosmos DB administra los datos a través de varias construcciones, incluidas las particiones *físicas*, *conjuntos de partición* y *conjuntos de réplicas*. Para obtener información más detallada sobre cómo funciona Azure Cosmos DB, consulte Distribución global de datos con Azure Cosmos DB: a fondo.

Resistencia a errores transitorios

Los errores transitorios son errores breves e intermitentes en los componentes. Se producen con frecuencia en un entorno distribuido como la nube y son una parte normal de las operaciones.

Los errores transitorios se corrigen después de un breve período de tiempo. Es importante que las aplicaciones puedan controlar errores transitorios, normalmente mediante el reintento de solicitudes afectadas.

Todas las aplicaciones hospedadas en la nube deben seguir las instrucciones de control de errores transitorios de Azure cuando se comunican con cualquier API, bases de datos y otros componentes hospedados en la nube. Para obtener más información, consulte Recomendaciones para controlar errores transitorios.

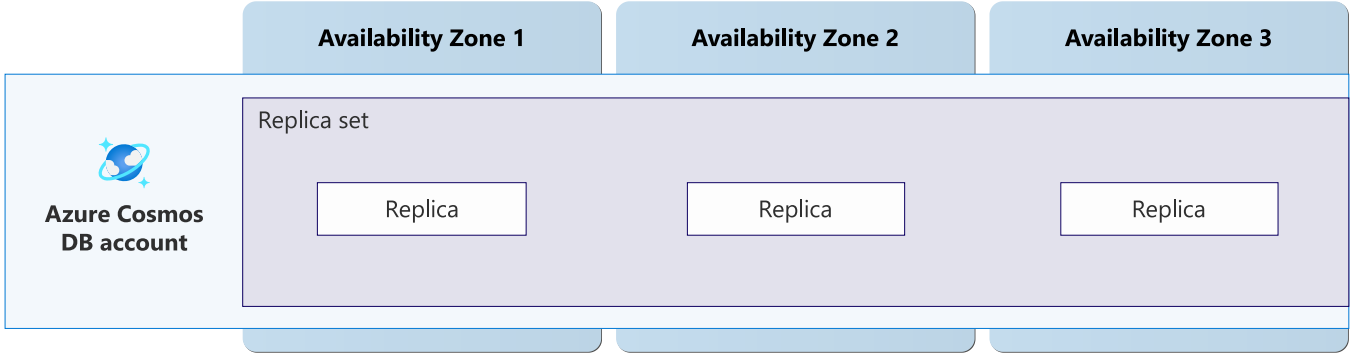
Use los SDK de Azure Cosmos DB. Los SDK implementan automáticamente la compatibilidad con una serie de consideraciones de resistencia, incluido el control de errores transitorios a través de reintentos automáticos, y respetan las respuestas de límite de velocidad enviadas por el servicio. Para obtener más información, consulte Design resilient applications with Azure Cosmos DB SDKs.

Al trabajar con una cuenta multirregional, el SDK también admite una estrategia de disponibilidad basada en un umbral, también denominada *hedging*, en la que se envían solicitudes de lectura en paralelo a varias regiones y se acepta la respuesta más rápida. Este enfoque puede mejorar el rendimiento de la aplicación cuando una región experimenta temporalmente una latencia mayor de lo habitual.

Resistencia a errores de zona de disponibilidad

Las zonas de disponibilidad son grupos físicamente independientes de centros de datos dentro de una región de Azure. Cuando una zona falla, los servicios pueden transferirse a una de las zonas restantes.

Azure Cosmos DB admite *redundancia de zona*. Al habilitar la redundancia de zona, Azure distribuye las réplicas de los datos en varias zonas de disponibilidad, lo que proporciona resistencia a problemas y interrupciones del centro de datos. Microsoft selecciona las zonas de disponibilidad que se van a usar.



Una cuenta de Azure Cosmos DB puede utilizar varias regiones (ubicaciones) para la distribución global, la escalabilidad y la conmutación por error. La redundancia de zona se configura por separado para cada región de la cuenta.

El uso de redundancia de zona en Azure Cosmos DB no tiene ningún impacto perceptible en el rendimiento o la latencia. No requiere ningún ajuste en el modo de coherencia seleccionado y no requiere ninguna modificación en el código de la aplicación.

Use la redundancia de zona en las regiones en las que se admite, especialmente para las cuentas de una sola región. Dado que las zonas de disponibilidad son físicamente independientes y proporcionan distintas fuentes de alimentación, redes y refrigeración, los acuerdos de nivel de servicio de disponibilidad para Azure Cosmos DB son más altos para las cuentas con redundancia de zona que las cuentas que no usan zonas de disponibilidad.

> **Sugerencia**
>
> Habilitar la redundancia de zona es una excelente manera de aumentar la resistencia de la base de datos de Azure Cosmos DB sin introducir complejidades de aplicación adicionales ni afectar al rendimiento. En función de la configuración de la cuenta, es posible que ni siquiera incurra en costos adicionales.

Si no habilita la redundancia de zona, la cuenta no es *zonal* en esa región. Las cuentas no zonales podrían localizar réplicas en una sola zona de disponibilidad, lo que provocaría un posible tiempo de inactividad si esa zona específica experimenta un problema.

Requirements

- **Region support:** Puede habilitar la redundancia de zona en las regiones de Azure que admiten zonas de disponibilidad. Para ver si la región admite zonas de disponibilidad, consulte la lista de regiones admitidas.

  La redundancia de zona no es una configuración para toda la cuenta. Azure Cosmos DB cuentas pueden abarcar varias regiones y cada región se puede configurar de forma independiente para usar zonas de disponibilidad. Las regiones que no admiten zonas de disponibilidad no le impiden habilitar la redundancia de zona en otras regiones dentro de la misma cuenta.

- **Cuentas sin servidor:** Solo puede configurar cuentas sin servidor con redundancia de zona al crearlas. No se pueden convertir cuentas sin servidor existentes sin zonas de

disponibilidad a una configuración de zona de disponibilidad. En el caso de las cargas de trabajo críticas, use el rendimiento aprovisionado.

Consideraciones

- **Varias interrupciones de zona simultáneas:** Una cuenta de una sola región con redundancia de zona puede mantener la disponibilidad de lectura y escritura cuando una interrupción afecta a una sola zona de disponibilidad. Sin embargo, si la interrupción afecta a varias zonas de disponibilidad o a toda la región, las cuentas de una sola región pierden el acceso de lectura y escritura hasta que se restaure el servicio. Considere desplegar una cuenta multirregional si necesita ser resiliente al fallo simultáneo de varias zonas.

- **Cuentas de varias regiones:** Si tiene una cuenta de varias regiones, opcionalmente puede habilitar la redundancia de zona en cualquiera o todas las regiones de cuenta que admiten zonas de disponibilidad. Se recomienda encarecidamente habilitar la redundancia de zona cuando la cuenta esté configurada para usar una sola región o si está configurada para usar una sola región de escritura con varias regiones de lectura.

Cost

Las regiones en las que se habilita la redundancia de zona se cobran a un precio superior. Sin embargo, se renuncia a los precios premium de las zonas de disponibilidad para las cuentas configuradas con escrituras de varias regiones y para las colecciones configuradas para usar el modo de rendimiento de escalabilidad automática. Para obtener más información, consulte Precios de Azure Cosmos DB.

Configurar soporte de zonas de disponibilidad

Para la mayoría de las cuentas, solo se habilita la redundancia de zona cuando se agrega una nueva región a una cuenta de Azure Cosmos DB. Para habilitar la compatibilidad con zonas de disponibilidad en una cuenta existente, agregue una región y habilite la redundancia de zona en ella. Puede seguir un proceso para agregar una región temporal para que pueda configurar la redundancia de zona en la región original. Para obtener pasos detallados, consulte Habilitar redundancia de zona en una cuenta de Azure Cosmos DB.

Para las cuentas sin servidor, debe habilitar la redundancia de zona al crear la cuenta.

Comportamiento cuando todas las zonas están en buen estado

En esta sección se describe qué esperar al configurar una cuenta de Azure Cosmos DB para la redundancia de zona y todas las zonas están operativas.

- **Operación entre zonas:** Azure Cosmos DB enruta automáticamente las solicitudes a las réplicas entre zonas de disponibilidad, por lo que cualquier réplica puede servir una solicitud.

- **Replicación de datos entre zonas:** Cuando un cliente realiza un cambio en los datos, ese cambio se aplica a varias réplicas en distintas zonas para lograr el cuórum. Este enfoque se conoce como *replicación sincrónica*. La replicación sincrónica garantiza un alto nivel de coherencia de datos, lo que reduce la probabilidad de pérdida de datos durante un error de zona. Las zonas de disponibilidad se encuentran relativamente cerca, lo que significa que hay un efecto mínimo en la latencia o el rendimiento.

Comportamiento durante un fallo de zona

En esta sección se describe qué esperar al configurar una cuenta de Azure Cosmos DB para la redundancia de zona y hay una interrupción en una de las zonas.

- **Detección y respuesta:** La plataforma Azure Cosmos DB es responsable de detectar un error en una zona de disponibilidad. No es necesario hacer nada para iniciar una conmutación por error de la zona.

- **Notificación**: Microsoft no le notifica automáticamente cuando una zona está inactiva. Sin embargo, puede usar Azure Resource Health para supervisar el estado de un recurso individual y puede configurar alertas de Resource Health para notificarle problemas. También puede usar Azure Service Health para comprender el estado general del servicio, incluidos los errores de zona, y puede configurar alertas de Service Health para notificarle problemas.

- **Active requests:** Cuando una zona de disponibilidad deja de estar disponible, Azure Cosmos DB finaliza las solicitudes en curso conectadas a las réplicas de la zona afectada y la aplicación debe reintentar esas solicitudes. Asegúrese de que la aplicación está preparada siguiendo las instrucciones de control de errores transitorios.

- **Pérdida de datos esperada:** No hay ninguna pérdida de datos esperada de un error de zona.

- **Tiempo de inactividad esperado:** Durante las interrupciones de zona, las conexiones pueden experimentar interrupciones breves que suelen durar unos segundos a medida que

se redistribuye el tráfico. Asegúrese de que las aplicaciones están preparadas siguiendo las instrucciones de control de errores transitorios.

- **Redistribution:** Azure Cosmos DB redirige automáticamente las solicitudes entrantes a réplicas saludables en otras zonas de disponibilidad. Cuando una zona de disponibilidad tiene una interrupción, la plataforma reasigna automáticamente el rendimiento aprovisionado a otras réplicas.

Recuperación de zona

Cuando se recupera la zona de disponibilidad, Azure Cosmos DB restaura automáticamente las réplicas en la zona de disponibilidad y vuelve a enrutar el tráfico entre réplicas como normal.

Prueba de fallos de zona

La conmutación por error y la recuperación de las zonas de disponibilidad de Azure Cosmos DB están totalmente administradas por Microsoft. No es necesario iniciar ni validar los procesos de fallo de las zonas de disponibilidad.

Resistencia a errores en toda la región

Al implementar una cuenta de Azure Cosmos DB en una sola región, una interrupción en toda la región que afecta a todos los nodos de Azure Cosmos DB normalmente no causa pérdida de datos, pero impide que la aplicación acceda a los datos. Azure Cosmos DB restaura el acceso a los datos después de que el servicio se recupere en la región afectada. La pérdida de datos solo se produce si la región experimenta un desastre irrecuperable.

Para prepararse para los casos excepcionales de interrupciones de regiones, puede configurar Azure Cosmos DB para admitir varios niveles de durabilidad y disponibilidad mediante uno de estos enfoques:

- Varias regiones de lectura con una sola región de escritura. Opcionalmente, puede habilitar la conmutación por error administrada por servicio o la conmutación automática por partición (PPAF).
- Regiones de escritura múltiples.

En la tabla siguiente se resumen las opciones de recuperación disponibles en función de la configuración de la cuenta y el tipo de interrupción. En las secciones posteriores de este artículo se proporcionan detalles exhaustivos de estas opciones y el comportamiento asociado.

Expandir tabla

Configuration	Tipo de interrupción	Impacto en la disponibilidad	Impacto de durabilidad	Qué hacer
Cuenta de una sola región	Interrupción regional	Se pierde el acceso de lectura y escritura hasta que se restaura el servicio.	No se pierden datos a menos que la región experimente un desastre irrecuperable.	Espere a que se restaure el servicio o solicite la restauración de la cuenta de copia de seguridad en otra región.
Región de escritura única, cuenta de varias regiones	Interrupción de la región de lectura	El SDK redirige a las regiones disponibles según la configuración de regiones preferidas. En el caso de las cuentas que utilizan consistencia fuerte con solo dos regiones o un retraso limitado que supere la ventana de retraso, también se pierde la disponibilidad de escritura a menos que se desconecte la región afectada.	No se produce pérdida de datos.	Asegúrese de que el rendimiento es suficiente en las regiones restantes. Para la consistencia fuerte o con caducidad limitada, considere la posibilidad de desconectar la región afectada.
Región de escritura única, cuenta de varias regiones	Interrupción de la región de escritura (con PPAF habilitado)	Conmutación por error automática a nivel de partición; no se requiere intervención manual.	Si la cuenta usa una coherencia fuerte, no se pierden datos. Si la cuenta no usa una coherencia fuerte, los datos no replicados podrían perderse en el improbable caso de que la región sufra una pérdida de datos permanente.	No es necesaria ninguna acción. PPAF administra automáticamente la conmutación por error.
Región de escritura única, cuenta de varias regiones	Interrupción de la región de escritura de datos (sin PPAF)	Se pierde la disponibilidad de escritura hasta que se complete una operación de desconexión de la	Si la cuenta usa una coherencia fuerte, no se pierden datos. Si la cuenta no usa una coherencia fuerte, los datos no	Realice una operación de desconexión de la región. Si la conmutación por error administrada por el servicio está habilitada, Azure Cosmos

Configuration	Tipo de interrupción	Impacto en la disponibilidad	Impacto de durabilidad	Qué hacer
		región o una conmutación por error administrada por el servicio. Las lecturas continúan desde regiones saludables.	replicados podrían perderse en el improbable caso de que la región sufra una pérdida de datos permanente.	DB inicia automáticamente la conmutación por error, pero esto puede tardar una hora o más. No cambie la región de escritura durante la interrupción.
Cuenta de región de escritura múltiple	Interrupción en cualquier región	Enrutamiento automático a regiones correctas a través de la configuración del SDK; no se requiere intervención manual.	Los datos actualizados recientemente en la región con errores podrían no estar disponibles en las regiones restantes. En el improbable caso de que la región sufre una pérdida permanente de datos, se podrían perder datos no replicados.	Asegúrese de que el rendimiento es suficiente en las regiones restantes. Después de la recuperación, Azure Cosmos DB recupera automáticamente los datos no replicados mediante el método de resolución de conflictos configurado.
Cualquier configuración de cuenta	Datos dañados o eliminación accidental	Sin impacto en la disponibilidad.	Posible pérdida de datos en función de cuándo se detectan daños o eliminaciones.	Restauración a un momento dado (copia de seguridad continua) o restauración a partir de una copia de seguridad periódica.

Nota:

Este artículo se centra en los aspectos de confiabilidad de las características de varias regiones de Azure Cosmos DB. Hay otras ventajas para varias regiones de lectura y escritura, como un mayor rendimiento y escalado para aplicaciones distribuidas globalmente. Debe evaluar toda la arquitectura de la solución y tener en cuenta todas las ventajas de usar estas funcionalidades.

SDK y resiliencia

Los SDK de Azure Cosmos DB son una parte importante de la estrategia de resistencia de la aplicación. Cuando tiene una cuenta de varias regiones, la configuración del SDK afecta a cómo

se enrutan las solicitudes entre regiones, incluidas las regiones preferidas a las que conectarse y las regiones a las que se debe excluir. Los SDK supervisan la disponibilidad de regiones y particiones, y pueden reconfigurarse dinámicamente para utilizar regiones y particiones saludables, como a través del disyuntor a nivel de partición.

Para obtener más información sobre cómo el SDK admite la alta disponibilidad, consulte la documentación de alta disponibilidad para el SDK que usa:

- SDK de Azure Cosmos DB .NET v3
- SDK v4 de Azure Cosmos DB Java
- SDK de Azure Cosmos DB para Python

Posible pérdida de datos durante fallos regionales

Al implementar una cuenta de Azure Cosmos DB en varias regiones, la durabilidad de los datos depende del nivel de coherencia que configure en la cuenta. En la tabla siguiente se detallan todos los niveles de coherencia, el objetivo de punto de recuperación (RPO) de una cuenta de Azure Cosmos DB que se implementa en al menos dos regiones. El RPO representa la pérdida potencial de datos durante una interrupción del servicio en una región.

Expandir tabla

Nivel de coherencia	RPO para la interrupción de la región
Sesión, prefijo coherente, eventual	Menos de 15 minutos
Uso vinculado	*K* y *T*
Alta	0

K = el número de versiones (es decir, actualizaciones) de un elemento.

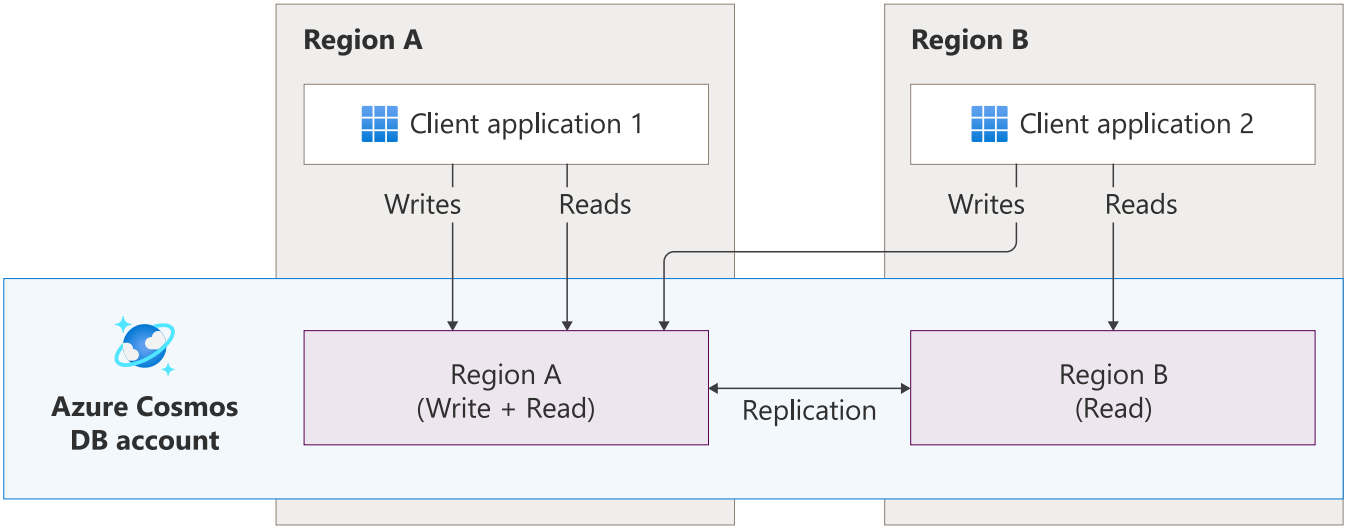
T = intervalo de tiempo desde la última actualización.

En el caso de cuentas de varias regiones, el valor mínimo de *K* y *T* es de 100 000 operaciones de escritura o de 300 segundos. Este valor define el RPO mínimo para los datos cuando se utiliza la obsolescencia limitada.

Para obtener más información sobre las diferencias entre los niveles de coherencia, consulte Niveles de coherencia en Azure Cosmos DB.

Varias regiones de lectura con una sola región de escritura

Si la solución requiere un tiempo de actividad continuo durante las interrupciones de la región, puede configurar Azure Cosmos DB para replicar los datos en varias regiones, con escrituras controladas por la región primaria. Opcionalmente, puede configurar las aplicaciones para conectarse a regiones de lectura específicas, lo que puede ayudar a mejorar su rendimiento. Si una región sufre una interrupción del servicio, la cuenta puede seguir funcionando desde las regiones que están en buen estado.



Conmutación por error entre regiones

Puede configurar el SDK de Azure Cosmos DB con una lista prioritaria de regiones de lectura. El SDK conecta la aplicación a la primera región disponible de la lista. Durante una interrupción de la región de lectura, el SDK detecta la interrupción de la región a través de códigos de respuesta de back-end, la marca como no disponible y enruta las operaciones futuras a la siguiente región disponible en la lista de preferencias. Asegúrese de que la lista de regiones preferidas se ha establecido correctamente y se alinea con los requisitos de negocio y latencia. Para obtener instrucciones detalladas, consulte Solución de problemas de disponibilidad del SDK de Azure Cosmos DB.

La conmutación por error es el proceso de hacer que una de las regiones de su cuenta no esté disponible, ya sea completamente o en parte. El efecto de una conmutación por error depende de si la región es una región de escritura o una región de lectura:

- Si una región de escritura deja de estar disponible, otra región se convierte en la región de escritura.
- Si una región de lectura deja de estar disponible, esa región no puede atender solicitudes de lectura y otras regiones se usan en su lugar para las operaciones de lectura.

Azure Cosmos DB proporciona varios tipos de conmutación por error:

- **Conmutación por error automática por partición (PPAF):** Internamente, Azure Cosmos DB distribuye los datos entre varias particiones físicas. Si se produce un problema con la infraestructura que admite una partición, es posible que otras particiones no se vean afectadas. PPAF permite a las cuentas de región de escritura única realizar automáticamente la conmutación por error de particiones individuales a una región secundaria, al tiempo que se mantienen las particiones en buen estado en la región principal. PPAF puede ayudar a minimizar el tiempo de inactividad y a permitir una recuperación más rápida durante un error de región parcial. Para obtener más información, consulte Cómo adoptar Per-Partition Automatic Failover (PPAF) para Azure Cosmos DB.

- **Conmutación por error forzada:** Puede desconectar una de las regiones de su cuenta. Esto también se conoce como *conmutación por error administrada por el cliente* o operación de *región fuera de línea*. Este es el enfoque recomendado para restaurar rápidamente la disponibilidad durante una interrupción. Usted es responsable de detectar la interrupción del servicio y de activar la conmutación por error. También puede utilizar las conmutaciones por error forzadas para simular escenarios de caída de una región con fines de prueba, como durante un simulacro de recuperación ante desastres.

  Si desconecta la región de escritura, la región de lectura con la siguiente prioridad más alta se convierte en la nueva región de escritura. Si deja sin conexión una región de lectura, las aplicaciones pueden conectarse a cualquier otra región de lectura de la cuenta.

  Una conmutación por error forzada de su región de escritura conlleva la posibilidad de pérdida de datos en el caso de las escrituras no replicadas.

  Tras una conmutación por error forzada, Microsoft debe volver a poner la región en línea. Para regiones saludables, este proceso está automatizado, pero puede tardar varios días. Si la región no vuelve a estar en línea en un día o dos, abra un caso de soporte técnico para solicitar asistencia.

- **Cambiar la región de escritura:** Cuando las regiones estén en buen estado, puede cambiar la región de escritura de la cuenta. Este cambio equivale, en la práctica, a una conmutación por error planificada de la región de escritura de su cuenta.

  Cambiar la región de escritura no produce ninguna pérdida de datos, ya que la replicación de datos se recupera antes de que se promueva la nueva región de escritura. Puede haber una breve interrupción, pero los clientes que usan lógica de reintento y otras técnicas de control de errores transitorios no suelen experimentar un impacto significativo.

Esta operación requiere que las regiones estén en buen estado, **por lo que no se puede usar durante una interrupción de la región.**

- **Conmutación por error administrada por el servicio:** Cuando la cuenta usa la conmutación por error administrada por el servicio, Microsoft es responsable de decidir cuándo realizar la conmutación por error entre regiones. Para habilitar la conmutación por error administrada por el servicio, especifique prioridades para cada región. Sin embargo, el proceso de declarar una interrupción del servicio y activar la conmutación por error administrada por el servicio puede llevar bastante tiempo, posiblemente una hora o más. Para una recuperación más rápida, realice una conmutación por error forzada en lugar de esperar a que se desencadene la conmutación por error administrada por el servicio.

  Si Microsoft activa una conmutación por error administrada por el servicio para la región de escritura de la cuenta, cualquier escritura no replicada podría perderse.

  Tras una conmutación por error administrada por el servicio, Microsoft debe volver a poner la región en línea. Microsoft trae automáticamente la región en línea, pero este proceso puede tardar varios días.

Requirements

Region support: Puede configurar cualquier región de Azure como región de lectura para la cuenta de Azure Cosmos DB.

Cost

Agregar una región de lectura adicional a una cuenta de Azure Cosmos DB aumenta los costos existentes para cada región. Para obtener más información, consulte Precios de Azure Cosmos DB.

Configuración de varias regiones de lectura

- **Agregue regiones de lectura a una cuenta:** Puede configurar varias regiones en su cuenta al crear la cuenta o en cualquier momento después de crearla. Para obtener más información, consulte Agregar o quitar regiones de la cuenta de base de datos.

- **Habilitación de la conmutación por error:** Algunos tipos de conmutación por error deben configurarse de antemano:

  - *Conmutación automática por error por partición*: Para obtener más información, consulte Cómo incorporar y adoptar la conmutación automática por error por partición (PPAF)

para Azure Cosmos DB.

- *Conmutación por error administrada por el servicio:* En primer lugar, habilite la conmutación por error administrada por el servicio. A continuación, establezca las prioridades de conmutación por error para cada región de su cuenta.

Planeamiento y administración de capacidad

Si la aplicación distribuye las solicitudes entre regiones y una región se queda sin conexión, las regiones restantes experimentan un volumen de solicitudes mayor. Utilice el escalado automático del rendimiento para adaptar dinámicamente la capacidad en función de la demanda. Si usa el rendimiento aprovisionado, planee la capacidad adecuada para controlar la pérdida de una región sin degradación del servicio y considere la posibilidad de sobreaprovisionar. Para más información, consulte Administración de la capacidad con sobreaprovisionamiento.

Comportamiento cuando todas las regiones están en buen estado

En esta sección se describe qué esperar al configurar una cuenta de Azure Cosmos DB con varias regiones de lectura y todas las regiones están operativas.

- **Operación entre regiones:** La aplicación configura la región que debe recibir operaciones de lectura. Puede configurar la aplicación con una lista prioritaria de regiones o excluir algunas regiones. Para obtener más información sobre cómo funciona la selección de regiones, consulte Diagnose y solucione la disponibilidad de los SDK de Azure Cosmos DB en entornos multirregionales.

  Todas las operaciones de escritura se dirigen a la región de escritura asignada a su cuenta.

- **Replicación de datos entre regiones:** Todas las operaciones de escritura se producen en la región primaria de la cuenta. Las operaciones de escritura se replican a las otras regiones de lectura según el nivel de coherencia configurado para la cuenta. Para obtener información sobre el retraso máximo de replicación, consulte Posible pérdida de datos durante las interrupciones de la región.

Comportamiento durante un error de región de lectura

En esta sección se describe qué esperar al configurar una cuenta de Azure Cosmos DB con varias regiones de lectura y se produce una interrupción en una de las regiones de lectura de la cuenta.

> **Importante**

Lo ideal es que las interrupciones de la región de lectura se controlen en el nivel de cliente configurando correctamente la lista de regiones preferidas en la configuración del SDK. Cuando se configura correctamente, el SDK detecta automáticamente la interrupción y vuelve a enrutar las operaciones de lectura a la siguiente región disponible sin necesidad de realizar ninguna conmutación por error del lado del servicio.

- **Detección y respuesta:** La responsabilidad de detectar la interrupción y responder depende del tipo de conmutación por error que usa su cuenta.
  - *PPAF:* PPAF normalmente no se aplica a interrupciones de la región de lectura. Sin embargo, en el caso de las cuentas con consistencia fuerte y solo dos regiones, la pérdida de la región de lectura reduce la cuenta a una única región, lo que impide mantener el quórum dinámico. En este escenario, PPAF puede activarse para conservar la disponibilidad cambiando las particiones afectadas a la región correcta.
  - *Conmutación por error forzada:* Usted es responsable de realizar una conmutación por error forzada. Para conocer los pasos detallados, consulte Realizar una conmutación por error forzada para su cuenta de Azure Cosmos DB.

    Si no realiza una conmutación por error, el comportamiento de su cuenta dependerá de su nivel de consistencia:

    - *Coherencia fuerte*: la coherencia fuerte requiere dos o más regiones para mantener el cuórum dinámico. Si hay menos de dos regiones disponibles y no realiza una conmutación por error, la cuenta perderá la disponibilidad de escritura hasta que se restablezca el servicio.
    - *Coherencia de obsolescencia limitada:* La coherencia de obsolescencia limitada se basa en el mantenimiento de un umbral de obsolescencia específico entre regiones. Si la duración de la interrupción de la región supera el umbral, el sistema no podrá mantener la consistencia entre las escrituras. Si no se realiza una conmutación por error, la cuenta perderá la disponibilidad de escritura hasta que se restablezca el servicio.
  - *Conmutación por error administrada por el servicio:* Si la conmutación por error administrada por el servicio está habilitada, Microsoft detectará la interrupción e iniciará una conmutación por error de la cuenta. Sin embargo, este proceso puede tardar mucho tiempo, potencialmente una hora o más. Para una recuperación más rápida, realice una

conmutación por error forzada en lugar de esperar a que se desencadene la conmutación por error administrada por el servicio.

- **Notificación**: Microsoft no le notifica automáticamente cuando una región está inactiva. Sin embargo:
  - Puede usar Azure Resource Health para supervisar el estado de un recurso individual y puede configurar alertas de Resource Health para notificarle problemas.
  - Puede usar Azure Service Health para comprender el estado general del servicio, incluidos los errores de región, y puede configurar alertas de Service Health para notificarle los problemas.
- **Solicitudes activas:** Es posible que cualquier solicitud activa se interrumpa y que el cliente tenga que volver a intentarla una vez finalizada la conmutación por error. Si los clientes controlan correctamente los errores transitorios mediante el reintento después de un breve período de tiempo, normalmente evitan un impacto significativo.
- **Pérdida de datos esperada:** Una interrupción en una región de lectura no causa pérdida de datos. Azure Cosmos DB sigue cumpliendo las garantías de coherencia de lectura.
- **Tiempo de inactividad esperado:** La cantidad de tiempo de inactividad que experimenta la cuenta depende del tipo de conmutación por error que usa la cuenta.
  - *PPAF:* Cuando PPAF está habilitado, el sistema detecta y recupera automáticamente el error, normalmente en un plazo de 3 minutos, sin intervención manual.
  - *Conmutación por error forzada:* El tiempo de inactividad depende de:
    - El tiempo que tardes en detectar la interrupción e iniciar la conmutación por error.
    - El tiempo que tarda la conmutación por error, que suele ser de unos segundos.

> ⚠ **Advertencia**
>
> No realice ninguna operación de configuración (plano de control) en la región afectada durante escenarios de interrupción, ya que dan como resultado una incoherencia de la cuenta y una recuperación retrasada. Algunos de los ejemplos de operaciones del plano de control que se deben evitar son:
>
> - Cambiar la región de escritura o modificar la prioridad de conmutación por error

- Actualice la cuenta a la configuración para escritura múltiple
- Actualización de los niveles de coherencia u otra configuración de la cuenta
- Actualización de configuraciones de punto de conexión privado o configuración de red
- Actualización del rendimiento de la cuenta u operaciones de escalado
- Cualquier otra operación que modifique la configuración de la cuenta o la configuración de región

  - *Conmutación por error administrada por el servicio:* Microsoft es responsable de iniciar la conmutación por error administrada por el servicio y el tiempo de inactividad que experimenta la cuenta se basa en el tiempo que tarda Microsoft declarar la interrupción e iniciar la conmutación por error. En algunas situaciones, puede tardar una hora o más. Si tu cuenta sufre una interrupción en las operaciones de escritura y necesitas restablecer rápidamente la disponibilidad de escritura, realiza una conmutación por error forzada.

- **Redistribución:** Para la conmutación por error forzada o la conmutación por error administrada por el servicio, la región afectada se desconecta y se marca como sin conexión.

  No es necesario realizar ningún cambio en el código de su aplicación para administrar la interrupción de la región de lectura. Los SDK Azure Cosmos DB redirigen las operaciones de lectura a la siguiente región disponible en la lista de regiones preferidas. Si ninguna de las regiones de la lista de regiones preferidas está disponible, las operaciones de lectura se revierten automáticamente a la región de escritura actual de la cuenta tal como está configurada en el servicio.

  > ⓘ **Nota:**
  >
  > Si usa puntos de conexión privados con una cuenta de Azure Cosmos DB, asegúrese de que el DNS privado se enruta correctamente después de la operación de región sin conexión. Para obtener instrucciones detalladas, consulte **Consideraciones sobre la conmutación por error para puntos de conexión privados**.

Comportamiento durante un fallo de región de escritura

En esta sección se describe qué esperar al configurar una cuenta de Azure Cosmos DB con varias regiones de lectura y se produce una interrupción en la región de escritura de la cuenta.

- **Detección y respuesta:** La responsabilidad de detectar la interrupción y responder depende del tipo de conmutación por error que usa su cuenta.
  - *PPAF:* Microsoft detecta automáticamente la interrupción e inicia una conmutación por error de algunas particiones, si procede. La aplicación no necesita realizar ninguna acción.
  - *Conmutación por error forzada:* Usted es responsable de realizar una conmutación por error forzada. Para conocer los pasos detallados, consulte Realizar una conmutación por error forzada para su cuenta de Azure Cosmos DB.

    Si no se realiza una conmutación por error, la cuenta perderá la disponibilidad de escritura hasta que se restablezca el servicio.

    Si hay una interrupción en la región de escritura de la cuenta, evite realizar una operación de *cambiar la región de escritura*. Los cambios en la región de escritura no se completan con éxito si se produce una interrupción en la región de origen o de destino. El motivo es que el procedimiento de cambio de región incluye una comprobación de coherencia que requiere conectividad entre las regiones.
  - *Conmutación por error gestionada por el servicio:* Microsoft detecta automáticamente la interrupción e inicia una conmutación por error de su cuenta. La aplicación no necesita realizar ninguna acción.
- **Notificación**: Microsoft no le notifica automáticamente cuando una región está inactiva. Sin embargo:
  - Puede usar Azure Resource Health para supervisar el estado de un recurso individual y puede configurar alertas de Resource Health para notificarle problemas.
  - Puede usar Azure Service Health para comprender el estado general del servicio, incluidos los errores de región, y puede configurar alertas de Service Health para notificarle los problemas.
- **Solicitudes activas:** Es posible que cualquier solicitud activa se interrumpa y que el cliente tenga que volver a intentarla una vez finalizada la conmutación por error. Si los clientes controlan correctamente los errores transitorios mediante el reintento después de un breve período de tiempo, normalmente evitan un impacto significativo.
- **Pérdida de datos esperada:** Si configura la cuenta con coherencia fuerte, no se produce ninguna pérdida de datos. De lo contrario, es posible que se pierdan las escrituras no

replicadas después de que se complete la conmutación por error. Para obtener información sobre la pérdida máxima de datos esperada durante una interrupción de la región, consulte Posibles pérdidas de datos durante las interrupciones de la región.

- **Tiempo de inactividad esperado:** La cantidad de tiempo de inactividad que experimenta la cuenta depende del tipo de conmutación por error que usa la cuenta.
  - *PPAF:* Cuando PPAF está habilitado, espere una breve interrupción, que normalmente es de aproximadamente 3 minutos.
  - *Conmutación por error forzada:* El tiempo de inactividad depende de:
    - El tiempo que tardes en detectar la interrupción e iniciar la conmutación por error.
    - El tiempo que tarda la conmutación por error, que suele ser de unos segundos.

    > ⚠ **Advertencia**
    >
    > No realice ninguna operación de plano de control en la región afectada durante los escenarios de interrupción, ya que provocan inconsistencias en la cuenta y retrasan la recuperación. Algunos de los ejemplos de operaciones del plano de control que se deben evitar son:
    > - Cambiar la región de escritura o modificar la prioridad de conmutación por error
    > - Actualice la cuenta a la configuración para escritura múltiple
    > - Actualización de los niveles de coherencia u otra configuración de la cuenta
    > - Actualización de configuraciones de punto de conexión privado o configuración de red
    > - Actualización del rendimiento de la cuenta u operaciones de escalado
    > - Cualquier otra operación que modifique la configuración de la cuenta o la configuración de región

  - *Conmutación por error administrada por el servicio:* Microsoft es responsable de iniciar la conmutación por error administrada por el servicio y el tiempo de inactividad que experimenta la cuenta se basa en el tiempo que tarda Microsoft declarar la interrupción e iniciar la conmutación por error. En algunas situaciones, puede tardar una hora o más. Para restaurar rápidamente la disponibilidad de escritura, realice una conmutación por error forzada.

- **Redistribución:** La redistribución del tráfico de escritura depende del tipo de conmutación por error que utilice su cuenta.

- *PPAF:* Azure Cosmos DB realiza automáticamente la conmutación por error de la partición defectuosa a una región en buen estado.
- *Conmutación por error forzada:* Al realizar una conmutación por error forzada, la región de escritura de la cuenta cambia a la región que especifique.

> ⓘ **Nota:**
>
> Si usa puntos de conexión privados con una cuenta de Azure Cosmos DB, asegúrese de que el DNS privado se enruta correctamente después de la operación de región sin conexión. Para obtener instrucciones detalladas, consulte **Consideraciones sobre la conmutación por error para puntos de conexión privados**.

- *Conmutación por error administrada por el servicio:* Azure Cosmos DB promueve automáticamente una de las regiones secundarias de la cuenta para que sea la nueva región de escritura principal. La conmutación por error se produce a otra región en el orden de prioridad de región que especifique.

Recuperación de regiones

Microsoft debe restaurar la operatividad de una región. Cuando una región se recupera después de una interrupción, Microsoft pone automáticamente la región en línea. Sin embargo, este proceso puede tardar varios días.

> ⓘ **Importante**
>
> Después de una conmutación por error forzada, Microsoft restablece automáticamente la conectividad de la región para las regiones saludables. Si la región no vuelve a estar en línea en un día o dos, abra un caso de soporte técnico para solicitar asistencia.

Una vez que la región está en línea, las acciones que realice son diferentes en función de si la interrupción se encontraba en una región de lectura o en una región de escritura.

- **Después de interrupciones de la región de lectura:** Cuando la región afectada vuelve a estar en línea, se sincroniza con la región de escritura actual y está disponible de nuevo para atender solicitudes de lectura después de haberse puesto al día por completo. Las siguientes lecturas se redirigen a la región recuperada sin necesidad de realizar cambios en el código de la aplicación. Tanto durante la conmutación por error como durante el

restablecimiento de una región con errores anteriores, Azure Cosmos DB sigue respetando las garantías de coherencia de lectura.

- **Después de interrupciones en la región de escritura:** Cuando la región afectada vuelve a estar en línea, la región se muestra como "en línea" en el portal de Azure y se vuelve disponible como una región de lectura. En este momento, es seguro cambiar la región de escritura a la región recuperada.

  > ⓘ **Importante**
  >
  > La región recuperada **no se promoverá automáticamente de nuevo como la región de escritura** una vez que se recupere. Es su responsabilidad **volver a cambiar a la región recuperada como región de escritura**, una vez que es seguro hacerlo.

  No hay *ninguna pérdida de datos o disponibilidad* antes, mientras, o después de cambiar la región de escritura. La aplicación sigue teniendo alta disponibilidad.

  Si alguna escritura no se replicó antes de que la región quedara fuera de línea, puede leer las escrituras no replicadas desde el canal de conflictos. La aplicación puede leer la fuente de conflictos, resolver los conflictos en función de la lógica específica de la aplicación y volver a escribir los datos actualizados en el contenedor según corresponda.

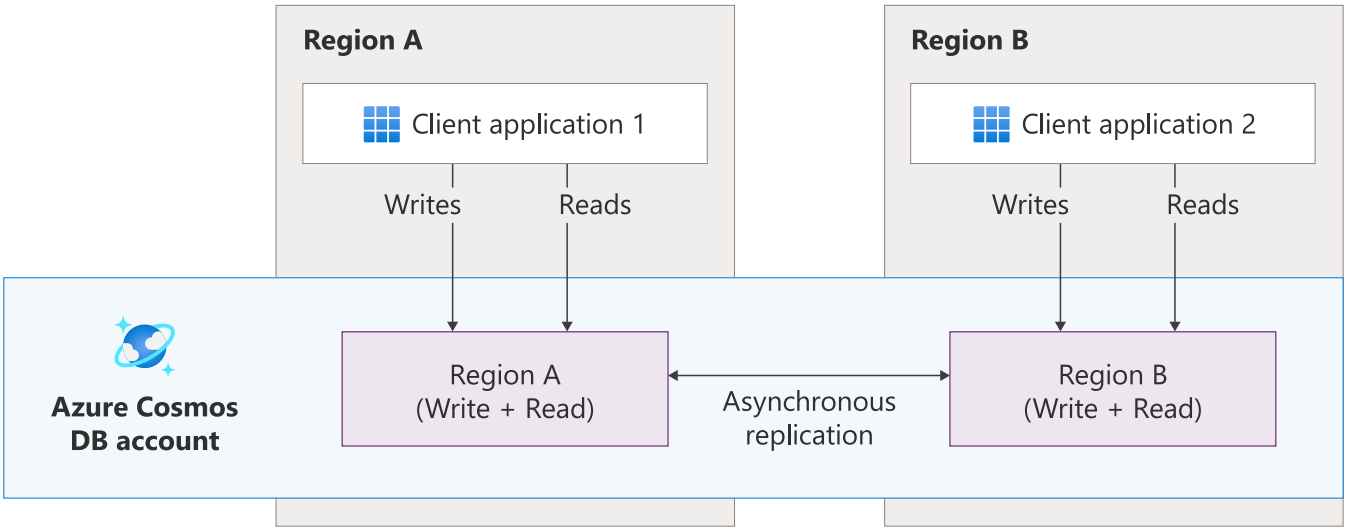
Prueba de fallos de región

Es posible que su aplicación no administre correctamente las conmutaciones por error de región, incluso si su cuenta de Azure Cosmos DB tiene alta disponibilidad. Para probar la alta disponibilidad de extremo a extremo de su aplicación como parte de sus pruebas de aplicaciones o simulacros de recuperación ante desastres (DR), deshabilite temporalmente la conmutación por error administrada por el servicio para la cuenta. Invoque la conmutación por error forzada mediante PowerShell, el CLI de Azure o el portal de Azure y, a continuación, supervise la aplicación. Una vez completada la prueba, puede volver a la región principal cuando esta vuelva a estar en línea automáticamente y, a continuación, restablecer la conmutación por error administrada por el servicio para la cuenta. Si la región no vuelve a estar en línea en un día o dos, abra un caso de soporte técnico para solicitar asistencia.

Si su cuenta usa PPAF, puede simular una conmutación por error de partición. Para obtener más información, consulte Probar la configuración de PPAF (simular error).

Varias regiones de escritura

Puede configurar Azure Cosmos DB para que acepte escrituras en varias regiones. Esta configuración puede proporcionar una resistencia muy alta a las interrupciones de la región. También es útil para reducir la latencia de escritura en aplicaciones distribuidas geográficamente.



Cuando se configura una cuenta de Azure Cosmos DB para varias regiones de escritura, no se admite la coherencia fuerte y pueden surgir conflictos de escritura. La región del hub actúa como árbitro en los conflictos de escritura. Para obtener más información sobre cómo resolver estos conflictos, consulte Tipos de conflictos y directivas de resolución al utilizar varias regiones de escritura.

Es importante tener en cuenta el diseño de la aplicación y cómo funciona con varias regiones de escritura. Revise los procedimientos recomendados para las escrituras multiregión.

Requirements

Region support: Puede configurar cualquier región de Azure como región de lectura o escritura para la cuenta de Azure Cosmos DB.

Cost

Agregar una región de escritura adicional a una cuenta de Azure Cosmos DB aumenta los costos existentes para cada región. Para obtener más información, consulte Precios de Azure Cosmos DB.

Configuración de varias regiones de escritura

Puede configurar varias regiones de escritura en su cuenta al crear la cuenta o en cualquier momento después de crearla. Para obtener más información, consulte Configuración de varias regiones de escritura.

Para usar eficazmente varias regiones de escritura, la aplicación también debe configurarse correctamente. Consulte Configuración de escrituras multiregión en aplicaciones que usan Azure Cosmos DB.

Planeamiento y administración de capacidad

Si la aplicación distribuye las solicitudes entre regiones y una región se queda sin conexión, las regiones restantes experimentan un volumen de solicitudes mayor. Utilice el escalado automático del rendimiento para adaptar dinámicamente la capacidad en función de la demanda. Si usa el rendimiento aprovisionado, planee la capacidad adecuada para controlar la pérdida de una región sin degradación del servicio y considere la posibilidad de sobreaprovisionar. Para más información, consulte Administración de la capacidad con sobreaprovisionamiento.

Comportamiento cuando todas las regiones están en buen estado

En esta sección se describe qué esperar al configurar una cuenta de Azure Cosmos DB con varias regiones de escritura y todas las regiones están operativas.

- **Operación entre regiones:** Cuando una cuenta está configurada con varias regiones de escritura, la aplicación configura la región que se debe usar para las operaciones de lectura y escritura. Puede configurar la aplicación con una lista prioritaria de regiones o excluir algunas regiones. Para obtener más información sobre cómo funciona la selección de regiones, consulte Diagnose y solucione la disponibilidad de los SDK de Azure Cosmos DB en entornos multirregionales. Para obtener información sobre cómo configurar la aplicación, consulte Configuración de escrituras multiregión en aplicaciones que usan Azure Cosmos DB.

- **Replicación de datos entre regiones:** Los datos se replican entre regiones de forma asincrónica. El retraso de replicación depende del nivel de coherencia de la cuenta. No se puede usar una coherencia fuerte para las escrituras de varias regiones. Para obtener más información, consulte Posibles pérdidas de datos durante interrupciones de región.

  Cuando una cuenta está configurada para varias regiones de escritura, las aplicaciones de diferentes regiones pueden realizar cambios que entren en conflicto entre sí. Azure Cosmos DB proporciona funcionalidades de resolución de conflictos. Para obtener más información, consulte Tipos de conflictos y directivas de resolución al usar varias regiones de escritura. Para obtener información sobre cómo configurar su propia directiva de resolución de conflictos, consulte Administrar directivas de resolución de conflictos en Azure Cosmos DB.

> ⓘ **Nota:**
>
> Actualizar el mismo identificador de documento con frecuencia o volver a crear el mismo identificador de documento con frecuencia después de que expire su TTL o se elimine, afecta negativamente al rendimiento de la replicación debido a un mayor número de conflictos generados en el sistema.

Comportamiento durante una falla de región

En esta sección se describe qué esperar al configurar una cuenta de Azure Cosmos DB con varias regiones de escritura y se produce una interrupción en una de las regiones de lectura o escritura de la cuenta.

- **Detección y respuesta:** La aplicación detecta la pérdida de la región. Azure Cosmos DB SDKs proporcionan funcionalidades automáticas de selección de regiones que enrutan las operaciones de lectura y escritura a regiones funcionales.
- **Notificación**: Microsoft no le notifica automáticamente cuando una región está inactiva. Sin embargo:
  - Puede usar Azure Resource Health para supervisar el estado de un recurso individual y puede configurar alertas de Resource Health para notificarle problemas.
  - Puede usar Azure Service Health para comprender el estado general del servicio, incluidos los errores de región, y puede configurar alertas de Service Health para notificarle los problemas.
- **Solicitudes activas:** Es posible que cualquier solicitud activa se interrumpa y que el cliente tenga que volver a intentarla una vez finalizada la conmutación por error. Si los clientes controlan correctamente los errores transitorios mediante el reintento después de un breve período de tiempo, normalmente evitan un impacto significativo.
- **Pérdida de datos esperada:** Los datos actualizados recientemente podrían dejar de estar disponibles en otras regiones. Para obtener información sobre la pérdida máxima de datos esperada durante una interrupción de la región, consulte Posibles pérdidas de datos durante las interrupciones de la región. En el improbable caso de que la región afectada sufre una pérdida de datos permanente, es posible que pierda datos no replicados.
- **Tiempo de inactividad esperado:** No hay tiempo de inactividad esperado en configuraciones de escritura múltiple, siempre que los SDK estén configurados

correctamente con `ApplicationRegions` o `PreferredRegions`.

> **Sugerencia**
>
> Para obtener los mejores resultados, las aplicaciones distribuidas globalmente deben estar delante de un servicio de equilibrio de carga global, como Azure Front Door o Azure Traffic Manager. Estos servicios pueden detectar la degradación regional y enrutar automáticamente el tráfico a instancias de aplicación en una región correcta.

- **Redistribution:** Los SDK de Azure Cosmos DB detectan automáticamente que la región es defectuosa y redireccionan las operaciones de lectura y escritura a la siguiente región disponible en la lista de regiones preferidas. No se requieren cambios en el código de la aplicación.

> **Sugerencia**
>
> Si la aplicación está gestionada por Azure Front Door o Traffic Manager, esos servicios también detectan la degradación regional y enrutan el tráfico a una región que esté funcionando correctamente.

Recuperación de regiones

Cuando la región afectada vuelve a estar en línea, la región se muestra como "en línea" en el portal de Azure y vuelve a estar disponible.

Cualquier dato de escritura que no se haya replicado cuando se produjo el error de la región queda disponible a través del canal de conflictos. Las aplicaciones pueden leer la fuente de conflictos, resolver los conflictos de acuerdo con la lógica específica de la aplicación y escribir los datos actualizados de nuevo en el contenedor de Azure Cosmos DB según corresponda.

Prueba de fallos de región

Para probar escenarios de conmutación por error de escritura multiregión, puede desconectar una región de escritura mediante una conmutación por error forzada. Este proceso simula una interrupción de la región y puede observar cómo responde la aplicación.

Copias de seguridad y restauración

Para la mayoría de las soluciones, no debe confiar exclusivamente en copias de seguridad. En su lugar, utilice las otras capacidades descritas en esta guía para apoyar los requisitos de resiliencia. Sin embargo, las copias de seguridad protegen contra algunos riesgos que otros enfoques no. Para más información, consulte ¿Qué son la redundancia, la replicación y la copia de seguridad?.

La pérdida de datos puede producirse debido a eliminaciones accidentales u otros problemas en la aplicación que provocan daños en los datos. Cuando se usa una cuenta de una sola región, la pérdida de datos también puede producirse debido a un desastre irrecuperable en la región de Azure Cosmos DB. Para ayudarle a protegerse contra la pérdida de datos, Azure Cosmos DB proporciona un conjunto de funcionalidades de copia de seguridad y restauración. Puede configurar copias de seguridad y retención en función de los requisitos de capacidad de recuperación y los requisitos de costo. Para más información, consulte Copia de seguridad en línea y restauración de datos a petición en Azure Cosmos DB.

Resistencia al mantenimiento del servicio

Azure Cosmos DB administra de forma transparente todos los detalles de los nodos de proceso individuales y realiza automáticamente la aplicación de revisiones y otros tipos de mantenimiento planeado. Los SLA de Azure Cosmos DB para la disponibilidad y la latencia se aplican a través de todas las operaciones de mantenimiento automáticas que realiza el sistema.

Acuerdo de nivel de servicio

El acuerdo de nivel de servicio (SLA) para Azure servicios describe la disponibilidad esperada de cada servicio y las condiciones que la solución debe cumplir para lograr esa expectativa de disponibilidad. Para obtener más información, consulte Acuerdos de Nivel de Servicio para servicios en línea.

Azure Cosmos DB proporciona acuerdos de nivel de servicio para una variedad de configuraciones y características de servicio, incluida la disponibilidad, la latencia, el rendimiento y la coherencia.

Los Acuerdos de Nivel de Servicio de disponibilidad son diferentes en función de si usa cualquiera de las siguientes funcionalidades de producto:

- Rendimiento aprovisionado
- Cuenta regional única con soporte para zona de disponibilidad (redundancia de zona)
- Cuentas que usan varias regiones de lectura
- Cuentas que usan varias regiones de escritura (nivel Crítico para la empresa)

Contenido relacionado

- confiabilidad Azure
- Información general de Azure Cosmos DB
- Niveles de coherencia en Azure Cosmos DB
- Distribución de datos global con Azure Cosmos DB
- Diagnose y solucione los problemas de disponibilidad de los SDK de Azure Cosmos DB en entornos multirregionales

ⓘ**Nota:** El autor ha creado este artículo con ayuda de inteligencia artificial. Más información

---

Last updated on 08/05/2026

Confiabilidad en Azure Database for MySQL

Azure Database for MySQL es un servicio de base de datos totalmente administrado que proporciona control y flexibilidad pormenorizados sobre las funciones de administración de bases de datos y las opciones de configuración. El servicio proporciona funcionalidades de alta disponibilidad (HA) y recuperación ante desastres (DR) en función de sus requisitos.

Al usar Azure, relibilidad es una responsabilidad compartida. Microsoft proporciona una variedad de capacidades para admitir resiliencia y recuperación. Es responsable de comprender cómo funcionan esas funcionalidades dentro de todos los servicios que usa y de seleccionar las funcionalidades que necesita para cumplir los objetivos empresariales y los objetivos de tiempo de actividad.

En este artículo se describe cómo hacer que Azure Database for MySQL sea resistente a diversas interrupciones y problemas potenciales, incluidos los errores transitorios, las interrupciones de zonas de disponibilidad, las interrupciones regionales y el mantenimiento del servicio. También se describe cómo puede usar copias de seguridad para recuperarse de otros tipos de problemas y se resalta cierta información clave sobre el acuerdo de nivel de servicio (SLA) de Azure Database for MySQL.

Recomendaciones de implementación de producción

Azure Well-Architected Framework proporciona recomendaciones sobre confiabilidad, seguridad, costo, operaciones y rendimiento. Para comprender cómo estas áreas influyen entre sí y contribuyen a una solución de Azure Database for MySQL confiable, consulte Procedimientos recomendados de arquitectura para Azure Database for MySQL.

Introducción a la arquitectura de confiabilidad

En esta sección se describen algunos de los aspectos importantes de cómo funciona el servicio que es más relevante desde una perspectiva de confiabilidad. En la sección se presenta la arquitectura lógica, que incluye algunos de los recursos y características que se implementan y usan. También se describe la arquitectura física, que proporciona detalles sobre cómo funciona el servicio en segundo plano.

Arquitectura lógica

Al trabajar con Azure Database for MySQL, se implementa un *server*, que representa los recursos de proceso y almacenamiento necesarios para admitir el servidor de bases de datos. Puede implementar una o varias *bases de datos* en el servidor.

Puede implementar servidores en los niveles de proceso ampliables, de uso general y *optimizados* para memoria. Cada nivel de proceso está optimizado para diferentes tipos de cargas de trabajo.

Para obtener más información sobre la arquitectura general del servicio y los modelos de implementación, consulte Azure Database for MySQL información general.

Arquitectura física

- **Separación entre proceso de cómputo y almacenamiento:** Azure Database for MySQL usa una arquitectura de separación entre proceso de cómputo y almacenamiento para ofrecer alta disponibilidad. El motor de base de datos se ejecuta en una máquina virtual (VM). Los archivos de datos se almacenan en Azure Storage, que mantiene sincrónicamente tres copias de los datos para protegerse frente a errores de hardware de almacenamiento. Según la configuración de alta disponibilidad del servidor, los archivos de datos se pueden almacenar en el almacenamiento con redundancia de zona (ZRS) o en el almacenamiento con redundancia local (LRS).

- **Alta disponibilidad:** Opcionalmente, puede habilitar una *configuración de alta disponibilidad* en el servidor. Al habilitar la configuración de alta disponibilidad, el servicio aprovisiona y mantiene un servidor de réplica en espera activa. Los cambios de datos en el servidor principal se replican sincrónicamente en el servidor de réplica en espera para garantizar una pérdida de datos cero durante un error del servidor principal.

  La arquitectura separa la capa de proceso de la capa de almacenamiento, lo que permite al servicio controlar los distintos tipos de errores de forma adecuada. Para lograr una mayor resistencia, puede distribuir los servidores entre zonas de disponibilidad.

  Un servidor de réplica en espera se implementa en la misma configuración de máquina virtual que el servidor principal, incluidos los núcleos virtuales, el almacenamiento y la configuración de red.

  Puede cambiar entre servidores mediante la realización de una *conmutación por error*. Use *conmutaciones por error no planeadas* cuando se produzca un error en el servidor principal y use *conmutaciones por error planeadas* cuando necesite minimizar el tiempo de inactividad de la aplicación durante una conmutación por error.

Para obtener más información, consulte HA en Azure Database for MySQL.

- **Backups:** Azure Database for MySQL crea automáticamente copias de seguridad del servidor. Para obtener más información, consulte Copia de seguridad y restauración.

Resistencia a errores transitorios

Los errores transitorios son errores breves e intermitentes en los componentes. Se producen con frecuencia en un entorno distribuido como la nube y son una parte normal de las operaciones. Los errores transitorios se corrigen después de un breve período de tiempo. Es importante que las aplicaciones puedan controlar errores transitorios, normalmente mediante el reintento de solicitudes afectadas.

Todas las aplicaciones hospedadas en la nube deben seguir las instrucciones de control de errores transitorios Azure cuando se comunican con cualquier API, bases de datos y otros componentes hospedados en la nube. Para obtener más información, consulte Recomendaciones para controlar errores transitorios.

Las aplicaciones deben controlar los errores de conectividad transitorios que pueden producirse durante el mantenimiento, las operaciones de escalado o las interrupciones de red. Siga estas recomendaciones:

- ✓ Cuando la aplicación encuentre errores transitorios, vuelva a intentar la operación mediante retroceso exponencial. Aumente el retraso entre reintentos y limite el número de intentos. Si la operación sigue fallando después del número máximo de reintentos, trátela como un fallo.
- ✓ Cuando sea posible, use bibliotecas cliente que gestionen automáticamente los reintentos.
- ✓ Los errores transitorios que se producen durante las operaciones de escritura requieren una consideración más cuidadosa. Considere la posibilidad de hacer que las operaciones de escritura sean idempotentes para poder ejecutarlas varias veces.

Resistencia a errores de zona de disponibilidad

Availability zones son grupos de centros de datos físicamente independientes dentro de una región de Azure. Cuando una zona falla, los servicios pueden transferirse a una de las zonas restantes.

Seleccione el tipo de soporte para zonas de disponibilidad mediante la configuración de alta disponibilidad. Al habilitar la alta disponibilidad, Azure Database for MySQL implementa un *servidor de réplica en espera* junto con el servidor principal. Este modelo de alta disponibilidad

ayuda a garantizar que los datos confirmados nunca se pierdan durante los errores. Sea cual sea el modelo de implementación de alta disponibilidad que elija, el servicio confirma los datos de forma sincrónica en ambos servidores de réplica, el principal y el en espera. Si se interrumpe el servidor principal, el sistema cambia automáticamente al servidor réplica en espera.

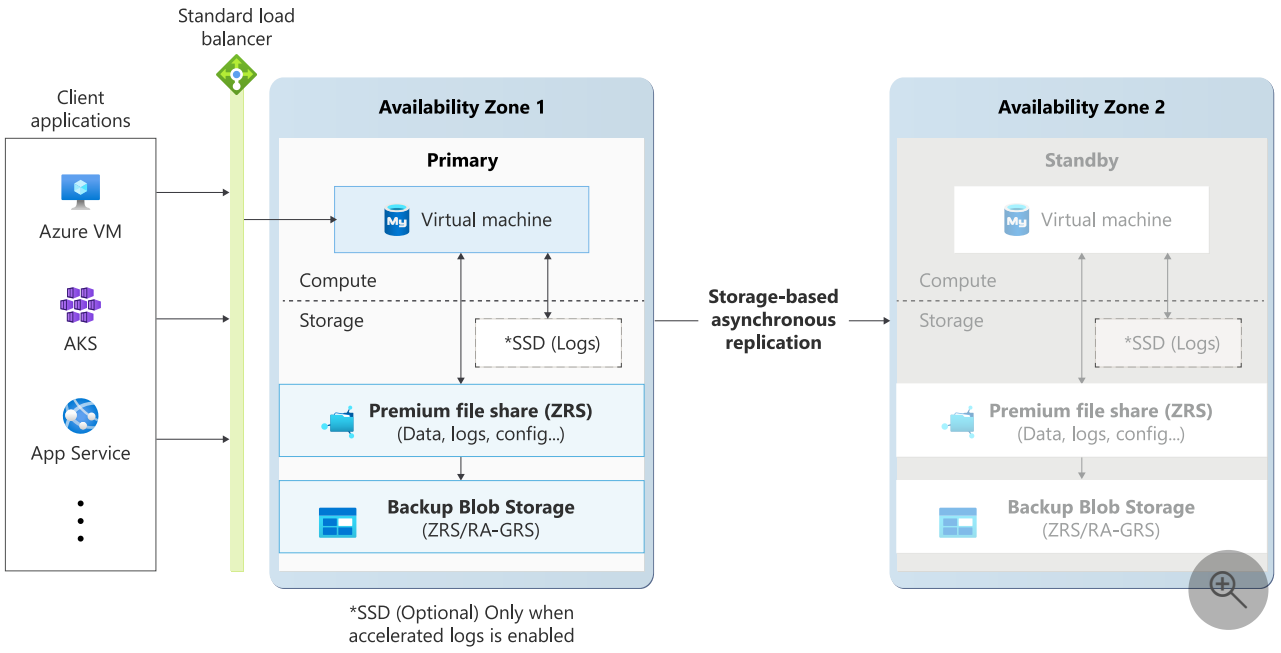
El servicio almacena datos en Azure Files Premium Storage. En función de la configuración de alta disponibilidad del servidor, usa ZRS o LRS, que almacena tres copias de datos dentro o entre zonas de disponibilidad.

Azure Database for MySQL admite dos tipos de configuración de zona de disponibilidad cuando se usa alta disponibilidad:

- **Alta disponibilidad con redundancia de zona:** La redundancia de zona proporciona el nivel más alto de resistencia de zona mediante la implementación de un servidor principal en una zona de disponibilidad y un servidor de réplica en espera en una zona de disponibilidad diferente. El servidor de réplicas en espera usa una configuración de red, almacenamiento y proceso similar al servidor principal. Una configuración con redundancia de zona proporciona aislamiento físico de toda la pila entre los servidores principal y en espera.

  Seleccione las zonas de disponibilidad para los servidores principales y en espera.

  Use implementaciones con redundancia de zona para servidores de producción.

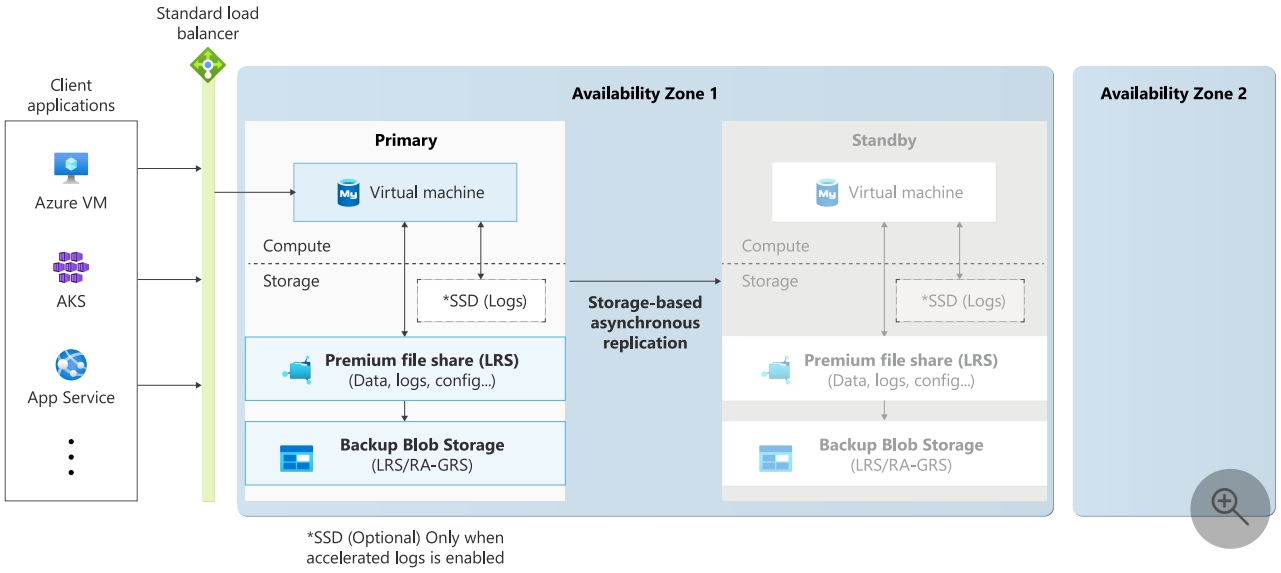


  Las operaciones de escritura pueden experimentar un pequeño aumento de la latencia de confirmación porque el servicio replica de forma sincrónica los datos en el servidor en espera. De media, cabe esperar un aumento de la latencia del 5 % al 10 % en las

operaciones de escritura y confirmación de la aplicación. El impacto varía según la carga de trabajo, la SKU seleccionada y la región.

- **Alta disponibilidad con redundancia local:** Los servidores principales y en espera usan la misma zona de disponibilidad. Si se produce una interrupción en el servidor principal, pero la zona sigue operativa, el servidor conmuta automáticamente al servidor en espera.

  Una implementación con redundancia local proporciona alta disponibilidad dentro de una sola zona de disponibilidad. Protege contra errores de nivel de nodo y también ayuda a reducir el tiempo de inactividad de la aplicación durante los eventos de tiempo de inactividad planeados y no planeados. Sin embargo, no protege contra una interrupción en esa zona. En las regiones que tienen zonas de disponibilidad, este tipo de configuración también se denomina a veces *zonal* o *de una sola zona*.



  Use HA con redundancia local solo en los siguientes escenarios:

  - Cuando se tienen aplicaciones inusualmente sensibles a la latencia, se valida la necesidad de minimizar la latencia entre la réplica principal y la secundaria, y se planifica la resiliencia entre zonas utilizando otros enfoques arquitectónicos.

  - Al implementar en una región que no admite zonas de disponibilidad. En este escenario, la región funciona como una sola zona, por lo que la HA con redundancia local es la única opción de alta disponibilidad.

  Los servidores están en la misma zona, lo que puede reducir la latencia de escritura a las aplicaciones que se implementan dentro de esa zona.

Si configura su servidor sin alta disponibilidad, funcionará en un solo servidor. Si se produce un error en ese servidor o su zona, el servidor no está disponible.

Requisitos

- **Compatibilidad con regiones:** Azure Database for MySQL admite diferentes configuraciones de zona de disponibilidad en función de la región de Azure. Para obtener una lista completa de las regiones, incluidos los tipos de compatibilidad con zonas de disponibilidad y las consideraciones específicas de cada región, consulte Regiones de Azure.
- **Nivel de servicio:** HA requiere los niveles General Purpose o Memory Optimized. El nivel de ráfaga no admite HA (alta disponibilidad), ni con redundancia de zona ni con redundancia local.

Costo

Cuando habilita la HA, crea y paga el servidor en espera a la misma tarifa que el servidor principal. La configuración de la zona de disponibilidad no afecta al costo. La replicación de datos no conlleva ningún cargo dentro o entre zonas de disponibilidad. Dependiendo del volumen de almacenamiento de copia de seguridad, también se le puede facturar por el almacenamiento de copia de seguridad. Para obtener información detallada sobre los precios, consulte los precios de Azure Database for MySQL.

Consideraciones

- **Claves principales:** Use claves principales en todas las tablas porque este enfoque reduce el tiempo de replicación y conmutación por error.
- **Limitaciones y problemas conocidos:** Revise la lista de limitaciones y problemas conocidos.

Configurar soporte de zonas de disponibilidad

Para configurar la compatibilidad con zonas de disponibilidad para un servidor, configure las opciones de alta disponibilidad.

> **Nota:**
>
> Cuando selecciona qué zonas de disponibilidad usar, en realidad está seleccionando la *zona de disponibilidad lógica*. Si implementa otros componentes de carga de trabajo en una suscripción de Azure diferente, pueden usar un número de zona de disponibilidad lógica diferente para acceder a la misma zona de disponibilidad física. Para obtener más información, consulte zonas de disponibilidad física y lógica.

- **Cree un servidor con redundancia de zona.** Para obtener información sobre cómo crear un servidor con alta disponibilidad y redundancia de zona habilitada, consulte los siguientes artículos:
  - Azure portal: administre alta disponibilidad con redundancia de zona en Azure Database for MySQL mediante el portal de Azure.
  - CLI de Azure: Administrar la alta disponibilidad con redundancia de zona en Azure Database for MySQL mediante CLI de Azure.
- **Cree un servidor con redundancia local.** Para crear un servidor con alta disponibilidad con redundancia local en una sola zona de disponibilidad, debe usar el CLI de Azure u otro método de implementación mediante programación. Para obtener las instrucciones de CLI de Azure, consulte Habilitación de alta disponibilidad durante la creación del servidor.
- **Cambie la configuración de zona de disponibilidad para los servidores existentes.** Si tiene un servidor existente, el enfoque que sigue para habilitar la compatibilidad con la zona de disponibilidad depende de la configuración inicial del servidor.
  - Para cambiar un servidor existente a alta disponibilidad con redundancia de zona, debe migrar a un nuevo servidor. Para obtener más información, consulte Migración de un servidor existente a un servidor con redundancia de zona.
  - Para cambiar un servidor existente a alta disponibilidad con redundancia local:
    1. Desactive HA, si está habilitada.
    2. Habilitar la HA con redundancia local. Debe usar el CLI de Azure u otro método de implementación mediante programación. Para obtener instrucciones sobre la CLI de Azure, consulte Administrar la HA con redundancia de zona en Azure Database for MySQL mediante la CLI de Azure.
- **Desactive HA.** Al deshabilitar la alta disponibilidad, se elimina el servidor réplica en espera, por lo que el servidor no es resiliente frente a interrupciones a nivel de zona. Sin embargo, si las copias de seguridad con redundancia geográfica están habilitadas, todavía puede recuperar el servidor en otra región mediante esas copias de seguridad. Para obtener más información, consulte Deshabilitar HA.

Comportamiento cuando todas las zonas están en buen estado

En esta sección se describe qué esperar al configurar servidores con compatibilidad con alta disponibilidad y zona de disponibilidad, y todas las zonas de disponibilidad están operativas.

- **Operación entre zonas:** Las aplicaciones cliente mySQL se conectan al servidor principal mediante el nombre de dominio completo (FQDN) del servidor de bases de datos. Evite usar la dirección IP del servidor principal porque la dirección IP puede cambiar, incluso durante las conmutaciones por error.

  Azure Database for MySQL usa una configuración activa-pasiva en la que el servidor principal controla todas las conexiones de base de datos y las consultas de la zona de disponibilidad principal. El servidor de réplicas en espera no atiende el tráfico de cliente durante las operaciones normales.

- **Replicación de datos entre zonas:** Las escrituras se confirman en el servidor principal y se escriben sincrónicamente en los registros del servidor en espera mediante ZRS. El servidor principal no espera a que el servidor en espera aplique los registros, pero dado que los registros están en ZRS, están disponibles incluso si se produce un error de réplica o zona.

  Los efectos de la replicación son diferentes en función de la configuración de zona de disponibilidad que use el servidor:

  - *Con redundancia de zona:* Dado que los servidores están en zonas independientes, este enfoque garantiza una pérdida de datos cero durante un error de zona. Esta situación también se conoce como lograr un objetivo de punto de recuperación (RPO) de cero en caso de fallos de zona.

    Sin embargo, la replicación entre zonas podría introducir una pequeña cantidad de latencia adicional. En promedio, puede esperar 5% a 10% mayor latencia para las escrituras y confirmaciones de la aplicación, pero el impacto varía según la carga de trabajo, la SKU seleccionada y la región.

  - *Redundancia local:* No se replica ningún tráfico entre zonas.

  > ⓘ **Nota:**
  >
  > El sistema replica todos los cambios en tiempo real en el servidor de réplicas en espera, incluidos los errores de usuario no deseados, como una eliminación accidental de una tabla o actualizaciones de datos incorrectas. Debido a la replicación inmediata, no puede usar la réplica en espera para la recuperación. Para recuperarse de errores de

usuario, debe realizar una restauración en un punto específico en el tiempo desde un respaldo. Para obtener más información, consulte Copia de seguridad y restauración.

Comportamiento durante un fallo de zona

En esta sección se describe qué cabe esperar cuando se configuran servidores con compatibilidad con HA y con zonas de disponibilidad, y se produce una interrupción en una zona de disponibilidad.

- **Detección y respuesta:** Azure comprueba periódicamente el estado de los servidores principal y en espera. Después de varios pings, si la supervisión de estado detecta que no se puede acceder a un servidor principal, el servicio inicia una conmutación automática al servidor en espera. El algoritmo de supervisión de estado usa varios puntos de datos para evitar situaciones de falsos positivos.

  Si se produce un error de zona, el comportamiento difiere en función de la configuración de zona de disponibilidad que use el servidor:

  - *Zone-redundant:* Azure Database for MySQL detecta automáticamente errores de zona de disponibilidad supervisando continuamente varios puntos de conexión de servidor. Para obtener más información, consulte Cómo funciona la detección automática de conmutación por error en servidores habilitados para alta disponibilidad.

    Para ver los posibles estados de HA, consulte Monitor HA. Cuando se produce un error en una zona, Azure inicia una conmutación por error no planificada al servidor en espera sin necesidad de que usted tome medidas.

  - *Con redundancia local:* Los servidores principales y en espera no están disponibles si la zona de disponibilidad que hospeda un servidor con redundancia local deja de estar disponible. En este escenario, el servicio no proporciona conmutación automática por error. Es responsable de detectar la interrupción de zona y realizar acciones de recuperación, como restaurar copias de seguridad con redundancia de zona en un servidor independiente en otra zona de disponibilidad o región.

- **Notificación:** Microsoft no le notifica automáticamente cuando una zona está inactiva. Sin embargo, puede usar Azure Resource Health para supervisar el estado de un recurso individual y puede configurar Resource Health alertas para notificarle problemas. También puede usar Azure Service Health para comprender el estado general del servicio, incluidos los errores de zona, y puede configurar alertas de Service Health para notificarle problemas.

Azure Database for MySQL genera un evento de Azure Resource Health cuando ocurre una conmutación por error no planificada.

- **Solicitudes activas:** Cuando una zona de disponibilidad deja de estar disponible, es posible que finalicen las solicitudes en curso a los servidores de la zona afectada. Las aplicaciones deben reintentar estas solicitudes. Si los clientes controlan correctamente los errores transitorios mediante el reintento después de un breve período de tiempo, normalmente evitan un impacto significativo.

- **Pérdida de datos esperada:** La cantidad de pérdida de datos depende de la configuración de la zona de disponibilidad del servidor.

  - *Redundante por zona:* Se anticipa una pérdida de datos cero durante el fallo de zona debido a la replicación sincrónica entre los servidores principal y en espera en distintas zonas.

  - *Con redundancia local:* Los datos de los servidores de la zona afectada no están disponibles hasta que se recupere la zona.

- **Tiempo de inactividad esperado:** La cantidad de tiempo de inactividad depende de la configuración de zona de disponibilidad que usa el servidor.

  - *Con redundancia de zona:* La conmutación por error se completa normalmente en un plazo de 60 a 120 segundos. Si los clientes controlan correctamente los errores transitorios mediante el reintento después de un breve período de tiempo, normalmente evitan un impacto significativo.

  - *Con redundancia local:* Los servidores de una zona afectada no están disponibles hasta que se recupere la zona de disponibilidad.

- **Redistribución:** El comportamiento de reenrutamiento del tráfico depende de la configuración de zona de disponibilidad que usa el servidor.

  - *Con redundancia entre zonas:* Después de la conmutación por error, el servidor en espera se convierte en el nuevo servidor principal y comienza a aceptar nuevas conexiones. Azure establece automáticamente un servidor en espera en la zona principal original después de recuperarse. Para obtener más información, consulte Conmutación por error no planeada.

  - *Con redundancia local:* Cuando una zona no está disponible, el servidor no está disponible. Si tiene un servidor independiente que ha creado previamente en otra zona

de disponibilidad o región, es responsable de redirigir el tráfico a ese servidor.

Recuperación de zona

El comportamiento de recuperación de zona depende de la configuración de zona de disponibilidad que usa el servidor.

- *Con redundancia de zona:* Cuando se recupera la zona de disponibilidad, Azure Database for MySQL vuelve a generar automáticamente el servidor en espera en la zona recuperada y lo sincroniza con el servidor principal actual. A continuación, la zona recuperada actúa como ubicación en espera. El servicio no mueve automáticamente el rol principal a la zona original para evitar interrupciones innecesarias. Si desea devolver el primario a la zona original, puede iniciar manualmente una conmutación por error planificada.

- *Con redundancia local:* Una vez que la zona está en buen estado, los servidores de la zona están disponibles de nuevo. Es responsable de los procedimientos de recuperación de zona y la sincronización de datos que requieren las cargas de trabajo.

Prueba de fallos de zona

Las opciones para probar errores de zona dependen de la configuración de zona de disponibilidad que usa la instancia.

- *Con redundancia de zona:* Puede probar la resistencia de su aplicación ante la conmutación por error iniciando una conmutación por error planeada. Utilice una conmutación por error planeada para simular una interrupción imprevista mientras la carga de trabajo se ejecuta y observe el tiempo de inactividad de su aplicación. Ejecutar simulaciones en entornos que no son de producción o en un momento silencioso. Para obtener más información, consulte Conmutación por error programada.

- *Con redundancia local:* No se puede simular una interrupción de zona completa, pero puede simular que el servidor no está disponible de forma similar a lo que ocurre durante una interrupción de zona. Para obtener más información, consulte los artículos siguientes:

  - Portal de Azure: Detener e iniciar una instancia de servidor flexible de Azure Database for MySQL.

  - CLI de Azure: reinicie, detenga e inicie una instancia de servidor flexible Azure Database for MySQL.

Resistencia a errores en toda la región

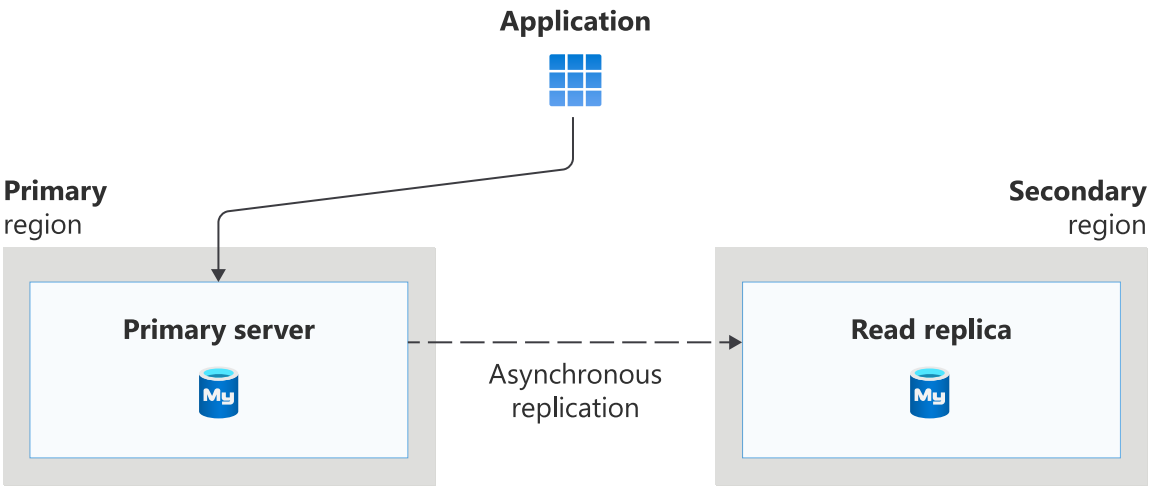
Azure Database for MySQL admite *réplicas de lectura entre regiones*, que puede usar para mantener una copia sincronizada de la base de datos en otra región para una recuperación más rápida.

También puede usar copias de seguridad con redundancia geográfica, en regiones admitidas, para proporcionar recuperación entre regiones. Sin embargo, las copias de seguridad suelen implicar más tiempo de inactividad y pérdida de datos que la replicación. Para obtener más información, consulte Copia de seguridad y restauración.

Réplicas de lectura entre regiones

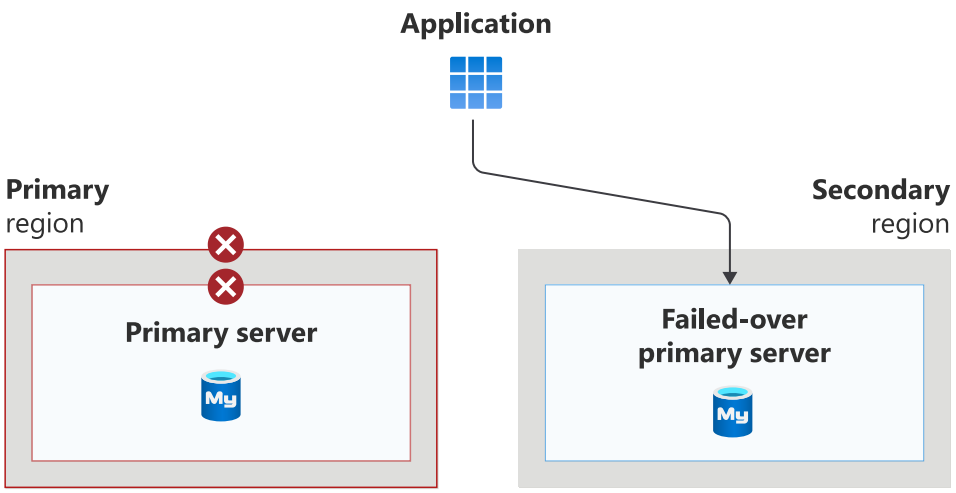
Implemente réplicas de lectura para proteger las bases de datos frente a errores de nivel de región. Cada réplica de lectura es un servidor Azure Database for MySQL independiente. Cuando colocas una réplica de lectura en una segunda región de Azure, tu servidor de bases de datos puede ofrecer resiliencia ante un problema que afecte a toda la región. Puede implementar hasta 10 réplicas de lectura, que opcionalmente pueden estar en diferentes regiones de Azure.

La tecnología de replicación física de MySQL actualiza las réplicas de lectura de forma asincrónica a partir del servidor de origen de la región principal, lo que significa que las réplicas pueden ir con retraso respecto al origen. Las réplicas de lectura entre regiones pueden servir opcionalmente cargas de trabajo de solo lectura para reducir la latencia de las aplicaciones distribuidas globalmente o descargar el tráfico de lectura desde el servidor de origen. Para obtener más información sobre las características y consideraciones relacionadas con las réplicas de lectura, consulte Réplicas de lectura.



Si la región primaria falla, puede realizar una conmutación por error manual para convertir la réplica secundaria en el servidor principal. Para realizar manualmente una conmutación por error, se detiene el proceso de replicación, lo que convierte la réplica de lectura en un servidor de

lectura/escritura. Debido a la replicación asincrónica, la conmutación por error puede provocar la pérdida de datos. La aplicación debe conectarse al nuevo servidor principal y es responsable de la reconfiguración de la aplicación.



> ⓘ **Nota:**
>
> En esta sección se resume alguna de la información importante sobre cómo las réplicas de lectura pueden admitir resiliencia ante fallos a nivel regional. También puede utilizar réplicas de lectura para mejorar el rendimiento y dar soporte a bases de usuarios geográficamente distribuidas y a gran escala.

Requisitos

- **Region support:** Puede crear réplicas de lectura entre regiones en cualquier región que admita Azure Database for MySQL. No se limita a las regiones emparejadas de Azure.
- **Niveles de cómputo:** Los niveles de cómputo de uso general y optimizados para memoria admiten réplicas de lectura. El nivel Ampliable no admite réplicas de lectura.

Consideraciones

- **Diferencias de configuración:** Al crear una réplica, hereda varias opciones de configuración del servidor de origen, incluida la generación de procesos, los núcleos virtuales y el almacenamiento. Puede personalizar estos valores en la réplica de lectura después de crearla, pero es mejor usar valores iguales o superiores para asegurarse de que la réplica pueda mantenerse al día con los cambios en la fuente.
- **Monitorización del retraso de replicación:** El proceso de replicación asíncrona implica un retraso de replicación, que puede variar en función de varios factores. Cuando el retraso de

replicación es muy alto, es posible que el servidor experimente problemas. Es importante supervisar el retraso de replicación para que pueda mitigar los problemas antes de que se escalen. Para más información, consulte Supervisión de la replicación.

- **HA:** Las réplicas de lectura no pueden tener HA habilitada y, cuando se realiza una conmutación por error para convertirlas en el servidor principal, tampoco tienen HA. Usted es responsable de configurar la alta disponibilidad tras la conmutación por error a una réplica.

Costo

Las réplicas de lectura incurren en costos de proceso y almacenamiento, así como cargos por transferencia de datos entre regiones para la replicación. Para obtener información detallada sobre los precios, consulte precios de Azure Database for MySQL y precios de ancho de banda.

Configuración de la compatibilidad con varias regiones

- **Cree una réplica de lectura:** Para obtener información sobre cómo crear una réplica de lectura, consulte los artículos siguientes:
  - Azure portal: Cómo crear y administrar réplicas de lectura en Servidor flexible de Azure Database for MySQL mediante el portal de Azure
  - CLI de Azure: Cómo crear y administrar réplicas de lectura en el servidor flexible de Azure Database for MySQL mediante CLI de Azure

  Puede configurar réplicas después de crear el servidor de origen solo si el servidor de origen está en ejecución y es accesible.

- **Detener la replicación:** Para obtener información sobre cómo detener la replicación, consulte Detener la replicación en un servidor de réplicas.

- **Eliminar una réplica de lectura:** Para obtener información sobre cómo eliminar una réplica de lectura, consulte Eliminación de un servidor de réplicas.

Comportamiento cuando todas las regiones están en buen estado

En esta sección se describe qué esperar al configurar el servidor con una réplica de lectura en otra región y todas las regiones están operativas:

- **Enrutamiento del tráfico entre regiones:** Durante las operaciones normales, la aplicación debe dirigir el tráfico de lectura y escritura al servidor de origen en la región primaria. Opcionalmente, puede dirigir las solicitudes de lectura a la réplica de lectura.

- **Replicación de datos entre regiones:** Las réplicas de lectura entre regiones usan la replicación asincrónica para minimizar el impacto en el rendimiento del servidor de origen. La cantidad de retraso de replicación depende de varios factores, incluida la carga de escritura y la latencia entre el servidor de origen y las réplicas. El retraso de replicación suele ser al menos varios minutos, pero puede ser mucho más largo. Para obtener más información, consulte Monitor replication y para obtener instrucciones detalladas, consulte Monitor replication in the Azure portal.

Comportamiento durante una falla de región

En esta sección se describe qué cabe esperar cuando se configura un servidor con compatibilidad con réplicas de lectura entre regiones y se produce una caída en la región primaria.

- **Detección y respuesta:** Eres responsable de detectar una caída en la región principal y activar manualmente una conmutación por error. Esta acción puede provocar la pérdida de datos no replicados.

  > ⓘ **Importante**
  >
  > Usted es responsable de desencadenar la conmutación por error. Azure no conmuta por error para leer las réplicas automáticamente, incluso si se produce un error en una región.

  La conmutación por error requiere que realice los pasos siguientes:

  1. Detenga la replicación. Este procedimiento es irreversible y el servidor no se puede volver a convertir en una réplica. El proceso da como resultado la pérdida de datos. Para obtener más información sobre las implicaciones de esta acción, vea Detener la replicación.

  2. Vuelva a configurar la aplicación para usar el nuevo servidor principal.

  Para obtener más información, consulte Conmutación por error.

- **Notification:** Microsoft no le notifica automáticamente cuando una región está inactiva. Sin embargo, puede usar Azure Service Health para comprender el estado general del servicio,

incluidos los errores de región, y puede configurar alertas de Service Health para notificarle problemas.

- **Solicitudes activas:** Todas las conexiones activas a la región de origen se quitan si el servidor de origen no está disponible. Las aplicaciones deben reintentar la realización de conexiones al nuevo servidor principal una vez completado el proceso de conmutación por error.

- **Pérdida de datos esperada:** Durante una interrupción en la región, debe realizar una conmutación por error para detener la replicación. Este proceso produce una pérdida permanente de datos no replicados.

  La cantidad de pérdida de datos depende del retraso de replicación en el momento de la interrupción. El retraso de replicación suele ser al menos varios minutos, pero puede ser mucho más largo. Para más información, consulte Supervisión de la replicación.

- **Tiempo de inactividad esperado:** La detención de la replicación se completa normalmente en dos minutos después de desencadenar la acción. Es responsable de volver a configurar las aplicaciones para conectarse al nuevo servidor principal. El tiempo que se tarda en realizar la reconfiguración también contribuye al tiempo de inactividad general.

- **Reenrutamiento del tráfico:** Es responsable de volver a configurar las aplicaciones para conectarse al nuevo servidor principal.

  > ⓘ **Nota:**
  >
  > Después de conmutar por error una réplica de lectura para convertirse en el servidor principal, el servidor no tiene habilitada la alta disponibilidad. Debe activar HA manualmente o incluirla en su automatización.

Recuperación de regiones

Cuando la región se recupere, usted será responsable de las actividades de retorno tras la recuperación para reanudar las operaciones en la región primaria. Microsoft no mueve automáticamente el servidor principal. Puede crear una nueva réplica de lectura en la región primaria y, a continuación, realizar otro proceso de conmutación por error para restaurar operaciones en la región primaria. Tenga en cuenta uno de los enfoques siguientes, en función de si la aplicación puede tolerar tiempo de inactividad o pérdida de datos:

- Ponga la aplicación fuera de servicio y espere a que la replicación se complete con todos los cambios. Este enfoque requiere tiempo de inactividad de la aplicación que sea aproximadamente el mismo que el retraso de replicación.
- Realice la conmutación por error y acepte la pérdida de datos no replicados.

Recuerde que también es responsable de volver a configurar las aplicaciones para conectarse al nuevo servidor principal según sea necesario.

Prueba de fallos de región

Pruebe periódicamente los procedimientos de conmutación por error de las réplicas de lectura para garantizar que sus procesos sean válidos y que las capacidades cumplan su objetivo de tiempo de recuperación (RTO) y sus requisitos de RPO.

Puede realizar la conmutación de una réplica de lectura para que se convierta en el servidor principal en cualquier momento, incluso cuando todas las regiones están en buen estado. Recomendamos realizar estas pruebas en un entorno de no producción, porque hacerlas puede provocar la pérdida de datos y requiere una conmutación por recuperación manual.

Como parte de la estrategia de recuperación ante desastres, ejecute periódicamente simulacros de recuperación completa. Estos simulacros incluyen la validación de datos, las pruebas de funcionalidad de la aplicación y los procedimientos documentados de reversión.

Copias de seguridad y restauración

Azure Database for MySQL realiza automáticamente una copia de seguridad de los datos, por lo que puede restaurarlos a cualquier momento dado dentro del período de retención de copia de seguridad. Esta protección le ayuda a evitar daños accidentales y la eliminación de datos. Microsoft administra completamente las copias de seguridad sin interrumpir la disponibilidad del servidor. Las copias de seguridad incluyen copias de seguridad completas y copias de seguridad del registro de transacciones.

- **Almacenamiento de copia de seguridad:** Si configura el servidor con alta disponibilidad con redundancia entre zonas, el sistema almacena las copias de seguridad en ZRS. En el caso de los servidores configurados sin alta disponibilidad o con alta disponibilidad con redundancia local, el sistema almacena copias de seguridad en LRS.

  En las regiones de Azure que tienen una región emparejada, puede configurar el almacenamiento con redundancia geográfica (GRS) para las copias de seguridad cuando cree el servidor. Este enfoque replica las copias de seguridad en la región emparejada de

Azure para una protección adicional frente a errores de región. El sistema replica las copias de seguridad de forma asincrónica.

El período de retención de copia de seguridad predeterminado es de 7 días, pero puede ampliar la retención hasta 35 días. Todas las copias de seguridad están cifradas.

- **Restaurar:** La restauración a un momento dado (PITR) le permite restaurar la base de datos en cualquier momento dentro del período de retención de copia de seguridad. El proceso de restauración crea un nuevo servidor de bases de datos con un nuevo nombre de servidor proporcionado por el usuario. Puede usar el nuevo servidor as-is o copiar datos de él.

  Al restaurar una copia de seguridad con redundancia geográfica, se crea un nuevo servidor en la región emparejada. En algunas regiones, puede usar Universal Geo-Restore para restaurar una copia de seguridad con redundancia geográfica en una región que no sea la emparejada con la primaria.

  Use esta funcionalidad para recuperarse de modificaciones accidentales de datos, errores de aplicación o escenarios de prueba.

Para la mayoría de las soluciones, no debe confiar exclusivamente en copias de seguridad. En su lugar, utilice las otras capacidades descritas en esta guía para apoyar los requisitos de resiliencia. Sin embargo, las copias de seguridad protegen contra algunos riesgos que otros enfoques no. Para más información, consulte ¿Qué son la redundancia, la replicación y la copia de seguridad?.

Para obtener más información, vea Backup and restore in Azure Database for MySQL.

Resistencia al mantenimiento del servicio

Azure Database for MySQL controla automáticamente las tareas de mantenimiento críticas, incluida la aplicación de revisiones del hardware subyacente, el sistema operativo y el motor de base de datos. El servicio incluye actualizaciones de seguridad, actualizaciones de software y actualizaciones de versiones secundarias como parte del mantenimiento planeado. Para obtener más información, consulte Scheduled maintenance in Azure Database for MySQL.

Para asegurarse de que el servidor permanece disponible durante las ventanas de mantenimiento, siga estas recomendaciones:

- ✓ **Evite las operaciones de administración durante los períodos de mantenimiento.** No realice operaciones de administración del servidor mientras el mantenimiento está en curso porque estas operaciones pueden afectar a la confiabilidad del servidor.

- **Utilice el mantenimiento con tiempo de inactividad casi nulo.** Si el servidor tiene habilitada la alta disponibilidad y cumple otros criterios de idoneidad, las operaciones de mantenimiento normalmente se completan en un plazo de 10 a 30 segundos. Si habilitas HA, las operaciones de mantenimiento suelen usar actualizaciones progresivas para minimizar el tiempo de inactividad. Las actividades de mantenimiento periódicas, como las actualizaciones de versiones secundarias, se producen primero en la réplica en espera. Para reducir el tiempo de inactividad, el servidor en espera se promueve a principal para que las cargas de trabajo puedan seguir ejecutándose mientras se aplican las tareas de mantenimiento en el nodo restante. Esta secuencia se aplica independientemente de si el servidor usa HA con redundancia de zona o con redundancia local. Para obtener más información, consulte Mantenimiento casi sin tiempo de inactividad.

- **Configurar ventanas de mantenimiento personalizadas.** Puede configurar la programación de mantenimiento para que se administre por el sistema o definir una ventana de mantenimiento personalizada para minimizar el impacto en las operaciones empresariales. También puede reprogramar las operaciones de mantenimiento planeado. Programe el mantenimiento durante períodos de baja actividad para minimizar el impacto empresarial. Para obtener más información, consulte Administrar la configuración de mantenimiento programado para Azure Database for MySQL.

- **Implemente la lógica de reintento.** Asegúrese de que las aplicaciones pueden controlar breves interrupciones de conectividad que pueden producirse durante los reinicios de mantenimiento. Para que las aplicaciones sean resistentes a estos tipos de problemas, consulte Resistencia a errores transitorios.

- **Habilite el mantenimiento de Virtual Canary en servidores de desarrollo y pruebas.** El mantenimiento de Virtual Canary proporciona acceso anticipado a las actualizaciones. Al habilitarlo en servidores de desarrollo y pruebas, puede comprobar que las próximas actualizaciones no afectan a la carga de trabajo antes de que lleguen a los servidores de producción. Para obtener más información, consulte Mantenimiento de Canario virtual.

Acuerdo de nivel de servicio

El acuerdo de nivel de servicio (SLA) para Azure servicios describe la disponibilidad esperada de cada servicio y las condiciones que la solución debe cumplir para lograr esa expectativa de disponibilidad. Para obtener más información, vea SLAs for servicios en línea.

Azure Database for MySQL proporciona acuerdos de nivel de servicio de disponibilidad diferentes en función de la configuración del servidor:

- Servidores configurados con alta disponibilidad con redundancia entre zonas.
- Servidores configurados con alta disponibilidad y redundancia local.
- Servidores configurados sin alta disponibilidad.

Contenido relacionado

- confiabilidad Azure
- mejores prácticas de arquitectura para Azure Database for MySQL
- Información general de la continuidad empresarial con Azure Database for MySQL

ⓘ**Nota:** El autor ha creado este artículo con ayuda de inteligencia artificial. Más información

---

Last updated on 18/06/2026

Confiabilidad en Azure Database for PostgreSQL

Azure Database for PostgreSQL es un servicio de base de datos totalmente administrado que proporciona control y flexibilidad pormenorizados sobre las funciones de administración de bases de datos y las opciones de configuración. El servicio proporciona funcionalidades de alta disponibilidad y recuperación ante desastres en función de sus requisitos.

Cuando se usa Azure, la confiabilidad es una responsabilidad compartida. Microsoft proporciona una variedad de funcionalidades para admitir resistencia y recuperación. Es responsable de comprender cómo funcionan esas funcionalidades dentro de todos los servicios que usa y de seleccionar las funcionalidades que necesita para cumplir los objetivos empresariales y los objetivos de tiempo de actividad.

En este artículo se describe cómo hacer que Azure Database for PostgreSQL sea resistente a diversas interrupciones y problemas potenciales, incluidos fallos transitorios, interrupciones de zonas de disponibilidad, interrupciones regionales y tareas de mantenimiento del servicio. También se describe cómo puede usar copias de seguridad para recuperarse de otros tipos de problemas y se resalta información clave sobre el acuerdo de nivel de servicio (SLA) de Azure Database for PostgreSQL.

Recomendaciones de implementación de producción

Para obtener información sobre cómo implementar Azure Database for PostgreSQL para admitir los requisitos de confiabilidad de la solución y cómo afecta la confiabilidad a otros aspectos de la arquitectura, consulte Procedimientos recomendados de arquitectura para Azure Database for PostgreSQL en Azure Well-Architected Framework.

Introducción a la arquitectura de confiabilidad

En esta sección se describen algunos de los aspectos importantes de cómo funciona el servicio que es más relevante desde una perspectiva de confiabilidad. En la sección se presenta la arquitectura lógica, que incluye algunos de los recursos y características que se implementan y usan. También se describe la arquitectura física, que proporciona detalles sobre cómo funciona el servicio en segundo plano.

Arquitectura lógica

Al trabajar con Azure Database for PostgreSQL, se implementa un *servidor*, que representa los recursos de proceso y almacenamiento necesarios para admitir las *bases de datos* que se implementan en el servidor.

Puede implementar servidores en varios *niveles de computación*: de ráfaga, de propósito general y optimizado para memoria. Cada nivel está optimizado para diferentes tipos de cargas de trabajo. En algunas regiones de Azure, puede implementar servidores con Azure Confidential Computing.

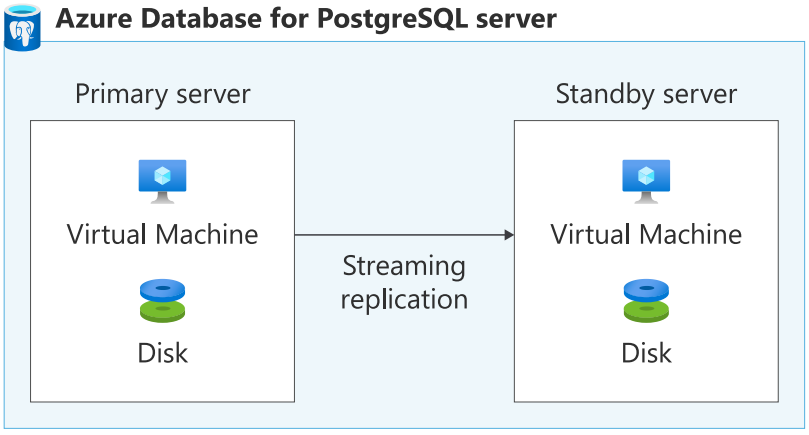
Para obtener más información sobre la arquitectura general del servicio y los modelos de implementación, consulte Azure Database for PostgreSQL información general.

Arquitectura física

- **Separación de proceso y almacenamiento:** Azure Database for PostgreSQL usa una arquitectura de separación de proceso y almacenamiento para admitir la alta disponibilidad. El motor de base de datos se ejecuta en una máquina virtual Linux, mientras que Azure Storage contiene los archivos de datos y mantiene tres copias sincrónicas con redundancia local de los archivos de base de datos para garantizar la durabilidad de los datos.

- **Alta disponibilidad:** Puede habilitar una configuración de alta disponibilidad en el servidor. Al habilitar la configuración de alta disponibilidad, el servicio aprovisiona y mantiene un servidor en espera activo. El servidor principal replica de forma sincrónica los cambios de datos en el servidor en espera para garantizar una pérdida de datos cero durante un error del servidor principal.

  La arquitectura separa la capa de proceso de la capa de almacenamiento, por lo que el servicio puede controlar los distintos tipos de errores de forma adecuada. Para lograr una mayor resistencia, puede distribuir los servidores entre zonas de disponibilidad.



Un servidor en espera se implementa en la misma configuración de máquina virtual que el servidor principal, incluidos los núcleos virtuales, el almacenamiento y la configuración de red.

Puede cambiar entre servidores mediante la realización de una *conmutación por error*. Existen dos tipos de conmutación por error: *las conmutaciones por error forzadas*, que se usan cuando se produce un error en el servidor principal y *las conmutaciones por error planeadas*, que se usan durante algunas operaciones de mantenimiento y en otros escenarios en los que es necesario minimizar el tiempo de inactividad de la aplicación durante una conmutación por error.

Al realizar operaciones como detener, iniciar y reiniciar, se producen en servidores de bases de datos principales y en espera al mismo tiempo. Los eventos planeados, como el escalado de proceso y el escalado de almacenamiento, se producen primero en espera y, a continuación, en el servidor principal. Actualmente, el servidor no conmuta por error para estas operaciones planeadas.

Para más información, consulte Alta disponibilidad en Azure Database for PostgreSQL.

- **Copias de seguridad:** Azure Database for PostgreSQL crea automáticamente copias de seguridad del servidor. Para obtener más información, consulte Copia de seguridad y restauración.

Resistencia a errores transitorios

Los errores transitorios son errores breves e intermitentes en los componentes. Se producen con frecuencia en un entorno distribuido como la nube y son una parte normal de las operaciones. Los errores transitorios se corrigen después de un breve período de tiempo. Es importante que las aplicaciones puedan controlar errores transitorios, normalmente mediante el reintento de solicitudes afectadas.

Todas las aplicaciones hospedadas en la nube deben seguir las instrucciones de control de errores transitorios de Azure cuando se comunican con cualquier API, bases de datos y otros componentes hospedados en la nube. Para obtener más información, consulte Recomendaciones para controlar errores transitorios.

Las aplicaciones deben controlar los errores de conectividad transitorios que pueden producirse durante el mantenimiento, las operaciones de escalado o las interrupciones de red. Siga estas recomendaciones:

- ✓ Cuando la aplicación encuentre errores transitorios, vuelva a intentar la operación mediante retroceso exponencial. Aumente el retraso entre reintentos y limite el número de intentos. Si la operación sigue produciendo un error después de los reintentos máximos, considérela un fracaso.
- ✓ Siempre que sea posible, use bibliotecas cliente (también denominadas controladores) que controlen automáticamente los reintentos.
- ✓ Los errores transitorios que se producen durante las operaciones de escritura requieren una consideración más cuidadosa. Considere la posibilidad de hacer que las operaciones de escritura sean idempotentes, por lo que se pueden ejecutar de forma segura varias veces.

Para más información, consulte Control de errores de conectividad transitorios en Azure Database for PostgreSQL.

Resistencia a errores de zona de disponibilidad

Las zonas de disponibilidad son grupos físicamente independientes de centros de datos dentro de una región de Azure. Cuando una zona falla, los servicios pueden transferirse a una de las zonas restantes.

Seleccione el tipo de compatibilidad con zonas de disponibilidad mediante la configuración de alta disponibilidad. Al habilitar la alta disponibilidad, el servicio implementa un servidor en espera junto con el servidor principal. Este modelo de alta disponibilidad ayuda a garantizar que los datos confirmados nunca se pierdan durante los errores. Sea cual sea el modelo de implementación de alta disponibilidad que utilice su servidor, confirma de forma síncrona los datos tanto en el servidor principal como en el servidor en espera. Si se produce una interrupción en el servidor principal, se conmuta automáticamente al servidor de respaldo.

Cada zona de disponibilidad almacena archivos de datos y registros de escritura anticipada (WAL) en discos administrados Premium con almacenamiento con redundancia local (LRS) que almacena automáticamente tres copias de datos dentro de cada zona.

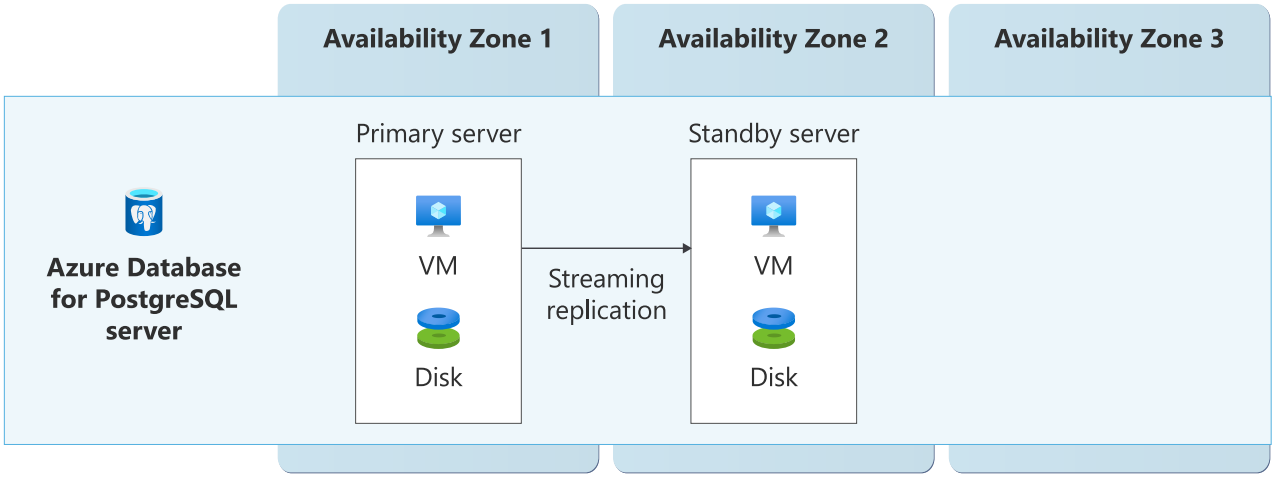
Azure Database for PostgreSQL admite dos tipos de configuración de zona de disponibilidad cuando se usa alta disponibilidad:

- **Alta disponibilidad con redundancia de zona:** La redundancia de zona proporciona el nivel más alto de resistencia de zona mediante la implementación de un servidor principal en una zona de disponibilidad y un servidor en espera en una zona de disponibilidad diferente. El servidor en espera usa la configuración de proceso, almacenamiento y red similar a la del

servidor principal. Una configuración con redundancia de zona proporciona aislamiento físico de toda la pila entre los servidores principal y en espera.

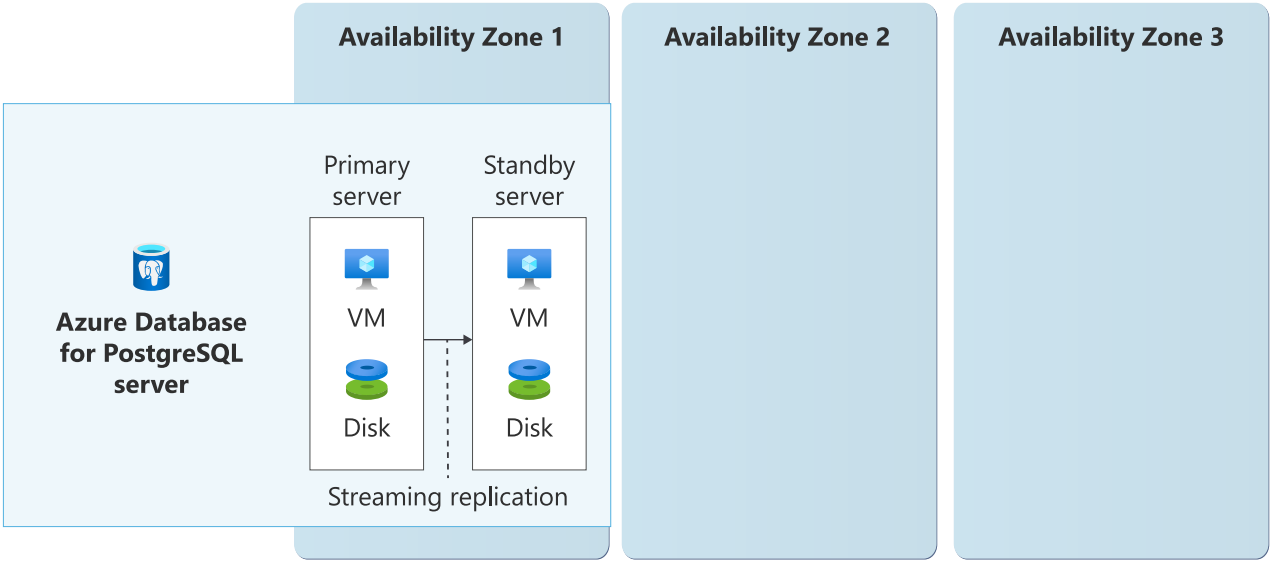
Puede seleccionar las zonas de disponibilidad para los servidores principales y en espera o permitir que Microsoft los elija.

Se recomiendan implementaciones con redundancia de zona para servidores de producción.



Las operaciones de escritura pueden experimentar un pequeño aumento de la latencia de confirmación porque el servicio replica de forma sincrónica los datos en el servidor en espera. El impacto varía según la carga de trabajo, la SKU seleccionada y la región.

- **Alta disponibilidad zonal (misma zona):** Los servidores principales y en espera usan la misma zona de disponibilidad. Si se produce una interrupción en el servidor principal, pero la zona sigue operativa, el servidor conmuta automáticamente al servidor en espera. Una implementación zonal proporciona alta disponibilidad dentro de una sola zona de disponibilidad. Protege contra errores de nivel de nodo y también ayuda a reducir el tiempo de inactividad de la aplicación durante los eventos de tiempo de inactividad planeados y no planeados. Sin embargo, no protege contra una interrupción en esa zona.



La alta disponibilidad zonal (misma zona) solo está disponible en las situaciones siguientes:

- La región no admite zonas de disponibilidad. La región funciona eficazmente como una sola zona, por lo que la única configuración de alta disponibilidad que puede seleccionar es la misma zona.
- Si una región no tiene capacidad suficiente para una implementación con redundancia de zona, el servicio puede colocar inicialmente ambos servidores en la misma zona de disponibilidad y, a continuación, migrarlos automáticamente a zonas independientes cuando la capacidad esté disponible. Esta opción está disponible cuando se usa Azure Portal o la CLI de Azure para implementar un servidor. Para obtener más información, consulte Configuración de opciones críticas para la empresa (alta disponibilidad).

Colocar los servidores en la misma zona puede reducir la latencia de escritura en las aplicaciones que implemente dentro de la misma zona.

Cuando los servidores están en la misma zona, se puede reducir la latencia de escritura en las aplicaciones que implemente dentro de la misma zona.

Si configura el servidor sin alta disponibilidad, se ejecuta en un solo servidor. Si ese servidor o su zona deja de funcionar, el servidor no está disponible. Para obtener más información, consulte Configuraciones sin zonas de disponibilidad.

Requisitos

- **Compatibilidad regional:** Azure Database for PostgreSQL admite configuraciones de zona de disponibilidad de manera diferente según las regiones de Azure. Para obtener una lista completa de las regiones, los tipos de compatibilidad con las zonas de disponibilidad y las consideraciones específicas de cada región, consulte Regiones de Azure.

- **Nivel de proceso:** En la tabla siguiente se muestra la compatibilidad con el nivel de proceso para cada tipo de compatibilidad con zonas de disponibilidad:

Expandir tabla

Nivel de computación	Con redundancia de zona	Zonal (misma zona)
Ampliable	No soportado	No soportado
General Purpose	Soportado	Soportado
Memoria optimizada	Soportado	Soportado

- **Nivel de servicio:** Ambos tipos de alta disponibilidad requieren niveles de uso general o Optimizado para memoria.

Consideraciones

Capacidad de región: Si una región no tiene capacidad suficiente para una implementación con redundancia de zona, el servicio puede colocar inicialmente ambos servidores en la misma zona de disponibilidad y migrarlos automáticamente a zonas independientes cuando la capacidad esté disponible. Esta opción está disponible cuando se usa Azure Portal o la CLI de Azure para implementar un servidor. Para obtener más información, consulte Configuración de opciones críticas para la empresa (alta disponibilidad).

Costo

Al habilitar la alta disponibilidad, se crea un servidor en espera y se factura a la misma velocidad que el servidor principal. La configuración de la zona de disponibilidad no afecta al costo. No hay cargos por la replicación de datos dentro o entre zonas de disponibilidad. Dependiendo del volumen de almacenamiento de copia de seguridad, también se le puede facturar por el almacenamiento de copia de seguridad. Para obtener información detallada sobre los precios, consulte Precios de Azure Database for PostgreSQL.

Configurar soporte de zonas de disponibilidad

Para configurar la compatibilidad con zonas de disponibilidad para un servidor, configure las opciones de alta disponibilidad.

- **Cree un servidor con redundancia de zona:** Para obtener información sobre cómo crear un servidor con alta disponibilidad y redundancia de zona habilitada, consulte Inicio rápido: Creación de un servidor de Azure Database for PostgreSQL.

- **Cambie la configuración de zona de disponibilidad para los servidores existentes:** Cambie la configuración de zona de disponibilidad para los servidores existentes cambiando la configuración de alta disponibilidad. Para ver los pasos detallados, consulte Habilitación de la alta disponibilidad para los servidores existentes.

  No se puede cambiar la zona usada para el servidor principal o en espera. Necesita volver a crear el servidor.

  > **Sugerencia**
  >
  > Se recomienda esperar hasta que la actividad del servidor sea baja antes de cambiar la configuración de alta disponibilidad.

- **Deshabilitar alta disponibilidad:** Al deshabilitar la alta disponibilidad, se quita el servidor en espera, por lo que el servidor no es resistente a interrupciones en su zona de disponibilidad. Para obtener más información, consulte Deshabilitación de la alta disponibilidad.

Comportamiento cuando todas las zonas están en buen estado

En esta sección se describe lo que cabe esperar al configurar servidores con alta disponibilidad y compatibilidad con zonas de disponibilidad, cuando todas las zonas de disponibilidad están operativas.

- **Operación entre zonas:** Las aplicaciones cliente de PostgreSQL se conectan al servidor principal mediante el nombre del servidor de base de datos. Azure Database for PostgreSQL usa una configuración activa-pasiva en la que el servidor principal de la zona de disponibilidad principal controla todas las conexiones y consultas de base de datos. El servidor en espera no atiende el tráfico de cliente durante las operaciones normales.

- **Replicación de datos entre zonas:** El servidor principal replica de forma sincrónica los cambios en el servidor en espera. Las transacciones no se consideran completas hasta que los servidores principal y en espera confirmen la escritura.

  Cuando una aplicación escribe y confirma los datos, PostgreSQL registra primero el cambio en el WAL en el servidor primario. El servidor principal transmite estos registros al servidor en espera mediante el protocolo de streaming de PostgreSQL. Después de que el servidor en espera almacene duraderamente el WAL, el servidor principal confirma la escritura. La aplicación confirma su transacción solo después de esta confirmación. Este proceso de confirmación no espera a que los registros se apliquen al servidor en espera.

Los efectos de la replicación son diferentes en función de la configuración de zona de disponibilidad que use el servidor:

- *Con redundancia de zona:* Dado que los servidores están en zonas independientes, este enfoque garantiza una pérdida de datos cero durante un error de zona. Esta situación también se denomina a veces lograr un objetivo de punto de recuperación (RPO) de cero para los errores de zona.

  Sin embargo, la replicación entre zonas podría introducir una pequeña cantidad de latencia adicional. El impacto de la latencia depende de la aplicación. Para la mayoría de las aplicaciones, la latencia adicional es insignificante.

- *Zonal*: dado que ambos servidores están en la misma zona, no se replica ningún tráfico entre zonas.

> ⓘ **Nota:**
>
> El sistema replica los datos de registro en tiempo real en el servidor en espera. Los errores de usuario en el servidor principal, como una eliminación accidental de una tabla o actualizaciones de datos incorrectas, se replican en el servidor en espera. No puede usar el modo de espera para recuperarse de estos tipos de errores y debe realizar una restauración a un momento dado desde la copia de seguridad. Para obtener más información, consulte **Copia de seguridad y restauración**.

Comportamiento durante un fallo de zona

En esta sección se describe qué puede esperar al configurar servidores con alta disponibilidad y compatibilidad con zonas de disponibilidad, si se produce una interrupción en una zona de disponibilidad.

- **Detección y respuesta:** Azure comprueba periódicamente el estado de los servidores principal y en espera. Después de varios pings, si la supervisión de estado detecta que no se puede acceder a un servidor principal, el servicio inicia una conmutación automática al servidor en espera. El algoritmo de supervisión de estado usa varios puntos de datos para evitar situaciones de falsos positivos.

  Si se produce un error en una zona de disponibilidad, el comportamiento es diferente en función de la configuración de la zona de disponibilidad que usa el servidor:

- *Con redundancia de zona:* Azure Database for PostgreSQL detecta automáticamente fallos en la zona de disponibilidad. Para ver los posibles tipos de estado de alta disponibilidad, consulte Supervisión del estado de alta disponibilidad (HA). Cuando se produce un error en una zona, Azure inicia una conmutación por error forzada al servidor en espera sin que sea necesario realizar ninguna acción.

- *Zonal:* Si la zona de disponibilidad que hospeda un servidor zonal deja de estar disponible, los servidores principal y en espera no están disponibles. En este escenario, el servicio no proporciona conmutación automática por error. Es responsable de detectar la interrupción de zona y realizar acciones de recuperación, como restaurar copias de seguridad con redundancia de zona en un servidor independiente en otra zona de disponibilidad o región.

- **Notificación:** La supervisión del estado de mantenimiento de alta disponibilidad en Azure Database for PostgreSQL proporciona información general continua sobre el estado y la preparación de instancias habilitadas para alta disponibilidad. La característica de supervisión se basa en Azure Resource Health y puede detectar y alertar sobre cualquier problema que pueda afectar a la preparación de la conmutación por error de la base de datos o a la disponibilidad general. Evalúe métricas clave como el estado de conexión, el estado de conmutación por error y el estado de replicación de datos para habilitar la solución de problemas proactiva y ayudar a mantener el tiempo de actividad y el rendimiento de la base de datos.

  Para obtener una guía detallada sobre la configuración y la interpretación de los estados de salud de alta disponibilidad, consulte Supervisión de los estados de salud de alta disponibilidad (HA).

- **Solicitudes activas:** Cuando una zona de disponibilidad deja de estar disponible, es posible que se finalicen las solicitudes en curso a los servidores de la zona afectada. Las aplicaciones deben reintentar estas solicitudes. Si los clientes controlan correctamente los errores transitorios mediante el reintento después de un breve período de tiempo, normalmente evitan un impacto significativo.

- **Pérdida de datos esperada:** La cantidad de pérdida de datos depende de la configuración de zona de disponibilidad que usa el servidor.

  - *Redundante por zona:* Se anticipa una pérdida de datos cero durante el fallo de zona debido a la replicación sincrónica entre los servidores principal y en espera en distintas zonas.

- *Zonal:* Los datos de los servidores de la zona afectada no están disponibles hasta que se recupere la zona.

- **Tiempo de inactividad esperado:** La cantidad de tiempo de inactividad depende de la configuración de zona de disponibilidad que usa el servidor.
  - *Con redundancia de zona:* La conmutación por error se completa normalmente en un plazo de 60 a 120 segundos. Si los clientes controlan correctamente los errores transitorios mediante el reintento después de un breve período de tiempo, normalmente evitan un impacto significativo.
  - *Zonal:* Los servidores de una zona afectada no están disponibles hasta que se recupere la zona de disponibilidad.

- **Redistribución:** El comportamiento de reenrutamiento del tráfico depende de la configuración de zona de disponibilidad que usa el servidor.
  - *Con redundancia de zona:* Después de la conmutación por error, el servidor en espera anterior se convierte en el nuevo servidor principal y comienza a aceptar nuevas conexiones. Azure establece automáticamente un nuevo servidor en espera en la zona principal original después de recuperarse. Para más información, consulte Conmutación por error forzada.
  - *Zonal:* Cuando una zona no está disponible, el servidor no está disponible. Si tiene un servidor independiente que creó con antelación en otra región o zona de disponibilidad, es responsable de redirigir el tráfico a ese servidor.

Recuperación de zona

El comportamiento de recuperación de zona depende de la configuración de zona de disponibilidad que usa el servidor.

- *Con redundancia de zona:* Cuando se recupera la zona de disponibilidad, Azure Database for PostgreSQL vuelve a generar automáticamente el servidor de reserva en la zona recuperada y lo sincroniza con el servidor principal actual. A continuación, la zona recuperada actúa como ubicación en espera. Para evitar interrupciones innecesarias, el servicio no traslada automáticamente de nuevo el rol principal a la zona original. Puede iniciar manualmente una conmutación por error planeada si desea devolver la principal a la zona original.

- *Zonal:* Una vez que la zona está en buen estado, los servidores de la zona están disponibles de nuevo. Es responsable de cualquier procedimiento de recuperación de zona y

sincronización de datos que requieran las cargas de trabajo.

Prueba de fallos de zona

Las opciones para probar errores de zona dependen de la configuración de zona de disponibilidad que usa la instancia.

- *Con redundancia de zona:* Puede probar la resistencia de la aplicación para la conmutación por error iniciando una *conmutación por error forzada*. Una conmutación por error forzada le permite simular un evento no planificado durante la ejecución de su carga de trabajo, permitiéndole observar el tiempo de inactividad de su aplicación. Se recomienda ejecutar simulaciones en un entorno que no sea de producción o en un momento silencioso. Para obtener más información, consulte Iniciar una conmutación por error forzada.

- *Zonal:* Aunque no se puede simular una interrupción de zona completa, puede simular que el servidor no está disponible de una manera similar a una interrupción de zona. Para obtener más información, vea Detener el proceso de un servidor.

Resistencia a errores en toda la región

Azure Database for PostgreSQL admite *réplicas de lectura entre regiones*, que puede usar para mantener una copia sincronizada de la base de datos en otra región para una recuperación más rápida.

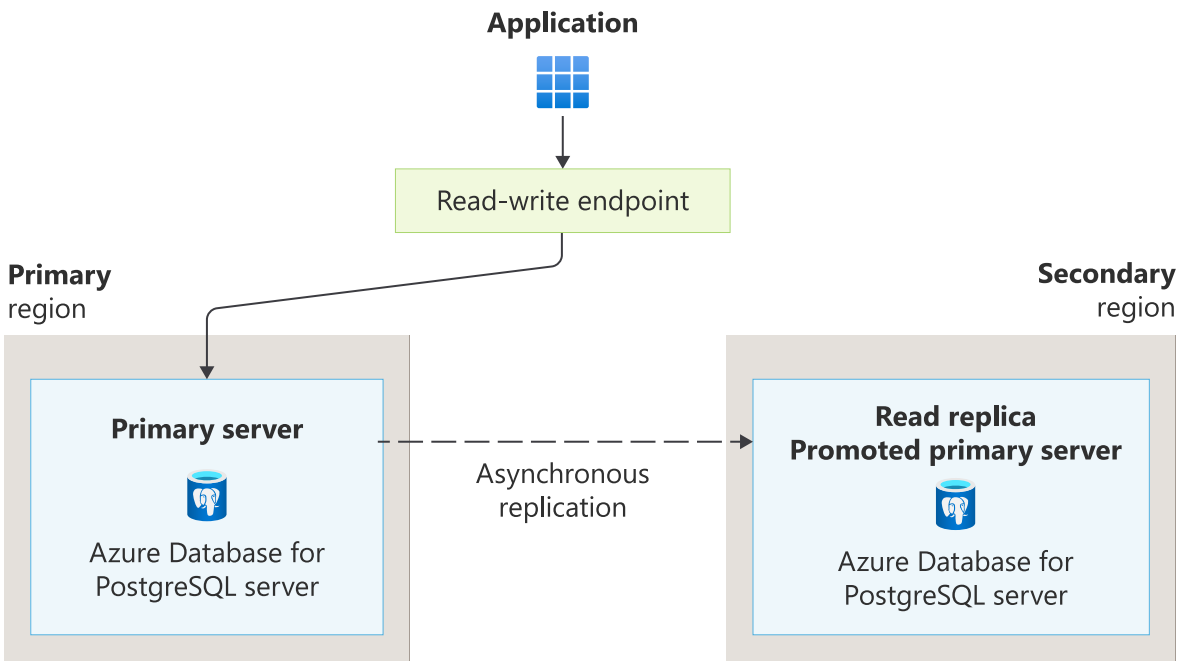
También puede usar copias de seguridad con redundancia geográfica, en regiones admitidas, para proporcionar recuperación entre regiones. Sin embargo, las copias de seguridad suelen implicar más tiempo de inactividad y pérdida de datos que la replicación. Para obtener más información, consulte Copia de seguridad y restauración.

Réplicas de lectura entre regiones

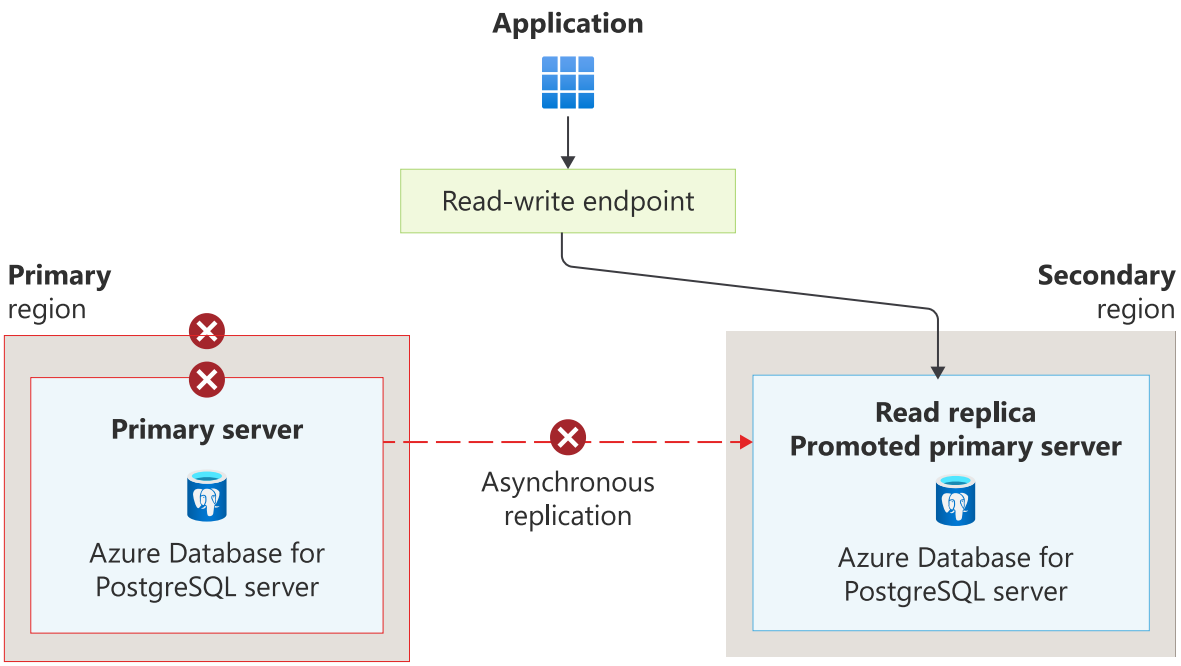
Puede implementar réplicas de lectura para proteger las bases de datos frente a errores de nivel de región. Cada réplica de lectura es un servidor independiente de Azure Database for PostgreSQL. Cuando coloca una réplica de lectura en una segunda región de Azure, el servidor de bases de datos puede proporcionar resiliencia ante un problema que afecte a toda la región. Puede implementar hasta cinco réplicas de lectura, las cuales opcionalmente pueden estar en distintas regiones de Azure. La tecnología de replicación física de PostgreSQL actualiza las réplicas de lectura de forma asincrónica y pueden retardar la principal. Las réplicas de lectura entre regiones pueden servir opcionalmente cargas de trabajo de solo lectura para reducir la latencia de las aplicaciones distribuidas globalmente o descargar el tráfico de lectura desde el

servidor principal. Para obtener más información sobre las características y consideraciones de réplica de lectura, consulte Réplicas de lectura.

Los puntos de conexión virtuales proporcionan puntos de conexión de lectura-escritura y de solo lectura, y redirigen automáticamente el tráfico cuando se promueve una réplica, lo que ayuda a minimizar el tiempo de inactividad durante las conmutaciones por error. Se recomienda encarecidamente usar puntos de conexión virtuales con réplicas de lectura entre regiones para mejorar la resistencia de la aplicación. Para más información, consulte Puntos de conexión virtuales para réplicas de lectura en Azure Database for PostgreSQL.



Si se produce un error en la región primaria, puede desencadenar una *promoción* para que la réplica secundaria se convierta en la principal. Es posible que distintos tipos de conmutación por error sean adecuados en función de cómo se usen las réplicas de lectura. Cuando se usan réplicas de lectura para proporcionar resistencia a errores de región, normalmente se usa el enfoque *de promoción al servidor principal* , que actualiza el punto de conexión virtual. Durante una interrupción de la región, debe realizar una *promoción forzada*, lo que puede dar lugar a una pérdida de datos para los datos no replicados. En escenarios planificados en los que la región primaria está en buen estado, puede optar por realizar una promoción planificada para evitar la pérdida de datos. Para más información, consulte Promoción de réplicas de lectura en Azure Database for PostgreSQL.



> **Nota:**
>
> En esta sección se resume información importante sobre cómo las réplicas de lectura pueden contribuir a la resiliencia frente a fallos que afectan a toda una región. También puede usar réplicas de lectura para mejorar el rendimiento y admitir bases de usuarios distribuidas geográficamente a gran escala. Para obtener más información, consulte Réplicas de lectura.

Requisitos

- **Compatibilidad con regiones:** Puede crear réplicas de lectura entre regiones en cualquier región que admita Azure Database for PostgreSQL. No está limitado a las regiones emparejadas de Azure.
- **Niveles de cómputo:** Los niveles de cómputo de uso general y optimizados para memoria admiten réplicas de lectura. El nivel Ampliable no admite réplicas de lectura.

Consideraciones

- **Diferencias de configuración:** Es posible que las réplicas de lectura no hereden todas las opciones de configuración del servidor principal. Planee configurar las opciones necesarias después de la conmutación por error. El servidor principal y las réplicas deben ser *simétricos*, lo que significa que deben tener los mismos niveles, tamaños de almacenamiento y valores para algunas configuraciones. Durante fallos en la región, se puede renunciar al requisito de tener un servidor simétrico para promociones forzadas,

pero se recomienda mantener una configuración simétrica siempre que sea posible para evitar problemas inesperados. Para obtener más información, consulte Administración de configuración.

- **Supervisión del retraso de replicación:** El proceso de replicación asíncrona conlleva un retraso de replicación, que puede variar en función de muchos factores. Cuando el retraso de replicación es alto, el servidor podría experimentar problemas. Es importante supervisar el retraso de replicación para que pueda mitigar los problemas antes de que se escalen. Para más información, consulte Supervisión de la replicación.

- **Alta disponibilidad:** Las réplicas de lectura no pueden tener alta disponibilidad habilitada y, cuando se promueven, tampoco tienen alta disponibilidad. Eres responsable de configurar la alta disponibilidad después de promover una réplica.

Para ver otros factores sobre el proceso de promoción que se debe tener en cuenta, consulte Consideraciones.

Costo

Las réplicas de lectura generan costes de computación y almacenamiento, además de cargos por transferencia de datos entre regiones para la replicación. Para obtener información detallada sobre los precios, consulte Precios de Azure Database for PostgreSQL y Precios de ancho de banda.

Configuración de la compatibilidad con varias regiones

- **Cree una réplica de lectura:** Para obtener información sobre cómo crear una réplica de lectura, consulte Creación de una réplica de lectura. Puede configurar réplicas después de crear el servidor principal, siempre que el servidor principal se ejecute y sea accesible.

  Para crear un punto de conexión virtual, consulte Creación de puntos de conexión virtuales.

- **Eliminar una réplica de lectura:** Para obtener información sobre cómo eliminar una réplica de lectura, consulte Eliminación de una réplica de lectura.

Comportamiento cuando todas las regiones están en buen estado

En esta sección se describe qué esperar cuando el servidor está configurado con una réplica de lectura en otra región y un punto de conexión virtual, y todas las regiones están operativas:

- **Enrutamiento del tráfico entre regiones:** En las operaciones normales, el punto de conexión virtual dirige el tráfico del punto de conexión de lectura y escritura al servidor principal de la región primaria. Si también utiliza el punto de conexión virtual de solo lectura, dirige el tráfico a la réplica que haya configurado.

- **Replicación de datos entre regiones:** Las réplicas de lectura entre regiones usan la replicación asincrónica para minimizar el impacto en el rendimiento del servidor principal. La cantidad de retraso de replicación depende de muchos factores, incluida la carga de escritura y la latencia entre el servidor principal y las réplicas. El retraso de replicación suele ser al menos varios minutos, pero puede ser más largo. Para más información, consulte Supervisión de la replicación.

Comportamiento durante una falla de región

En esta sección se describe qué esperar cuando el servidor está configurado con una réplica de lectura en otra región y un punto de conexión virtual, y hay una interrupción en la región primaria:

- **Detección y respuesta:** Usted es responsable de detectar una interrupción en la región primaria y promover manualmente una réplica de lectura como el nuevo servidor principal. Durante una interrupción regional, debe realizar una promoción forzada, lo que resulta en la pérdida de cualquier dato no replicado.

  > **Importante**
  >
  > Usted es responsable de activar la promoción. Azure no promueve las réplicas de lectura automáticamente, incluso si se produce una falla en una región.

  Para obtener pasos detallados para comenzar una promoción, consulte Convertir réplica de lectura en principal.

- **Notificación:** Microsoft no le notifica automáticamente cuando una región está inactiva. Sin embargo, puede usar Azure Service Health para comprender el estado general del servicio, incluidos los errores de región, y puede configurar alertas de Service Health para notificarle problemas.

- **Peticiones activas:** El proceso de promoción interrumpe todas las conexiones activas con la región principal. Una vez completado el proceso de promoción, las aplicaciones deben reintentar la realización de conexiones a la réplica promocionada.

- **Pérdida de datos esperada:** Durante una interrupción de la región, debe realizar una promoción obligatoria, lo que da lugar a la pérdida permanente de los datos no replicados.

  La cantidad de pérdida de datos depende del retraso de replicación en el momento de la interrupción. El retraso de replicación suele ser al menos varios minutos, pero puede ser más largo. Para más información, consulte Supervisión de la replicación.

- **Tiempo de inactividad esperado:** La promoción forzada se completa normalmente en un plazo de 1 a 3 minutos después de desencadenarse. Es posible que las aplicaciones también necesiten volver a conectarse al punto de conexión correcto. Los puntos de conexión virtuales se actualizan como parte del proceso de promoción forzada. Las aplicaciones deben respetar el período de vida (TTL) de los registros DNS del punto de conexión para asegurarse de que se vuelven a conectar rápidamente a la réplica correcta una vez completada la promoción.

- **Reenrutamiento del tráfico:** El punto de conexión virtual del servidor redirige automáticamente el tráfico de la aplicación a la nueva réplica principal.

  > ⓘ **Nota:**
  >
  > Después de promover una réplica de lectura para que sea el servidor principal, no tiene habilitada la configuración de alta disponibilidad. Debe habilitar la configuración de alta disponibilidad manualmente o agregarla a sus propios procesos de automatización.

Recuperación de regiones

Cuando se usan puntos de conexión virtuales, una vez recuperada la región primaria, el servidor principal antiguo se configura automáticamente como una réplica de lectura. Puede realizar otra promoción para devolver las operaciones principales a la región primaria preferida.

Prueba de fallos de región

Pruebe periódicamente los procedimientos de promoción de réplica de lectura para asegurarse de que los procesos son válidos y que las funcionalidades cumplen los requisitos del objetivo de tiempo de recuperación (RTO) y del objetivo de punto de recuperación (RPO).

Puede promover una réplica de lectura para convertirse en el servidor principal en cualquier momento, incluso cuando todas las regiones estén en buen estado. Para las pruebas:

- Puede realizar pruebas de promoción forzadas. Se recomienda realizar estas pruebas en un entorno que no sea de producción, ya que puede provocar la pérdida de datos. Las pruebas de promoción forzadas ayudan a simular el comportamiento que se observa durante una interrupción de una región.
- Para el mantenimiento planeado o los escenarios de prueba en los que desea evitar la pérdida de datos, use una promoción planeada en su lugar. Sin embargo, la promoción planeada sigue un proceso diferente al de la promoción durante una interrupción de la región, por lo que es posible que no refleje el comportamiento durante una interrupción de región verdadera.

Para obtener instrucciones paso a paso, consulte Conmutar réplica de lectura a principal.

Como parte de la estrategia de recuperación ante desastres, ejecute periódicamente los simulacros de recuperación completos. Estos simulacros deben incluir la validación de datos, las pruebas de funcionalidad de la aplicación y los procedimientos de reversión documentados.

Copias de seguridad y restauración

Azure Database for PostgreSQL realiza automáticamente una copia de seguridad de los datos. Estas copias de seguridad proporcionan funcionalidades de recuperación a un momento dado y le ayudan a protegerse frente a daños accidentales y la eliminación de datos. Microsoft administra completamente las copias de seguridad. No interrumpen la disponibilidad del servidor e incluyen copias de seguridad completas y copias de seguridad del registro de transacciones.

- **Almacenamiento de copia de seguridad:** Si implementa el servidor en una región con zonas de disponibilidad, el servicio almacena copias de seguridad en el almacenamiento con redundancia de zona (ZRS), independientemente de la configuración de alta disponibilidad del servidor. En el caso de los servidores implementados en regiones sin zonas de disponibilidad, el servicio almacena copias de seguridad en el almacenamiento con redundancia local (LRS).

  En las regiones de Azure con pares, puede configurar el almacenamiento de copia de seguridad con redundancia geográfica al crear el servidor para replicar las copias de seguridad en la región emparejada de Azure y obtener protección adicional frente a fallos regionales. El servicio replica las copias de seguridad de forma asincrónica.

  El período de retención de copia de seguridad predeterminado es de siete días, pero puede ampliar la retención hasta 35 días. También puede usar Azure Backup para el

almacenamiento a largo plazo de copias de seguridad manuales durante hasta 10 años. Todas las copias de seguridad están cifradas.

- **Restaurar:** La recuperación a un momento dado le permite restaurar la base de datos a cualquier momento dentro del período de retención de copia de seguridad. El proceso de restauración crea un nuevo servidor de bases de datos con un nuevo nombre de servidor proporcionado por el usuario. Puede usar el nuevo servidor as-is o copiar datos de él.

  Al restaurar una copia de seguridad con redundancia geográfica, se crea un nuevo servidor en la región emparejada.

  Esta funcionalidad es útil para recuperarse de modificaciones accidentales de datos, errores de aplicación o escenarios de prueba.

Para la mayoría de las soluciones, no debe confiar exclusivamente en copias de seguridad. En su lugar, utilice las otras capacidades descritas en esta guía para apoyar los requisitos de resiliencia. Sin embargo, las copias de seguridad protegen contra algunos riesgos que otros enfoques no. Para más información, consulte ¿Qué son la redundancia, la replicación y la copia de seguridad?.

Para más información, consulte Copia de seguridad y restauración en Azure Database for PostgreSQL.

Resistencia al mantenimiento del servicio

Azure Database for PostgreSQL controla automáticamente las tareas de mantenimiento críticas, incluida la aplicación de revisiones al hardware subyacente, el sistema operativo y el motor de base de datos. El servicio incluye actualizaciones de seguridad, actualizaciones de software y actualizaciones de versiones secundarias como parte del mantenimiento planeado.

Para asegurarse de que el servidor permanece disponible durante las ventanas de mantenimiento, siga estas recomendaciones:

- ✓ **Habilitar alta disponibilidad:** Durante el mantenimiento, es posible que el servidor tenga que reiniciarse como parte del proceso de actualización. Si habilita alta disponibilidad, las operaciones de mantenimiento suelen usar actualizaciones graduales para minimizar el tiempo de inactividad. Las actividades de mantenimiento periódicas, como las actualizaciones de versiones secundarias, se producen primero en la réplica en espera. Para reducir el tiempo de inactividad, el modo de espera se promueve a principal para que las cargas de trabajo puedan continuar en el nodo promocionado mientras se aplican las tareas

de mantenimiento al otro nodo. Esta secuenciación se aplica tanto si el servidor utiliza alta disponibilidad con redundancia de zona como si utiliza alta disponibilidad zonal.

En el caso de los servidores sin alta disponibilidad habilitada, espere un breve tiempo de inactividad durante las operaciones de mantenimiento. Con alta disponibilidad habilitada, las operaciones de mantenimiento normalmente se completan con un tiempo de inactividad mínimo o sin tiempo de inactividad.

- ✓ **Configurar ventanas de mantenimiento personalizadas:** Puede configurar la programación de mantenimiento para administrar el sistema o definir una ventana de mantenimiento personalizada para minimizar el impacto en las operaciones empresariales. Programe el mantenimiento durante períodos de baja actividad para minimizar el impacto empresarial. Para obtener más información, consulte Programación del mantenimiento.
- ✓ **Implemente la lógica de reintento:** Asegúrese de que las aplicaciones pueden controlar breves interrupciones de conectividad que pueden producirse durante los reinicios de mantenimiento. Para que las aplicaciones sean resistentes a estos tipos de problemas, consulte Guía sobre resistencia a errores transitorios .

Acuerdo de nivel de servicio

El acuerdo de nivel de servicio (SLA) para Azure servicios describe la disponibilidad esperada de cada servicio y las condiciones que la solución debe cumplir para lograr esa expectativa de disponibilidad. Para obtener más información, consulte Acuerdos de Nivel de Servicio para servicios en línea .

Azure Database for PostgreSQL proporciona diferentes acuerdos de nivel de servicio de disponibilidad, en función de la configuración del servidor:

- Los servidores configurados con redundancia de zona de alta disponibilidad ofrecen un acuerdo de nivel de servicio (SLA) de tiempo de actividad del 99,99%.
- Los servidores configurados con alta disponibilidad zonal ofrecen un SLA con un tiempo de actividad del 99,95%.
- Los servidores configurados sin alta disponibilidad ofrecen un SLA del 99,9%.

Contenido relacionado

- confiabilidad Azure
- Procedimientos recomendados de arquitectura para Azure Database for PostgreSQL
- Introducción a la continuidad empresarial con Azure Database for PostgreSQL

- Recuperación ante desastres geográfica en Azure Database for PostgreSQL
- Scripts de Terraform para implementar principios de resiliencia y recuperación

ⓘ**Nota:** El autor ha creado este artículo con ayuda de inteligencia artificial. Más información

---

Last updated on 19/06/2026

Confiabilidad en Azure DocumentDB

Este artículo contiene información detallada sobre la resistencia regional con zonas de disponibilidad y recuperación ante desastres entre regiones y continuidad empresarial para Azure DocumentDB.

Para obtener información general sobre la confiabilidad de la arquitectura en Azure, consulte confiabilidad de Azure.

Soporte para zonas de disponibilidad

Las zonas de disponibilidad son grupos físicamente independientes de centros de datos dentro de una región de Azure. Cuando una zona falla, los servicios pueden transferirse a una de las zonas restantes.

Para obtener compatibilidad con zonas de disponibilidad, debe habilitar alta disponibilidad (ALTA DISPONIBILIDAD).

La alta disponibilidad evita el tiempo de inactividad de la base de datos manteniendo las réplicas en espera de todas las particiones de un clúster. Si una partición deja de funcionar, Azure DocumentDB cambia las conexiones entrantes de la partición con errores a su réplica en espera.

Cuando la alta disponibilidad está habilitada en una región que admite zonas de disponibilidad, las particiones de réplica de alta disponibilidad se aprovisionan en una zona de disponibilidad diferente de sus particiones principales. Las réplicas de alta disponibilidad no reciben solicitudes de los clientes a menos que su extensión principal falle.

Si la alta disponibilidad está deshabilitada, cada partición tiene su propio almacenamiento con redundancia local (LRS) con tres réplicas sincrónicas mantenidas por el servicio Azure Storage. Si hay un error de una sola réplica, el servicio Azure Storage detecta el error y vuelve a crear de forma transparente los datos relevantes. Para obtener durabilidad del almacenamiento LRS, consulte Resumen de las opciones de redundancia. Sin embargo, si se produce un error en una región, se corre el riesgo de un tiempo de inactividad extenso y posible pérdida de datos.

Creación de un recurso con zonas de disponibilidad habilitadas

Para habilitar las zonas de disponibilidad, debe habilitar alta disponibilidad (HA) al Crear un clúster o en la sección**Escalado** de un clúster existente en Azure Portal.

Recuperación ante desastres entre regiones y continuidad empresarial

La recuperación ante desastres (DR) hace referencia a las prácticas que las organizaciones usan para recuperarse de eventos de alto impacto, como desastres naturales o implementaciones con errores que producen tiempo de inactividad y pérdida de datos. Independientemente de la causa, el mejor remedio para un desastre es un plan de recuperación ante desastres bien definido y probado y un diseño de aplicaciones que apoye activamente la recuperación ante desastres. Antes de empezar a crear el plan de recuperación ante desastres, consulte Recomendaciones para diseñar una estrategia de recuperación ante desastres.

Para la recuperación ante desastres, Microsoft usa el modelo de responsabilidad compartida. En este modelo, Microsoft garantiza que la infraestructura de línea base y los servicios de plataforma estén disponibles. Sin embargo, muchos servicios de Azure no replican automáticamente datos ni se revierten de una región con errores para realizar la replicación cruzada en otra región habilitada. En esos servicios, es responsable de configurar un plan de recuperación ante desastres válido para su carga de trabajo. La mayoría de los servicios que se ejecutan en ofertas de plataforma como servicio (PaaS) de Azure proporcionan características e instrucciones para admitir la recuperación ante desastres. Puede usar características específicas del servicio para admitir la recuperación rápida con el fin de contribuir al desarrollo del plan de recuperación ante desastres.

Azure DocumentDB no proporciona conmutación automática por error ni recuperación ante desastres integrada. La planificación de la alta disponibilidad es un paso fundamental a medida que se escala la solución.

Recuperación ante desastres en una sola región geográfica

Para maximizar el tiempo de actividad, planee el mantenimiento de la continuidad empresarial y prepárese para la recuperación ante desastres con Azure DocumentDB.

Aunque los servicios de Azure están diseñados para maximizar el tiempo de actividad, es posible que se produzcan interrupciones de servicio no planeadas. Un plan de recuperación ante desastres garantiza que tiene una estrategia para abordar las interrupciones de servicio en regiones.

Azure DocumentDB realiza automáticamente copias de seguridad de los datos a intervalos regulares. Las copias de seguridad automáticas se crean sin afectar al rendimiento o a la disponibilidad de las operaciones de las bases de datos. Todas las copias de seguridad se realizan de forma automática en segundo plano y se almacenan por separado de los datos de origen en un servicio de almacenamiento. Estas copias de seguridad automáticas son útiles en

escenarios en los que se eliminan o modifican recursos de manera accidental y, posteriormente, se necesitan las versiones originales.

Las copias de seguridad automáticas se conservan en varios intervalos en función de si el clúster está activo o se ha eliminado recientemente.

Expandir tabla

	Período de retención
Clústeres activos	`35` días
Clústeres eliminados	`7` días

Diseño para lograr alta disponibilidad

La alta disponibilidad (HA) debe estar habilitada para clústeres críticos de Azure DocumentDB que ejecutan cargas de trabajo de producción. En un clúster habilitado para alta disponibilidad, cada partición actúa como principal junto con una partición en espera activa aprovisionada en otra zona de disponibilidad. La replicación entre la partición principal y la secundaria es sincrónica de forma predeterminada. Cualquier modificación de la base de datos se conserva en las particiones principal y secundaria (en espera activa) antes de recibir una respuesta de la base de datos.

El servicio mantiene comprobaciones de estado y latidos en cada partición principal y secundaria del clúster. Si una partición principal deja de estar disponible debido a una interrupción regional o de zona, la partición secundaria se promueve automáticamente para convertirse en la nueva principal y se crea una partición secundaria posterior para la nueva principal. Además, si una partición secundaria deja de estar disponible, el servicio crea automáticamente una nueva partición secundaria con una copia completa de datos de la principal.

Si el servicio desencadena una conmutación por error de la partición principal a la secundaria, las conexiones se enrutan sin problemas en segundo plano a la nueva partición principal.

La replicación sincrónica entre las particiones principal y secundaria garantiza ninguna pérdida de datos si hay una conmutación por error.

Pasos siguientes

- Obtenga más información sobre la compatibilidad de características con MongoDB.
- Revisión de las opciones para migrar de MongoDB a Azure DocumentDB

- Para empezar, cree una cuenta.

Last updated on 22/01/2026

Confiabilidad en Azure Managed Redis

Azure Managed Redis es un servicio de Azure totalmente administrado basado en Redis Enterprise. Proporciona almacenamiento de datos en memoria de alto rendimiento para aplicaciones y está diseñado para cargas de trabajo empresariales que requieren una latencia ultra baja, un alto rendimiento y estructuras de datos avanzadas.

Cuando se usa Azure, la confiabilidad es una responsabilidad compartida. Microsoft proporciona una variedad de funcionalidades para admitir resistencia y recuperación. Es responsable de comprender cómo funcionan esas funcionalidades dentro de todos los servicios que usa y de seleccionar las funcionalidades que necesita para cumplir los objetivos empresariales y los objetivos de tiempo de actividad.

En este artículo se describe cómo hacer que Azure Managed Redis sea resistente a una variedad de posibles interrupciones y problemas, incluidos errores transitorios, interrupciones de zona de disponibilidad y interrupciones en regiones. También se describen las estrategias de copia de seguridad y el acuerdo de nivel de servicio (SLA).

Recomendaciones de implementación de producción

Para garantizar una alta confiabilidad para las instancias de Azure Managed Redis de producción, se recomienda:

- ✓ **Use la alta disponibilidad**, que implementa varios nodos para la memoria caché.
- ✓ **Use la redundancia de zona** mediante la implementación de una caché de alta disponibilidad en una región que tenga zonas de disponibilidad.
- ✓ **Considere la posibilidad de implementar la replicación geográfica activa** para cargas de trabajo críticas que requieren conmutación por error entre regiones.

Introducción a la arquitectura de confiabilidad

En esta sección se describen algunos de los aspectos importantes de cómo funciona el servicio que es más relevante desde una perspectiva de confiabilidad. En la sección se presenta la arquitectura lógica, que incluye algunos de los recursos y características que se implementan y usan. También se describe la arquitectura física, que proporciona detalles sobre cómo funciona el servicio en segundo plano.

Arquitectura lógica

Azure Managed Redis se basa en Redis Enterprise y proporciona confiabilidad a través de configuraciones de alta disponibilidad y funcionalidades de replicación.

Puede implementar una *instancia* de Azure Managed Redis, que también se conoce como *una instancia de caché* o una *caché*. Las aplicaciones cliente almacenan e interactúan con los datos de la caché mediante las API de Redis.

Arquitectura física

Al planear la resistencia en Azure Managed Redis, debe comprender los conceptos clave de los nodos y las particiones.

- **Nodos:** Cada instancia de caché consta de *nodos*, que son máquinas virtuales (VM). Cada máquina virtual actúa como una unidad de proceso independiente en el clúster. No ve ni administra las máquinas virtuales directamente. La plataforma crea automáticamente instancias, supervisa el estado de las instancias y reemplaza las instancias que se vuelven no saludables. Este conjunto de máquinas virtuales forma un *clúster*.

  Puede configurar la instancia para lograr una alta disponibilidad. Cuando se usa alta disponibilidad, Azure Managed Redis garantiza que la instancia tenga al menos dos nodos y replique automáticamente los datos entre ellos. En las regiones que tienen zonas de disponibilidad, el servicio coloca los nodos en diferentes zonas de disponibilidad. Para obtener más información, consulte Resilience to availability zone failures (Resistencia a errores de zona de disponibilidad).

  El servicio abstrae el número específico de nodos de cada configuración para evitar la complejidad y garantizar configuraciones óptimas.

- **Particiones:** Cada nodo ejecuta varios procesos de servidor de Redis *conocidos como particiones*. Cada partición administra un subconjunto de los datos de la memoria caché. Al establecer la memoria caché para alta disponibilidad, las particiones se distribuyen y replican automáticamente entre los nodos. Se especifica una *directiva de clúster*, que determina cómo se distribuyen los fragmentos entre los nodos.

Para más información, consulte Arquitectura de Azure Managed Redis y Conmutación y aplicación de revisiones para Azure Managed Redis.

Resistencia a errores transitorios

Los errores transitorios son errores breves e intermitentes en los componentes. Se producen con frecuencia en un entorno distribuido como la nube y son una parte normal de las operaciones. Los errores transitorios se corrigen después de un breve período de tiempo. Es

importante que las aplicaciones puedan controlar errores transitorios, normalmente mediante el reintento de solicitudes afectadas.

Todas las aplicaciones hospedadas en la nube deben seguir las instrucciones de control de errores transitorios de Azure cuando se comunican con cualquier API, bases de datos y otros componentes hospedados en la nube. Para obtener más información, consulte Recomendaciones para controlar errores transitorios.

Siga estas recomendaciones para administrar errores transitorios al usar Azure Managed Redis:

- **Use configuraciones del SDK** que vuelvan a intentarlo automáticamente cuando se produzcan errores transitorios y apliquen los períodos de retroceso y tiempo de espera adecuados. Considere la posibilidad de usar el patrón Retry y el patrón Circuit Breaker en las aplicaciones.

- **Diseñe patrones de reserva de caché**, donde la aplicación puede seguir funcionando con un rendimiento degradado al revertir al almacén de datos principal cuando Redis deja de estar disponible temporalmente.

Resistencia a errores de zona de disponibilidad

Las zonas de disponibilidad son grupos físicamente independientes de centros de datos dentro de una región de Azure. Cuando una zona falla, los servicios pueden transferirse a una de las zonas restantes.

Puede hacer que las instancias de caché de Redis administradas de Azure *tengan redundancia de zona*, que distribuye automáticamente los nodos de caché entre varias zonas de disponibilidad dentro de una región. La redundancia de zona reduce el riesgo de que un centro de datos o una interrupción de la zona de disponibilidad haga que la memoria caché no esté disponible.

Para que una zona de caché sea redundante, debe implementarla en una región admitida y establecerla para usar la configuración de alta disponibilidad. En regiones sin zonas de disponibilidad, la configuración de alta disponibilidad todavía crea al menos dos nodos, pero no se colocan en zonas independientes.

En el diagrama siguiente se muestra una caché con redundancia de zona con dos nodos, cada una en una zona independiente.

Requisitos

- **Compatibilidad con regiones:** Puede implementar cachés de Azure Managed Redis con redundancia de zona en cualquier región que admita zonas de disponibilidad y donde el servicio esté disponible. Para obtener la lista de regiones que admiten zonas de disponibilidad, consulte Regiones de Azure con zonas de disponibilidad. Para obtener la lista de regiones que admiten Azure Managed Redis, consulte Disponibilidad del producto por región.
- **Configuración de alta disponibilidad:** Debe usar la configuración de alta disponibilidad en la memoria caché para que sea redundante de zona.
- **Niveles:** Todos los niveles de Redis administrados de Azure admiten zonas de disponibilidad.

Cost

La redundancia de zona requiere configurar la memoria caché para lograr una alta disponibilidad, lo que implementa un mínimo de dos nodos para la memoria caché. La configuración de alta disponibilidad cuesta más que una configuración sin alta disponibilidad. Para más información, consulte Precios de Azure Managed Redis.

Configurar soporte de zonas de disponibilidad

- **Cree una nueva instancia con redundancia de zona.** Al crear una nueva instancia de Azure Managed Redis, use la configuración de alta disponibilidad e impléntela en una región que admita zonas de disponibilidad. La instancia incluye la redundancia de zona de forma predeterminada, sin que se requiera ninguna configuración adicional.

  Para más información, consulte Creación de una instancia de Azure Managed Redis.

- **Haga que una zona de instancia existente sea redundante.** Para que una instancia de Azure Managed Redis existente tenga redundancia de zona, impléntela en una región

que admita zonas de disponibilidad y configure alta disponibilidad en la memoria caché.

- **Desactivar la redundancia de zona.** No se puede desactivar la redundancia de zona en instancias existentes porque no se puede revertir la alta disponibilidad después de aplicarla a una memoria caché.

Planeamiento y administración de capacidad

Durante un evento de zona inactiva, es posible que la instancia tenga menos recursos disponibles para atender la carga de trabajo. Si la instancia a menudo experimenta presión de recursos y necesita prepararse para una falla de zona de disponibilidad, considere uno de los siguientes enfoques:

- **Sobreaprovisionar la instancia.** Seleccione un nivel de rendimiento superior al necesario para que la instancia pueda tolerar cierta pérdida de capacidad y seguir funcionando sin degradar el rendimiento. Para más información, consulte Administración de la capacidad mediante sobreaprovisionamiento y Escalado de una instancia de Azure Managed Redis.

- **Use la replicación geográfica activa.** Puede implementar varias instancias en diferentes regiones y configurar la replicación geográfica activa para distribuir la carga entre esas instancias independientes.

Comportamiento cuando todas las zonas están en buen estado

En esta sección se describe qué esperar cuando una caché administrada de Redis es con redundancia de zona y todas las zonas de disponibilidad funcionan normalmente:

- **Enrutamiento de tráfico entre zonas:** Azure Managed Redis distribuye particiones entre nodos en función de la directiva de clúster. La directiva de clúster también determina cómo se enruta el tráfico a cada nodo. La redundancia de zona no cambia el modo en que el servicio enruta el tráfico.

- **Replicación de datos entre zonas:** Azure Managed Redis replica automáticamente particiones entre nodos mediante la replicación asincrónica. Normalmente, se mide el retraso de replicación entre particiones en segundos, pero la duración exacta depende de la carga de trabajo de la memoria caché. Los escenarios con alta carga de escritura y de red suelen experimentar una mayor latencia de replicación.

Comportamiento durante un fallo de zona

En esta sección se describe qué esperar cuando una caché administrada de Redis es redundante de zona y una o varias zonas de disponibilidad no están disponibles:

- **Detección y respuesta:** Azure Managed Redis es responsable de detectar un error en una zona de disponibilidad. No es necesario hacer nada para iniciar una conmutación por error de la zona.
- **Notificación:** Microsoft no le notifica automáticamente cuando una zona está inactiva. Sin embargo, puede usar Azure Service Health para comprender el estado general del servicio, incluidos los errores de zona, y puede configurar alertas de Service Health para notificarle problemas.
- **Solicitudes activas:** El servicio podría descartar las solicitudes en proceso y las aplicaciones deben intentar de nuevo. Las aplicaciones deben implementar lógica de reintento para controlar estas interrupciones temporales.
- **Pérdida de datos esperada:** Es posible que los datos que no se hayan replicado en particiones de otra zona se pierdan durante un error de zona. Normalmente, se mide la cantidad de pérdida de datos en segundos, pero depende del retraso de replicación.
- **Tiempo de inactividad esperado:** Una pequeña cantidad de tiempo de inactividad, normalmente de 10 a 15 segundos, puede producirse mientras los fragmentos realizan la conmutación por error a los nodos en zonas saludables. Para obtener más información sobre el proceso de conmutación por error no planeado, consulte Explicación de una conmutación por error. Al diseñar aplicaciones, siga los procedimientos para el control de errores transitorios.
- **Reenrutamiento del tráfico:** Azure Managed Redis redirige automáticamente el tráfico a los nodos de zonas correctas.

Recuperación de zona

Cuando se recupera la zona de disponibilidad afectada, Azure Managed Redis restaura automáticamente las operaciones a esa zona. La plataforma Azure administra completamente este proceso y no requiere ninguna intervención del cliente.

Prueba de fallos de zona

Azure Managed Redis administra completamente el enrutamiento del tráfico, la conmutación por error y la conmutación por restablecimiento para las fallas de zona, de modo que no es necesario validar los procesos de fallas en la zona de disponibilidad ni aportar más información.

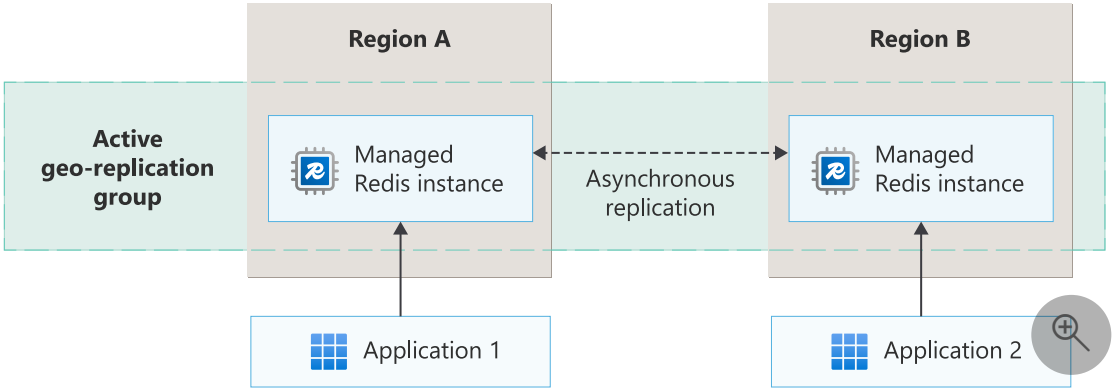
Resistencia a errores en toda la región

Azure Managed Redis proporciona compatibilidad nativa con varias regiones a través de *la replicación geográfica activa*, que permite vincular varias instancias de Azure Managed Redis entre distintas regiones de Azure en un único grupo de replicación. Después, puede configurar su propio enfoque de conmutación por error entre las instancias.

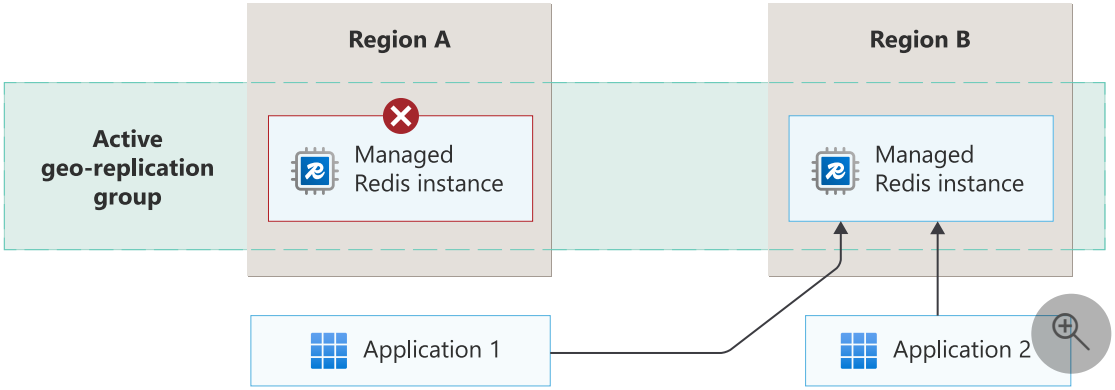
Replicación geográfica activa

Cuando se usa la replicación geográfica activa, las aplicaciones pueden leer y escribir en cualquier instancia de caché del grupo, con cambios sincronizados automáticamente en todas las regiones. El servicio admite patrones de replicación activo-activo en los que cada región puede controlar las operaciones de lectura y escritura simultáneamente. Cuando se producen conflictos debido a escrituras simultáneas en regiones diferentes, el servicio los resuelve automáticamente mediante algoritmos de resolución de conflictos predeterminados sin necesidad de intervención manual. Este enfoque proporciona resistencia a errores de región al tiempo que mantiene el acceso de baja latencia para las aplicaciones distribuidas globalmente.

En el diagrama siguiente se muestran dos instancias de caché en regiones diferentes dentro del mismo grupo de replicación geográfica activa, junto con las aplicaciones cliente que se conectan a cada instancia de caché.



Tú eres responsable de configurar las aplicaciones cliente para redirigir las solicitudes a una instancia operativa si falla alguna instancia regional. En el diagrama siguiente se muestra cómo una aplicación puede redirigir sus solicitudes a una instancia de caché correcta cuando se produce un error en la instancia que suelen usar.



Requisitos

- **Compatibilidad con regiones:** Puede configurar la replicación geográfica activa de Azure Managed Redis entre cualquier región de Azure en la que el servicio esté disponible.
- **Configuración de instancia:** La replicación geográfica activa requiere instancias de Azure Managed Redis que usen el mismo nivel y tamaño en cada región participante. Todas las instancias de caché de un grupo de replicación deben usar una configuración idéntica, incluidas las opciones de persistencia, los módulos y las directivas de agrupación en clústeres.
- **Otros requisitos:** Las instancias de caché deben cumplir otros requisitos, incluidos los módulos que use. Estos requisitos afectan a cómo puede escalar las instancias de caché. Para más información, consulte Requisitos previos de replicación geográfica activa.

Consideraciones

- **Responsabilidad de conmutación por error:** Al usar la replicación geográfica activa, **es responsable de la conmutación por error entre instancias de caché**. Prepare su aplicación para gestionar la conmutación por error. La conmutación por error requiere preparación y puede implicar realizar varios pasos. Para obtener más información, consulte Cómo forzar la desvinculación durante una interrupción regional.
- **Coherencia final:** Diseñe aplicaciones para controlar escenarios de coherencia finales, ya que los cambios pueden tardar tiempo en propagarse en todas las regiones, en función de las condiciones de red y la distancia geográfica. Durante las interrupciones de la región, es posible que experimente más incoherencias de datos hasta que finalice la conectividad y finalice la sincronización.

Cost

Al configurar la replicación geográfica activa, paga por cada instancia de Azure Managed Redis en cada región del grupo de replicación. También puede incurrir en cargos de transferencia de datos para el tráfico de replicación entre regiones diferentes. Para más información, consulte Precios de Azure Managed Redis y Detalles de precios de ancho de banda.

Configuración de la compatibilidad con varias regiones

- **Cree una nueva instancia de caché con replicación geográfica.** Configure la replicación geográfica activa al aprovisionar la memoria caché especificando un grupo de replicación y vinculando varias instancias. Para obtener más información, consulte Creación o unión de un grupo de replicación geográfica activa.

- **Agregue una instancia de caché existente a un grupo de replicación geográfica.** Puede agregar una instancia de caché existente a un grupo de replicación geográfica activo.

  Al agregar una instancia existente a un grupo de replicación geográfica activa, la plataforma debe borrar los datos de la instancia y experimentar una pequeña cantidad de tiempo de inactividad. Si es posible, planee configurar la replicación geográfica activa al crear instancias de caché.

  Para obtener más información, consulte Incorporación de una instancia existente a un grupo de replicación geográfica activo.

- **Desactivar la replicación geográfica en una instancia de caché.** Quite una instancia de un grupo de replicación geográfica mediante la eliminación de la instancia de caché. Las instancias restantes ajustan automáticamente su configuración.

Planeamiento y administración de capacidad

Durante un evento de región inactiva, las otras instancias pueden experimentar una mayor carga. Si una instancia se ejecuta a menudo cerca de sus límites de recursos y debe prepararse para aumentar los requisitos de capacidad durante un error de región, considere la posibilidad de sobreaprovisionar la instancia. Para más información, consulte Escalado de una instancia de Azure Managed Redis.

Comportamiento cuando todas las regiones están en buen estado

En esta sección se describe qué esperar al configurar instancias para usar la replicación geográfica activa y todas las regiones funcionan normalmente.

- **Enrutamiento del tráfico entre regiones:** Es responsable de configurar las aplicaciones para conectarse a una instancia de caché específica. Las aplicaciones pueden conectarse a

cualquier instancia de caché del grupo de replicación y realizar operaciones de lectura y escritura. La aplicación controla el enrutamiento de tráfico, lo que permite dirigir a los clientes a la región más cercana para una latencia mínima. Azure Managed Redis no enruta automáticamente el tráfico entre regiones.

- **Replicación de datos entre regiones:** El servicio usa la replicación asincrónica entre regiones para mantener la coherencia final. Confirma inmediatamente las operaciones de escritura en la región local y, a continuación, las propaga a otras regiones en segundo plano. Los tipos de datos replicados sin conflictos (CRDT) garantizan que el servicio combina automáticamente escrituras simultáneas en diferentes regiones.

Comportamiento durante una falla de región

En esta sección se describe qué esperar al configurar instancias para usar la replicación geográfica activa y se produce una interrupción en una región.

- **Detección y respuesta:** Eres responsable de detectar el fallo de una instancia de caché y decidir cuándo realizar una conmutación por error. Puede supervisar el estado de salud de un clúster con replicación geográfica, lo que puede ayudarle a decidir cuándo comenzar el proceso de failover. Para más información, consulte Métrica de replicación geográfica.

  La conmutación por error requiere que realice varios pasos. Para obtener más información, consulte Desvinculación forzada durante una interrupción regional.

- **Notificación:** Microsoft no le notifica automáticamente cuando una región está inactiva. Sin embargo, puede usar Azure Service Health para comprender el estado general del servicio, incluidos los errores de región, y puede configurar alertas de Service Health para notificarle problemas.

  También puede supervisar la salud de cada instancia.

  Para supervisar el estado de la relación de replicación geográfica, utilice la métrica *estado saludable de replicación geográfica*. La métrica siempre tiene un valor de `1` (correcto) o `0` (incorrecto). Configure alertas de Azure Monitor en esta métrica para identificar cuándo las instancias podrían no estar sincronizadas.

- **Solicitudes activas:** El servicio finaliza las solicitudes a la región que ha fallado, y la lógica de conmutación de la aplicación debe controlarlas. Las aplicaciones deben implementar directivas de reintento que puedan redirigir el tráfico a cachés correctas.

- **Pérdida de datos esperada:** La replicación asincrónica entre regiones puede resultar en la pérdida de escrituras recientes en la región fallida si esas escrituras no se han replicado

en otras regiones. La cantidad de pérdida de datos potencial depende del retraso de replicación en el momento del error. Para obtener más información, consulte Redis distribuido geográficamente activo y consideraciones sobre la coherencia y la pérdida de datos en un error regional de base de datos replicada (CRDB) sin conflictos.

- **Tiempo de inactividad esperado:** Las aplicaciones experimentan tiempo de inactividad solo durante el tiempo necesario para detectar el error y redirigir el tráfico a regiones correctas. Este tiempo de inactividad suele oscilar entre segundos y unos minutos y depende de cómo configure la comprobación de estado y la configuración de conmutación por error de la aplicación.
- **Reenrutamiento del tráfico:** Es responsable de implementar lógica en las aplicaciones para detectar errores de región y enrutar el tráfico a regiones correctas. Puede implementar esta lógica a través de comprobaciones de estado, patrones de disyuntor o soluciones de equilibrio de carga externas.

Recuperación de regiones

Cuando se recupera una región con errores, Azure Managed Redis reintegra automáticamente las instancias de esa región en el grupo de replicación geográfica activa sin su intervención. El servicio sincroniza automáticamente datos de instancias saludables. Durante este proceso, las instancias recuperadas sincronizan gradualmente los cambios que se produjeron durante la interrupción. Una vez finalizada la sincronización, las instancias recuperadas se vuelven totalmente activas y pueden controlar las operaciones de lectura y escritura.

Usted es responsable de volver a configurar su aplicación para enrutar el tráfico de regreso a la instancia de la región recuperada.

Prueba de fallos de región

Prueben periódicamente los procedimientos de conmutación por error de su aplicación. La aplicación debe poder conmutar por error entre instancias y cumplir los requisitos empresariales para el tiempo de inactividad durante la transición. Pruebe también los procesos de respuesta generales, incluida cualquier reconfiguración de firewalls y otras infraestructuras, y el proceso de recuperación.

Resistencia al mantenimiento del servicio

Azure Managed Redis controla las actualizaciones de servicio normales y otras tareas de mantenimiento.

Durante el mantenimiento, Azure Managed Redis crea automáticamente nuevos nodos y realiza una conmutación. Las aplicaciones cliente pueden experimentar interrupciones de conexión y otros errores transitorios. Las aplicaciones deben implementar lógica de reintento para controlar estas interrupciones temporales.

Para obtener más información sobre los procesos de mantenimiento de Azure Managed Redis, consulte Conmutación por error y aplicación de revisiones para Azure Managed Redis.

Copias de seguridad y restauración

Azure Managed Redis proporciona funcionalidades de persistencia y copia de seguridad de datos para protegerse frente a escenarios de pérdida de datos que podrían no abordar otras características de confiabilidad. Las copias de seguridad protegen frente a escenarios como daños en los datos, eliminación accidental o errores de configuración.

- **Persistencia de datos:** De forma predeterminada, Azure Managed Redis almacena todos los datos de caché en memoria. Opcionalmente, el servicio puede escribir instantáneas de los datos en el disco mediante la persistencia de datos. Si un error de hardware afecta al nodo, Azure Managed Redis restaura automáticamente los datos. Puede elegir entre diferentes tipos de persistencia de datos, que proporcionan diferentes ventajas entre la frecuencia de instantánea y los efectos de rendimiento en la memoria caché.

  No puede restaurar archivos de datos en otra instancia y no puede acceder a los archivos. La persistencia de datos no le protege de daños en los datos ni de la eliminación accidental.

- **Importación y exportación:** Azure Managed Redis admite la copia de seguridad de los datos cuando se usa la funcionalidad de importación y exportación, que guarda los archivos de copia de seguridad en Azure Blob Storage. Puede configurar el almacenamiento con redundancia geográfica (GRS) en la cuenta de Azure Storage en regiones admitidas, o bien puede copiar o mover los blobs de copia de seguridad a otras ubicaciones para mayor protección.

  Puede restaurar archivos exportados a la misma instancia de caché o a una instancia de caché diferente.

  El servicio no incluye un programador de importación o exportación integrado, pero puede desarrollar sus propios procesos de automatización que usen la CLI de Azure o Azure PowerShell para iniciar operaciones de exportación.

Acuerdo de nivel de servicio

El contrato de nivel de servicio (SLA) para los servicios de Azure describe la disponibilidad esperada de cada servicio y las condiciones que la solución deberá cumplir para lograr esa expectativa de disponibilidad. Para obtener más información, consulte Acuerdos de Nivel de Servicio para servicios en línea.

El Acuerdo de Nivel de Servicio para Azure Managed Redis cubre la conectividad con los puntos de conexión de caché. No cubre la protección contra la pérdida de datos.

Para ser apto para los SLA de disponibilidad para Azure Managed Redis, debe cumplir los siguientes requisitos:

- Debe usar una configuración de alta disponibilidad.
- No debe iniciar ninguna característica de producto ni acciones de administración documentadas para producir una falta de disponibilidad temporal.

Se aplican acuerdos de nivel de servicio (SLA) de mayor disponibilidad cuando la instancia tiene redundancia de zona. En algunos niveles, puede optar a un Acuerdo de Nivel de Servicio de mayor disponibilidad al implementar instancias con redundancia de zona en al menos tres regiones mediante la replicación geográfica activa.

Contenido relacionado

- Confiabilidad en Azure

ⓘ **Nota:** El autor ha creado este artículo con ayuda de inteligencia artificial. Más información

---

Last updated on 06/02/2026

Confiabilidad en Azure SQL Database

Azure SQL Database es un motor de base de datos de plataforma como servicio (PaaS) totalmente administrado que se encarga de la mayoría de las funciones de administración de bases de datos, como actualizar, aplicar revisiones, crear copias de seguridad y supervisar sin intervención del usuario.

Cuando se usa Azure, la confiabilidad es una responsabilidad compartida. Microsoft proporciona una variedad de funcionalidades para admitir resistencia y recuperación. Es responsable de comprender cómo funcionan esas funcionalidades dentro de todos los servicios que usa y de seleccionar las funcionalidades que necesita para cumplir los objetivos empresariales y los objetivos de tiempo de actividad.

En este artículo se describe cómo hacer que Azure SQL Database sea resistente a una variedad de posibles interrupciones y problemas, incluidos errores transitorios, interrupciones de zona de disponibilidad y interrupciones de regiones. También se describe cómo puede usar copias de seguridad para recuperarse de otros tipos de problemas, cómo controlar el mantenimiento del servicio y resaltar cierta información clave sobre el acuerdo de nivel de servicio (SLA) de Azure SQL Database.

Recomendaciones de implementación de producción

Para obtener información sobre cómo implementar Azure SQL Database para admitir los requisitos de confiabilidad de la solución y cómo afecta la confiabilidad a otros aspectos de la arquitectura, consulte Procedimientos recomendados de arquitectura para Azure SQL Database en Azure Well-Architected Framework.

Introducción a la arquitectura de confiabilidad

SQL Database se ejecuta en el motor de base de datos de SQL Server estable más reciente del sistema operativo Windows, incluidas todas las revisiones aplicables.

SQL Database logra redundancia almacenando tres copias de los datos en un único centro de datos de la región primaria de forma predeterminada. Este enfoque protege los datos si se produce un error localizado, como un error de red a pequeña escala o un error de energía, y también durante los siguientes eventos:

- Operaciones de administración iniciadas por el cliente que dan lugar a un breve tiempo de inactividad.
- Operaciones de mantenimiento de servicio
- Problemas y interrupciones del centro de datos, donde el centro de datos tiene los siguientes componentes:
  - Bastidores, donde se ejecutan las máquinas que impulsan el servicio
  - Máquinas físicas que hospedan la máquina virtual (VM) que ejecuta el motor de SQL Database
- Otros problemas con el motor de SQL Database
- Otras posibles interrupciones localizadas no planeadas

SQL Database usa Azure Service Fabric para administrar la replicación de la base de datos.

La redundancia se implementa de maneras diferentes para distintos niveles de servicio de SQL Database. Para obtener más información, consulte Disponibilidad mediante redundancia: SQL Database.

Resistencia a errores transitorios

Los errores transitorios son errores breves e intermitentes en los componentes. Se producen con frecuencia en un entorno distribuido como la nube y son una parte normal de las operaciones. Los errores transitorios se corrigen después de un breve período de tiempo. Es importante que las aplicaciones puedan controlar errores transitorios, normalmente mediante el reintento de solicitudes afectadas.

Todas las aplicaciones hospedadas en la nube deben seguir las instrucciones de control de errores transitorios de Azure cuando se comunican con cualquier API, bases de datos y otros componentes hospedados en la nube. Para obtener más información, consulte Recomendaciones para controlar errores transitorios.

SQL Database controla automáticamente las tareas de mantenimiento críticas, como la aplicación de revisiones, las copias de seguridad, Windows y las actualizaciones del motor de SQL Database. También controla automáticamente eventos no planeados, como errores de red, software o hardware subyacentes. SQL Database está diseñado para recuperarse rápidamente de errores críticos, lo que garantiza que los datos estén siempre disponibles. La mayoría de los usuarios no observan que las actualizaciones se realizan continuamente.

Cuando una base de datos se revisa o transfiere, el tiempo de inactividad no es disruptivo si emplea lógica de reintento en su aplicación.

Puede probar la resistencia de la aplicación a errores transitorios siguiendo las instrucciones de Prueba de resistencia de errores de la aplicación.

Resistencia a errores de zona de disponibilidad

Las zonas de disponibilidad son grupos físicamente independientes de centros de datos dentro de una región de Azure. Cuando una zona falla, los servicios pueden transferirse a una de las zonas restantes.

Para ver información sobre la compatibilidad de zona de disponibilidad con otros niveles de servicio, asegúrese de seleccionar el nivel de servicio adecuado al principio de esta página.

Requisitos

Los niveles de servicio Básico y Estándar no admiten la redundancia de zona.

La redundancia de zona está disponible para las bases de datos en los niveles de servicio Crítico para negocios, Propósito general e Hiperescala del modelo de compra basado en vCore, y únicamente el nivel de servicio Premium del modelo de compra basado en DTU.

Resistencia a errores en toda la región

En esta sección se proporciona información general sobre dos características relacionadas pero independientes que se pueden usar para la replicación geográfica en varias regiones de SQL Database:

- La replicación geográfica activa replica una base de datos única en una base de datos secundaria sincronizada.
- Los grupos de conmutación por error se basan en la replicación geográfica activa y permiten realizar una conmutación por error de un grupo de bases de datos.

Replicación geográfica activa

La replicación geográfica activa crea una base de datos secundaria legible sincronizada continuamente (que a veces se conoce como réplica geográfica *secundaria* o *geográfica*) en cualquier región para una base de datos principal única. La replicación geográfica activa puede crear bases de datos secundarias en la misma región, pero esta configuración no proporciona

protección contra una interrupción de la región. Al usar la replicación geográfica activa para lograr redundancia geográfica, busque la base de datos secundaria en una región diferente a la base de datos principal.

Requisitos

Al usar la replicación geográfica activa, tenga en cuenta los siguientes requisitos:

- **Compatibilidad con regiones:** la replicación geográfica activa se puede habilitar en todas las regiones de Azure y no requiere que use pares de regiones de Azure.

  > **Sugerencia**
  >
  > SQL Database sigue una práctica de implementación segura en la que Azure se esfuerza por no implementar actualizaciones en regiones emparejadas al mismo tiempo. Si configura la replicación geográfica activa para usar regiones no emparejadas, establezca diferentes ventanas de mantenimiento para los servidores de cada región. Este enfoque ayuda a reducir la posibilidad de que ambas regiones experimenten problemas de conectividad causados por el mantenimiento al mismo tiempo.

- **Configuración:** Tanto el principal como el secundario geográfico deben tener el mismo nivel de servicio y deben tener el mismo nivel de proceso, tamaño de proceso y redundancia de almacenamiento de copia de seguridad.

- **Firewall:** Tanto el servidor principal como el secundario geográfico deben tener las mismas reglas de firewall de direcciones IP.

- **Suscripciones de Azure:** La replicación geográfica activa es compatible con bases de datos en distintas suscripciones de Azure.

Consideraciones

- La replicación geográfica activa está diseñada para proporcionar tolerancia a fallos de una única base de datos. Si necesita conmutar por error varias bases de datos, considere la posibilidad de usar en su lugar grupos de conmutación por error.

- Dado que la replicación geográfica es asincrónica, la pérdida de datos es posible cuando se produce una conmutación por error. Si necesita eliminar la pérdida de datos de la replicación asincrónica durante los fallos, configure la aplicación para bloquear el

subproceso de llamada hasta que se transmita y endurezca la última transacción confirmada en el registro de transacciones de la base de datos secundaria. Este enfoque requiere desarrollo personalizado y reduce el rendimiento de la aplicación. Para obtener más información, consulte Evitar la pérdida de datos críticos.

- Para más información sobre las limitaciones y consideraciones, consulte Replicación geográfica activa.

Cost

Las bases de datos secundarias se facturan como bases de datos independientes.

Si no usa una base de datos secundaria para cargas de trabajo de lectura o escritura, considere si puede designarla como réplica en espera para reducir los costos.

Configuración de la compatibilidad con varias regiones

- **Habilite la replicación geográfica activa:** Para más información sobre cómo habilitar la replicación geográfica activa en Azure Portal, consulte Configuración de la replicación geográfica activa para SQL Database o Replicación geográfica activa.

  Después de habilitar la replicación geográfica activa, un paso de propagación inicial puede tardar algún tiempo.

- **Deshabilitar la replicación geográfica activa:** Para obtener más información sobre cómo deshabilitar la replicación geográfica activa en una base de datos, consulte Eliminación de la base de datos secundaria.

Comportamiento cuando todas las regiones están en buen estado

En esta sección se describe qué esperar cuando una base de datos está configurada para usar la replicación geográfica activa y todas las regiones están operativas.

- **Enrutamiento del tráfico entre regiones:** La aplicación debe configurarse para enviar solicitudes de lectura y escritura a la base de datos principal. Opcionalmente, puede enviar solicitudes de solo lectura a una base de datos secundaria, lo que ayuda a reducir el impacto de las cargas de trabajo de solo lectura en la base de datos principal.

- **Replicación de datos entre regiones:** La replicación entre las bases de datos principal y secundaria se produce de forma asincrónica, lo que significa que puede haber un retraso

entre el momento en que se aplica un cambio a la base de datos principal y cuando se replica en la base de datos secundaria.

Al realizar una conmutación por error, decide cómo gestionar el riesgo de pérdida de datos.

Comportamiento durante una falla de región

En esta sección se describe qué esperar cuando una base de datos está configurada para usar la replicación geográfica activa y hay una interrupción en la región primaria:

- **Detección y respuesta:** Usted es responsable tanto de detectar una interrupción en una base de datos o región como de activar la conmutación por error.
- **Notificación:** Microsoft no le notifica automáticamente cuando una región está inactiva. Sin embargo, puede usar Azure Service Health para comprender el estado general del servicio, incluidos los errores de región, y puede configurar alertas de Service Health para notificarle problemas.
- **Solicitudes activas:** Las solicitudes activas durante la conmutación por error se finalizan y se deben reintentar.
- **Pérdida de datos esperada:** Si la base de datos principal está disponible, opcionalmente puede realizar una conmutación por error sin pérdida de datos. El proceso de conmutación por error sincroniza los datos entre las bases de datos principal y secundaria antes de cambiar de roles.

  Si la base de datos principal no está disponible, es posible que tenga que realizar una *conmutación por error forzada*, lo que da lugar a la pérdida de datos. Puede calcular la cantidad de pérdida de datos mediante la supervisión del retraso de replicación. Para obtener más información, consulte Supervisión de SQL Database con métricas y alertas.
- **Tiempo de inactividad esperado:** Normalmente, hay hasta 60 segundos de tiempo de inactividad durante una conmutación por error. Asegúrese de que la aplicación controla los errores transitorios para que pueda recuperarse de breves períodos de tiempo de inactividad. Además, la rapidez con la que reconfigura la aplicación para conectarse a la nueva base de datos principal afecta al tiempo de inactividad que experimenta.
- **Reenrutamiento del tráfico:** Es responsable de volver a configurar la aplicación para enviar solicitudes a la nueva base de datos principal. Si necesita tener conmutación por error transparente, considere la posibilidad de usar grupos de conmutación por error.

Recuperación de regiones

Una vez recuperada la región primaria, puede realizar manualmente una conmutación por restablecimiento en la región primaria realizando otra conmutación por error.

Prueba de fallos de región

Puede simular una interrupción de región desencadenando una conmutación por error manual en cualquier momento. Puede desencadenar una conmutación por error (sin pérdida de datos) o una conmutación por error forzada.

Grupos de conmutación por error

Los grupos de conmutación por error se basan en la replicación geográfica activa. Con los grupos de conmutación por error, puede realizar las siguientes operaciones:

- Replique un conjunto de bases de datos desde un único servidor lógico en cualquier combinación de regiones de Azure.
- Realice la conmutación por error en las bases de datos como grupo.
- Use puntos de conexión que dirijan automáticamente las conexiones a la principal.

Requisitos

- **Compatibilidad regional:** Los grupos de conmutación de emergencia se pueden crear en todas las regiones de Azure y no es necesario utilizar pares de regiones de Azure.

  > **Sugerencia**
  >
  > Si configura un grupo de conmutación por error con regiones no emparejadas, considere la posibilidad de establecer diferentes ventanas de mantenimiento para los servidores de cada región. Este enfoque ayuda a reducir la posibilidad de que ambas regiones experimenten problemas de conectividad causados por el mantenimiento al mismo tiempo.

- **Configuración:** Las bases de datos secundarias de un grupo de conmutación por error deben tener el mismo nivel de servicio, nivel de proceso, tamaño de proceso, reglas de firewall de direcciones IP y redundancia de almacenamiento de copia de seguridad que la base de datos principal.

Consideraciones

- **Redundancia de zona:** Las bases de datos secundarias no tienen habilitada la redundancia de zona de forma predeterminada, pero puede habilitarla más adelante.
- **Posible pérdida de datos:** Dado que la replicación geográfica es asincrónica, es posible que se pierdan datos cuando se produce una conmutación por error. Si necesita eliminar la pérdida de datos de la replicación asincrónica durante las conmutaciones por error, puede configurar la aplicación para bloquear el subproceso de llamada hasta que la última transacción confirmada se ha transmitido y protegido en el registro de transacciones de la base de datos secundaria. Este enfoque requiere desarrollo personalizado y reduce el rendimiento de la aplicación. Para obtener más información, consulte Evitar la pérdida de datos críticos.
- Para obtener más información sobre las limitaciones y consideraciones, consulte Grupos de conmutación por error.

Cost

Las bases de datos secundarias se facturan como bases de datos independientes.

Si no usa una base de datos secundaria para cargas de trabajo de lectura o escritura, considere si puede designarla como réplica en espera para reducir los costos.

Configuración de la compatibilidad con varias regiones

- **Habilite los grupos de conmutación por error:** Usted configura un grupo de conmutación por error en un servidor lógico. Puede agregar todas las bases de datos del servidor lógico al grupo de conmutación por error, o bien puede seleccionar un subconjunto de bases de datos que se van a agregar.

  Al crear un grupo de conmutación por error, seleccione la directiva de conmutación por error, que especifica quién es responsable de detectar una interrupción y realizar una conmutación por error. Puede configurar la conmutación por error administrada por el cliente, que se recomienda, o bien la conmutación por error administrada por Microsoft.

  > **Importante**
  >
  > Es probable que se produzca una conmutación por error automatizada iniciada por Microsoft después de un retraso significativo y se lleva a cabo de la mejor manera posible. La conmutación por error de las bases de datos puede producirse en un

momento diferente a cualquier conmutación por error de otros servicios de Azure. Para obtener más información, consulte [Configuración de un grupo de conmutación por error para SQL Database](#).

Después de configurar el grupo de conmutación por error, el paso de inicialización puede tardar algún tiempo.

- **Deshabilitar grupos de conmutación por error:** Puede quitar una base de datos individual de un grupo de conmutación por error, quitar un grupo de conmutación por error completo o mover una base de datos a otro grupo de conmutación por error.

Comportamiento cuando todas las regiones están en buen estado

En esta sección se describe qué esperar cuando una base de datos está configurada dentro de un grupo de conmutación por error y todas las regiones están operativas.

- **Enrutamiento del tráfico entre regiones:** En el caso de las cargas de trabajo de lectura y escritura y de solo lectura, los grupos de conmutación por error proporcionan dos puntos de conexión de escucha donde puede conectar sus aplicaciones. Utilice los puntos de conexión del escucha del grupo de conmutación para minimizar el tiempo de inactividad durante las conmutaciones. Durante las operaciones normales, se produce el siguiente comportamiento de enrutamiento:
  - El punto de conexión del escuchador de lectura y escritura enruta todas las solicitudes a las bases de datos primarias.
  - El punto de conexión del agente de escucha de solo lectura enruta todas las solicitudes a una base de datos secundaria legible.
- **Replicación de datos entre regiones:** La replicación geográfica entre las bases de datos principal y secundaria se produce de forma asincrónica. Esta latencia significa que puede haber un retraso entre un cambio que se aplica a la base de datos principal y cuando se replica en la base de datos secundaria.

  Al realizar una conmutación por error, decide cómo gestionar el riesgo de pérdida de datos.

Comportamiento durante una falla de región

En esta sección se describe qué esperar cuando una base de datos está configurada dentro de un grupo de conmutación por error y hay una interrupción en la región primaria:

- **La detección y la respuesta** dependen de la política de conmutación por error que use.
  - *Conmutación por error iniciada por el cliente:* Es responsable de detectar la interrupción de una base de datos o región y de desencadenar la conmutación por error.
  - *Conmutación por error iniciada por Microsoft:* Microsoft desencadena la conmutación por error para todos los grupos de conmutación por error de la región afectada. Microsoft espera realizar este tipo de conmutación por error solo en situaciones excepcionales. No confíe en la conmutación por error administrada por Microsoft para la mayoría de las soluciones. Para obtener más información, consulte Directiva de conmutación por error: administrada por Microsoft.
- **Notificación:** Microsoft no le notifica automáticamente cuando una región está inactiva. Sin embargo, puede usar Azure Service Health para comprender el estado general del servicio, incluidos los errores de región, y puede configurar alertas de Service Health para notificarle problemas.
- **Solicitudes activas:** Las solicitudes activas durante la conmutación por error se finalizan y se deben reintentar.
- **Pérdida de datos esperada:** La pérdida de datos depende de la directiva de conmutación por error que use.
  - *Conmutación por error iniciada por el cliente:* Si la base de datos principal está disponible, opcionalmente puede realizar una conmutación por error sin pérdida de datos. El proceso de conmutación por error sincroniza los datos entre las bases de datos principal y secundaria antes de cambiar de roles.

    Si la base de datos principal no está disponible, es posible que tenga que realizar una *conmutación por error forzada*, lo que da lugar a la pérdida de datos. Puede calcular la cantidad de pérdida de datos mediante la supervisión del retraso de replicación. Para obtener más información, consulte Supervisión de SQL Database con métricas y alertas.
  - *Conmutación por error iniciada por Microsoft:* Una conmutación por error administrada por Microsoft solo se desencadena en situaciones excepcionales. Las conmutaciones por error administradas por Microsoft son conmutaciones por error forzadas, lo que significa que se espera una pérdida de datos. No puede controlar la cantidad de pérdida de datos que puede experimentar.
- **Tiempo de inactividad esperado:** El tiempo de inactividad depende de la directiva de conmutación por error que use.

- *Conmutación por error iniciada por el cliente:* Normalmente, hay hasta 60 segundos de tiempo de inactividad durante una conmutación por error. Asegúrese de que la aplicación controla los errores transitorios para que pueda recuperarse de breves períodos de tiempo de inactividad.
- *Conmutación por error iniciada por Microsoft:* Puede especificar un *período de gracia* que determine cuánto tiempo debe esperar Microsoft antes de iniciar la conmutación por error. El período de gracia mínimo es una hora. Sin embargo, es probable que el tiempo de respuesta de Microsoft sea de varias horas al menos.

- **Reenrutamiento del tráfico:** durante la conmutación por error, SQL Database actualiza los puntos de conexión de escucha de lectura y escritura y de solo lectura para dirigir el tráfico a la nueva base de datos primaria y secundaria según sea necesario.

  Si la nueva base de datos secundaria (anteriormente la base de datos principal) no está disponible, el punto de conexión del agente de escucha de solo lectura no se puede conectar hasta que esté disponible la nueva base de datos secundaria.

Recuperación de regiones

Usted es responsable de restablecer la región primaria si es necesario. Puede realizar manualmente una conmutación por recuperación en la región primaria mediante la realización de una conmutación por error administrada por el cliente.

Incluso si Microsoft inició la conmutación por error original, usted sigue siendo responsable de la conmutación por recuperación a la región anterior, si elige hacerlo.

Prueba de fallos de región

Puede simular una interrupción de región desencadenando una conmutación por error manual en cualquier momento. Puede desencadenar una conmutación por error (sin pérdida de datos) o una conmutación por error forzada.

Resistencia a la eliminación accidental del servidor lógico

Además de la resistencia frente a errores de infraestructura, Azure SQL Database proporciona protección contra la eliminación accidental o no deseada de los servidores lógicos.

Azure SQL Database admite la retención de eliminación temporal para servidores lógicos (versión preliminar). Al habilitar esta función, un servidor lógico eliminado no se elimina permanentemente de inmediato. En su lugar, se conserva en un estado eliminado temporalmente durante un período de retención configurable, durante el cual puede detectarlo y restaurarlo.

La retención del servidor lógico ayuda a mitigar los riesgos de:

- Eliminación accidental a través del portal de Azure, las API o los scripts
- Errores de automatización o implementación
- Errores operativos durante las pruebas o las actividades de limpieza

Azure conserva los metadatos cuando se elimina temporalmente un servidor lógico. Esta retención le permite restaurar el servidor a su estado anterior y recuperar sus bases de datos, siempre que sigan dentro del período de retención de copia de seguridad.

La retención de eliminación temporal complementa las funcionalidades de copia de seguridad y restauración existentes, como la restauración a un momento dado y la retención a largo plazo, proporcionando una ruta de recuperación de nivel de servidor lógico. Esta ruta de recuperación puede reducir el tiempo de recuperación y la complejidad operativa en comparación con la recreación de recursos y la restauración de bases de datos individuales.

> ⓘ **Nota:**
>
> La capacidad de configurar un período de retención para eliminación suave y restaurar un servidor lógico eliminado está actualmente en **versión preliminar**. Cualquier servidor lógico con más de dos años tiene automáticamente un período de retención de eliminación suave de siete días. Los servidores lógicos menores de dos años tienen deshabilitada la retención de eliminación suave de forma predeterminada.

Copias de seguridad y restauración

Realice copias de seguridad de las bases de datos para protegerse frente a diversos riesgos, incluida la pérdida de datos. Las copias de seguridad se pueden restaurar para recuperarse de pérdida accidental de datos, daños u otros problemas. Las copias de seguridad difieren de la redundancia de zona, la replicación geográfica activa o los grupos de conmutación por error, y tienen diferentes propósitos. Para obtener más información, consulte Redundancia, replicación y copia de seguridad.

SQL Database proporciona copias de seguridad automáticas de las bases de datos. Para obtener más información sobre la frecuencia de copia de seguridad, que puede afectar a la cantidad de pérdida de datos si necesita restaurar desde una copia de seguridad, consulte Copias de seguridad automatizadas en SQL Database.

Almacenamiento de copia de seguridad

Puede elegir almacenar las copias de seguridad automatizadas en LRS o ZRS. Si usa una región emparejada, puede optar por replicar las copias de seguridad automatizadas en la región emparejada mediante el almacenamiento con redundancia geográfica. Esta funcionalidad permite la restauración geográfica de las copias de seguridad en la región emparejada. Para obtener más información, consulte Copias de seguridad automatizadas en SQL Database.

Si usa una región no emparejada o si necesita replicar copias de seguridad en una región distinta de la región emparejada, considere la posibilidad de exportar la base de datos y almacenar el archivo exportado en una cuenta de almacenamiento que use la replicación de objetos de blob para replicar en una cuenta de almacenamiento de otra región. Para obtener más información, consulte Exportación de una base de datos.

Para una protección más sólida, puede habilitar la inmutabilidad de copia de seguridad para las copias de seguridad de retención a largo plazo (LTR). La inmutabilidad de copia de seguridad garantiza que las copias de seguridad de LTR para una base de datos se almacenan en un estado **Write Once, Read Many (WORM),** por lo que no se pueden modificar y no se pueden borrar durante un período de retención definido por el usuario. La inmutabilidad de copia de seguridad protege contra la eliminación o modificación accidental o malintencionada, incluso por parte de los administradores con privilegios. Para obtener más información, consulte Inmutabilidad de las copias de seguridad de LTR en Azure SQL Database.

Resistencia al mantenimiento del servicio

Cuando SQL Database realiza el mantenimiento en tus bases de datos y en los grupos elásticos, puede conmutar automáticamente las bases de datos para usar una réplica secundaria. Las aplicaciones cliente pueden observar breves interrupciones de conectividad cuando se produce una conmutación por error. Las aplicaciones cliente deben seguir las instrucciones de resistencia a errores transitorios para minimizar los efectos.

SQL Database permite especificar una ventana de mantenimiento que se usa normalmente para las actualizaciones de servicio y otras operaciones de mantenimiento. La configuración de una ventana de mantenimiento puede ayudarle a minimizar los efectos secundarios, como las

conmutaciones por error automáticas, durante el horario comercial. También puede recibir notificaciones anticipadas del mantenimiento planeado.

La plataforma mantiene automáticamente las puertas de enlace usadas para procesar las conexiones a SQL Database. Las actualizaciones o las operaciones de mantenimiento también pueden provocar interrupciones breves de conectividad que los clientes pueden reintentar.

Para obtener más información, vea Ventana de mantenimiento en SQL Database.

Acuerdo de nivel de servicio

El contrato de nivel de servicio (SLA) para los servicios de Azure describe la disponibilidad esperada de cada servicio y las condiciones que la solución deberá cumplir para lograr esa expectativa de disponibilidad. Para obtener más información, consulte Acuerdos de Nivel de Servicio para servicios en línea.

El acuerdo de nivel de servicio (SLA) para SQL Database describe la disponibilidad esperada del servicio y el punto de recuperación esperado y el tiempo de recuperación para la replicación geográfica activa.

Al implementar una base de datos con redundancia de zona o un grupo elástico y usar un nivel de servicio compatible, el SLA de tiempo de actividad es mayor.

Contenido relacionado

- Introducción a la continuidad empresarial con SQL Database
- Lista de comprobación de alta disponibilidad y recuperación ante desastres: SQL Database
- Confiabilidad en Azure

ⓘ**Nota:** El autor ha creado este artículo con ayuda de inteligencia artificial. Más información

---

Last updated on 27/05/2026

Confiabilidad en Azure SQL Managed Instance

Azure SQL Managed Instance es un motor de base de datos de plataforma como servicio (PaaS) totalmente administrado. Proporciona casi 100% compatibilidad de características con SQL Server. Azure SQL Managed Instance controla la mayoría de las funciones de administración de bases de datos, como la actualización, la aplicación de revisiones, las copias de seguridad y la supervisión sin intervención del usuario. Se ejecuta en la versión estable más reciente del motor de base de datos de SQL Server y un sistema operativo revisado con alta disponibilidad integrada.

Cuando se usa Azure, la confiabilidad es una responsabilidad compartida. Microsoft proporciona una variedad de funcionalidades para admitir resistencia y recuperación. Es responsable de comprender cómo funcionan esas funcionalidades dentro de todos los servicios que usa y de seleccionar las funcionalidades que necesita para cumplir los objetivos empresariales y los objetivos de tiempo de actividad.

En este artículo se describe cómo hacer que Azure SQL Managed Instance sea resistente a una variedad de posibles interrupciones y problemas, incluidos errores transitorios, interrupciones de zona de disponibilidad y interrupciones de regiones. También se describe cómo puede usar copias de seguridad para recuperarse de otros tipos de problemas, cómo controlar el mantenimiento del servicio y resaltar cierta información clave sobre el contrato de nivel de servicio (SLA) de Azure SQL Managed Instance.

Recomendaciones de implementación de producción para la confiabilidad

Para la mayoría de las implementaciones de producción de SQL Managed Instance, tenga en cuenta las siguientes recomendaciones:

- Siga las instrucciones proporcionadas en la lista de comprobación de alta disponibilidad y recuperación ante desastres (DR).
- Habilitación de la redundancia de zona.
- Configure copias de seguridad automatizadas y use almacenamiento con redundancia de zona (ZRS) o almacenamiento con redundancia geográfica (GRS) para las copias de seguridad.
- Planee probar periódicamente las copias de seguridad y el proceso de restauración.

Introducción a la arquitectura de confiabilidad

Las instancias administradas de SQL de uso general se ejecutan en un único nodo que administra Azure Service Fabric . Cada vez que se actualiza el motor de base de datos o el sistema operativo, o se detecta un error, SQL Managed Instance funciona con Service Fabric para mover el proceso del motor de base de datos sin estado a otro nodo de proceso sin estado que tenga suficiente capacidad libre. Los archivos de base de datos se almacenan en Azure Blob Storage, que tiene características de redundancia integradas. Los archivos de datos y registro se desasocian del nodo de proceso original y se adjuntan al proceso del motor de base de datos recién inicializado.

Redundancia

De forma predeterminada, SQL Managed Instance logra redundancia mediante la propagación de nodos de proceso y datos a lo largo de un único centro de datos de la región primaria. Este enfoque protege los datos durante los siguientes tiempos de inactividad esperados e inesperados:

- Operaciones de administración iniciadas por el cliente que dan lugar a un breve tiempo de inactividad
- Operaciones de mantenimiento de servicio
- Errores de red o energía a pequeña escala
- Problemas y interrupciones del centro de datos que implican los siguientes componentes:
  - *Bastidor* en el que se ejecutan las máquinas que impulsan el servicio
  - La *máquina física* que hospeda la máquina virtual (VM) que ejecuta el motor de base de datos SQL.
  - Máquina *virtual* que ejecuta el motor de base de datos de SQL
- Problemas con el motor de base de datos sql
- Posibles interrupciones localizadas no planeadas

Para obtener más información sobre cómo SQL Managed Instance proporciona redundancia, consulte Disponibilidad a través de redundancia de zona y local.

Resistencia a errores transitorios

Los errores transitorios son errores breves e intermitentes en los componentes. Se producen con frecuencia en un entorno distribuido como la nube y son una parte normal de las operaciones. Los errores transitorios se corrigen después de un breve período de tiempo. Es importante que las aplicaciones puedan controlar errores transitorios, normalmente mediante el reintento de solicitudes afectadas.

Todas las aplicaciones hospedadas en la nube deben seguir las instrucciones de control de errores transitorios de Azure cuando se comunican con cualquier API, bases de datos y otros componentes hospedados en la nube. Para obtener más información, consulte Recomendaciones para controlar errores transitorios.

Sql Managed Instance controla automáticamente las tareas de mantenimiento críticas, como la aplicación de revisiones, las copias de seguridad y las actualizaciones del motor de base de datos de SQL y Windows. También controla eventos no planeados, como errores de red, software o hardware subyacentes. Sql Managed Instance puede recuperarse rápidamente incluso en las circunstancias más críticas, lo que garantiza que los datos estén siempre disponibles. La mayoría de los usuarios no observan que las actualizaciones se realizan continuamente.

Cuando se revisa una instancia o se produce una conmutación por error, el tiempo de inactividad tiene un efecto mínimo si emplea un mecanismo de reintento en la aplicación. Puede probar la resistencia de la aplicación a errores transitorios.

Resistencia a errores de zona de disponibilidad

> ⓘ **Nota:**
>
> La redundancia de zona no está disponible actualmente para el nivel de servicio de próxima generación de uso general.

Las zonas de disponibilidad son grupos físicamente independientes de centros de datos dentro de una región de Azure. Cuando una zona falla, los servicios pueden transferirse a una de las zonas restantes.

Al habilitar una configuración con redundancia de zona, puede asegurarse de que la instancia administrada de SQL sea resistente a un gran conjunto de errores, incluidas interrupciones catastróficas del centro de datos, sin cambios en la lógica de la aplicación.

SQL Managed Instance logra la redundancia de zona al colocar nodos de cómputo sin estado en diferentes zonas de disponibilidad. Se basa en ZRS con estado que está asociado al nodo que actualmente contiene el proceso activo del motor de base de datos SQL. Si se produce una

interrupción, el proceso del motor de base de datos SQL se activa en uno de los nodos de cómputo sin estado, que luego accede a los datos en el almacenamiento con estado.

Requisitos

- **Compatibilidad con regiones:** La redundancia de zona para SQL Managed Instance se admite en regiones seleccionadas. Para obtener más información, consulte Regiones admitidas.

- **Redundancia de almacenamiento de copia de seguridad:** Para habilitar la redundancia de zona para la instancia administrada de SQL, establezca la redundancia del almacenamiento de copia de seguridad en *ZRS* o *almacenamiento con redundancia de zona geográfica* (GZRS).

Cost

Al habilitar la redundancia de zona, hay un costo adicional para la instancia administrada de SQL y para las copias de seguridad con redundancia de zona. Para obtener más información, consulte el apartado Precios.

Configurar soporte de zonas de disponibilidad

En esta sección se explica cómo configurar la compatibilidad con zonas de disponibilidad para las instancias administradas de SQL:

- **Habilite la redundancia de zona:** Para obtener información sobre cómo configurar la redundancia de zona en instancias nuevas y existentes, consulte Configuración de la redundancia de zona.

  Todas las operaciones de escalado de SQL Managed Instance, incluida la habilitación de la redundancia de zona, son operaciones en línea y requieren un tiempo de inactividad mínimo o nulo. Para obtener más información, consulte Duración de las operaciones de administración.

- **Deshabilitar la redundancia de zona:** puede deshabilitar la redundancia de zona siguiendo los mismos pasos para habilitar la redundancia de zona. Este proceso es una operación en línea similar a una actualización normal del objetivo de nivel de servicio. Al final del proceso, la instancia se migra de la infraestructura con redundancia de zona a la infraestructura de una sola zona.

Comportamiento cuando todas las zonas están en buen estado

En esta sección se describe qué esperar cuando la instancia administrada de SQL está configurada para que sea redundante de zona y todas las zonas de disponibilidad estén operativas:

- **Enrutamiento de tráfico entre zonas:** Durante las operaciones normales, las solicitudes se enrutan al nodo que ejecuta la capa de proceso de SQL Managed Instance.

- **Replicación de datos entre zonas:** Los archivos de base de datos se almacenan en Azure Storage mediante ZRS, que se adjunta al nodo que contiene actualmente el proceso activo del motor de base de datos de SQL.

  Las operaciones de escritura son sincrónicas y no se consideran completadas hasta que los datos se replican correctamente en todas las zonas de disponibilidad. Esta replicación sincrónica garantiza una fuerte coherencia y cero pérdida de datos durante los errores de zona. Sin embargo, podría dar lugar a una latencia de escritura ligeramente mayor en comparación con el almacenamiento con redundancia local.

Comportamiento durante un fallo de zona

En esta sección se describe qué esperar cuando la instancia administrada de SQL está configurada para que sea redundante de zona y una o varias zonas de disponibilidad no están disponibles:

- **Detección y respuesta:** Instancia administrada de SQL es responsable de detectar y responder a un error en una zona de disponibilidad. No es necesario hacer nada para iniciar una conmutación por error de la zona.

- **Notificación**: Microsoft no le notifica automáticamente cuando una zona está inactiva. Sin embargo, puede usar Azure Resource Health para supervisar el estado de un recurso individual y puede configurar alertas de Resource Health para notificarle problemas. También puede usar Azure Service Health para comprender el estado general del servicio, incluidos los errores de zona, y puede configurar alertas de Service Health para notificarle problemas.

- **Solicitudes activas:** Cuando una zona de disponibilidad no está disponible, las solicitudes que se procesan en la zona de disponibilidad defectuosa se finalizan y se deben reintentar. Para que las aplicaciones sean resistentes a estos tipos de problemas, consulte Guía sobre resistencia a errores transitorios .

- **Reenrutamiento del tráfico:** Sql Managed Instance funciona con Service Fabric para mover el motor de base de datos a un nodo de proceso sin estado adecuado que se encuentra en una zona de disponibilidad diferente y tiene suficiente capacidad libre. Una vez completada la conmutación por error, las nuevas conexiones se redirigen automáticamente al nuevo nodo de proceso principal.

  Una carga de trabajo pesada podría experimentar cierta degradación del rendimiento durante la transición de un nodo de proceso al otro porque el nuevo proceso del motor de base de datos comienza con una caché inactiva.

- **Tiempo de inactividad esperado:** puede haber una pequeña cantidad de tiempo de inactividad durante una conmutación por error de zona de disponibilidad. El tiempo de inactividad suele ser inferior a 30 segundos, que la aplicación debe tolerar si sigue las instrucciones de resistencia a errores transitorios .

- **Pérdida de datos esperada:** no se espera ninguna pérdida de datos para las transacciones confirmadas durante una conmutación por error de zona de disponibilidad. Es necesario reintentar las transacciones en curso.

Recuperación de zona

Cuando se recupera la zona de disponibilidad, SQL Managed Instance funciona con Service Fabric para restaurar operaciones en la zona recuperada. No se requiere intervención del cliente.

Prueba de fallos de zona

La plataforma instancia administrada de SQL administra el enrutamiento del tráfico, la conmutación por error y la conmutación por recuperación para las instancias con redundancia de zona. Dado que esta característica está totalmente administrada, no es necesario iniciar ni validar los procesos de error de zona de disponibilidad. Sin embargo, puede validar la gestión de fallos de su aplicación.

Resistencia a errores en toda la región

Una instancia administrada de SQL individual se implementa dentro de una sola región. Sin embargo, puede implementar una instancia administrada de SQL secundaria en una región de Azure independiente y configurar un *grupo de conmutación por error*.

Grupos de conmutación por error

Los grupos de conmutación por error replican automáticamente de forma geográfica sus datos y pueden conmutar por error automáticamente o manualmente si se produce un error regional, en función de la directiva de conmutación por error.

En esta sección se resume la información clave sobre los grupos de conmutación por error, pero es importante revisar la introducción a los grupos de conmutación por error y los procedimientos recomendados para obtener más información sobre cómo funcionan y cómo configurarlos.

Directivas de conmutación por error

Al crear un grupo de conmutación por error, seleccione la directiva de conmutación por error, que especifica quién es responsable de detectar una interrupción y realizar una conmutación por error. Puede configurar dos tipos de políticas de conmutación por error:

- **Conmutación por error administrada por el cliente (recomendada):** Cuando se usa una directiva de conmutación por error administrada por el cliente, puede decidir si se debe realizar una *conmutación* por error, que no incurre en pérdida de datos o una *conmutación por error forzada*, lo que podría suponer una pérdida de datos. La conmutación por error forzada se utiliza como método de recuperación durante las interrupciones del servicio cuando no se puede acceder a la instancia principal.
- **Conmutación por error administrada por Microsoft:** La conmutación por error administrada por Microsoft solo se usa en situaciones excepcionales para desencadenar una conmutación por error forzada.

> **Importante**
>
> Utilice las opciones de conmutación por error administradas por el cliente para desarrollar, probar e implementar tus planes de recuperación ante desastres. **No confíe en la conmutación por error administrada por Microsoft**, que solo se puede utilizar en circunstancias extremas. Es probable que se haya iniciado una conmutación por error administrada por Microsoft para toda una región. No se puede iniciar para grupos de conmutación por error individuales, instancias administradas de SQL, suscripciones o clientes. La conmutación por error puede producirse en momentos diferentes para distintos servicios de Azure. Recomendamos que utilice la conmutación por error administrada por el cliente.

Soporte para regiones

Puede seleccionar cualquier región de Azure para las instancias administradas de SQL en el grupo de conmutación por error. Debido a la alta latencia de las redes de área extensa, la replicación geográfica usa un mecanismo de replicación asincrónico. Para reducir los retrasos de red, seleccione regiones que tengan conexiones de baja latencia. Para más información sobre la latencia entre regiones de Azure, consulte Estadísticas de latencia de ida y vuelta de red de Azure.

Cost

Al crear varias instancias administradas de SQL en regiones diferentes, se le factura cada instancia administrada de SQL.

Sin embargo, si la instancia secundaria no tiene ninguna carga de trabajo de lectura o aplicaciones conectadas a ella, puede ahorrar en costes de licencias mediante la designación de la réplica como una instancia en espera. Para más información, consulte Configuración de una réplica en espera sin licencia para SQL Managed Instance.

Para más información sobre los precios de SQL Managed Instance, consulte Información de precios del servicio.

Configuración de la compatibilidad con varias regiones

Para obtener información sobre cómo configurar un grupo de conmutación por error, consulte Configuración de un grupo de conmutación por error para SQL Managed Instance.

Planeamiento y administración de capacidad

Durante una conmutación por error, el tráfico se redirige a una instancia administrada de SQL secundaria. Es importante que la instancia administrada de SQL secundaria esté lista para recibir tráfico. Cree una instancia administrada de SQL secundaria con el mismo nivel de servicio, generación de hardware y tamaño de proceso que la instancia principal.

Para más información sobre el escalado de instancias administradas de SQL en un grupo de conmutación por error, consulte Escalado de instancias.

Comportamiento cuando todas las regiones están en buen estado

En esta sección se describe qué esperar cuando las instancias administradas de SQL están configuradas para usar grupos de conmutación por error de varias regiones y todas las regiones están operativas:

- **Enrutamiento del tráfico entre regiones:** durante las operaciones normales, las solicitudes de lectura y escritura van a la instancia principal única de la región primaria.

  Los grupos de conmutación por error también proporcionan un punto de conexión de agente de escucha de solo lectura independiente. Durante las operaciones normales, este punto de conexión se conecta a la instancia secundaria para enrutar el tráfico de solo lectura especificado en la cadena de conexión.

  Para obtener más información sobre cómo los grupos de conmutación por error envían tráfico a cada instancia y cómo puede dirigir el tráfico a un punto de conexión del oyente de solo lectura, consulte Introducción a los grupos de conmutación por error y procedimientos recomendados.

- **Replicación de datos entre regiones:** de manera predeterminada, los datos se replican de forma asincrónica desde la instancia principal a la instancia administrada de SQL secundaria.

  Dado que la replicación geográfica es asincrónica, si realiza una conmutación por error forzada, es posible experimentar la pérdida de datos. Puede supervisar el retraso de replicación para comprender la posible pérdida de datos durante una conmutación por error forzada. Para más información, consulte lista de comprobación de recuperación ante desastres.

  Si necesita eliminar la pérdida de datos de la replicación asincrónica durante las conmutaciones por error, configure la aplicación para bloquear el subproceso de llamada hasta que confirme que la última transacción confirmada se ha transmitido y protegido en el registro de transacciones de la base de datos secundaria. Este enfoque requiere desarrollo personalizado y reduce el rendimiento de la aplicación. Para obtener más información, consulte Evitar la pérdida de datos críticos.

Comportamiento durante una falla de región

En esta sección se describe qué esperar cuando las instancias administradas de SQL están configuradas para usar grupos de conmutación por error de varias regiones y hay una interrupción en la región primaria:

- **Detección y respuesta:** la responsabilidad de la detección y respuesta depende de la directiva de conmutación por error utilizada por el grupo de conmutación por error.

  - *Directiva de conmutación por error administrada por el cliente:* es responsable de detectar el error en una región y desencadenar una conmutación por error o una conmutación por error forzada a la instancia secundaria del grupo de conmutación por error.

Si realiza una conmutación por error, SQL Managed Instance espera a que los datos se sincronicen con la instancia secundaria antes de realizar el procedimiento de conmutación por error.

Si realiza una conmutación por error forzada, SQL Managed Instance cambia inmediatamente la instancia secundaria al rol principal sin esperar a que los cambios recientes se propaguen desde la principal. Este tipo de conmutación por error puede incurrir en pérdida de datos.

- *Directiva de conmutación por error administrada por Microsoft:* las conmutaciones por error administradas por Microsoft se realizan en circunstancias excepcionales. Cuando Microsoft desencadena una conmutación por error, el grupo de conmutación por error realiza automáticamente una conmutación por error forzada a la instancia secundaria del grupo de conmutación por error. Sin embargo, se recomienda usar una directiva de conmutación por error administrada por el cliente para cargas de trabajo de producción para que pueda controlar cuándo se produce la conmutación por error.

- **Notificación**: Microsoft no le notifica automáticamente cuando una zona está inactiva. Sin embargo, puede usar Azure Resource Health para supervisar el estado de un recurso individual y puede configurar alertas de Resource Health para notificarle problemas. También puede usar Azure Service Health para comprender el estado general del servicio, incluidos los errores de zona, y puede configurar alertas de Service Health para notificarle problemas.

- **Solicitudes activas:** Cuando se produce una conmutación por error, se finalizan las solicitudes que se están procesando y se deben reintentar. Para que las aplicaciones sean resistentes a estos tipos de problemas, consulte Resistencia a errores transitorios.

- **Pérdida de datos esperada:** La cantidad de pérdida de datos depende de cómo configure la aplicación. Para más información, consulte Introducción a los grupos de conmutación por error y procedimientos recomendados.

- **Tiempo de inactividad esperado:** Puede haber una breve interrupción durante una conmutación por error de un grupo. El tiempo de inactividad suele ser inferior a 60 segundos.

- **Reenrutamiento del tráfico:** Una vez que el grupo de conmutación por error completa el proceso de conmutación por error, el tráfico de lectura y escritura se enruta automáticamente a la nueva instancia principal. Si las aplicaciones usan los puntos de conexión del grupo de conmutación por error en sus cadenas de conexión, no necesitan modificar sus cadenas de conexión después de la conmutación por error.

Recuperación de regiones

Los grupos de conmutación por error no conmutan automáticamente por recuperación a la región primaria cuando se restauran y, por tanto, es su responsabilidad iniciar una conmutación por recuperación.

Prueba de fallos de región

Puede probar la conmutación por error de un grupo de conmutación por error.

Probar un grupo de conmutación por error es solo una parte de realizar un simulacro de recuperación ante desastres. Para más información, consulte Realización de simulacros de DR.

Copias de seguridad y restauración

Realice copias de seguridad de las bases de datos para protegerse frente a diversos riesgos, incluida la pérdida de datos. Las copias de seguridad se pueden restaurar para recuperarse de pérdida accidental de datos, daños u otros problemas. Las copias de seguridad no son lo mismo que la replicación geográfica, y tienen diferentes propósitos y mitigan distintos riesgos.

Instancia administrada de SQL realiza automáticamente copias de seguridad completas, diferenciales y del registro de transacciones de las bases de datos. Para obtener más información sobre los tipos de copias de seguridad, su frecuencia, funcionalidades de restauración, costos de almacenamiento y cifrado de copia de seguridad, consulte Copias de seguridad automatizadas en SQL Managed Instance.

SQL Managed Instance proporciona copias de seguridad automatizadas integradas y también admite copias de seguridad de solo copia iniciadas por el usuario para bases de datos de usuario. Para más información, consulte Copias de seguridad de solo copia.

Replicación de copia de seguridad

Al configurar copias de seguridad automatizadas para la instancia administrada de SQL, puede especificar cómo se deben replicar las copias de seguridad. Las copias de seguridad configuradas para almacenarse en ZRS tienen un mayor nivel de resistencia. Se recomienda configurar las copias de seguridad para que usen uno de los siguientes tipos de almacenamiento:

- ZRS para la resistencia dentro de la región, si la región tiene zonas de disponibilidad
- GZRS para mejorar la resistencia de las copias de seguridad entre regiones si la región tiene zonas de disponibilidad y está emparejada con otra región
- GRS si la región no admite zonas de disponibilidad, pero tiene una región emparejada

Para más información sobre los distintos tipos de almacenamiento y sus funcionalidades, consulte Redundancia de almacenamiento de respaldo.

Restauración geográfica

La funcionalidad de restauración geográfica es una solución básica de recuperación ante desastres que permite restaurar copias de seguridad en otra región de Azure. La copia de seguridad geográfica normalmente requiere una cantidad significativa de tiempo de inactividad y pérdida de datos. Para lograr niveles más altos de capacidad de recuperación si se produce una interrupción regional, debe configurar grupos de conmutación por error.

Si usa la restauración geográfica, debe tener en cuenta cómo hacer que las copias de seguridad estén disponibles en la región secundaria:

- Si la región primaria está emparejada, use el almacenamiento de copia de seguridad de GZRS o GRS para admitir la restauración geográfica en la región emparejada.
- Si la región primaria no está emparejada, puede crear una solución personalizada para replicar las copias de seguridad en otra región. Considere la posibilidad de usar copias de seguridad de solo copia iniciadas por el usuario y almacenarlas en una cuenta de almacenamiento que use la replicación de objetos de blob para replicar en una cuenta de almacenamiento de otra región.

Resistencia al mantenimiento del servicio

Cuando SQL Managed Instance realiza el mantenimiento en la instancia, la instancia administrada de SQL permanece totalmente disponible, pero puede estar sujeta a reconfiguraciones breves. Las aplicaciones cliente pueden observar breves interrupciones de conectividad cuando se produce un evento de mantenimiento. Las aplicaciones cliente deben seguir las instrucciones de resistencia a errores transitorios para minimizar los efectos.

Instancia administrada de SQL permite especificar una ventana de mantenimiento que se usa generalmente para las actualizaciones del servicio y otras operaciones de mantenimiento. La configuración de una ventana de mantenimiento puede ayudarle a minimizar los efectos secundarios, como las conmutaciones automáticas por error, durante el horario comercial. También puede recibir notificaciones anticipadas del mantenimiento planeado.

Para obtener más información, consulte Ventana de mantenimiento en SQL Managed Instance.

Acuerdo de nivel de servicio

El contrato de nivel de servicio (SLA) para los servicios de Azure describe la disponibilidad esperada de cada servicio y las condiciones que la solución deberá cumplir para lograr esa expectativa de disponibilidad. Para obtener más información, consulte Acuerdos de Nivel de Servicio para servicios en línea.

Para SQL Managed Instance, el Acuerdo de Nivel de Servicio de disponibilidad solo se aplica cuando la red virtual de Azure está configurada correctamente para que no impida el tráfico de administración. Esta configuración incluye el tamaño de subred, los grupos de seguridad de red (NSG), las rutas definidas por el usuario (UDR), la configuración de DNS y otros recursos que afectan a la administración y el uso de recursos de red. Para obtener más información sobre la configuración de red necesaria para SQL Managed Instance, consulte Requisitos de red.

Contenido relacionado

- Introducción a la continuidad empresarial con SQL Managed Instance
- Lista de comprobación de alta disponibilidad y recuperación ante desastres
- Confiabilidad en Azure

Last updated on 22/01/2026

Confiabilidad en el Centro de API de Azure

En este artículo se describe el soporte de fiabilidad en el Centro de API de Azure, incluidas las zonas de disponibilidad, la redundancia de zona, la residencia de datos y lo que los clientes deben esperar durante las caídas de zona o región. Para obtener información general más detallada sobre la confiabilidad de Azure, consulte Confiabilidad de Azure.

Soporte para zonas de disponibilidad

Las zonas de disponibilidad son grupos físicamente independientes de centros de datos dentro de una región de Azure. Cuando una zona falla, los servicios pueden transferirse a una de las zonas restantes.

En cada región donde está disponible, el Centro de API de Azure admite la redundancia de zona de forma predeterminada. El servicio de Centro de API de Azure se ejecuta en un entorno multiinquilino en componentes habilitados para zonas de disponibilidad. No es necesario configurarlo ni volver a configurarlo para la compatibilidad con las zonas de disponibilidad.

Experiencia a nivel de zona

Durante una interrupción en toda la zona, el cliente debe esperar una breve degradación del rendimiento hasta que la recuperación automática del servicio vuelva a equilibrar la capacidad subyacente para ajustarse a las zonas en estado correcto. Este período de degradación del rendimiento no depende de la restauración de zonas; Se espera que el estado de recuperación automática del servicio administrado por Microsoft compense una zona perdida mediante el uso de la capacidad de otras zonas.

Limitaciones de recuperación ante desastres entre regiones

Actualmente, el Centro de API de Azure solo admite una configuración de una sola región. No hay capacidad de conmutación por error automática ni activada por el cliente entre regiones para el Centro de API de Azure. Si se produce un desastre regional, el servicio no está disponible hasta que se restaure la región.

Residencia de datos

Todos los datos del cliente permanecen en la región donde se implementa un centro de API. El Centro de API de Azure no replica datos entre regiones.

Pasos siguientes

Confiabilidad en Azure

Last updated on 22/01/2026

Información general sobre la protección de datos

Azure DevOps Services

Azure DevOps Services es una aplicación hospedada en la nube para los proyectos de desarrollo, desde la planeación hasta la implementación. En función de las prestaciones locales, con más servicios en la nube, administramos el código fuente, los elementos de trabajo, las compilaciones y las pruebas. Azure DevOps usa la infraestructura de plataforma como servicio (PaaS) y muchos servicios de Azure, incluidos Azure SQL, para ofrecer un servicio confiable y disponible globalmente para los proyectos de desarrollo.

En este artículo se describen los pasos que Microsoft realiza para ayudar a mantener los proyectos seguros, disponibles, seguros y privados. Describe el rol que desempeña el cliente para mantener los proyectos seguros y protegidos.

Este artículo está destinado a los administradores de organizaciones y a los profesionales de TI que administran sus recursos de proyecto a diario. Resulta más útil a los usuarios que ya conocen Azure DevOps y desean obtener más información sobre cómo Microsoft protege los recursos almacenados en Azure DevOps.

Nuestro Compromiso

Microsoft ayuda a garantizar que los proyectos permanezcan seguros y seguros, sin excepción. Cuando se almacenan proyectos en Azure DevOps, estos se benefician de varias capas de tecnologías de seguridad y gobernanza, prácticas operativas y directivas de cumplimiento. Aplicamos la privacidad y la integridad de los datos tanto en reposo como en tránsito.

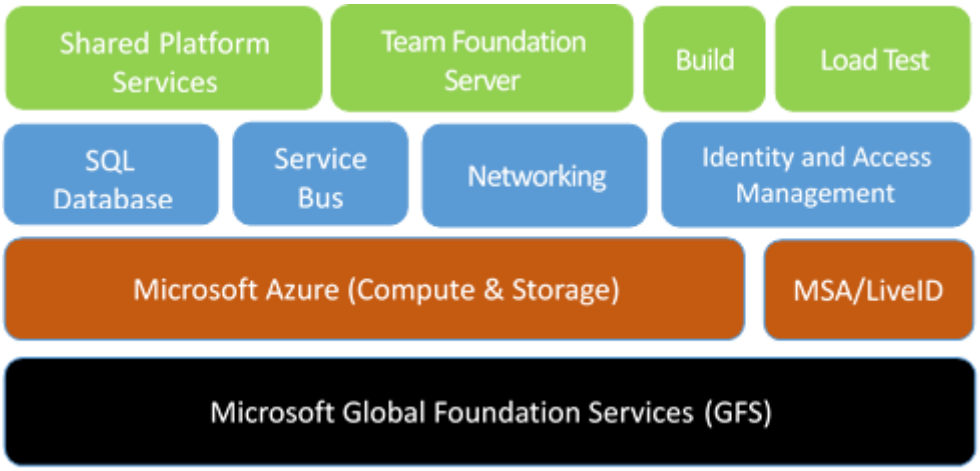
Las amenazas a las que se enfrenta pertenecen a cuatro categorías básicas: disponibilidad de datos, disponibilidad del servicio, seguridad del servicio y privacidad de los datos. En este artículo se exploran amenazas específicas pertenecientes a cada categoría y se explica qué hace Azure DevOps para abordarlas. El artículo comienza con una descripción de cómo se almacenan los datos y de cómo administra Azure DevOps el acceso a ellos.

Para proteger los datos, se requiere la participación activa de administradores y usuarios que sepan qué pasos se deben seguir para proteger los recursos frente a la divulgación y la manipulación no autorizadas. Cuando conceda permisos a los puntos de acceso de usuarios,

hágalo de manera explícita, de tal forma que únicamente las personas adecuadas puedan obtener acceso a los datos contenidos en Azure DevOps.

Debe considerar que todos los datos pueden estar en riesgo, independientemente de dónde se encuentren o cómo se usen. Esto se aplica a los datos almacenados tanto en la nube como en centros de datos privados. Es importante clasificar los datos, su nivel de confidencialidad y de riesgo, y el daño que podría derivarse en caso de que se viesen comprometidos. Además, hay que clasificar los datos de conformidad con una directiva general de administración de la seguridad de la información.

Basado en Azure



Hospedamos Azure DevOps íntegramente en centros de datos de Azure. Azure DevOps usa muchos servicios principales de Azure, como proceso, almacenamiento, redes, Azure SQL, administración de identidades y acceso, y Azure Service Bus.

Azure DevOps usa Azure Storage como repositorio principal para los metadatos del servicio y los datos del cliente. Según el tipo de datos y los requisitos de almacenamiento y recuperación, Azure DevOps usa Azure Blob Storage y Azure SQL Database Storage.

Para ayudarle a comprender el enfoque de Azure DevOps Services respecto a la protección de datos, aquí presentamos información de referencia sobre los servicios de almacenamiento:

- Azure Blob Storage almacena grandes fragmentos de datos no estructurados. Todos los proyectos usan este servicio. Los datos incluyen información potencialmente confidencial o privada, como el contenido de archivos de origen y datos adjuntos para elementos de trabajo. Para casi todos los proyectos, la mayoría del almacenamiento en uso es este tipo de almacenamiento de blobs no estructurado.

- Azure SQL Database almacena los aspectos estructurados y transaccionales de su organización, incluidos los metadatos de los proyectos, el historial de control de código fuente con versiones y los detalles de los elementos de trabajo. El almacenamiento de bases de datos proporciona acceso rápido a los elementos importantes del proyecto. Proporciona índices en Blob Storage para buscar archivos y datos adjuntos.

Los administradores pueden administrar el acceso a los recursos concediéndole o restringiendo permisos en las identidades o grupos de usuarios. Azure DevOps usa la autenticación federada de identidades de usuario a través de Microsoft Entra ID y cuentas de Microsoft.

Durante la autenticación, el usuario se enruta al proveedor de autenticación, donde proporciona sus credenciales. Una vez que el proveedor de autenticación comprueba las credenciales del usuario, Azure DevOps envía a este último una cookie de autenticación. Esta cookie permite al usuario permanecer autenticado en Azure DevOps.

De esta manera, la información de credenciales del usuario nunca se comparte directamente con Azure DevOps. Para cada recurso de Azure DevOps al que el usuario intenta obtener acceso, la validación se basa en los permisos explícitos del usuario y en los permisos que este heredó a través de la pertenencia a grupos.

Los administradores pueden usar controles de acceso para ayudar a proteger el acceso a la organización, las colecciones de proyectos y los proyectos de equipo, así como los datos y la funcionalidad de ámbito de equipo. Los administradores también pueden usar controles de acceso para recursos específicos, como carpetas para el control de versiones y rutas de área para los elementos de trabajo.

Disponibilidad de datos

Azure DevOps usa numerosas características de Azure Storage para ayudar a garantizar la disponibilidad de los datos en caso de producirse un error de hardware, una interrupción del servicio o un desastre regional. Además, el equipo de Azure DevOps sigue los procedimientos destinados a ayudar a proteger los datos contra la eliminación accidental o malintencionada.

Redundancia de datos

Con el fin de ayudar a proteger los datos en caso de errores de hardware o de servicio, Azure Storage replica geográficamente los datos del cliente entre dos regiones de la misma ubicación geográfica. Por ejemplo, Azure Storage puede replicar geográficamente los datos entre el norte y el oeste de Europa o entre norte y el sur de los Estados Unidos.

En el caso de Azure Blob Storage, los datos del cliente se replican tres veces dentro de una sola región. Los datos del cliente se replican de forma asincrónica en una segunda región de la misma ubicación geográfica. Por lo tanto, Azure siempre mantiene el equivalente de seis copias de los datos.

Contar con varias copias le permite conmutar por error a una región independiente si se produce una interrupción o un desastre de gran calado, además de ofrecer redundancia local para errores de hardware dentro de una región. Para Azure SQL almacenamiento de base de datos, las copias de seguridad diarias se mantienen fuera del sitio si se produce un desastre regional.

Con respecto a la redundancia de datos y la conmutación por error:

- Hay una diferencia inherente, medida en minutos, cuando Microsoft replica los datos entre la región primaria y la secundaria.
- La conmutación por error a la región secundaria es una decisión que Microsoft debe tomar de forma centralizada, ya que afecta a todos los clientes de una unidad de escalado determinada. Excepto en circunstancias extremas, Microsoft prefiere evitar realizar una conmutación por error para que no se pierdan los datos del cliente.
- Azure DevOps ofrece un acuerdo de nivel de servicio (SLA) del tiempo de actividad del 99,9 %. Azure DevOps devuelve una parte de los cargos mensuales si el SLA se incumple en un mes determinado.
- Dado que solo hay una región en Brasil, los datos de los clientes de esta nación se pueden replicar en la región Centro-sur de EE. UU. con fines de recuperación ante desastres.

Se producen errores

Para proteger contra la pérdida accidental de datos, Microsoft utiliza las copias de seguridad de un momento dado para los blobs almacenados en Azure Blob Storage y las bases de datos contenidas en Azure SQL Database. En cada cuenta de almacenamiento, se mantiene una copia independiente de todos los blobs, cuyos cambios se anexan a los datos existentes. Estas copias de seguridad son inmutables, lo que elimina la necesidad de reescribir los almacenamientos existentes durante los procedimientos de copia de seguridad.

Azure SQL Database incluye características de copia de seguridad estándar usadas por Azure DevOps. Los datos se conservan durante 28 días y estas copias de seguridad también se replican en una región emparejada para facilitar la recuperación en caso de interrupción regional.

Puede recuperar organizaciones o proyectos eliminados dentro del período de 28 días posterior a la eliminación. Sin embargo, una vez transcurrido este período, estas entidades se eliminan de forma permanente y ya no se pueden restaurar. Aunque las copias de seguridad sirven como

componente fundamental en la recuperación ante desastres, es esencial que los clientes apliquen las estrategias adecuadas de copia de seguridad y administración de datos para garantizar una protección íntegra de sus datos.

> **Importante**
>
> - Por *eliminación accidental* se entienden los escenarios derivados de un incidente en nuestros servicios. No incluye la eliminación accidental de recursos por parte de los clientes (por ejemplo, repositorios, elementos de trabajo, datos adjuntos o artefactos).
> - No admitimos la restauración de recursos que los clientes hayan eliminado accidentalmente. Estas copias de seguridad están pensadas únicamente para ofrecer continuidad empresarial y para facilitar la recuperación en escenarios de interrupción o de desastre.
> - En raras ocasiones, el proceso de eliminación puede tardar hasta 70 días debido a reintentos de back-end y la necesidad de eliminar datos de varios orígenes.

La práctica es crítica

Contar con varias copias de seguridad de los datos es positivo, pero si no se tiene práctica, la restauración puede resultar impredecible. Se suele decir que "lo que falla no son las copias de seguridad, sino las restauraciones". Aunque la afirmación es técnicamente incorrecta, la opinión sí es acertada.

En Microsoft, practicamos con regularidad la restauración de conjuntos de datos a partir de copias de seguridad. Sometemos a pruebas periódicas el almacenamiento con redundancia geográfica de Azure. Hay muchas combinaciones de escenarios de desastres y de daños en los datos. Continuamos planeando y llevando a cabo nuevas pruebas para estos escenarios con frecuencia.

Disponibilidad del servicio

Azure DevOps ofrece protecciones de denegación de servicio distribuidas (DDoS) y respuesta al sitio activo para ayudar a garantizar que tiene acceso a la organización y a los recursos asociados.

Protecciones de DDoS

En algunos casos, un ataque DDoS malintencionado puede afectar a la disponibilidad del servicio. Azure tiene un sistema de defensa DDoS que ayuda a evitar ataques contra nuestro servicio. Usa

técnicas de detección y mitigación estándar, como cookies SYN, limitación de velocidad y límites de conexión. El sistema está diseñado para resistir ataques no solo desde el exterior, sino también desde Dentro de Azure.

En el caso de los ataques específicos de aplicaciones que pueden penetrar en los sistemas de defensa de Azure, Azure DevOps establece cuotas de nivel de aplicación y de nivel de organización, así como limitaciones. Esta práctica ayuda a evitar el uso excesivo de los recursos de servicios esenciales durante un ataque o un uso incorrecto accidental de los recursos.

Respuesta del sitio activo

En raras circunstancias, es posible que necesite una respuesta del sitio activo a un problema con la disponibilidad del servicio. Nuestro equipo de operaciones está disponible en todo momento para identificar enseguida el problema e interactuar con el personal necesario del equipo de desarrollo.

A continuación, el equipo de desarrollo soluciona el problema. También trata de actualizar la página de estado del servicio en los minutos siguientes desde que se detecta un problema que afecta al servicio. Una vez que el equipo de desarrollo ha solucionado un problema, identifica la causa principal y realiza un seguimiento de los cambios necesarios para evitar que se reproduzcan otros similares en el futuro.

Los procesos de Azure DevOps de administración de sitios activos se centran en la experiencia del cliente y en el estado del servicio. Estos procesos minimizan el tiempo para detectar, responder y mitigar los problemas. Todas las disciplinas de ingeniería están implicadas y asumen su responsabilidad, por lo que la evolución de las mejoras continuas se basa en la experiencia directa. Así, se mejoran con el tiempo los procesos de supervisión, diagnóstico, resiliencia y control de calidad.

La administración de sitios activos en Azure DevOps tiene tres pistas distintas: telemetría, administración de incidentes y revisión del sitio activo. Esto es lo que conllevan estas pistas:

Telemetry	Incident management	Live-site review
• **Alerts** – define health alerts for failure modes • **Diagnostics** – deliver instrumentation data and operational reports • **Troubleshooting guides** – guidance for investigating an alert is defined by the feature, and then refined by the Service Engineer • **Failure mode testing** – the Service Delivery (SD) team performs failure testing to ensure alerts fire as expected • **Onboarding** – the feature team works with their Service Engineer (SE) to onboard new alerts to the 24 x 7 team	• **Detection** – product alerts detect health issues and start the Live Site Incident (LSI) process • **Triage** – The 24 x 7 team receives all critical alerts and confirms impact using Azure DevOps guidance • **Escalation** – both Dev and Ops have individuals in an on-call rotation. SE is the initial escalation path. The SE calls Dev, as needed • **Incident management** – a bridge is managed by the SE who engages Dev. and Partners to troubleshoot • **Resolution** - communication and service restoration are actively driven until customer impact is eliminated	• **Goal** – weekly review of LSI ensure that leadership has visibility into live site health and repeat issues • **Cadence** – Incidents from prior week have root cause documented, then reviewed on weekly basis • **Audience** – VS Leadership. Partner team when they drive impact. Developer attends to provide details on Service incident • **Ownership** – Dev. Owns reviews for App and Deploy issues. SD owns for Platform issues. • **Driving improvements** – Bugs and problem work items are logged for gaps (e.g. – missing alerts) and repeat root cause

El equipo de operaciones también supervisa las métricas de disponibilidad de las organizaciones individuales. Estas métricas proporcionan información sobre condiciones específicas que podrían afectar solo a algunos de nuestros clientes. Las investigaciones sobre estos datos a menudo pueden dar lugar a mejoras dirigidas para solucionar problemas específicos del cliente. En algunos casos, Microsoft puede incluso ponerse en contacto con usted directamente para comprender su experiencia y trabajar con usted para mejorar el servicio.

Microsoft publica un SLA y proporciona una garantía financiera a fin de velar por que cumplamos este acuerdo cada mes. Para más información, consulte Acuerdo de Nivel de Servicio para Azure DevOps.

A veces, los equipos asociados o las dependencias tienen incidentes que afectan a Azure DevOps. Todos los equipos asociados siguen enfoques similares para identificar, resolver y aprender de estas interrupciones del servicio.

Seguridad del servicio

La seguridad del servicio requiere vigilancia constante, desde técnicas de diseño y codificación adecuadas a factores operativos. Microsoft invierte activamente en la prevención de agujeros de

seguridad y en la detección de infracciones. Si se produce una infracción, Microsoft usa planes de respuesta de seguridad para minimizar la pérdida de datos, la pérdida o los daños. Para obtener más información, consulte Acerca de la seguridad, la autenticación y la autorización.

Seguridad por diseño

Azure DevOps está diseñado para ser seguro. Azure DevOps usa el ciclo de vida de desarrollo de seguridad de Microsoft en el núcleo de su proceso de desarrollo. El programa Microsoft Operational Security Assurance sirve para orientar los procedimientos de operación en la nube en Azure DevOps.

El equipo de Azure DevOps presenta requisitos de formación anual para todos los ingenieros y personal de operaciones. El equipo también patrocina reuniones informales moderadas por ingenieros de Microsoft. Una vez que el equipo ha resuelto un problema presentado durante una reunión, comparte las conclusiones extraídas con los demás equipos.

En las metodologías siguientes, se especifican los requisitos de formación:

- Modelar las amenazas durante el diseño del servicio
- Aplicar procedimientos recomendados para el diseño y la codificación
- Comprobar la seguridad con herramientas y pruebas estándar
- Limitar el acceso a los datos operativos y de los clientes
- Controlar la implementación de nuevas funciones a través de un rígido proceso de aprobación

Un servicio en la nube solo es tan seguro como la plataforma host. Azure DevOps usa PaaS para gran parte de su infraestructura. PaaS proporciona automáticamente actualizaciones periódicas de vulnerabilidades de seguridad conocidas.

Las máquinas virtuales hospedadas en Azure usan la infraestructura como servicio (IaaS), por ejemplo, para un servicio de compilación hospedado. Estas imágenes reciben actualizaciones periódicas para incluir las revisiones de seguridad más recientes disponibles en Windows Update. El mismo rigor de actualización se aplica a las máquinas locales, incluidas aquellas que se usan con fines de implementación, supervisión y generación de informes.

El equipo de Azure DevOps realiza pruebas de penetración periódicas centradas en la seguridad de Azure DevOps. Las pruebas de penetración tratan de aprovechar los servicios de producción activos y la infraestructura de Azure DevOps aplicando las mismas técnicas y mecanismos que usan los atacantes malintencionados. El objetivo es identificar vulnerabilidades del mundo real, configuraciones, errores u otras brechas de seguridad en un proceso controlado.

El equipo revisa los resultados de estas pruebas para identificar otras áreas de mejora y aumentar la calidad de los sistemas preventivos, así como la formación. Puede revisar los resultados de las pruebas de penetración recientes de Azure DevOps en el Portal de confianza del servicio de Microsoft.

Seguridad de credenciales

Microsoft se compromete a garantizar que todos sus proyectos permanezcan seguros y protegidos, sin excepción. En Azure DevOps, los proyectos se benefician de varias capas de tecnologías de seguridad y gobernanza, prácticas operativas y directivas de cumplimiento. Aplicamos la privacidad y la integridad de los datos tanto en reposo como en tránsito. Además, cumplimos los procedimientos siguientes con respecto a las credenciales o secretos que se almacenan en Azure DevOps. Para obtener más información sobre cómo elegir el mecanismo de autenticación adecuado, consulte nuestra guía sobre autenticación.

> **Importante**
>
> Azure DevOps no admite la autenticación de credenciales alternativas. Si sigue usando credenciales alternativas, cambie a un método de autenticación más seguro.

Tokens de acceso personal (PAT)

- Almacenamos un hash del PAT.
- Los PAT sin procesar se generan en memoria del lado del servidor. Se generan 32 bytes aleatoriamente a través de RNGCryptoServiceProvider y se comparten con el autor de la llamada como una cadena codificada en base 32. Este valor NO se almacena.
- El hash PAT se genera en memoria en el lado del servidor como un *HMACSHA256Hash* del PAT sin procesar mediante una clave de firma simétrica de 64 bytes almacenada en nuestra bóveda de claves.
- El hash se almacena en nuestra base de datos.

> **Importante**
>
> Considere la posibilidad de usar los tokens de Microsoft Entra más seguros sobre tokens de acceso personal de mayor riesgo. Para obtener más información, consulte Reducir el uso de PAT. Revise las instrucciones de autenticación para elegir el mecanismo de autenticación adecuado para sus necesidades.

Claves de Secure Shell (SSH)

- Almacenamos un hash del identificador de la organización contenedora y la clave pública SSH.
- El autor de la llamada proporciona las claves públicas sin procesar, directamente a través de SSL.
- La hash SSH se genera en la memoria del lado del servidor como un *HMACSHA256Hash* del identificador de la organización y la clave pública en bruto, utilizando una clave de firma simétrica de 64 bytes almacenada en nuestro almacén de claves.
- El hash se almacena en nuestra base de datos.

Credenciales de OAuth (JWT)

- Las credenciales de OAuth emitidas son tokens web JSON (JWT) autodescriptivos que no se guardan en nuestro servicio.
- Las notificaciones de JWT emitidas y presentadas a nuestro servicio se validan mediante un certificado almacenado en nuestro almacén de claves.

> **Importante**
>
> Azure DevOps cifra los tokens de autenticación. Una vez que el cifrado surte efecto, los clientes ya no pueden leer las cargas de token. Las aplicaciones que descodifican tokens para extraer reclamaciones dejarán de funcionar. Siempre trate los tokens como opacos: úselos solo para la autorización y llame a las APIs REST de Azure DevOps para recuperar datos de usuario u organización. Para obtener más información, consulte ¿Por qué no debería descodificar o leer afirmaciones en los tokens de autenticación?

Informes de errores de seguridad

Si considera que las pruebas de penetración han revelado un posible fallo de seguridad relacionado con el servicio Azure DevOps, comuníqueselo a Microsoft en el plazo de 24 horas. Para obtener más información, consulte la página web de Microsoft para notificar vulnerabilidades de seguridad informática.

> **Importante**
>
> Aunque no es preciso notificar a Microsoft las actividades de pruebas de penetración, debe cumplir las reglas de realización de las pruebas de penetración de Microsoft.

Programa de recompensas

Azure DevOps participa en el Programa de recompensas de Microsoft por la detección de errores. Este programa recompensa a los investigadores de seguridad que nos comunican problemas y anima a que otras personas contribuyan a mantener la seguridad de Azure DevOps. Para obtener más información, consulte el programa de recompensas de Microsoft Azure DevOps.

Restricción del acceso

Microsoft mantiene un control estricto sobre quién obtiene acceso a nuestros datos de cliente y entorno de producción. Concedemos acceso en el nivel de mínimos privilegios requeridos y solo tras haber comprobado que están justificados para el usuario. Si un miembro del equipo necesita acceso para resolver un problema urgente o implementar un cambio de configuración, debe solicitar acceso cuando es necesario. El acceso se revoca en cuanto se resuelve la situación.

Las solicitudes y aprobaciones de acceso se someten a seguimiento y se supervisan en un sistema independiente. Todo el acceso al sistema está correlacionado con estas aprobaciones. Si detectamos un acceso no aprobado, alertamos al equipo de operaciones para que lo investigue.

Usamos la autenticación en dos fases para todo el acceso remoto al sistema. Si se roba el nombre de usuario y la contraseña de uno de nuestros desarrolladores o miembro del personal de operaciones, los datos permanecen protegidos. Se deben realizar más comprobaciones de autenticación a través de una tarjeta inteligente o una llamada telefónica a un número previamente aprobado antes de permitir el acceso remoto al servicio.

Para administrar y mantener el servicio, Microsoft usa secretos como contraseñas RDP, certificados SSL y claves de cifrado. Estos secretos se administran, almacenan y transmiten de forma segura a través de Azure Portal. Para tener acceso a cualquiera de estos secretos, se requiere un permiso específico que se graba y registra de forma segura. Todos los secretos se rotan con una cadencia regular y podemos rotarlos a petición si hay un evento de seguridad.

El equipo de operaciones de Azure DevOps usa estaciones de trabajo de administrador protegidas para administrar el servicio. Estas máquinas ejecutan un número mínimo de aplicaciones y funcionan en un entorno segmentado lógicamente.

Los miembros del equipo de operaciones deben proporcionar credenciales específicas con autenticación en dos fases para acceder a las estaciones de trabajo. Todo el acceso se supervisa y se registra de forma segura. Para aislar el servicio contra manipulaciones externas, no permitimos

aplicaciones como Outlook y Office, ya que a menudo son blanco de ataques, como los de phishing de objetivo definido, entre otros.

Protección y respuesta de intrusiones

Ciframos los datos a través de HTTPS y SSL para ayudar a garantizar que no se intercepta ni modifica mientras está en tránsito entre usted y Azure DevOps. Los datos que almacenamos en su nombre en Azure DevOps se cifran de la siguiente manera:

- Los datos almacenados en bases de datos de Azure SQL se cifran a través del cifrado de datos transparent. Esta característica ayuda a proteger contra actividades maliciosas mediante el cifrado en tiempo real de la base de datos, las copias de seguridad asociadas y los archivos de registro de transacciones en reposo.
- Las conexiones de Azure Blob Storage se cifran para ayudar a proteger los datos en tránsito. Para los datos en reposo almacenados en Azure Blob Storage, Azure DevOps usa el cifrado en el lado del servicio.

El equipo de Azure DevOps usa la infraestructura de Azure para registrar y supervisar los aspectos esenciales del servicio. El registro y la supervisión ayudan a garantizar que las actividades que se realizan en el seno del servicio sean legítimas y a detectar infracciones o intentos de infracciones.

Todas las actividades de implementación y administración se registran de forma segura, al igual que los accesos de los operadores al almacenamiento de producción. La información de registro se analiza automáticamente y se generan alertas en tiempo real en caso de cualquier comportamiento potencialmente malintencionado o no autorizado.

Cuando el equipo de Azure DevOps identifica una posible vulnerabilidad de seguridad de alta prioridad o intrusiones, cuenta con un plan de respuesta claro. En este plan, se describe quiénes son los responsables, qué pasos deben darse para proteger los datos de los clientes y cuáles son las instrucciones sobre cómo interactuar con los expertos en seguridad de Microsoft. El equipo también notifica a los propietarios de la organización si los datos se han divulgado o dañado, para que puedan tomar los pasos adecuados para solucionar la situación.

Para ayudar a combatir las amenazas emergentes, Azure DevOps emplea una estrategia de *suposición de infracciones*. Un equipo altamente especializado de expertos en seguridad de Microsoft asume el rol de adversarios sofisticados. Este equipo prueba la detección y respuesta de infracciones para medir con precisión la preparación y los impactos de los ataques reales. Esta

estrategia refuerza la detección de amenazas, la respuesta y la defensa del servicio. También permite al equipo validar y mejorar la eficacia de todo el programa de seguridad.

Protección contra ataques por ransomware

Azure DevOps usa controles de Azure para ayudar a prevenir ataques por ransomware, detectarlos y responder ante ellos. Para obtener más información sobre cómo ayuda Azure a proteger a los clientes frente a ataques por ransomware, consulte Protección contra ransomware en Azure.

Privacidad de datos

Debe confiar en que controlamos sus datos de la forma adecuada y para usos legítimos. Una parte de esa garantía consiste en restringir cuidadosamente su uso.

Reglamento general de protección de datos

El Reglamento general de protección de datos (RGPD) es el mayor cambio en las leyes de protección de datos en Europa desde la introducción de la Directiva de protección de datos de la Unión Europea (UE) 95/46/CE. Para obtener más información sobre el RGPD, consulte la página de información general del Centro de confianza de Microsoft.

Soberanía y residencia de datos

Azure DevOps está disponible en las ocho ubicaciones geográficas siguientes en todo el mundo: Estados Unidos, Canadá, Europa, Reino Unido, India, Australia, Asia Pacífico y Brasil. De forma predeterminada, se asigna a cada organización la ubicación más próxima. Sin embargo, puede elegir otra ubicación al crear la organización. Si más adelante cambia de opinión, puede migrar la organización a otra ubicación con la ayuda del soporte técnico de Microsoft.

Azure DevOps no mueve ni replica los datos del cliente fuera de la ubicación elegida. En lugar de eso, los datos se replican geográficamente en una segunda región que pertenece a la misma ubicación. La única excepción es Brasil, cuyos datos se replican en la región Centro-sur de EE. UU. con fines de recuperación ante desastres.

> **Nota:**
>
> En el caso de compilaciones y versiones que se ejecutan en agentes macOS proporcionados por Microsoft, los datos se transfieren a un centro de datos de GitHub en los Estados

Unidos.

Para obtener más información, consulte Ubicaciones de datos para Azure DevOps.

Acceso al cumplimiento de la ley

En algunos casos, puede que terceros, por ejemplo, entidades encargadas de la aplicación de la ley, soliciten a Microsoft que les facilite acceso a datos de clientes almacenados en Azure DevOps. Intentamos redirigir las solicitudes al propietario de la organización para su resolución. Cuando una orden judicial obliga a Microsoft a revelar datos de clientes a terceros, Microsoft se esfuerza dentro de lo razonable para notificárselo al propietario de la organización por adelantado, a menos que la ley prohíba hacerlo.

Algunos clientes requieren que su almacenamiento de datos se encuentre en una ubicación geográfica determinada para pertenecer a una jurisdicción legal concreta en términos de actividades de aplicación de la ley. Todos los datos de clientes, como el código fuente, los elementos de trabajo, los resultados de las pruebas, así como los reflejos con redundancia geográfica y las copias de seguridad fuera del sitio, se conservan dentro de una de las ubicaciones mencionadas anteriormente.

Microsoft Access

En ocasiones, los empleados de Microsoft necesitan obtener acceso a los datos de clientes almacenados en Azure DevOps. Como medida de precaución, todos los empleados que tienen (o pueden tener) acceso a datos de clientes deben superar una verificación de antecedentes tanto profesionales como penales. Permitimos el acceso a los sistemas de producción únicamente cuando se produce un incidente en un sitio activo u otra actividad de mantenimiento aprobada, que se registra y supervisa.

Dado que no todos los datos de nuestro sistema se tratan de la misma manera, los clasificamos en estos tipos:

- **Datos del cliente**: los que carga en Azure DevOps.
- **Datos de la organización**: información que envía cuando se registra o administra su organización.
- **Datos de Microsoft**: información necesaria para el funcionamiento del servicio o recopilada a través de este.

En función de la clasificación, controlamos escenarios de uso, requisitos de ubicación geográfica, restricciones de acceso y requisitos de retención.

Uso promocional de Microsoft

En ocasiones, Microsoft desea ponerse en contacto con los clientes para informarles sobre más características y servicios que podrían resultarles útiles. Dado que no todos los clientes quieren ponerse en contacto con estas ofertas, puede optar por no recibir comunicaciones por correo electrónico de marketing.

Microsoft nunca usa datos de cliente para dirigirse a ofertas específicas para usuarios o organizaciones específicos. En su lugar, usamos datos de la organización y agregamos estadísticas de uso en el nivel de organización para determinar los grupos que deben recibir ofertas específicas.

Administración de directivas de privacidad para que los administradores controle la recopilación de comentarios de los usuarios

La función de conmutador de comentarios permite a los propietarios de la organización de Azure DevOps controlar si se pide a los usuarios que proporcionen comentarios y los envíen. Esta característica es esencial para garantizar que las prácticas de comentarios se alineen con las directivas de privacidad y gobernanza de su organización.

Creación de confianza

Puede confiar en los demás esfuerzos que Microsoft realiza en nombre de Azure DevOps. Estos esfuerzos incluyen directivas de adopción internas de Microsoft, el nivel de transparencia del estado de nuestro servicio y los avances para recibir la certificación de nuestros sistemas destinados a administrar la seguridad de la información.

Adopción interna

Los equipos de Microsoft están adoptando internamente Azure DevOps. El equipo de Azure DevOps se trasladó a una organización en 2014 y lo usa ampliamente. Hemos establecido directrices a fin de habilitar los planes de adopción para otros equipos.

Los equipos grandes avanzan de manera más gradual que los pequeños, a causa de sus inversiones en sistemas de DevOps existentes. Para los equipos que avanzan con rapidez, hemos

establecido un enfoque de clasificación de proyectos. Evalúa la tolerancia a riesgos, en función de las características del proyecto, para determinar si el proyecto es adecuado para Azure DevOps. En el caso de los equipos más grandes, la adopción suele producirse en fases, con más planeamiento.

Entre los requisitos adicionales que se requieren en los proyectos internos, se incluye la asociación de la organización con Microsoft Entra ID para garantizar el ciclo de vida de la identidad de los usuarios y la complejidad adecuada de las contraseñas. Los proyectos con mayor nivel de confidencialidad requieren además la autenticación en dos fases.

Certificaciones de cumplimiento

Es posible que le interese comprender cómo terceros realizan la evaluación de nuestros procedimientos para la seguridad de los datos. Azure DevOps ha obtenido las certificaciones siguientes:

- ISO 27001:2022
- ISO 27018:2019
- ISO 26262:2023
- Ley de transferencia y responsabilidad de seguros de salud (HIPAA)
- Contrato de socio comercial (BAA)
- Cláusulas modelo de la UE
- Controles de sistema y organización (SOC) 1, tipo 2
- SOC 2 tipo 2
- C5 de Alemania
- IRAP en Australia
- ENS de España

La auditoría de SOC para Azure DevOps abarca los controles de seguridad de datos, disponibilidad, integridad de procesamiento y confidencialidad. Los informes de SOC para Azure DevOps están disponibles a través del Portal de confianza de servicios de Microsoft.

El cuestionario Consensus Assessment Initiative Questionnaire (CAIQ) ayuda a las organizaciones a determinar y evaluar las prácticas y las prestaciones de seguridad de los proveedores de servicios en la nube. En consonancia con nuestro compromiso con la seguridad y la transparencia, recientemente hemos llevado a cabo la evaluación del CAIQ para Azure DevOps. Le invitamos a consultar el informe completo en el Portal de confianza de servicios de Microsoft.

Pasos que puede realizar

Para proteger correctamente los datos, es preciso que se impliquen tanto usted como sus administradores y usuarios. Sus datos de proyectos almacenados en Azure DevOps estarán protegidos en tanto en cuanto los puntos de acceso de usuario sean seguros. Coincida con el nivel de estrictoidad y granularidad de los permisos para esas organizaciones con el nivel de confidencialidad del proyecto.

Clasificación de los datos

El primer paso es clasificar los datos. Clasifique los datos en función de su confidencialidad y del horizonte de riesgo, así como de los perjuicios que sufriría en caso de que se viesen comprometidos. Muchas empresas ya poseen métodos de clasificación que reutilizan cuando migran los proyectos a Azure DevOps. Para obtener más información, puede descargar el documento sobre clasificación de datos a fin de prepararlos para la nube en Microsoft Trustworthy Computing.

Adopción de Microsoft Entra ID

Use Microsoft Entra ID para administrar el acceso de su organización a Azure DevOps. Microsoft Entra ID proporciona otra manera de mejorar la seguridad de las credenciales de los usuarios.

Microsoft Entra ID permite al departamento de TI administrar la directiva de acceso de los usuarios, la complejidad de las contraseñas, las actualizaciones de las contraseñas y la expiración cuando los usuarios abandonan la organización. A través de la federación de Active Directory, puede vincular directamente Microsoft Entra ID al directorio central de la organización, de modo que haya una sola ubicación en la que administrar estos datos para su empresa.

En la tabla siguiente, se comparan las características de Microsoft Entra y de las cuentas de Microsoft en términos de acceso a Azure DevOps:

Expandir tabla

Propiedad	Cuenta Microsoft	Microsoft Entra ID
Creador de identidades	Usuario	Organización
Nombre de usuario y contraseña únicos para todos los activos de trabajo	No	Sí
Control de la vigencia y de la complejidad de las contraseñas	Usuario	Organización

Propiedad	Cuenta Microsoft	Microsoft Entra ID
Límites de pertenencia a Azure DevOps	Cualquier cuenta de Microsoft	Directorio de la organización
Identidad rastreable	No	Sí
Quién posee la organización y la propiedad intelectual	Confuso	Organización
Inscripción de autenticación en dos fases	Usuario	Organización
Acceso condicional basado en dispositivos	No	Organización

Obtenga más información sobre cómo configurar esta compatibilidad con su organización.

Requerir la autenticación en dos fases.

Es posible que quiera restringir el acceso a su organización al requerir más de un factor para iniciar sesión. Puede exigir varios factores mediante Microsoft Entra ID. Por ejemplo, puede requerir autenticación telefónica, además de un nombre de usuario y una contraseña, para todas las solicitudes de autenticación.

Usar BitLocker

En el caso de proyectos confidenciales, puede usar BitLocker en su equipo portátil o de escritorio Windows. BitLocker cifra toda la unidad en la que residen Windows y los datos. Cuando BitLocker está habilitado, cifra automáticamente cualquier archivo que guarde en esa unidad. Si el equipo cae en las manos equivocadas, BitLocker impide el acceso no autorizado a copias locales de datos de los proyectos.

Limitar el uso de credenciales de autenticación alternativas

El mecanismo de autenticación predeterminado de las herramientas relacionadas con Git es la autenticación alternativa (a veces denominada *autenticación básica*). Este mecanismo permite que el usuario configure un nombre de usuario y una contraseña alternativos para usarlos durante las operaciones de línea de comandos con Git. El usuario puede utilizar esta combinación de nombre de usuario y contraseña para obtener acceso a cualquier otro dato para el que ese usuario tenga permisos. Por su naturaleza, las credenciales de autenticación alternativas son menos seguras que la autenticación federada predeterminada.

Todavía puede tomar decisiones para aumentar la seguridad. Toda la comunicación se envía a través de HTTPS y se aplican requisitos de complejidad de contraseñas. Su organización debe

seguir evaluando si necesita más directivas para cumplir los requisitos de seguridad de los proyectos.

Puede deshabilitar las credenciales de autenticación alternativas si decide que esto no cumple los requisitos de seguridad de su organización. Para más información, consulte Cambio de directivas de seguridad de conexión & de aplicaciones para su organización.

Protección del acceso a su organización

Los administradores pueden usar Microsoft Entra ID para controlar el acceso a los recursos y aplicaciones de Azure, como Azure DevOps. Con el control de acceso condicional instaurado, Microsoft Entra ID comprueba si se dan las condiciones específicas que ha establecido para que un usuario acceda a una aplicación. Una vez que el usuario cumple los requisitos de acceso, se lo autentica y se le permite obtener acceso a la aplicación.

Azure DevOps admite la aplicación de determinados tipos de directivas de acceso condicional (por ejemplo, barreras IP) para mecanismos de autenticación personalizados de Azure DevOps. Estos mecanismos incluyen tokens de acceso personal, autenticación alternativa, OAuth y claves de Secure Shell (SSH). Si los usuarios obtienen acceso a Azure DevOps a través de un cliente externo, solo se respetan las directivas basadas en IPv4.

Más recursos

- Página principal de Azure DevOps
- Ubicaciones de los datos para Azure DevOps
- Declaración de privacidad de Microsoft
- Compatibilidad de Azure DevOps
- Características y servicios incluidos en Azure DevOps
- Centro de confianza de Azure
- Ciclo de vida de desarrollo de seguridad de Microsoft

Last updated on 18/03/2026

Confiabilidad en registros de Azure Monitor

Azure Monitor Logs es una plataforma centralizada de software como servicio (SaaS) para recopilar, analizar y tomar medidas en función de los datos generados por el sistema procedentes de recursos y aplicaciones de Azure y no Azure.

Al usar Azure, relibilidad es una responsabilidad compartida. Microsoft proporciona una variedad de capacidades para admitir resiliencia y recuperación. Es responsable de comprender cómo funcionan esas funcionalidades dentro de todos los servicios que usa y de seleccionar las funcionalidades que necesita para cumplir los objetivos empresariales y los objetivos de tiempo de actividad.

Azure Monitor Registros proporciona características de resistencia integradas que protegen los datos y minimizan las interrupciones. En este artículo se describe cómo puede usar esas características para que las áreas de trabajo de Log Analytics sean resistentes a varias posibles interrupciones y problemas, incluidos errores transitorios, interrupciones de zona de disponibilidad y interrupciones de regiones.

Recomendaciones de implementación de producción

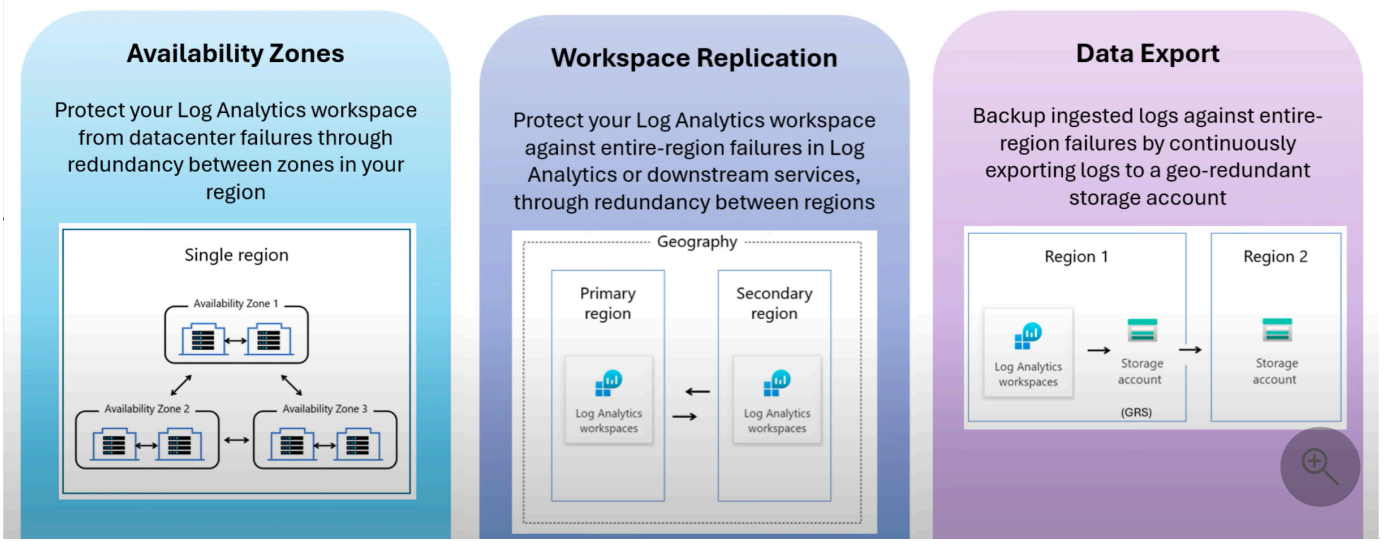
Azure Monitor Logs ofrece varias características que puede usar individualmente o en combinación para ayudar a mejorar la resistencia del espacio de trabajo frente a diversos tipos de problemas.

Expandir tabla

Capacidad	Descripción	Costo
Zonas de disponibilidad	Use redundancia entre zonas de su región para proteger el área de trabajo de Log Analytics frente a errores del centro de datos.	Sin cargo adicional. Incluido en el precio estándar del espacio de trabajo. Los clústeres dedicados deben cumplir los requisitos mínimos del nivel de compromiso.
Replicación del área de trabajo	Use la redundancia entre regiones para proteger el área de trabajo de Log Analytics contra fallos de toda una región en Log Analytics o en servicios dependientes.	Todos los datos ingeridos se replican, lo que agrega cargos de replicación. Para reducir los costos, elija las reglas de recopilación de datos (DCR) que participan en la replicación.

Capacidad	Descripción	Costo
Exportación de datos	Realice una copia de seguridad de los registros ingeridos y expórtelos continuamente al almacenamiento con redundancia geográfica para protegerse frente a errores de toda la región.	Los costos de almacenamiento varían según el volumen y el nivel de redundancia (GRS). No se aplica ningún cargo adicional de Azure Monitor por la exportación en sí.

Para obtener información detallada sobre los precios, consulte Azure Monitor precios.



Algunas características de Log Analytics no son compatibles con todas las características de confiabilidad. Por ejemplo, la replicación del área de trabajo no admite tablas auxiliares. Para más información sobre la compatibilidad de características, consulte Reliability best practices for Log Analytics.

Introducción a la arquitectura de confiabilidad

En esta sección se describen algunos de los aspectos importantes de cómo funciona el servicio que es más relevante desde una perspectiva de confiabilidad. En la sección se presenta la arquitectura lógica, que incluye algunos de los recursos y características que se implementan y usan. También se describe la arquitectura física, que proporciona detalles sobre cómo funciona el servicio en segundo plano.

Arquitectura lógica

Cuando se usan registros de Azure Monitor, se implementa un *área de trabajo de Log Analytics*, que es un almacén de datos que puede recopilar cualquier tipo de datos de registro.

Log Analytics áreas de trabajo están asociadas a un *cluster*, que proporciona los recursos computacionales para el área de trabajo y otras funcionalidades. La mayoría de las áreas de

trabajo se ejecutan en un *clúster compartido*, que varios clientes usan al mismo tiempo. Microsoft implementa y administra clústeres compartidos. Opcionalmente, puede implementar un clúster dedicado, que proporciona recursos dedicados para el área de trabajo. Los clústeres compartidos y dedicados tienen diferentes características de confiabilidad. Las diferencias se describen en este artículo.

Azure Monitor Registros tiene dos rutas de acceso de datos distintas, cada una con sus propias características de confiabilidad:

- **Ingestion** es la ruta de acceso a través de la cual fluyen los datos de registro a un área de trabajo de Log Analytics. El servicio Azure Monitor confirma la recepción de datos solo después de escribirlos en el almacenamiento de forma duradera. La ruta de acceso de ingestión incluye varios componentes:
  - *Los orígenes de datos* generan telemetría. Algunas fuentes, como los registros de plataforma de Azure, envían datos directamente al servicio Azure Monitor a través de *la configuración de diagnóstico*. Otros orígenes requieren que un *agente* recopile y reenvíe datos. Las aplicaciones también pueden usar la API de ingesta de Logs para enviar registros directamente a Azure Monitor.
  - *Las reglas de recopilación de datos (DCR)* definen qué datos se van a recopilar, cómo transformarlos y dónde enviarlos. Las DCR enrutan los datos al servicio Azure Monitor a través de un punto de conexión de recopilación de datos (DCE).
- **La consulta** es la ruta de acceso a través de la que se recuperan y analizan los datos almacenados en el área de trabajo. Las consultas de registro usan el lenguaje de consulta kusto (KQL) y se ejecutan en los datos almacenados del área de trabajo.

La ingesta y la consulta se controlan como operaciones de servicio distintas, por lo que una interrupción de una no afecta necesariamente a la otra. Por ejemplo, durante una degradación del proceso de ingesta, es posible que todavía pueda consultar los datos ingeridos previamente. Del mismo modo, un problema del lado de la consulta no impide que se ingieren y almacenan nuevos datos.

Para obtener más información sobre los componentes principales de los registros de Azure Monitor, consulte Azure Monitor Logs overview.

Arquitectura física

Internamente, los registros de Azure Monitor usan varias *réplicas* de los componentes computacionales y de almacenamiento de los datos de registro. Microsoft administra estas réplicas y no es necesario configurarlas ni administrarlas.

Es responsable de implementar y administrar agentes, así como de la confiabilidad de los recursos de computación en los que se ejecutan.

Resistencia a errores transitorios

Los errores transitorios son errores breves e intermitentes en los componentes. Se producen con frecuencia en un entorno distribuido como la nube y son una parte normal de las operaciones. Los errores transitorios se corrigen después de un breve período de tiempo. Es importante que las aplicaciones puedan controlar errores transitorios, normalmente mediante el reintento de solicitudes afectadas.

Todas las aplicaciones hospedadas en la nube deben seguir las instrucciones de control de errores transitorios Azure cuando se comunican con cualquier API, bases de datos y otros componentes hospedados en la nube. Para obtener más información, consulte Recomendaciones para controlar errores transitorios.

Tenga en cuenta los siguientes tipos de errores transitorios en los registros de Azure Monitor:

- **Fallos transitorios en el servicio Azure Monitor:** El servicio Azure Monitor verifica que cada entrada de registro se ingesta correctamente en el área de trabajo antes de eliminarla del proceso de ingesta. Si la ingesta no está disponible o limita las solicitudes, la configuración de diagnóstico y los agentes de Azure Monitor almacenan temporalmente los datos en un búfer local y vuelven a intentar enviarlos durante varias horas, con retroceso exponencial para evitar sobrecargar el servicio. Por el contrario, una aplicación personalizada que envía solicitudes de ingesta o consultas debe implementar su propia lógica de reintento.

- **Fallos transitorios que afectan a la conectividad de ingestión:** Los agentes y la configuración de diagnóstico de Azure almacenan los datos en búfer localmente y reintentan la entrega cuando el punto de conexión de ingestión no está disponible temporalmente, lo que hace que la ruta de acceso de ingestión sea más resistente a los fallos transitorios.

- **Errores transitorios que afectan al registro de aplicaciones personalizado:** Para validar que la aplicación personalizada controla correctamente los errores transitorios, supervise estas métricas:
  - Latencia de ingesta

    - Recuento de ingestión fallida
    - Errores de exportación
    - Tasa de errores de consulta

    La latencia de ingesta sostenida que dura más de cinco minutos podría indicar un problema más significativo y notransciente. Revise las instrucciones de las otras secciones de este documento para comprender cómo ser resistentes a otros tipos de errores.

- **Errores transitorios durante las consultas y otras operaciones:** Si se producen errores transitorios durante las consultas o al realizar otras operaciones, los clientes son responsables de reintentar.

Resistencia a errores de zona de disponibilidad

Availability zones son grupos de centros de datos físicamente independientes dentro de una región de Azure. Cuando una zona falla, los servicios pueden transferirse a una de las zonas restantes.

Azure Monitor Registros ofrece dos tipos de compatibilidad con zonas de disponibilidad, en función de la región y el tipo de clúster del área de trabajo:

- **La resistencia de los datos** proporciona redundancia de zona para los datos de registro mediante la replicación en varias zonas de disponibilidad.

  La resistencia de los datos está disponible en todas las regiones que admiten zonas de disponibilidad. La mayoría de las regiones que admiten zonas de disponibilidad requieren que el área de trabajo se implemente en un clúster dedicado. Sin embargo, algunas regiones lo admiten mediante la configuración predeterminada del área de trabajo de un clúster compartido. Al pasar a un clúster dedicado en una región que admite zonas de disponibilidad, se protegen los datos ingeridos después del traslado, no los datos históricos.

- **La resistencia del servicio** proporciona redundancia de zona para la ingesta y la continuidad de las consultas durante una interrupción de la zona de disponibilidad.

  Solo algunas regiones ofrecen resiliencia del servicio.

Si un incidente afecta a una zona, Microsoft administra la conmutación por error a otra zona de disponibilidad de la región automáticamente. No es necesario realizar ninguna acción, ya que el cambio de zona se realiza de forma sencilla.

Requisitos

- **Compatibilidad con regiones:** Consulte las regiones admitidas de la zona de disponibilidad para obtener la lista de regiones que admiten la resistencia de datos, la resistencia del servicio y la resistencia de datos en clústeres compartidos y dedicados.
- **Clúster dedicado:** Algunas regiones requieren un clúster dedicado para la resistencia de los datos.

Consideraciones

- **Ingesta de Event Hubs:** En algunas regiones, la ingesta de Event Hubs en el espacio de trabajo no es resiliente a una interrupción de zona. Para obtener una lista de estas regiones, consulte las notas al pie en la lista de regiones admitidas. Evalúe vías de ingesta alternativas para los datos críticos.
- **Traslado a clústeres dedicados:** Los clústeres dedicados solo protegen los nuevos datos. Cuando un área de trabajo se mueve a un clúster dedicado, los datos ingeridos previamente permanecen en el clúster compartido. Todo está disponible en operaciones normales. Sin embargo, durante una interrupción de zona, es posible que solo pueda acceder a los nuevos datos del clúster dedicado.

Costo

La redundancia de zona no afecta a los precios. Las áreas de trabajo en clústeres compartidos se cobran según el precio estándar del área de trabajo y las áreas de trabajo de los clústeres dedicados deben cumplir el nivel de compromiso del clúster. Ninguno de los dos modelos de precios cambia al habilitar la redundancia de zona. Para obtener más información sobre los precios, consulte los precios de Azure Monitor .

Configurar soporte de zonas de disponibilidad

En el caso de las áreas de trabajo que usan clústeres compartidos, al crear un área de trabajo en una región que admita zonas de disponibilidad para registros de Azure Monitor, el servicio habilita automáticamente la redundancia de zona.

Si la región requiere un clúster dedicado para la compatibilidad con zonas, implemente el clúster dedicado y, a continuación, vincule el área de trabajo al clúster dedicado.

Planeamiento y administración de capacidad

Durante una falla de zona, el área de trabajo podría experimentar una mayor carga debido a los reintentos de ingestión y consulta. Si la solución es sensible a los retrasos en la ingesta de registros o las consultas, incluso durante y después de un error de zona, considere la posibilidad de sobreaprovisionar la capacidad de los límites diarios que configure en el área de trabajo. En el caso de los clústeres dedicados, seleccione un nivel de compromiso que tenga espacio para los picos de ingesta temporales. Para más información, consulte Administración de la capacidad mediante el sobreaprovisionamiento.

Comportamiento cuando todas las zonas están en buen estado

En esta sección se describe lo que cabe esperar cuando un área de trabajo de Log Analytics tiene redundancia de zona y todas las zonas están operativas.

- **Operación entre zonas:** Los clústeres que tienen resiliencia de servicio permiten que la ingesta y las consultas usen la infraestructura en cualquier zona. El servicio distribuye automáticamente el trabajo entre zonas, optimizando la localidad y la carga.
- **Replicación de datos entre zonas:** Para los clústeres que tienen resistencia de datos, el servicio Azure Monitor confirma todas las escrituras en varias réplicas en zonas independientes antes de confirmar.

Comportamiento durante un fallo de zona

En esta sección se describe qué cabe esperar cuando un área de trabajo de Log Analytics tiene redundancia de zona y se produce una interrupción en una de las zonas.

- **Detección y respuesta:** La plataforma detecta un error de zona y responde automáticamente. En el caso de los clústeres que tienen resistencia de servicio, la plataforma reasigna la ingesta y el proceso de consulta a zonas supervivientes. En el caso de los clústeres que tienen resiliencia de datos, la plataforma cambia para usar réplicas en zonas saludables. No es necesario iniciar una conmutación por error de zona.
- **Notification:** Microsoft no le notifica automáticamente cuando una zona está inactiva. Sin embargo, puede usar Azure Service Health para comprender el estado general del servicio, incluidos los errores de zona, y puede configurar alertas de Service Health para notificarle problemas.

  También puede supervisar las métricas de salud del área de trabajo y buscar una latencia de ingesta sostenida o una tasa elevada de fallos en las consultas. Puede configurar alertas en estas métricas.

- **Solicitudes activas:** Durante una interrupción de zona, es posible que se produzca un error en las solicitudes activas.

  Es posible que se finalicen las consultas activas. Los clientes pueden reintentar las consultas.

  En los clústeres que no tienen resistencia del servicio, es posible que Azure Monitor deje de procesar los datos no confirmados. Los agentes o las aplicaciones cliente pueden volver a intentarlo cuando la zona está en buen estado.

- **Pérdida de datos esperada:** En el caso de los clústeres que tienen resistencia a los datos, no se espera ninguna pérdida de datos para los datos totalmente confirmados. Las aplicaciones cliente o agente deben reintentar lotes no reconocidos.

- **Tiempo de inactividad esperado:** En el caso de los clústeres que tienen resistencia de servicio, no se espera ningún tiempo de inactividad. Es posible que se produzcan incrementos temporales de latencia o restricciones mientras el servicio reequilibra la infraestructura a través de zonas.

- **Rerouting:** Para los clústeres que tienen resiliencia del servicio, los equilibradores de carga internos dentro del servicio Azure Monitor Logs redirigen el tráfico a nodos de zona saludables. No se requiere ningún cambio de punto de conexión.

Recuperación de zona

Cuando se recupera una zona de disponibilidad, Azure Monitor Logs reintegra automáticamente la zona en la topología de servicio activa. La zona recuperada comienza a procesar las solicitudes de ingesta y consulta en paralelo con las demás zonas. Los datos que se replican en zonas supervivientes durante la interrupción permanecen intactos y la replicación sincrónica normal se reanuda en todas las zonas. No es necesario tomar medidas para la recuperación y la reintegración de zonas.

Prueba de fallos de zona

Azure administra el enrutamiento del tráfico, la conmutación por error y la recuperación de zona en caso de fallos, por lo que no es necesario validar los procesos de fallo de la zona de disponibilidad ni proporcionar más datos.

Resistencia a errores en toda la región

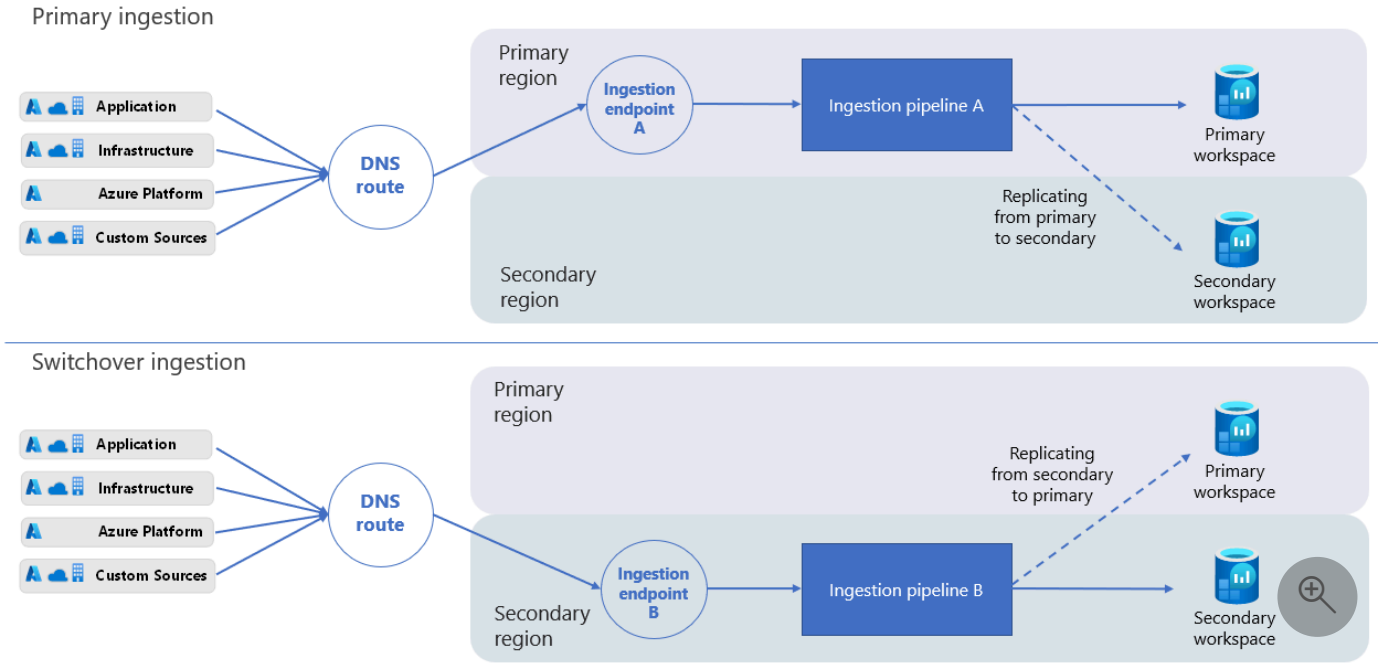
Azure Monitor Registros admite escenarios de varias regiones a través de la replicación del área de trabajo.

Replicación del área de trabajo

La replicación del área de trabajo crea un área de trabajo secundaria en una región secundaria admitida y, a continuación, replica de forma asincrónica los nuevos registros, así como los cambios de esquema y configuración. Los datos que existían antes de configurar la replicación no se rellenan.

No ve el área de trabajo secundaria como un recurso independiente y administrable. Puede acceder a él a través del mismo punto de conexión de recursos del área de trabajo.

Si se produce un error en la región primaria, debe desencadenar una conmutación por medio de una API. La conmutación redirige la ingesta y las consultas al Área de trabajo secundaria. No hay conmutación por error automática. La conmutación es una decisión manual. Debe iniciarse una conmutación de retorno por parte del cliente para restaurar el sistema principal.



En esta sección se resumen aspectos importantes de la replicación del área de trabajo. Revise la documentación completa para comprender exactamente cómo funciona. Para obtener más información, consulte Mejore la resiliencia mediante la replicación del espacio de trabajo de Log Analytics entre regiones.

> ⓘ **Nota:**

La replicación de registros del área de trabajo de Azure Monitor utiliza el término *switchover* porque representa mejor el proceso de conmutación manual al área de trabajo secundaria. También podría ver el término *conmutación por error* utilizado para describir el proceso general.

Requisitos

- **Compatibilidad con regiones:** La replicación del área de trabajo requiere seleccionar una región secundaria dentro del mismo *grupo de regiones*. Algunas combinaciones de regiones no pueden vincularse directamente entre ellas. Para ver los pares primarios y secundarios permitidos y los grupos de regiones, consulte Regiones admitidas. Si la región del área de trabajo no aparece en la lista, no admite actualmente la replicación.

- **Clústeres dedicados:** Si el área de trabajo está vinculada a un clúster dedicado, habilite primero la replicación en el clúster.

Consideraciones

- **Reglas de alertas de búsqueda de registros:** Después de cambiar entre regiones, las reglas de alertas de búsqueda de registros siguen funcionando a menos que el servicio de alertas de la región activa no funcione correctamente o las reglas de alerta no estén disponibles. Esta condición puede ocurrir, por ejemplo, si la región en la que se crearon las reglas de alerta está completamente inactiva. La replicación de reglas de alerta entre regiones no se realiza automáticamente como parte de la replicación del área de trabajo, pero puede replicarlas. Por ejemplo, puede exportar reglas de alertas de la región primaria a la región secundaria.

- **Cambio manual y cambio de vuelta:** Usted es responsable de decidir cuándo cambiar y volver atrás, y de desencadenar las acciones de cambio y cambio de vuelta.

  Para prepararse para un fallo de región, debe:
  - Establezca la monitorización (estado de replicación, latencia de ingesta, métricas de salud del área de trabajo) para la región activa. Audite la región del área de trabajo inactiva antes de una conmutación.
  - Considere si se deben crear runbooks para definir criterios de conmutación, que podrían formar parte del planeamiento de la recuperación ante desastres. Por ejemplo, podría supervisar la latencia de ingesta de toda la región, los incumplimientos del objetivo de nivel de servicio (SLO), las tasas sostenidas de errores en las consultas o los avisos de Azure Service Health. Pruebe los runbooks (guías de ejecución) con regularidad.

- Documente los criterios de conmutación y retorno.
- Compruebe que los usuarios o entidades de servicio que desencadenen la conmutación tienen los permisos de rol necesarios.

- **Tablas auxiliares:** Las tablas auxiliares no se replican. Como procedimiento recomendado, no habilite la replicación en áreas de trabajo que incluyan tablas auxiliares.

- **Microsoft Sentinel:** Las listas de seguimiento y las entradas de inteligencia sobre amenazas de Microsoft Sentinel solo se replican cuando se actualizan, por lo que pueden tardar hasta 12 días en lograr una paridad completa después de la activación inicial.

- **Operaciones de purga:** Las operaciones de purga quitan registros de las áreas de trabajo principales y secundarias. Si alguna de las áreas de trabajo no está disponible, se produce un error en la purga. En este caso, debe planificar los plazos de cumplimiento en consecuencia.

Para más consideraciones, consulte Consideraciones de implementación.

Costo

Al habilitar la replicación del área de trabajo, paga por la replicación de todos los datos que ingiere en el área de trabajo. Las DCR asociadas a la replicación determinan directamente los costos de replicación. Para obtener más información sobre los precios, consulte los precios de Azure Monitor.

Configuración de la compatibilidad con varias regiones

- **Habilitar la replicación del área de trabajo:** Los pasos generales necesarios para habilitar la replicación del área de trabajo son:

  1. Habilite la replicación y especifique la región secundaria.

  2. Valide el estado de aprovisionamiento.

> **Sugerencia**
>
> Incluso cuando el estado de aprovisionamiento es *Correcto*, los esquemas de tabla podrían seguir replicándose. Para comprobar cuándo se replican completamente los esquemas de tabla, **audite el área de trabajo secundaria**.

3. Asocie los DCR con el punto de conexión de recopilación de datos del área de trabajo (DCE) después de habilitar la replicación de esos flujos durante la conmutación.

- **Deshabilitar la replicación del área de trabajo:** Para obtener pasos detallados para deshabilitar la replicación del área de trabajo, consulte Habilitación y deshabilitación de la replicación del área de trabajo.

  Al deshabilitar la replicación, se impide que se repliquen nuevos registros, pero no se quitan las copias replicadas hasta que la característica esté totalmente deshabilitada. Para supervisar el progreso de la operación de deshabilitación, consulte Comprobar el estado de aprovisionamiento del espacio de trabajo.

Planeamiento y administración de capacidad

Durante un fallo en la región u otro evento de conmutación de ida o de vuelta, tu área de trabajo podría experimentar una mayor carga debido a la ingesta y los reintentos de consulta. Si su solución es sensible a los retrasos en la ingesta de registros o en las consultas, incluso durante y después de un fallo de una región, considere sobreaprovisionar la capacidad asociada a cualquier límite diario de su área de trabajo. En el caso de los clústeres dedicados, seleccione un nivel de compromiso con margen para los picos de ingesta temporales. Para más información, consulte Administración de la capacidad mediante el sobreaprovisionamiento.

Comportamiento cuando todas las regiones están en buen estado

En esta sección se describe qué esperar al configurar un área de trabajo de Log Analytics para la replicación del área de trabajo y todas las regiones están operativas.

- **Operación entre regiones:** Todas las consultas y ingestas tienen como destino el área de trabajo de la región primaria mientras está en buen estado. El espacio de trabajo de la región secundaria contiene una réplica pasiva de datos, esquemas y configuración hasta la conmutación.

- **Replicación de datos entre zonas:** Los nuevos registros, esquemas y actualizaciones de configuración se replican de forma asincrónica en lotes en la base de datos secundaria. Dado que es asincrónico, este proceso de replicación no afecta a la latencia de ingesta. Normalmente, los datos se replican en unos dos minutos, pero este período de tiempo no está garantizado.

La replicación del área de trabajo se aplica en el nivel de área de trabajo, por lo que no puede seleccionar tablas individuales para replicar. Sin embargo, puedes controlar la replicación

asociando solo las DCR que contienen flujos de datos críticos con el punto de conexión de recopilación de datos del área de trabajo. Las DCRs que no están asociadas con el DCE del entorno de trabajo no replican sus datos. Para obtener más información, consulte Asocia reglas de recopilación de datos.

Comportamiento durante una falla de región

En esta sección se describe qué esperar al configurar un área de trabajo de Log Analytics para la replicación del área de trabajo y cuando ocurre una interrupción en una de las regiones.

- **Detección y respuesta:** Es responsable de decidir cuándo cambiar el área de trabajo secundaria para convertirse en la nueva área de trabajo principal. Microsoft no toma esta decisión ni inicia el proceso, incluso si hay una interrupción en la región. La alternancia y la reversión son acciones manuales.

  Para obtener información sobre cómo decidir cuándo cambiar, vea ¿Cuándo debo cambiar?. Para obtener información sobre cómo supervisar el estado del área de trabajo durante una posible conmutación, consulte Supervisión del rendimiento del área de trabajo mediante consultas.

  Para obtener más información sobre cómo promover un área de trabajo secundaria a la nueva área de trabajo principal, consulte Cambio a su área de trabajo secundaria.

- **Notifications:** Microsoft no le notifica automáticamente cuando una región está inactiva. Sin embargo, puede usar Azure Service Health para comprender el estado general del servicio, incluidos los errores de región, y puede configurar alertas de Service Health para notificarle problemas.

  Para validar la actualización del área de trabajo secundaria, puede supervisar el estado de replicación y también auditar el área de trabajo secundaria. Para obtener información sobre cómo supervisar el estado del área de trabajo al tomar la decisión de cambiar, consulte Supervisión del rendimiento del área de trabajo mediante consultas.

- **Solicitudes activas:** Es posible que se produzca un error en cualquier ingesta activa en la región con errores y también se produzca un error en las consultas en curso en la región con errores.

  Una vez que el proceso de conmutación cambia las entradas DNS del área de trabajo, el área de trabajo secundaria está disponible para la ingesta y la consulta. Los agentes de Microsoft vuelven a intentar automáticamente los intentos fallidos de ingesta de registros. Las aplicaciones que usan la API de ingesta de registros también deben reintentar.

- **Pérdida de datos esperada:** Los registros u otros cambios que aún no se han replicado no están disponibles en la región secundaria. Normalmente, los datos se replican en unos dos minutos, pero este período de tiempo no está garantizado.

  Después de cambiar a la región secundaria, si la región primaria no puede procesar los datos de registro entrantes, Azure Monitor almacena en búfer los datos de la región secundaria durante un máximo de 11 días. Durante los primeros cuatro días, Azure Monitor vuelve a intentar automáticamente replicar los datos periódicamente.

  La cantidad total de pérdida de datos a veces se denomina objetivo de punto de recuperación (RPO).

- **Tiempo de inactividad esperado:** El proceso de conmutación implica activar la ingesta en el área de trabajo secundaria y actualizar los registros DNS del área de trabajo.

  Aunque el cambio de registro DNS se produce rápidamente, la propagación de DNS puede tardar más tiempo. Si los agentes u otros clientes no respetan el período de vida de DNS (TTL), es posible que sigan intentando la ingesta en el área de trabajo principal. Asegúrese de que los clientes no almacenan en caché las entradas DNS durante más tiempo que su TTL.

  Las consultas se reanudan en el área de trabajo secundaria una vez completada la conmutación.

  La cantidad de tiempo que pasa antes de que el área de trabajo secundaria esté disponible a veces se denomina objetivo de tiempo de recuperación (RTO).

- **Redistribución:** El proceso de conmutación actualiza los registros DNS de los puntos de conexión de ingesta para que apunten a la región secundaria. La plataforma enruta las consultas internamente a la región activa.

  Mientras el espacio de trabajo se encuentra en estado de conmutación, los registros se replican de vuelta a la región primaria mediante replicación asincrónica.

Recuperación de regiones

Switchback es manual. Usted es responsable de decidir cuándo volver. Después de que el servidor principal esté estable y sincronizado, puede iniciar una conmutación de retorno. Compruebe que no queda ninguna cola de registros en el secundario y que la replicación se reanude correctamente desde el principal restaurado.

Para obtener información sobre cómo decidir cuándo volver a cambiar, vea ¿Cuándo debo volver? Para ver los pasos detallados de la conmutación, consulte Cambio al área de trabajo principal.

Prueba de fallos de región

Puede desencadenar una conmutación y una reinversión en cualquier momento, incluidas pruebas o simulacros de recuperación ante desastres.

Si realiza simulacros, se recomienda seguir estos procedimientos recomendados:

- ✓ Use un entorno que no sea de producción o, si prueba en producción, ejecute el proceso en un período de tiempo de bajo riesgo.
- ✓ Si es posible, simule los criterios que desencadenan los criterios de conmutación por error de la región en función de sus propias directivas. Este enfoque le permite probar la detección y automatización, así como el proceso de conmutación y conmutación por recuperación.
- ✓ Mida el RTO (finalización de conmutación) y el RPO (intervalo máximo no replicado). Use conjuntos de datos exportados para validar la paridad de registros de ejemplo entre las áreas de trabajo principal y secundaria.

Soluciones personalizadas de varias regiones para la resistencia

Si la replicación del área de trabajo no está disponible para su región o si debe usar un tipo de tabla que la replicación del área de trabajo no admite, tenga en cuenta estos enfoques alternativos para la resistencia de varias regiones:

- **Escritura dual:** Configura orígenes, como la configuración de diagnóstico, los agentes y la ingestión basada en DCR, para enviar logs a dos áreas de trabajo independientes en regiones diferentes.

  Este enfoque proporciona análisis casi en tiempo real en ambas regiones. Sin embargo, duplica el costo de ingesta e introduce el riesgo de desfase de configuración entre las áreas de trabajo.

- **Exportar y rehidratar:** Si la región está emparejada con otra región de Azure, puede usar data export para exportar continuamente los registros a Azure Blob Storage. Configure la cuenta de almacenamiento para usar uno de los tipos de almacenamiento con redundancia geográfica (GRS). Azure Storage replica automáticamente y de forma asincrónica los registros exportados en la región emparejada.

Durante un desastre, implemente una nueva área de trabajo y ingiera de forma selectiva datos críticos recientes. También puede consultar manualmente el historial a largo plazo directamente desde Blob Storage mediante Azure Data Explorer. Para obtener más información, consulte Consulta de datos exportados.

Copias de seguridad y restauración

Azure Monitor Registros no proporciona una funcionalidad tradicional de copia de seguridad o restauración para un punto específico en el tiempo. En su lugar, la durabilidad y la recuperación dependen de la replicación intraregión (incluida la redundancia de zona), la replicación del área de trabajo opcional y la exportación de datos para mantener una copia externa. Para una recuperación rápida de análisis después de un error, confíe en la replicación de la plataforma en lugar de en las operaciones de restauración.

Puede habilitar data export, que exporta automáticamente copias de los registros a Azure Storage. Si la región de Azure está emparejada, puede considerar la posibilidad de habilitar el almacenamiento con redundancia geográfica (GRS) para replicar los datos de registro en la región emparejada.

Resistencia a eliminaciones accidentales

Para cumplir los estrictos requisitos de cumplimiento y la protección contra alteraciones, considere la posibilidad de usar directivas de inmutabilidad para evitar la eliminación de datos de registro de la cuenta de almacenamiento.

Resistencia al mantenimiento del servicio

Microsoft aplica periódicamente actualizaciones de servicio y realiza otro mantenimiento. La plataforma Azure controla estas actividades automáticamente, lo que garantiza que el mantenimiento sea transparente y sin problemas. No se espera ningún tiempo de inactividad durante los eventos de mantenimiento, a menos que se le haya notificado a través de Azure Service Health de mantenimiento programado.

Para minimizar el efecto del mantenimiento, Azure Monitor realiza el mantenimiento gradual en las réplicas del área de trabajo y las zonas de disponibilidad.

Si tu solución es sensible a la latencia de ingesta y consulta, supervisa la latencia de ingesta y las métricas de salud del área de trabajo tanto durante como después de los eventos de mantenimiento programados.

Algunos cambios de configuración de ejecución prolongada pueden consumir muchos recursos, como habilitar la replicación, vincular clústeres y agregar esquemas grandes. Si necesita realizar estos cambios, evite programarlos cuando también se programe el mantenimiento de la plataforma. Utilice Azure Service Health para el mantenimiento programado y comprenda cuándo se programa el mantenimiento.

Acuerdo de nivel de servicio

El acuerdo de nivel de servicio (SLA) para Azure servicios describe la disponibilidad esperada de cada servicio y las condiciones que la solución debe cumplir para lograr esa expectativa de disponibilidad. Para obtener más información, vea SLAs for servicios en línea.

Registros de Azure Monitor es único entre los servicios en línea porque es la plataforma recomendada para recopilar los datos esenciales y la información necesaria para presentar una reclamación al soporte técnico de Microsoft por un problema relacionado con el SLA.

Dos acuerdos de nivel de servicio cubren los registros de Azure Monitor:

- El **SLA de disponibilidad de consultas de Log Analytics** define las expectativas de disponibilidad para consultar los datos de un área de trabajo de Log Analytics.
- El Acuerdo de Nivel de **Servicio de Azure Monitor** define las expectativas de disponibilidad para las reglas de alertas y los grupos de acciones, incluidas las expectativas para analizar las señales de telemetría y para la entrega de notificaciones.

Contenido relacionado

- Los procedimientos recomendados de fiabilidad para registros de Azure Monitor
- Zonas de disponibilidad en Azure Monitor
- Mejore la resiliencia mediante la replicación del espacio de trabajo de Log Analytics
- Exportación de datos del área de trabajo de Log Analytics
- Reglas de recopilación de datos
- Creación y administración de un clúster dedicado

ⓘ**Nota:** El autor ha creado este artículo con ayuda de inteligencia artificial. Más información

---

Last updated on 24/06/2026

Confiabilidad en Azure API Management

Azure API Management es un servicio totalmente administrado que ayuda a las organizaciones a publicar, proteger, transformar, mantener y supervisar las API. Como servicio de Azure, API Management proporciona una variedad de funcionalidades para satisfacer los requisitos de confiabilidad.

Cuando se usa Azure, la confiabilidad es una responsabilidad compartida. Microsoft proporciona una variedad de funcionalidades para admitir resistencia y recuperación. Es responsable de comprender cómo funcionan esas funcionalidades dentro de todos los servicios que usa y de seleccionar las funcionalidades que necesita para cumplir los objetivos empresariales y los objetivos de tiempo de actividad.

En este artículo se describe cómo hacer que API Management sea resistente a una variedad de posibles interrupciones y problemas, incluidos errores transitorios, interrupciones de zona de disponibilidad, interrupciones de regiones y mantenimiento del servicio. También se describe cómo puede usar copias de seguridad para recuperarse de otros tipos de problemas y se resalta cierta información clave sobre el acuerdo de nivel de servicio (SLA) de API Management.

Introducción a la arquitectura de confiabilidad

API Management usa una arquitectura basada en unidades de escalado para proporcionar redundancia y escalabilidad integradas. Al implementar una instancia de API Management, se configuran una o varias *unidades de escalado* o *unidades*. Cada unidad es una representación lógica de la capacidad que contiene los recursos de proceso necesarios para controlar las solicitudes de API.

Unidades

Cada unidad consta de dos recursos de proceso (máquinas virtuales o servidores similares, en función del nivel de servicio) que controlan las solicitudes de API juntas. Usted no ve estas máquinas virtuales ni otros servidores. La plataforma administra automáticamente su creación y monitorización del estado. Si se produce un error en un recurso de proceso, la unidad continúa funcionando pero con una capacidad reducida, lo que proporciona cierta protección de confiabilidad integrada.

Al configurar una instancia con dos o más unidades, las unidades disponibles funcionan juntas para procesar solicitudes y proporcionar equilibrio de carga automático. Si una de las unidades deja de estar disponible, las unidades restantes siguen controlando el tráfico, pero con capacidad potencialmente reducida.

Para obtener mayores niveles de confiabilidad, API Management admite la distribución de unidades entre zonas de disponibilidad dentro de una región y entre varias regiones.

> ⓘ **Nota:**
>
> API Management usa unidades para los componentes del gateway. Las unidades no son aplicables al portal para desarrolladores ni al plano de administración.

Niveles de servicio

Los niveles de servicio de API Management proporcionan distintos niveles de confiabilidad:

- **Nivel Premium (clásico):** Admite varias unidades que se pueden distribuir entre zonas de disponibilidad y regiones para lograr una máxima resistencia.
- **Nivel Premium v2**: admite varias unidades que se pueden distribuir entre zonas de disponibilidad. Actualmente no admite implementaciones de varias regiones.
- **Niveles Básico v2, Estándar y Estándar v2:** Todas admiten varias unidades dentro de un único centro de datos. No admiten zonas de disponibilidad ni implementaciones de varias regiones.
- **Nivel de desarrollador:** Solo admite una sola unidad y no proporciona compatibilidad con zonas de disponibilidad ni con varias regiones. Este nivel está diseñado para escenarios de desarrollo y pruebas. No es adecuado para cargas de trabajo de producción.
- **Nivel de consumo:** Tiene funcionalidades de resistencia integradas y es resistente a una serie de errores dentro de un único centro de datos de Azure. Sin embargo, el nivel Consumo no proporciona compatibilidad con zonas de disponibilidad ni implementaciones de varias regiones. Para comprender el tiempo de actividad esperado de una instancia de API Management de nivel de consumo, revise el acuerdo de nivel de servicio (SLA).

> ⓘ **Nota:**
>
> Los niveles Desarrollador y Premium de API Management admiten puertas de enlace autohospedada, que puede ejecutar en su propia infraestructura. Al usar puertas de enlace autohospedadas, es responsable de configurarlas para cumplir los requisitos de confiabilidad. Las puertas de enlace autohospedados están fuera del ámbito de este artículo.

Recomendaciones de implementación de producción

Azure Well-Architected Framework proporciona recomendaciones sobre confiabilidad, rendimiento, seguridad, costo y operaciones. Para comprender cómo estas áreas influyen entre sí y contribuyen a una solución confiable de API Management, consulte Procedimientos recomendados de arquitectura para API Management.

Resistencia a errores transitorios

Los errores transitorios son errores breves e intermitentes en los componentes. Se producen con frecuencia en un entorno distribuido como la nube y son una parte normal de las operaciones. Los errores transitorios se corrigen después de un breve período de tiempo. Es importante que las aplicaciones puedan controlar errores transitorios, normalmente mediante el reintento de solicitudes afectadas.

Todas las aplicaciones hospedadas en la nube deben seguir las instrucciones de control de errores transitorios de Azure cuando se comunican con cualquier API, bases de datos y otros componentes hospedados en la nube. Para obtener más información, consulte Recomendaciones para controlar errores transitorios.

Al usar API Management delante de una API, es posible que tenga que reintentar solicitudes que produzcan un error debido a errores transitorios. Para proteger la API de back-end de ser sobrecargada por demasiadas solicitudes, API Management proporciona directivas de reintento, límite de velocidad y cuota. También puede configurar las funcionalidades de equilibrio de carga y disyuntor mediante el uso de recursos de backend.

Resistencia a errores de zona de disponibilidad

Las zonas de disponibilidad son grupos físicamente independientes de centros de datos dentro de una región de Azure. Cuando una zona falla, los servicios pueden transferirse a una de las zonas restantes.

API Management proporciona dos tipos de compatibilidad con zonas de disponibilidad al implementar una instancia de API Management Premium (clásica) en una región admitida:

- **Automático** (recomendado): API Management proporciona compatibilidad con zonas de disponibilidad automáticas cuando no se especifican las zonas de disponibilidad que se van a usar.

- **Manual:** API Management proporciona compatibilidad con la zona de disponibilidad manual cuando se especifican explícitamente las zonas de disponibilidad que se van a usar.

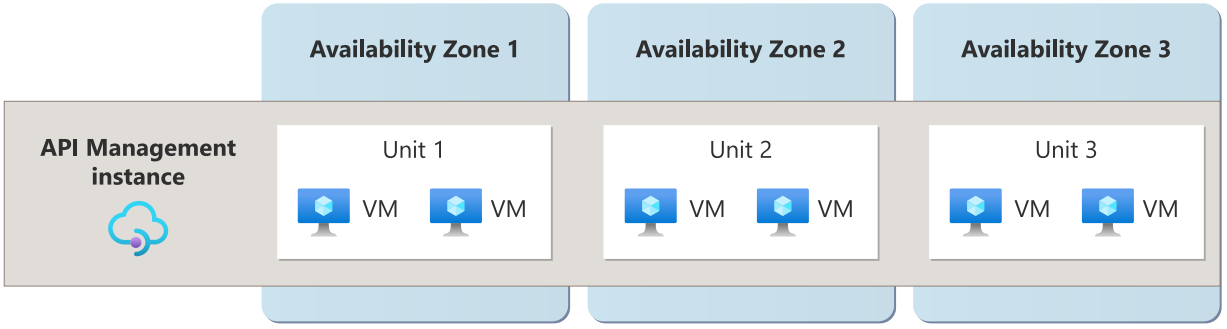
Con la compatibilidad con zonas de disponibilidad, API Management replica componentes de servicio entre zonas para lograr alta disponibilidad. En la región primaria, estos componentes incluyen la puerta de enlace (unidades de escalado), el plano de administración y el portal para desarrolladores. En regiones secundarias, solo se replican las unidades de puerta de enlace. Para más información sobre las regiones secundarias, consulte Resistencia a errores en toda la región.

Compatibilidad con zonas de disponibilidad automáticas

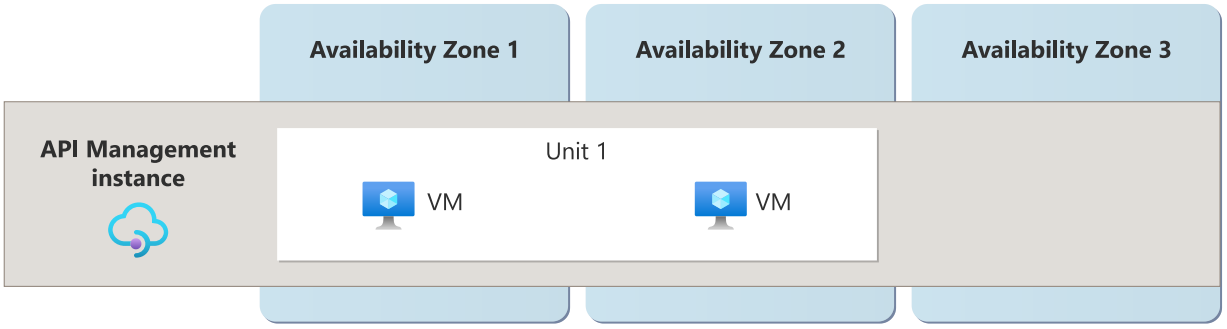
Puede usar la compatibilidad automática con zonas de disponibilidad para elegir una configuración de instancia de una sola unidad o de varias unidades con el fin de lograr la redundancia de zonas:

- **Configuración** de varias unidades (recomendada): si la instancia tiene dos o más unidades, API Management realiza un mejor intento de distribuir las unidades de la instancia entre las zonas de disponibilidad de la región. No puede determinar en qué zonas de disponibilidad se colocan las unidades. Implemente un mínimo de dos unidades, que se pueden distribuir entre dos zonas.

  En el diagrama siguiente se muestra una instancia de API Management con tres unidades configuradas para la compatibilidad automática con zonas de disponibilidad:



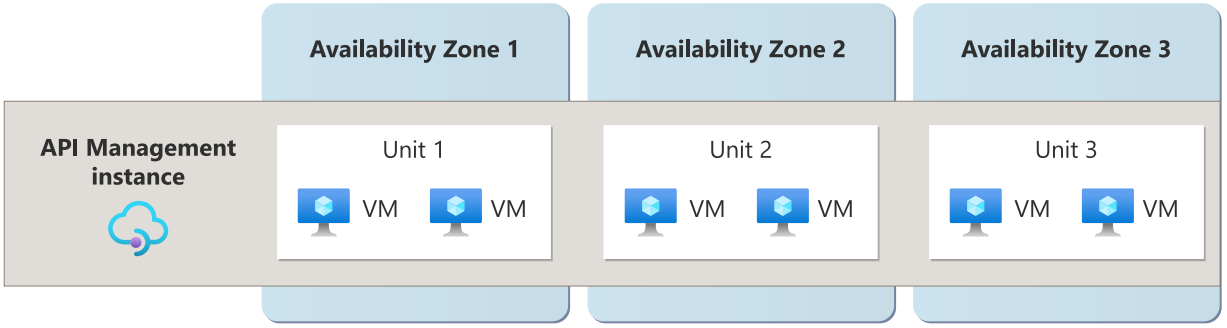
- **Configuración de una sola unidad:** Si la instancia tiene una sola unidad, los recursos de proceso subyacentes de la unidad se distribuyen a dos zonas de disponibilidad. No puede determinar en qué zonas de disponibilidad se colocan los recursos de proceso de la unidad.



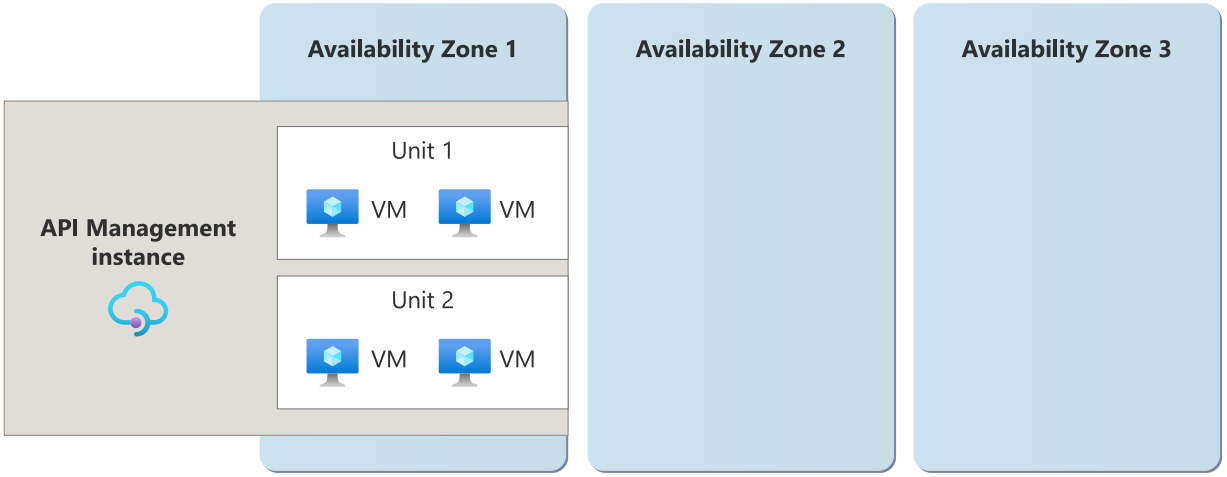
Compatibilidad con zonas de disponibilidad configuradas manualmente

Si desea seleccionar explícitamente las zonas de disponibilidad que se van a usar, puede elegir entre configuraciones con redundancia de zona y zonales:

- **Redundancia de zona:** Configure manualmente la redundancia de zona para una instancia de API Management en una región compatible con el fin de proporcionar redundancia a los componentes del servicio. Al seleccionar dos o más zonas de disponibilidad que se van a usar, Azure replica automáticamente los componentes de servicio en las zonas seleccionadas.



- **Zonal:** Los componentes del servicio API Management se implementan en una sola zona que se selecciona dentro de una región de Azure. Todas las unidades se colocan en la misma zona de disponibilidad.



> **Importante**
>
> Ancle a una sola zona de disponibilidad solo si la latencia entre zonas es demasiado alta para sus necesidades y después de comprobar que la latencia no cumple sus requisitos. Por sí mismo, una instancia zonal no proporciona resistencia a una interrupción de zona de disponibilidad. Para mejorar la resistencia de una implementación de API Management zonal, debe implementar explícitamente instancias independientes en varias zonas de disponibilidad y configurar el enrutamiento y la conmutación por error del tráfico.

Requisitos

- **Compatibilidad con regiones:** API Management admite zonas de disponibilidad para los niveles Premium (clásico) y Premium v2 en regiones en las que el nivel api Management está disponible y la región admite zonas de disponibilidad.
- **Requisito de nivel:** Debe usar el nivel Premium (clásico) o Premium v2 para configurar la compatibilidad con zonas de disponibilidad. API Management no admite actualmente zonas de disponibilidad en los niveles consumo clásico, desarrollador, básico y estándar o en los niveles Básico v2 y Estándar v2. Para ver las opciones de actualización, consulte Actualización y escalado de una instancia de API Management.

Consideraciones

- **Número de unidades para instancias con redundancia de zona:** Si configura manualmente la redundancia de zona para una instancia, también debe configurar una serie de unidades de API Management que se pueden distribuir uniformemente en todas las zonas de disponibilidad seleccionadas. Por ejemplo, si selecciona dos zonas, debe configurar al menos dos unidades. En su lugar, puede configurar cuatro unidades u otras

varias de dos unidades. Si selecciona tres zonas de disponibilidad, debe configurar tres unidades, seis unidades u otro múltiplo de tres unidades.

Si usa la compatibilidad con la zona de disponibilidad automática, no es necesario usar un número específico de unidades. Las unidades que implemente se distribuyen entre las zonas de disponibilidad de una manera más adecuada. Para obtener redundancia máxima de zona, use al menos dos unidades para que una interrupción de zona de disponibilidad no afecte al rendimiento de la puerta de enlace.

Para determinar el número de unidades que proporcionan el rendimiento necesario de su puerta de enlace, use las métricas de capacidad y sus propias pruebas. Para más información sobre el escalado y la actualización de la instancia de servicio, consulte Actualización y escalado de una instancia de API Management.

- **Escalado automático:** Si configura manualmente zonas de disponibilidad en una instancia de API Management configurada con el escalado automático, es posible que tenga que ajustar la configuración de escalado automático después de la configuración. En este caso, el número de unidades de API Management en reglas y límites de escalado automático debe ser un múltiplo del número de zonas. Si usa la compatibilidad con la zona de disponibilidad automática, no es necesario ajustar la configuración de escalado automático.

- **Requisitos de dirección IP:** Cuando habilita la compatibilidad con zonas de disponibilidad en una instancia de API Management implementada en una red virtual externa o interna, debe especificar un recurso de dirección IP pública para que lo utilice la instancia. En una red virtual interna, la dirección IP pública solo se usa para las operaciones de administración, no para las solicitudes de API. Para más información, consulte Direcciones IP en API Management.

Cost

Independientemente de la configuración de la zona de disponibilidad, si agrega más unidades, incurrirá en más costos. Para más información, consulte Precios de API Management.

Configurar soporte de zonas de disponibilidad

En esta sección se explica cómo configurar la compatibilidad con zonas de disponibilidad para la instancia de API Management. Para obtener más información, consulte Habilitación de la compatibilidad con zonas de disponibilidad en instancias de API Management.

- **Cree una instancia de API Management que admita zonas de disponibilidad:** Al crear una instancia de API Management Premium (clásica) en una región que admita zonas de

disponibilidad, la instancia admite zonas de disponibilidad de forma predeterminada. Puede seleccionar la compatibilidad automática con zonas de disponibilidad o configurar manualmente el soporte zonal o el soporte con redundancia de zona.

> ⓘ **Nota:**
>
> Cuando selecciona qué zonas de disponibilidad usar, en realidad está seleccionando la *zona de disponibilidad lógica*. Si implementa otros componentes de la carga de trabajo en una suscripción de Azure diferente, podrían usar un número de zona de disponibilidad lógica distinto para acceder a la misma zona de disponibilidad física. Para obtener más información, consulte zonas de disponibilidad física y lógica.

- **Habilite o vuelva a configurar la compatibilidad con la zona de disponibilidad:** Puede cambiar la configuración de zona de disponibilidad de una instancia de API Management, incluida la adición de zonas de disponibilidad y el traslado de una instancia zonal entre zonas de disponibilidad. Para obtener información sobre cómo configurar la compatibilidad con zonas de disponibilidad en una instancia de API Management, consulte Habilitar la compatibilidad con zonas de disponibilidad en instancias de API Management. Ninguna de las opciones de configuración requiere tiempo de inactividad.

  Al cambiar la configuración de la zona de disponibilidad, los cambios pueden tardar entre 15 y 45 minutos o más en aplicarse. La puerta de enlace de API Management puede seguir controlando las solicitudes de API durante ese tiempo.

  Cambiar la configuración de la zona de disponibilidad desencadena un cambio de dirección IP pública y privada.

Planeamiento y administración de capacidad

En un escenario de reducción de zona, no hay ninguna garantía de que las solicitudes de más capacidad en otra zona de disponibilidad tengan éxito. El reposición de unidades perdidas se produce de forma más adecuada. Si necesita capacidad garantizada cuando se produce un error en una zona de disponibilidad, cree y configure la instancia de API Management para que tenga en cuenta la pérdida de una zona mediante la realización de todas las acciones siguientes:

- Aprovisionamiento excesivo de las unidades de la instancia de API Management.
- Use la configuración de zona de disponibilidad automática o redundante por zonas.

Para más información, consulte Administración de la capacidad con sobreaprovisionamiento.

Use métricas de capacidad y sus propias pruebas para determinar el número de unidades que proporcionan el rendimiento de la puerta de enlace necesaria. Para más información sobre cómo escalar y actualizar la instancia de servicio, consulte Actualización y escalado de una instancia de API Management.

Comportamiento cuando todas las zonas están en buen estado

En esta sección se describe qué esperar cuando las instancias de API Management están configuradas con compatibilidad con zonas de disponibilidad y todas las zonas de disponibilidad están operativas.

- **Enrutamiento de tráfico entre zonas:** Durante las operaciones normales, el tráfico se enruta entre todas las unidades de API Management disponibles en todas las zonas de disponibilidad seleccionadas.
- **Replicación de datos entre zonas:** API Management almacena y replica los siguientes datos.
  - *La configuración de la puerta de enlace*, como las API y las definiciones de directiva, se sincroniza periódicamente entre las zonas de disponibilidad que seleccione para la instancia. La propagación de actualizaciones entre las zonas de disponibilidad normalmente tarda menos de 10 segundos.
  - *Datos de la memoria caché interna*, si usa la memoria caché interna que proporciona API Management. Las entradas de caché se distribuyen entre zonas de disponibilidad. La caché interna es volátil y no se garantiza que los datos se conserven. Considere la posibilidad de usar una caché externa si necesita conservar los datos almacenados en caché.
  - *Contadores de límite de velocidad*, si usa las funcionalidades de limitación de velocidad que proporciona API Management. Los contadores de límite de velocidad se replican de forma asíncrona entre las zonas de disponibilidad que seleccione para la instancia.

Comportamiento durante un fallo de zona

En esta sección se describe qué esperar cuando las instancias de API Management están configuradas con compatibilidad con la zona de disponibilidad y hay una interrupción de zona de disponibilidad.

- **Detección y respuesta:** La responsabilidad de la detección y la respuesta depende de la configuración de zona de disponibilidad que usa la instancia.

- *Automático y con redundancia de zona:* En el caso de las instancias configuradas para utilizar la compatibilidad automática con zonas de disponibilidad o configuradas manualmente para utilizar la redundancia de zona, la plataforma API Management se encarga de detectar los fallos en una zona de disponibilidad y responder a ellos. No es necesario hacer nada para iniciar una conmutación por error de la zona.
- *Zonal:* para que las instancias configuradas sean zonales, debe detectar la pérdida de una zona de disponibilidad e iniciar una conmutación por error a una instancia secundaria que cree en otra zona de disponibilidad.

- **Solicitudes activas:** Cuando una zona de disponibilidad no está disponible, las solicitudes en curso que están conectadas a una unidad de API Management en la zona de disponibilidad defectuosa se finalizan y deben reintentarse.
- **Notificación**: Microsoft no le notifica automáticamente cuando una zona está inactiva. Sin embargo, puede usar Azure Resource Health para supervisar el estado de un recurso individual y puede configurar alertas de Resource Health para notificarle problemas. También puede usar Azure Service Health para comprender el estado general del servicio, incluidos los errores de zona, y puede configurar alertas de Service Health para notificarle problemas.
- **Pérdida de datos esperada:** API Management almacena los siguientes datos.
  - *Cambios en la configuración de la puerta de enlace*, y que se replican en cada zona de disponibilidad seleccionada en aproximadamente 10 segundos. Si se produce una interrupción de una zona de disponibilidad, puede perder los cambios de configuración que no se replican.
  - *Datos de la memoria caché interna*, si usa la característica de caché interna. La caché interna es volátil y no se garantiza que los datos se conserven. Durante una interrupción del servicio en una zona de disponibilidad, es posible que se pierda parte o la totalidad de los datos almacenados en caché. Considere la posibilidad de usar una caché externa si necesita conservar los datos almacenados en caché.
  - *Contadores de límite de velocidad*, si usa la característica de límite de velocidad. Durante una interrupción de zona de disponibilidad, es posible que los contadores de límite de frecuencia no estén actualizados en las zonas supervivientes.
- **Tiempo de inactividad esperado:** El tiempo de inactividad esperado depende de la configuración de zona de disponibilidad que usa la instancia.
  - *Automático:* Puede esperar que las instancias que usen compatibilidad con zonas de disponibilidad automáticas no tengan tiempo de inactividad durante una interrupción

de la zona de disponibilidad. Las unidades de la zona o zonas no afectadas siguen funcionando.

También puede esperar instancias que usen compatibilidad automática con zonas de disponibilidad, pero que tengan una sola unidad, sin tiempo de inactividad. En este caso, API Management distribuye los recursos de proceso subyacentes de la unidad a dos zonas. El recurso de la zona no afectada sigue funcionando.

- *Con redundancia de zona:* Puede esperar que las instancias con redundancia de zona no sufran interrupciones durante una interrupción de la zona de disponibilidad.
- *Zonal:* En el caso de las instancias zonales, cuando una zona no está disponible, la instancia no está disponible hasta que se recupere la zona de disponibilidad.

- **Reenrutamiento del tráfico:** El comportamiento de reenrutamiento del tráfico depende de la configuración de zona de disponibilidad que usa la instancia.
  - *Automático y con redundancia de zona:* En el caso de las instancias que están configuradas para utilizar la compatibilidad automática con zonas de disponibilidad o que están configuradas manualmente para utilizar la redundancia de zona, cuando una zona no está disponible, ninguna de las unidades de la zona afectada estará disponible. Puede optar por escalar la instancia para agregar más unidades.
  - *Zonal:* En el caso de las instancias zonales, cuando una zona no está disponible, la instancia no está disponible. Si tiene una instancia secundaria en otra zona de disponibilidad, es responsable de redirigir el tráfico a esa instancia secundaria.

Recuperación de zona

- **Automático y con redundancia de zona:** En el caso de las instancias que están configuradas para utilizar la compatibilidad automática con zonas de disponibilidad o que están configuradas manualmente para utilizar la redundancia de zona, cuando la zona de disponibilidad se recupera, API Management restaura automáticamente las unidades de la zona de disponibilidad y redirige el tráfico entre sus unidades con normalidad.
- **Zonal:** En el caso de las instancias zonales, usted es responsable de redirigir el tráfico a la instancia en la zona de disponibilidad original una vez que la zona de disponibilidad se haya recuperado.

Prueba de fallos de zona

- **Automático y con redundancia de zona:** En el caso de las instancias que están configuradas para utilizar la compatibilidad automática con zonas de disponibilidad o que están configuradas manualmente para utilizar la redundancia de zona, la plataforma API Management administra el enrutamiento del tráfico, la conmutación por error y la conmutación por recuperación. Esta característica está totalmente administrada, por lo que no es necesario iniciar ni validar los procesos de error de zona de disponibilidad.

- **Zonal:** En el caso de las instancias zonales, no se puede simular una interrupción de la zona de disponibilidad que contiene la instancia de API Management. Sin embargo, puede configurar manualmente puertas de enlace ascendentes o equilibradores de carga para redirigir el tráfico a una instancia diferente en una zona de disponibilidad diferente.

Resistencia a errores en toda la región

Mediante una implementación de varias regiones, puede agregar puertas de enlace de API regionales a una instancia de API Management existente en una o varias regiones de Azure admitidas. La implementación en varias regiones ayuda a reducir la latencia de las solicitudes que perciben los consumidores de API distribuidos geográficamente. Una implementación de varias regiones también mejora la disponibilidad del servicio si una región se queda sin conexión.

> ⓘ **Importante**
>
> Las implementaciones de varias regiones solo se admiten en el nivel Premium (clásico) de API Management.

Implementación multirregión administrada por Microsoft

Al agregar una región, configure:

- Número de unidades que hospeda esa región.

- Resistencia a errores de zona de disponibilidad, si esa región proporciona zonas de disponibilidad.

- La configuración de red virtual en la región agregada, si las redes están configuradas en la región o regiones existentes.

Requisitos

- **Compatibilidad con regiones:** Puede crear implementaciones de varias regiones en el nivel Premium (clásico) con cualquier región de Azure que admita API Management. Para ver qué regiones admiten implementaciones de varias regiones, consulte Disponibilidad del producto por región.

- **Requisito de nivel:** Debe usar el nivel Premium (clásico) para configurar la compatibilidad con varias regiones. Para actualizar la instancia al nivel Premium (clásico), consulte Actualización al nivel Premium.

Consideraciones

- **Solo puerta de enlace:** Únicamente el componente de puerta de enlace de su instancia de gestión de API se replica en múltiples regiones. El plano de administración y el portal para desarrolladores de la instancia permanecen hospedados solo en la región primaria donde implementó originalmente el servicio.

- **Requisitos de red:** Si desea configurar una ubicación secundaria para su instancia de API Management cuando se implementa (inyecta) en una red virtual, la red virtual y la región de la subred deben coincidir con la ubicación secundaria que configure. Si agrega, quita o habilita la zona de disponibilidad en la región primaria, o si cambia la subred de la región primaria, cambia la dirección IP virtual (VIP) de la instancia de API Management. Para obtener más información, consulte Cambios en las direcciones IP. Sin embargo, si se agrega una región secundaria, el VIP de la región principal no cambia, ya que cada región tiene su propio VIP privado.

- **Nombres del sistema de nombres de dominio (DNS):** La puerta de enlace de cada región, incluida la región primaria, tiene un nombre DNS regional que sigue el patrón de dirección URL de `https://<service-name>-<region>-01.regional.azure-api.net`, por ejemplo `https://contoso-westus2-01.regional.azure-api.net`.

Cost

Agregar regiones implica costos. Para más información, consulte Precios de API Management.

Configuración de la compatibilidad con varias regiones

Para configurar la compatibilidad con varias regiones en una instancia de API Management, consulte Implementar una instancia de API Management en varias regiones de Azure.

Para quitar una región de una instancia de API Management, consulte Eliminación de una región del servicio API Management.

Planeamiento y administración de capacidad

En un escenario de caída regional, no hay garantía de que las solicitudes de más capacidad en otra región se realicen con éxito. Si necesita capacidad garantizada cuando se produce un error en una región, debe crear y configurar la instancia de API Management para que tenga en cuenta la pérdida de una región. Puede hacerlo mediante el aprovisionamiento excesivo de la capacidad de la instancia de API Management. Para más información sobre el principio de sobreaprovisionamiento, consulte Gestionar la capacidad con sobreaprovisionamiento.

En las implementaciones de varias regiones, el escalado automático solo se aplica a la región primaria. Las regiones secundarias requieren ajustes de escalado manual o herramientas personalizadas que controle.

Comportamiento cuando todas las regiones están en buen estado

En esta sección se describe qué se puede esperar cuando las instancias de API Management se configuran con compatibilidad con varias regiones y todas las regiones están operativas.

- **Enrutamiento del tráfico entre regiones:** API Management enruta automáticamente las solicitudes recibidas a una puerta de enlace regional. Una solicitud se enruta a la puerta de enlace regional con la latencia más baja del cliente. Si necesita usar un enfoque de enrutamiento diferente, puede configurar su propio Administrador de tráfico con reglas de enrutamiento personalizadas. Para obtener más información, consulte Uso del enrutamiento personalizado en las puertas de enlace regionales de API Management.

  Cuando una solicitud llega a una puerta de enlace regional de API Management, se enruta a la API backend, a menos que una directiva devuelva una respuesta directamente desde la puerta de enlace, como una respuesta almacenada en caché o un código de error. En una solución de varias regiones, debe tener cuidado de enrutar a una instancia de la API de back-end que cumpla los requisitos de rendimiento. Para obtener más información, consulte Enrutamiento de llamadas API a servicios back-end regionales.

- **Replicación de datos entre regiones:** La configuración de puerta de enlace, como las API y las definiciones de directiva, se sincronizan periódicamente entre las regiones primarias y secundarias que agregue. La propagación de actualizaciones a las puertas de enlace regionales normalmente tarda menos de 10 segundos.

  Los contadores de límite de velocidad y los datos de la caché interna son específicos de la región, por lo que no se replican entre regiones.

Comportamiento durante una falla de región

En esta sección se describe qué puede ocurrir cuando las instancias de API Management se configuran con compatibilidad con varias regiones y se produce una interrupción del servicio en una de las regiones que utiliza.

- **Detección y respuesta:** La API Management se encarga de detectar un error en una región y de realizar automáticamente una conmutación por error a una puerta de enlace en una de las otras regiones que usted configure.

- **Solicitudes activas:** Las solicitudes activas que se están procesando en la región defectuosa se pueden quitar y el cliente debe reintentarlas.

- **Pérdida de datos esperada:** API Management no almacena datos, excepto los contadores de configuración, caché y límite de velocidad.

  Los cambios de configuración se replican en cada región en aproximadamente 10 segundos. Si se produce una interrupción del servicio en su región principal, es posible que pierda los cambios de configuración que no se hayan replicado.

  Los contadores de límite de velocidad y los datos de la caché interna son específicos de la región, por lo que no se replican entre regiones.

- **Tiempo de inactividad esperado:** No se espera ningún tiempo de inactividad de puerta de enlace.

  Si la región primaria se queda sin conexión, el plano de administración de API Management y el portal para desarrolladores no estarán disponibles. Sin embargo, las regiones secundarias siguen atendiendo solicitudes de API mediante la configuración de puerta de enlace más reciente.

- **Reenrutamiento del tráfico:** Si una región se queda sin conexión, las solicitudes de API se enrutan automáticamente evitando la región fallida hacia la siguiente puerta de enlace más cercana.

Recuperación de regiones

Cuando se recupera la región primaria, API Management restaura automáticamente las unidades de la región y vuelve a enrutar el tráfico entre las unidades.

Prueba de fallos de región

Para estar preparado para interrupciones inesperadas de regiones, realice pruebas periódicas de sus respuestas a los fallos de la región. Puede simular algunos aspectos de un error de región deshabilitando el enrutamiento a una puerta de enlace regional.

Copias de seguridad y restauración

API Management no almacena la mayoría de los datos en tiempo de ejecución. Sin embargo, puede realizar una copia de seguridad de la configuración del servicio API Management. También puede usar operaciones de copia de seguridad y restauración para replicar las configuraciones del servicio API Management entre entornos operativos, como el desarrollo y el almacenamiento provisional.

> **Importante**
>
> En un procedimiento de copia de seguridad, se incluyen datos en tiempo de ejecución, como usuarios y suscripciones, que podrían no ser siempre deseables.

La copia de seguridad se admite en los niveles Desarrollador, Básico, Estándar y Premium.

Para obtener más información, consulte Cómo implementar la recuperación ante desastres mediante la copia de seguridad y restauración del servicio en API Management.

Para realizar copias de seguridad o restaurar algunos componentes o recursos del servicio, también puede considerar opciones administradas por el cliente, como las herramientas APIOps y las soluciones de infraestructura como código (IaC).

Resistencia al mantenimiento del servicio

API Management realiza actualizaciones de servicio periódicas y otras formas de mantenimiento.

En los niveles Básico, Estándar y Premium (clásico), puede personalizar cuándo en el proceso de actualización la instancia recibe una actualización. Si necesita validar el efecto de las actualizaciones en la carga de trabajo, considere la posibilidad de configurar una instancia de prueba para recibir actualizaciones al principio de un ciclo de actualización y establecer la instancia de producción para recibir actualizaciones en un momento posterior al ciclo. También puede especificar una ventana de mantenimiento, que es la hora del día en la que desea que la instancia aplique las actualizaciones del servicio.

Para obtener más información, consulte Configurar los ajustes de actualización del servicio para sus instancias de API Management.

Acuerdo de nivel de servicio

El contrato de nivel de servicio (SLA) para los servicios de Azure describe la disponibilidad esperada de cada servicio y las condiciones que la solución deberá cumplir para lograr esa expectativa de disponibilidad. Para obtener más información, consulte Acuerdos de Nivel de Servicio para servicios en línea.

Al implementar una instancia de API Management en varias zonas de disponibilidad o regiones, aumenta el porcentaje de tiempo de actividad definido en el Acuerdo de Nivel de Servicio.

El servicio proporciona su propio Acuerdo de Nivel de Servicio, pero también debe tener en cuenta la confiabilidad prevista de otros componentes de carga de trabajo, como los back-end de API.

Contenido relacionado

- Recuperación ante desastres y continuidad empresarial para API Management
- Uso de zonas de disponibilidad en API Management
- Implementación en varias regiones de API Management
- Supervisar API Management
- Planeamiento de la capacidad de API Management
- Procedimientos recomendados de arquitectura para API Management

ⓘ **Nota:** El autor ha creado este artículo con ayuda de inteligencia artificial. Más información

---

Last updated on 22/01/2026

Confiabilidad en Azure Data Factory

Azure Data Factory le permite crear canalizaciones de datos flexibles y eficaces para la integración de datos sin servidor y la transformación de datos. Como servicio de Azure, Data Factory proporciona una variedad de funcionalidades para satisfacer los requisitos de fiabilidad.

Cuando se usa Azure, la confiabilidad es una responsabilidad compartida. Microsoft proporciona una variedad de funcionalidades para admitir resistencia y recuperación. Es responsable de comprender cómo funcionan esas funcionalidades dentro de todos los servicios que usa y de seleccionar las funcionalidades que necesita para cumplir los objetivos empresariales y los objetivos de tiempo de actividad.

En este artículo se describe cómo hacer que Data Factory sea resistente a una variedad de posibles interrupciones y problemas, incluidos errores transitorios, interrupciones de zona de disponibilidad y interrupciones de regiones. También se describe cómo puede usar copias de seguridad para recuperarse de otros tipos de problemas y se resalta cierta información clave sobre el acuerdo de nivel de servicio (SLA) de Data Factory.

> ⓘ **Nota:**
>
> Al considerar la confiabilidad de la factoría de datos, también debe tener en cuenta la confiabilidad de los almacenes de datos a los que se conecta. La mejora de la resistencia de la factoría de datos por sí sola podría tener un impacto limitado si los almacenes de datos no son igualmente resistentes. En función de los requisitos de resistencia, es posible que tenga que realizar cambios de configuración en varias áreas. Para ayudar a garantizar que los almacenes de datos cumplan los requisitos de continuidad empresarial, consulte la documentación y las instrucciones de confiabilidad del producto.

Introducción a la arquitectura de confiabilidad

Data Factory consta de varios componentes de infraestructura. Cada componente admite la confiabilidad de la infraestructura de maneras diferentes.

Los componentes de Data Factory incluyen:

- **El servicio principal de Data Factory**, que administra los desencadenadores de canalización y supervisa la coordinación de las actividades de canalización. El servicio principal también administra los metadatos de cada componente de la factoría de datos. Microsoft administra el servicio principal.

- **Entornos de ejecución de integración (IR)**, que se conectan a almacenes de datos y realizan actividades definidas en la canalización. Hay diferentes tipos de infrarrojos.
    - *IR administrados por Microsoft*, que incluyen Azure IR y Azure-SQL Server Integration Services (Azure-SSIS) IR. Microsoft administra los componentes que componen estos entornos de ejecución. En algunos escenarios, se configuran las opciones que afectan a la resiliencia del IR.
    - *Entornos de ejecución de integración autohospedados (SHIR)*. Microsoft proporciona software que puede ejecutar en su propia infraestructura informática para realizar algunas partes de sus propias canalizaciones de Data Factory. Es responsable de la implementación y administración de los recursos de computación y de la resiliencia de esos recursos de computación.

Resistencia a errores transitorios

Los errores transitorios son errores breves e intermitentes en los componentes. Se producen con frecuencia en un entorno distribuido como la nube y son una parte normal de las operaciones. Los errores transitorios se corrigen después de un breve período de tiempo. Es importante que las aplicaciones puedan controlar errores transitorios, normalmente mediante el reintento de solicitudes afectadas.

Todas las aplicaciones hospedadas en la nube deben seguir las instrucciones de control de errores transitorios de Azure cuando se comunican con cualquier API, bases de datos y otros componentes hospedados en la nube. Para obtener más información, consulte Recomendaciones para controlar errores transitorios.

Cuando se usa Data Factory, es importante prepararse para errores transitorios, especialmente al diseñar canalizaciones y actividades.

Idempotencia

Las actividades de canalización deben ser *idempotentes*, lo cual significa que se puedan volver a ejecutar varias veces sin causar efectos adversos. Si se produce un error transitorio como un error de red o una interrupción de zona de disponibilidad, Data Factory podría volver a ejecutar actividades de canalización. Esta repetición puede crear registros duplicados.

Para evitar la inserción de registros duplicados debido a un error transitorio, implemente los procedimientos recomendados siguientes:

- *Use identificadores únicos* para cada registro antes de escribir en la base de datos. Este enfoque puede ayudarle a encontrar y eliminar registros duplicados.

- *Use una estrategia upsert* para los conectores que admiten upsert. Antes de que se produzca la inserción de registros duplicados, use este enfoque para comprobar si ya existe un registro. Si existe, actualícela. Si no existe, insértelo. Por ejemplo, comandos SQL como `MERGE` o `ON DUPLICATE KEY UPDATE` usan este enfoque de upsert.

- *Use estrategias de acción de copia.* Para obtener más información, consulte Comprobación de coherencia de datos en la actividad de copia.

Directivas de reintentos

Puede usar directivas de reintento para configurar partes de la canalización para reintentar si hay un problema, como errores transitorios en los recursos conectados. En Data Factory, puede configurar directivas de reintento en los siguientes tipos de objetos de canalización:

- Desencadenadores periódicos
- Actividades de ejecución

Para obtener más información sobre cómo cambiar o deshabilitar las directivas de reintento para los desencadenadores y actividades de Data Factory, consulte Ejecuciones y desencadenadores de canalización.

Resistencia a errores de zona de disponibilidad

Las zonas de disponibilidad son grupos físicamente independientes de centros de datos dentro de una región de Azure. Cuando una zona falla, los servicios pueden transferirse a una de las zonas restantes.

Data Factory admite *redundancia de zona*, lo que proporciona resistencia a errores en las zonas de disponibilidad.

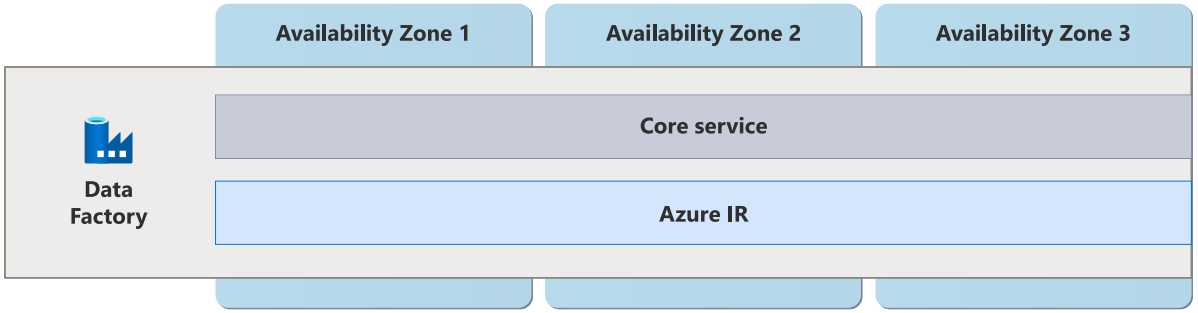
Cada parte de Data Factory admite redundancia de zona:

- **Servicio principal:** Microsoft administra los componentes en el servicio principal de Data Factory y los distribuye entre zonas de disponibilidad.

  Pero después de un error de zona, Microsoft no garantiza el estado de los desencadenadores de ventana de saltos de tamaño constante.

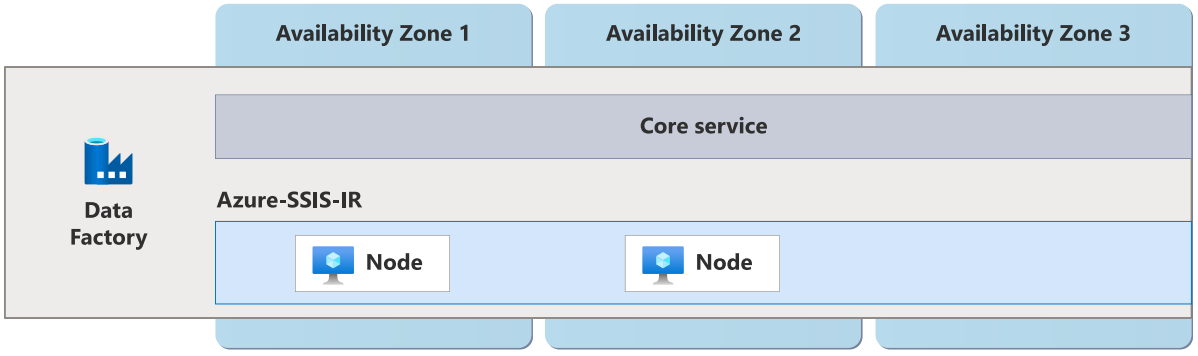
- **IRs:** La compatibilidad con la redundancia de zona depende del tipo de IR que se utilice.

  - *Azure IR* admite redundancia de zona y Microsoft administra esta funcionalidad.



- *Un Azure-SSIS IR* requiere la implementación de al menos dos nodos. Estos nodos se asignan automáticamente a diferentes zonas de disponibilidad.

  En el diagrama siguiente se muestra una canalización con redundancia de zona y un entorno de ejecución de integración de Azure-SSIS con dos nodos que se implementan en zonas diferentes:



- *Un SHIR* le ofrece la responsabilidad de implementar la infraestructura de proceso para hospedar el entorno de ejecución. Puede implementar varios nodos, como máquinas virtuales individuales y configurarlos para alta disponibilidad. Después, puede distribuir esos nodos en varias zonas de disponibilidad. Para más información, consulte Alta disponibilidad y escalabilidad.

Requisitos

Los recursos de Data Factory con redundancia de zona se pueden implementar en cualquier región que admita zonas de disponibilidad.

Cost

- **Servicio principal:** No se aplica ningún costo adicional para la redundancia de zona.
- **IR:** El costo de la redundancia de zona varía en función del tipo de IR que use.
  - *Azure IR* incluye redundancia de zona sin costo adicional.

- *Un Azure-SSIS IR* requiere que despliegue al menos dos nodos para lograr redundancia de zona. Para obtener más información sobre cómo se factura cada nodo, vea Ejemplo de precios: Ejecutar paquetes SSIS en un IR de Azure-SSIS.
- *Un SHIR* requiere que se implemente y administre la infraestructura de proceso. Para lograr la resiliencia de la zona, debe distribuir los recursos informáticos entre varias zonas. Dependiendo del número de nodos que implemente y cómo los configure, puede incurrir en costos adicionales de los servicios de proceso subyacentes y otros servicios auxiliares. No hay ningún cargo adicional para ejecutar el SHIR en varios nodos.

Configurar soporte de zonas de disponibilidad

- **Servicio principal:** No se requiere ninguna configuración. El servicio principal de Data Factory admite automáticamente la redundancia de zona.
- **IRs:**
  - *Una instancia de Azure IR:* No se requiere ninguna configuración. Azure IR habilita automáticamente la redundancia de zona.
  - *Un Azure-SSIS IR:* no se requiere configuración. Un IR de nivel Azure-SSIS habilita automáticamente la redundancia de zona cuando se implementa con dos o más nodos.
  - *Una SHIR* requiere que configure su propia resistencia, lo que incluye la propagación de los nodos entre varias zonas de disponibilidad.

Planeamiento y administración de capacidad

- **Servicio principal:** El servicio principal de Data Factory se escala automáticamente en función de la demanda y no es necesario planear ni administrar la capacidad.
- **IRs:**
  - *Una instancia de Azure IR* se escala automáticamente en función de la demanda y no es necesario planear ni administrar la capacidad.
  - *Un Azure-SSIS IR* requiere que configure específicamente el número de nodos que use. Para prepararse para errores de zona de disponibilidad, considere la posibilidad de sobreaprovisionar la capacidad del IR. El sobreaprovisionamiento permite a la solución tolerar cierto grado de pérdida de capacidad y seguir funcionando sin degradar el

rendimiento. Para obtener más información, consulte Administración de la capacidad mediante el aprovisionamiento excesivo.

- *Un SHIR* requiere que configure su propia capacidad y escalamiento. Considere la posibilidad de sobreaprovisionar al implementar una SHIR.

Comportamiento cuando todas las zonas están en buen estado

En esta sección se describe qué esperar cuando los recursos de Data Factory están configurados para la redundancia de zona y todas las zonas de disponibilidad están operativas.

- **Enrutamiento de tráfico entre zonas:** Durante las operaciones normales, Data Factory distribuye automáticamente las actividades de canalización, los desencadenadores y otros trabajos entre instancias correctas en cada zona de disponibilidad.

- **Replicación de datos entre zonas:** En general, Data Factory es un servicio sin estado, por lo que no es necesario replicar ningún estado entre zonas.

  Sin embargo, los desencadenadores de ventanas deslizantes contienen estado, que podría no estar completamente replicado entre zonas.

Comportamiento durante un fallo de zona

En esta sección se describe qué puede ocurrir cuando los recursos de Data Factory están configurados para la redundancia de zona y se produce una interrupción en la zona de disponibilidad.

- **Detección y respuesta:** La plataforma data Factory es responsable de detectar un error en una zona de disponibilidad y responder. No es necesario hacer nada para iniciar una conmutación por error de zona en las canalizaciones u otros componentes.

- **Notificación**: Microsoft no le notifica automáticamente cuando una zona está inactiva. Sin embargo, puede usar Azure Resource Health para supervisar el estado de un recurso individual y puede configurar alertas de Resource Health para notificarle problemas. También puede usar Azure Service Health para comprender el estado general del servicio, incluidos los errores de zona, y puede configurar alertas de Service Health para notificarle problemas.

- **Solicitudes activas:** cualquier canalización y desencadenador en proceso sigue ejecutándose, y no experimentará ninguna interrupción inmediata debido a un error en la zona. Sin embargo, las actividades en curso durante un error de zona pueden producir un error y reiniciarse. Es importante diseñar actividades para que sean idempotentes, lo que

les ayuda a recuperarse de errores de zona y de otros errores. Para obtener más información, consulte Resistencia a errores transitorios.

- **Tiempo de inactividad esperado:** No se espera ningún tiempo de inactividad durante un error de zona.
- **Pérdida de datos esperada:** En general, Data Factory es un servicio sin estado, por lo que no se espera ninguna pérdida de datos durante un error de zona.

  Pero si usa un desencadenador periódico, el estado del desencadenador podría perderse después de un error de zona. Debe reiniciar o volver a ejecutar los triggers que estaban en ejecución en el momento del fallo de zona.

Recuperación de zona

Cuando se recupera la zona de disponibilidad, Data Factory vuelve automáticamente a la zona original. No es necesario hacer nada para iniciar una conmutación por recuperación de zona en las canalizaciones u otros componentes.

Sin embargo, si usa una SHIR, es posible que tenga que reiniciar los recursos de cómputo si se han detenido.

Prueba de fallos de zona

Para el servicio principal, así como para los Azure y los Azure-SSIS IR, Data Factory administra el enrutamiento del tráfico, la conmutación por error y la conmutación por recuperación para los recursos con redundancia de zona. Dado que esta característica está totalmente administrada, no es necesario iniciar ni validar los procesos de error de zona de disponibilidad.

En el caso de los SHIR, puede usar Azure Chaos Studio para simular un error de zona de disponibilidad en una máquina virtual de Azure.

Resistencia a errores en toda la región

Los recursos de Data Factory se implementan en una sola región de Azure. Si la región deja de estar disponible, la factoría de datos tampoco está disponible. Sin embargo, hay enfoques que puede usar para ayudar a garantizar la resistencia a las interrupciones de la región. Estos enfoques dependen de si la factoría de datos está en una región emparejada o no emparejada y en sus requisitos y configuración específicos.

Conmutación por error administrada por Microsoft en una región emparejada

Data Factory admite la conmutación por error administrada por Microsoft para factorías de datos en regiones emparejadas, excepto Sur de Brasil y Sudeste de Asia. En el improbable caso de un error prolongado en la región, Microsoft podría iniciar una conmutación por error regional de la instancia de Data Factory.

Debido a los requisitos de residencia de datos en el Sur de Brasil y sudeste de Asia, los datos de Data Factory solo se almacenan en la región local mediante el almacenamiento con redundancia de zona de Azure Storage. Para sudeste asiático, todos los datos se almacenan en Singapur. Para el Sur de Brasil, todos los datos se almacenan en Brasil.

En el caso de las factorías de datos en regiones no emparejadas o en el Sur de Brasil o Sudeste de Asia, Microsoft no realiza la conmutación por error regional en su nombre.

> **ⓘ Importante**
>
> Microsoft inicia la conmutación por error administrada por Microsoft. Es probable que se produzca después de un retraso significativo y se realice con el mejor esfuerzo. También hay algunas excepciones a este proceso. Es posible que experimente alguna pérdida de los metadatos de la factoría de datos. La conmutación por error de los recursos de Data Factory puede producirse en un momento diferente del tiempo de conmutación por error de otros servicios de Azure.
>
> Si necesita ser resistente a interrupciones de regiones, considere la posibilidad de usar una de las **soluciones personalizadas de varias regiones para lograr resistencia**.

Conmutación por error de los IR

Para prepararse para una conmutación por error, puede haber algunas consideraciones adicionales, en función del entorno de ejecución de integración utilizado.

- Puede configurar *Azure IR* para resolver automáticamente la región que usa. Si la región está establecida para *resolverse automáticamente* y hay una interrupción en la región principal, el Azure IR cambia automáticamente a la región emparejada. Esta conmutación por error está sujeta a limitaciones. Para configurar la región de Azure IR para la implementación de la actividad o el envío en la configuración de IR, establezca la región para *resolverse automáticamente*.
- La conmutación por error de *Azure-SSIS IR* se administra por separado en la conmutación por error administrada por Microsoft de Data Factory. Para obtener más información,

consulte Soluciones personalizadas de varias regiones para lograr resistencia.

- *Una SHIR* se ejecuta en la infraestructura por la que es responsable, por lo que una conmutación por error administrada por Microsoft no se aplica a las SHIR. Para obtener más información, consulte Soluciones personalizadas de varias regiones para lograr resistencia.

Reconfiguración posterior a la conmutación por error

Una vez completada una conmutación por error administrada por Microsoft, puede acceder a la canalización de Data Factory en la región emparejada. Sin embargo, una vez completada la conmutación por error, es posible que tenga que realizar alguna reconfiguración para los IR u otros componentes. Este proceso incluye volver a establecer la configuración de red.

Soluciones personalizadas de varias regiones para la resistencia

Si necesita que las canalizaciones sean resistentes a las interrupciones regionales y necesita controlar el proceso de conmutación por error, considere la posibilidad de usar una canalización controlada por metadatos.

- **Configure el control de código fuente para que Data Factory** realice un seguimiento y audite los cambios realizados en los metadatos. Puede usar este enfoque para acceder a los archivos JSON de metadatos para canalizaciones, conjuntos de datos, servicios vinculados y desencadenadores. Data Factory admite diferentes tipos de repositorio de Git, como Azure DevOps y GitHub. Para más información, consulte Control de código fuente en Data Factory.

- **Utilice un sistema de integración continua y entrega continua (CI/CD),** como Azure DevOps, para gestionar los metadatos de la canalización y las implementaciones. Puede usar CI/CD para restaurar rápidamente las operaciones en una instancia de otra región. Si una región no está disponible, puede aprovisionar una nueva factoría de datos manualmente o a través de la automatización. Una vez creada la nueva factoría de datos, puede restaurar las canalizaciones, los conjuntos de datos y los servicios vinculados JSON desde el repositorio de Git existente. Para más información, consulte Continuidad empresarial y recuperación ante desastres (BCDR) para canalizaciones de Data Factory y Azure Synapse Analytics.

Dependiendo del IR que utilice, puede haber otras consideraciones.

- Un *Azure-SSIS IR* usa una base de datos almacenada en Azure SQL Database o Azure SQL Managed Instance. Puede configurar la replicación geográfica o un grupo de

conmutación por error para esta base de datos. La base de datos Azure-SSIS se encuentra en una región primaria de Azure que tiene acceso de lectura y escritura. La base de datos se replica continuamente en una región secundaria que tiene acceso de solo lectura. Si la región primaria no está disponible, se desencadena una conmutación por error, lo que hace que las bases de datos principales y secundarias intercambien roles.

También puede configurar un par de Azure-SSIS IR en espera dual que funcione en sincronización con el grupo de conmutación por error de Azure SQL Database o SQL Managed Instance.

Para obtener más información, consulte Configurar un IR de Azure-SSIS para BCDR.

- *Una SHIR* se ejecuta en la infraestructura que administra. Si el SHIR se implementa en una máquina virtual de Azure, puede usar Azure Site Recovery para desencadenar la conmutación por error de la máquina virtual en otra región.

Copias de seguridad y restauración

Data Factory admite CI/CD a través de la integración del control de código fuente, de modo que pueda realizar una copia de seguridad de los metadatos de una instancia de Data Factory. Las canalizaciones de CI/CD implementan estos metadatos sin problemas en un nuevo entorno. Para más información, consulte CI/CD en Data Factory.

Acuerdo de nivel de servicio

El contrato de nivel de servicio (SLA) para los servicios de Azure describe la disponibilidad esperada de cada servicio y las condiciones que la solución deberá cumplir para lograr esa expectativa de disponibilidad. Para obtener más información, consulte Acuerdos de Nivel de Servicio para servicios en línea.

Data Factory proporciona acuerdos de nivel de servicio de disponibilidad independientes para:

- La tasa de éxito de las llamadas API que hagas, como aquellas para gestionar tu fábrica de datos.
- Número de ejecuciones de actividad que comienzan a ejecutarse.

Las ejecuciones de actividad se pueden retrasar brevemente y requieren que se cumplan todas las dependencias para ejecutar el trabajo.

Contenido relacionado

- Confiabilidad en Azure

Last updated on 22/01/2026

Confiabilidad en Azure Event Grid

Azure Event Grid es un servicio de mensajería totalmente administrado que permite la comunicación basada en eventos entre servicios y aplicaciones. Normalmente se usa para crear arquitecturas controladas por eventos e integrar servicios de Azure con aplicaciones personalizadas.

Cuando se usa Azure, la confiabilidad es una responsabilidad compartida. Microsoft proporciona una variedad de funcionalidades para admitir resistencia y recuperación. Es responsable de comprender cómo funcionan esas funcionalidades dentro de todos los servicios que usa y de seleccionar las funcionalidades que necesita para cumplir los objetivos empresariales y los objetivos de tiempo de actividad.

En este artículo se describe cómo hacer que Azure Event Grid sea resistente a varias posibles interrupciones y problemas, incluidos errores transitorios, errores de zona de disponibilidad y errores en toda la región. También resalta información clave sobre el acuerdo de nivel de servicio (SLA) de Azure Event Grid.

Recomendaciones de implementación de producción

Azure Well-Architected Framework proporciona recomendaciones sobre confiabilidad, seguridad, costo, operaciones y rendimiento. Para comprender cómo estas áreas influyen entre sí y contribuir a una solución confiable de Azure Event Grid, consulte Guía de Azure Well-Architected Framework para Azure Event Grid.

Introducción a la arquitectura de confiabilidad

En esta sección se describen algunos de los aspectos importantes de cómo funciona el servicio que es más relevante desde una perspectiva de confiabilidad. En la sección se presenta la arquitectura lógica, que incluye algunos de los recursos y características que se implementan y usan. También se describe la arquitectura física, que proporciona detalles sobre cómo funciona el servicio en segundo plano.

Arquitectura lógica

Azure Event Grid enruta eventos de *publicadores* de eventos a *consumidores de eventos*. Las aplicaciones de cliente y los servicios de Azure las usan para emitir y consumir eventos, como las notificaciones cuando se crean, actualizan o eliminan recursos.

Event Grid admite varios tipos de recursos y modelos de implementación:

- **Los temas** son las entidades principales que reciben y almacenan eventos.

  Los servicios de Azure crean automáticamente **los temas del sistema** para emitir eventos para tipos de recursos específicos de Azure. **Los temas personalizados** son creados y administrados por usted.

  Los temas pueden admitir la entrega push y pull.

- **Los dominios** de eventos agrupan varios temas personalizados en un único punto de conexión, lo que simplifica la publicación de eventos. Para más información, consulte Dominios de eventos para administrar temas de Event Grid.

- **Namespaces** se utilizan con el nivel estándar y proporcionan un contenedor para varios recursos de Event Grid. Para más información, consulte Conceptos del espacio de nombres de Azure Event Grid.

Event Grid admite varios niveles, incluido el nivel básico y el nivel estándar. Estos niveles proporcionan diferentes funcionalidades y difieren en cómo se implementan y administran los recursos. Para obtener más información, consulte Elegir el nivel correcto de Event Grid para la solución.

Arquitectura física

Azure Event Grid es un servicio totalmente administrado. Microsoft administra la infraestructura subyacente, incluidos los recursos de proceso y almacenamiento. En las regiones admitidas, Event Grid distribuye automáticamente los recursos entre zonas de disponibilidad para proporcionar redundancia de zona integrada.

Resistencia a errores transitorios

Los errores transitorios son errores breves e intermitentes en los componentes. Se producen con frecuencia en un entorno distribuido como la nube y son una parte normal de las operaciones. Los errores transitorios se corrigen después de un breve período de tiempo. Es importante que las aplicaciones puedan controlar errores transitorios, normalmente mediante el reintento de solicitudes afectadas.

Todas las aplicaciones hospedadas en la nube deben seguir las instrucciones de control de errores transitorios de Azure cuando se comunican con cualquier API, bases de datos y otros componentes hospedados en la nube. Para obtener más información, consulte Recomendaciones para controlar errores transitorios.

Al usar Event Grid, tenga en cuenta los procedimientos siguientes para asegurarse de que la solución es resistente a errores transitorios:

- **Publicadores de eventos:** Cuando una aplicación cliente publica eventos en Event Grid, es responsable de controlar errores transitorios. Las aplicaciones deben implementar lógica de reintento al publicar eventos. Para obtener más información, consulte Solución de problemas de conexión transitorios.

  Se recomienda usar los SDK del plano de datos de Event Grid, que proporcionan automáticamente el control de errores transitorios.

- **Consumidores de eventos:** Event Grid entrega eventos a destinos configurados. Para estas conexiones salientes, puede configurar políticas de reintento en las suscripciones de eventos. Estas directivas definen la frecuencia y el tiempo durante el que Event Grid reintenta la entrega cuando se producen errores, incluidos los errores transitorios. Para obtener más información, consulte Entrega push de mensajes y reintento con temas bajo espacios de nombres.

- **Idempotency:** Se recomienda diseñar la arquitectura de eventos para *la idempoencia*, lo que significa que los eventos se pueden recibir y procesar de forma segura varias veces. Por ejemplo, si se produce un error transitorio u otro problema cuando la aplicación procesa un evento, un enfoque idempotente significa que la aplicación puede volver a procesar el mensaje y recuperarse.

  Usted es responsable de diseñar la arquitectura de eventos y la aplicación para admitir la idempotencia. Para obtener información general, consulte Idempotency.

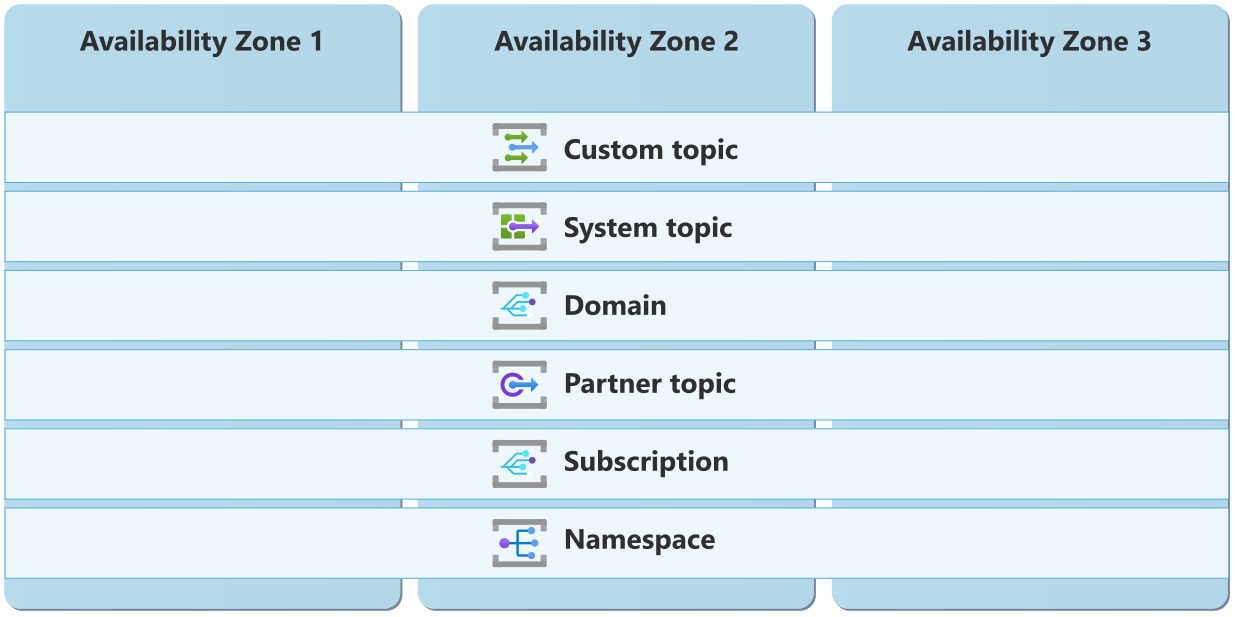
- **Mensajes fallidos:** Event Grid admite *mensajes fallidos* para eventos que no se pueden entregar, lo que ayuda a conservar los datos durante errores más duraderos en los consumidores de eventos. Para más información, consulte Mensajes fallidos para suscripciones de eventos a temas de espacios de nombres en Azure Event Grid.

Resistencia a errores de zona de disponibilidad

Las zonas de disponibilidad son grupos físicamente independientes de centros de datos dentro de una región de Azure. Cuando una zona falla, los servicios pueden transferirse a una de las zonas restantes.

Los recursos de Azure Event Grid tienen redundancia de zona en regiones que admiten zonas de disponibilidad. La redundancia de zona significa que incluso cuando una zona de disponibilidad tiene un problema, los recursos de Event Grid siguen funcionando mediante la infraestructura de otras zonas. Los datos de eventos se replican automáticamente en tres zonas de disponibilidad para garantizar la resiliencia dentro de la región, y Event Grid se restaura

automáticamente durante una interrupción a nivel de zona. No es necesario habilitar ni configurar esta funcionalidad.



Requisitos

Compatibilidad con regiones: La redundancia de zona está disponible en todas las regiones de Azure que admiten zonas de disponibilidad.

Costo

No hay ningún costo adicional para la redundancia de zona. No se puede habilitar ni deshabilitar esta característica; se incluye de forma predeterminada en las regiones admitidas.

Configurar soporte de zonas de disponibilidad

No se requiere ninguna configuración. Todos los recursos de Event Grid en las regiones admitidas son automáticamente con redundancia de zona.

Comportamiento cuando todas las zonas están en buen estado

En esta sección se describe qué esperar cuando un recurso de Event Grid tiene redundancia de zona y todas las zonas están plenamente operativas.

- **Operación entre zonas:** Event Grid funciona en un modelo activo-activo entre zonas de disponibilidad. Las conexiones de cliente se equilibran automáticamente entre zonas y el

servicio enruta las operaciones a la infraestructura de mensajería disponible independientemente de la zona.

- **Replicación de datos entre zonas:** Event Grid replica automáticamente los metadatos y los datos de eventos en las zonas de disponibilidad para mantener la resistencia.

Comportamiento durante un fallo de zona

En esta sección se describe qué esperar cuando un recurso de Event Grid tiene redundancia de zona y hay una interrupción en una de las zonas.

- **Detección y respuesta:** Event Grid detecta automáticamente errores de zona e inicia la conmutación por error a zonas correctas. No es necesario hacer nada para iniciar una conmutación por error de la zona.

- **Notificación:** Microsoft no le notifica automáticamente cuando una zona está inactiva. Sin embargo, puede usar Azure Service Health para comprender el estado general del servicio, incluidos los errores de zona, y puede configurar alertas de Service Health para notificarle problemas.

- **Solicitudes activas:** Durante un error de zona, Event Grid podría quitar las solicitudes activas. Si los clientes controlan correctamente los errores transitorios, siempre y cuando reintenten tras un breve intervalo, normalmente evitan un impacto significativo.

- **Pérdida de datos esperada:** El modelo de redundancia de zona de Event Grid está diseñado para permitir la resistencia a los errores de zona con un impacto mínimo. Sin embargo, durante un fallo de zona, existe cierta posibilidad de pérdida de datos.

  Si necesita asegurarse de que la aplicación no pierde datos incluso durante un error de zona, debe hacer lo siguiente:
  - Diseñe sus productores y consumidores de eventos para que sigan las recomendaciones de control de errores transitorios, incluidos los reintentos y la idempotencia.
  - Planee la durabilidad de eventos en el origen o en un almacén de eventos duradero.

- **Tiempo de inactividad esperado:** Un error de zona puede provocar unos segundos de tiempo de inactividad. Si los clientes controlan correctamente los errores transitorios, siempre y cuando reintenten tras un breve intervalo, normalmente evitan un impacto significativo.

- **Reenrutamiento del tráfico:** Event Grid detecta la pérdida de la zona y redirige automáticamente las nuevas solicitudes a la infraestructura en una de las zonas de disponibilidad correctas.

Recuperación de zona

Cuando se recupera la zona afectada, Event Grid lo vuelve a reinsertar automáticamente en el servicio sin necesidad de la acción del cliente. A continuación, la zona recuperada acepta nuevas conexiones y procesa mensajes junto con las demás zonas. Los datos que se replican en zonas supervivientes durante la interrupción permanecen intactos y la replicación normal se reanuda en todas las zonas. No es necesario tomar medidas para la recuperación o la reintegración de zonas.

Prueba de fallos de zona

Event Grid administra el enrutamiento de tráfico, la conmutación por error y la recuperación de zona para los fallos de zona, por lo que no es necesario validar los procesos de fallos de zona de disponibilidad ni proporcionar más información adicional.

Resistencia a errores en toda la región

Los recursos de Event Grid se implementan en una sola región. Si se produce un error en toda la región, los recursos de Event Grid no están disponibles.

En regiones de Azure emparejadas, Event Grid proporciona una recuperación ante desastres geográfica limitada para los metadatos de los recursos de Event Grid. También puede diseñar y crear su propia solución multi-región, que puede respaldar la planificación de la recuperación ante desastres. En la tabla siguiente se muestra cómo diferentes tipos de recursos de Event Grid admiten cada modelo:

Expandir tabla

Recurso de Event Grid	Admite la recuperación ante desastres geográficos	Admite la solución personalizada.
Temas personalizados	Compatible	Compatible
Temas del sistema	Habilitada automáticamente	No está soportado
Dominios	Compatible	Compatible
Espacios de nombres	No está soportado	Compatible
Espacios de nombres de asociados	No está soportado	Compatible

Recuperación de desastres geográficos

La recuperación ante desastres geográfica replica los metadatos del Event Grid en la región emparejada de la región primaria para los recursos compatibles. Los datos de eventos no se replican.

La recuperación ante desastres geográficas está diseñada como una reserva administrada por Microsoft para interrupciones regionales graves y no está diseñada para proporcionar tiempos de recuperación rápidos o predecibles. La conmutación por error iniciada por Microsoft la ejecuta Microsoft en situaciones poco frecuentes para conmutar por error recursos de Event Grid de una región afectada a la región emparejada geográficamente correspondiente. Microsoft se reserva el derecho de determinar cuándo se ejercerá esta opción. Este mecanismo no implica el consentimiento del cliente antes de que se redirija el tráfico.

> **ⓘ Importante**
>
> Microsoft inicia la conmutación por error administrada por Microsoft. Es probable que se produzca después de un retraso significativo y se realice con el mejor esfuerzo. La conmutación por error de los recursos de Event Grid puede producirse en un momento diferente del tiempo de conmutación por error de otros servicios de Azure.
>
> Si necesita ser resistente a interrupciones de regiones, considere la posibilidad de usar una de las soluciones personalizadas de varias regiones para lograr resistencia.

Opcionalmente, puede deshabilitar la recuperación ante desastres geográfica y usar su propia solución personalizada de varias regiones que cumpla sus requisitos para la selección de regiones, el tiempo de conmutación por error, etc. Cuando deshabilitas la recuperación ante desastres geográfica, Microsoft no replica datos de eventos a otra región.

Esta característica no está disponible en regiones sin una región emparejada.

Requisitos

- **Compatibilidad con regiones:** La recuperación ante desastres geográfica solo está disponible en regiones de Azure que tienen una región emparejada.
- **Tipos de recursos:** Los temas y dominios personalizados admiten la recuperación ante desastres geográfica. Automáticamente se habilitan los tópicos del sistema para la recuperación de desastres geográfica. No se admiten otros tipos de recursos, como espacios de nombres y espacios de nombres asociados.

Costo

No hay ningún costo adicional para la recuperación ante desastres geográfica.

Configuración de la compatibilidad con varias regiones

En las regiones admitidas, los temas del sistema se configuran automáticamente para la recuperación ante desastres geográficos. Para otros tipos de recursos de Event Grid:

- **Para habilitar la recuperación ante desastres geográfica:** Actualice la configuración del tema o dominio y seleccione **Cross-Geo** (valor predeterminado).
- **Para deshabilitar la recuperación de diasteres geográficos:** Actualice la configuración del tema o dominio y seleccione **Regional**.

Comportamiento cuando todas las regiones están en buen estado

En esta sección se describe qué esperar cuando un recurso de Event Grid está configurado para la recuperación ante desastres geográficas y todas las regiones están operativas.

- **Operación entre regiones:** Todo el tráfico se enruta a la región primaria.
- **Replicación de datos entre regiones:** Los metadatos se replican sincrónicamente en la región emparejada. Los datos de eventos no se replican.

Comportamiento durante una falla de región

En esta sección se describe qué esperar cuando un recurso de Event Grid está configurado para la recuperación ante desastres geográfica y hay una interrupción en la región primaria.

- **Detección y respuesta:** Microsoft detecta errores de región y determina si y cuándo iniciar la conmutación por error.
- **Notificación:** Microsoft no le notifica automáticamente cuando una región está inactiva. Sin embargo, puede usar Azure Service Health para comprender el estado general del servicio, incluidos los errores de región, y puede configurar alertas de Service Health para notificarle problemas.
- **Solicitudes activas:** Se finalizan las solicitudes activas a la región primaria. Las aplicaciones cliente deben reintentar estas solicitudes una vez completada la conmutación por error.
- **Pérdida de datos esperada:**
  - *Metadatos:* Event Grid conserva los metadatos durante la conmutación por error. Dado que todos los cambios de metadatos se replican sincrónicamente, no se espera ninguna pérdida de metadatos.

- *Datos de eventos:* Los datos de eventos de la región primaria no están disponibles y podrían perderse si la región no es recuperable.

  Una vez que se produce una conmutación por error, los nuevos datos se procesan desde la región emparejada. Los eventos no procesados se envían desde la región primaria en cuanto se mitiga la interrupción. Si la recuperación de la región primaria requiere un tiempo mayor que el valor de período de vida establecido en eventos, es posible que se quiten los datos de la región primaria. Para mitigar esta pérdida de datos, se recomienda configurar un destino de mensajes fallidos para una suscripción de eventos.

  Si la región afectada se pierde y no se puede recuperar, habrá cierta pérdida de datos. En el mejor de los casos, el consumidor se actualiza conforme al ritmo de publicación y solo se pierden unos segundos de datos. El peor escenario sería cuando el consumidor no procesa activamente eventos y con un tiempo máximo de vida de 24 horas, la pérdida de datos puede ser de hasta 24 horas.

  > ⓘ **Nota:**
  >
  > Event Grid no puede garantizar la retención de datos durante una interrupción en la región. Si necesita una retención garantizada, debe diseñar la aplicación para almacenar eventos de forma duradera en otro almacén de datos.

- **Tiempo de inactividad esperado:** La cantidad de tiempo de inactividad depende de la gravedad de la interrupción y del tiempo necesario para que Microsoft evalúe e inicie la conmutación por error. Debe esperar que el tiempo de inactividad sea al menos una hora y quizás más tiempo.

  Una vez iniciada la conmutación por error, en un plazo de cinco minutos, Event Grid comienza a aceptar el tráfico de temas y suscripciones, incluidas las operaciones de creación, actualización y eliminación.

- **Redistribución:** Una vez completada la conmutación por error, el tráfico se enruta automáticamente a la región secundaria.

Recuperación de regiones

Microsoft administra la recuperación de regiones y el proceso de recuperación depende del escenario de interrupción específico. Por lo general, la conmutación por error se considera una operación unidireccional.

Prueba de fallos de región

Azure Event Grid administra el enrutamiento del tráfico, la conmutación por error y la recuperación en caso de desastres geográficos. No es necesario iniciar nada. Dado que esta característica está totalmente administrada, no es necesario validar los procesos de error de región.

Soluciones personalizadas de varias regiones para la resistencia

Es posible que desee deshabilitar o no confiar en la conmutación por error iniciada por Microsoft por cualquiera de estos motivos:

- Necesita que los datos de eventos se repliquen entre regiones y no solo los metadatos.
- Debe garantizar un enfoque o un tiempo de conmutación por error específicos. La conmutación por error iniciada por Microsoft se realiza de la mejor manera posible.
- La región no está emparejada con otra región de Azure.
- El emparejamiento de tu región no cumple los requisitos de localización de datos de tu organización.

Para obtener niveles más altos de control y previsibilidad, puede implementar arquitecturas personalizadas de varias regiones. Este enfoque implica implementar recursos independientes de Event Grid en varias regiones y administrar la conmutación por error en el nivel de aplicación. Al usar este modelo, es responsable de implementar, configurar y mantener los recursos sincronizados entre regiones.

Tenga en cuenta los siguientes factores al diseñar una solución de varias regiones:

- **Replicación:** Debe implementar un proceso personalizado para replicar los recursos de Event Grid y su configuración entre las regiones primarias y secundarias. Recuerde replicar identidades de cliente, certificados de autoridad de certificación, grupos de clientes, espacios de temas y vinculaciones de permisos, si procede. Puede decidir si implementar la replicación manual o automatizada.

- **Enfoques de conmutación por error:** Puede elegir si desea crear una solución activa-activa o activa-pasiva:
  - *Las soluciones activa-activa* se pueden lograr mediante la replicación de los metadatos y el equilibrio de carga entre los espacios de nombres.
  - *Las soluciones activas-pasivas* se pueden lograr mediante la replicación de los metadatos para mantener el espacio de nombres secundario listo para que cuando el espacio de nombres principal no esté disponible, el tráfico se puede dirigir al espacio de nombres secundario.

- **Supervisión del estado:** Puede usar las API de mantenimiento integradas proporcionadas por Event Grid para supervisar el estado de los temas.

  Las aplicaciones cliente deben detectar errores de una región y enrutar eventos a otra región adecuada.

  Como alternativa, puede implementar un *servicio de conserjería* que dirija a los clientes a los puntos de conexión principales o secundarios para sus temas o espacios de nombres realizando comprobaciones de estado en esos puntos de conexión. El servicio de conserjería puede ser una aplicación web que se replica geográficamente y se mantiene accesible mediante técnicas de redirección de DNS o servicios como Azure Traffic Manager.

Para más información sobre un enfoque, incluido el código de ejemplo, consulte Implementación de conmutación por error del lado cliente en Azure Event Grid.

Copias de seguridad y restauración

Event Grid es principalmente un servicio de enrutamiento de eventos y no tiene características nativas de copia de seguridad o restauración.

Si necesita implementar funcionalidades de copia de seguridad o si tiene necesidades de retención a largo plazo, se recomienda realizar el archivado en la aplicación. Para ello, debe construir una lógica para enrutar o copiar los eventos en un almacén duradero, como Azure Blob Storage, en paralelo con la ruta de acceso de entrega principal. Si los sistemas posteriores no están disponibles, su aplicación puede usar el archivo para reproducir los eventos.

Resistencia al mantenimiento del servicio

Microsoft aplica periódicamente actualizaciones de servicio y realiza otro mantenimiento. La plataforma Azure controla estas actividades automáticamente, lo que garantiza que el mantenimiento sea transparente y sin problemas. No se espera ningún tiempo de inactividad durante los eventos de mantenimiento a menos que se le haya informado a través del mantenimiento planeado de Azure Service Health.

Acuerdo de nivel de servicio

El acuerdo de nivel de servicio (SLA) para Azure servicios describe la disponibilidad esperada de cada servicio y las condiciones que la solución debe cumplir para lograr esa expectativa de disponibilidad. Para obtener más información, consulte Acuerdos de Nivel de Servicio para servicios en línea.

El Acuerdo de Nivel de Servicio de disponibilidad de Event Grid abarca la publicación de eventos.

Contenido relacionado

- Confiabilidad en Azure
- Guía de Azure Well-Architected Framework para Azure Event Grid
- Documentación de Azure Event Grid

ⓘ **Nota:** El autor ha creado este artículo con ayuda de inteligencia artificial. Más información

---

Last updated on 17/03/2026

Confiabilidad en Azure Event Hubs

Azure Event Hubs es un servicio en la nube nativo que puede transmitir millones de eventos por segundo con baja latencia, desde cualquier origen a cualquier destino. Use Event Hubs para ingerir y almacenar datos de streaming e integrarlos con aplicaciones cliente creadas para Apache Kafka o aplicaciones que usan los SDK de cliente de Event Hubs.

Cuando se usa Azure, la confiabilidad es una responsabilidad compartida. Microsoft proporciona una variedad de funcionalidades para admitir resistencia y recuperación. Es responsable de comprender cómo funcionan esas funcionalidades dentro de todos los servicios que usa y de seleccionar las funcionalidades que necesita para cumplir los objetivos empresariales y los objetivos de tiempo de actividad.

En este artículo se describe cómo Event Hubs es resistente a una variedad de posibles interrupciones y problemas, y cómo puede configurarlo para que sea resistente a otros usuarios, incluidos errores transitorios, interrupciones de zona de disponibilidad y interrupciones de región. También se describen las opciones de copia de seguridad y recuperación, y se resalta cierta información clave sobre el contrato de nivel de servicio (SLA) de Azure Event Hubs.

Recomendaciones de implementación de producción

Para obtener información sobre cómo implementar Event Hubs para admitir los requisitos de confiabilidad de la solución y comprender cómo afecta la confiabilidad a otros aspectos de la arquitectura, consulte Procedimientos recomendados de arquitectura para Event Hubs en Azure Well-Architected Framework.

Introducción a la arquitectura de confiabilidad

En esta sección se describen aspectos importantes sobre cómo funciona Event Hubs desde una perspectiva de confiabilidad. Presenta la arquitectura lógica, que incluye recursos y características que se implementan y usan. También se explica la arquitectura física, que proporciona detalles sobre cómo el servicio administra las operaciones internamente.

Arquitectura lógica

Un *espacio de nombres* de Event Hubs actúa como contenedor de administración para uno o varios centros de eventos. Configure el servicio, como asignar capacidad de streaming,

configurar la seguridad de red y habilitar la resistencia geográfica y la recuperación ante desastres geográficas, en el nivel de espacio de nombres.

Dentro de un espacio de nombres, puede organizar eventos en un *centro de eventos*. El ecosistema de Apache® Kafka hace referencia a este tipo de entidad como *tema*. El centro de eventos o tema es un registro distribuido de solo adición de tus eventos.

Cada centro de eventos contiene una o varias *particiones*, que son registros de eventos secuenciales. Un centro de eventos puede usar varias particiones para realizar el procesamiento paralelo y el escalado horizontal. Event Hubs solo garantiza la ordenación dentro de una sola partición. La creación de particiones desempeña un papel clave en el diseño de confiabilidad de la aplicación. Al diseñar la aplicación, realice un equilibrio entre maximizar la disponibilidad y la coherencia. Para maximizar el tiempo de actividad de la mayoría de las aplicaciones, evite direccionar particiones directamente desde las aplicaciones cliente. Para más información, consulte Disponibilidad y coherencia en Event Hubs.

Los consumidores que leen desde el centro de eventos pueden leer sus eventos secuencialmente manteniendo su propio *punto de control*, que identifica el último evento que reciben.

Para más información sobre las particiones y otros conceptos fundamentales de Event Hubs, consulte Características y terminología en Event Hubs.

Arquitectura física

En la arquitectura física, un espacio de nombres de Event Hubs se ejecuta dentro de un *clúster*. Un clúster proporciona los recursos de proceso y almacenamiento subyacentes. La mayoría de los espacios de nombres se ejecutan en clústeres compartidos por otros clientes de Azure. Cuando se usa el nivel Premium, el espacio de nombres se asigna a recursos dedicados dentro de un clúster compartido. Cuando usas el nivel Dedicado, un clúster se asigna exclusivamente a tus espacios de nombres. Para más información sobre los clústeres dedicados, consulte Introducción al nivel dedicado de Event Hubs. Independientemente del tipo de clúster y del nivel, Microsoft administra los clústeres y sus máquinas virtuales y almacenamiento subyacentes.

Para lograr redundancia, cada clúster tiene varias réplicas que procesan solicitudes de lectura y escritura. Para optimizar la alta disponibilidad y el rendimiento, todos los datos se almacenan en tres réplicas de almacenamiento. Para escalar los recursos de proceso del espacio de nombres, implemente unidades de rendimiento (TUs), unidades de procesamiento (PUs) o unidades de capacidad (CUs), en función del nivel. Para más información, consulte Escalado con Event Hubs.

Los clústeres abarcan varias máquinas físicas y bastidores, lo que reduce el riesgo de errores catastróficos que afectan al espacio de nombres. En las regiones que tienen zonas de disponibilidad, los clústeres se extienden entre centros de datos físicos independientes. Para obtener más información, consulte Resilience to availability zone failures (Resistencia a errores de zona de disponibilidad).

Resistencia a errores transitorios

Los errores transitorios son errores breves e intermitentes en los componentes. Se producen con frecuencia en un entorno distribuido como la nube y son una parte normal de las operaciones. Los errores transitorios se corrigen después de un breve período de tiempo. Es importante que las aplicaciones puedan controlar errores transitorios, normalmente mediante el reintento de solicitudes afectadas.

Todas las aplicaciones hospedadas en la nube deben seguir las instrucciones de control de errores transitorios de Azure cuando se comunican con cualquier API, bases de datos y otros componentes hospedados en la nube. Para obtener más información, consulte Recomendaciones para controlar errores transitorios.

Event Hubs implementa mecanismos transparentes de detección de errores y conmutación por error para que el servicio siga funcionando dentro de los niveles de servicio garantizados, normalmente sin interrupciones perceptibles cuando se produzcan errores.

Al diseñar aplicaciones cliente para trabajar con Event Hubs, siga estas instrucciones:

- **Utilice políticas de reintento integradas.** Los SDK de Event Hubs y Apache Kafka reintentan automáticamente las operaciones para errores que se pueden reintentar, como tiempos de espera en la red, respuestas de limitación, o cuando el servidor está ocupado. Para evitar sobrecargar innecesariamente el servicio, implementan el retroceso exponencial de forma predeterminada.

- **Configure los valores de tiempo de espera adecuados** en función de los requisitos de la aplicación. El tiempo de espera predeterminado suele ser de 60 segundos, pero puede ajustarlo en función de su escenario.

- **Implemente puntos de control** en el procesador de eventos para realizar un seguimiento del progreso y habilitar la recuperación desde la última posición procesada después de errores transitorios.

- **Use el procesamiento por lotes para las operaciones de envío** para mejorar el rendimiento y reducir el impacto de los problemas de red transitorios en mensajes individuales.

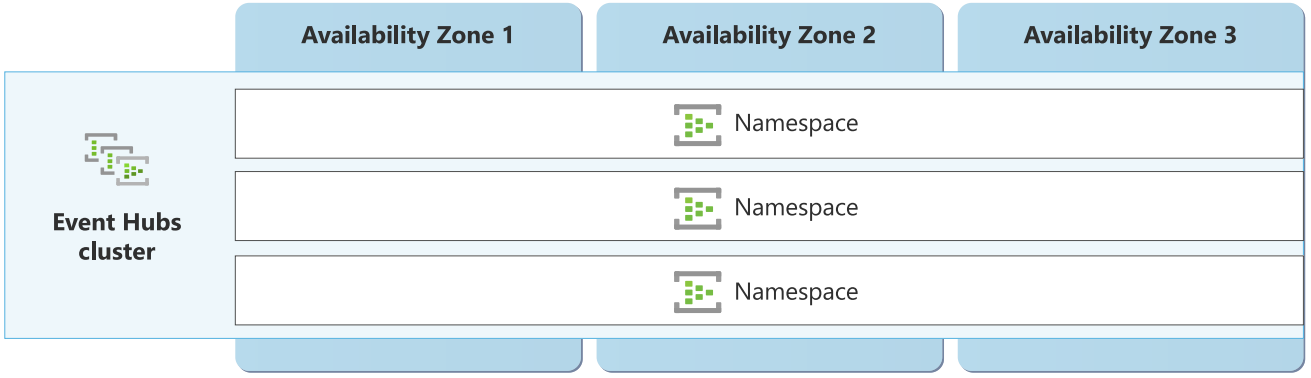
- **Use los SDK de Apache Kafka** si trabaja con el protocolo Kafka. Los SDK de Kafka también implementan directivas de reintento y otros procedimientos recomendados que ayudan a controlar errores transitorios.

Resistencia a errores de zona de disponibilidad

Las zonas de disponibilidad son grupos físicamente independientes de centros de datos dentro de una región de Azure. Cuando una zona falla, los servicios pueden transferirse a una de las zonas restantes.

Event Hubs admite implementaciones con redundancia de zona en todos los niveles de servicio. Al crear un espacio de nombres de Event Hubs en una región admitida, la redundancia de zona se habilita automáticamente sin costo adicional. Pero con el nivel Dedicado, las zonas de disponibilidad solo son compatibles con un mínimo de tres CUs. El modelo de implementación con redundancia de zona se aplica a todas las características de Event Hubs, incluida la compatibilidad con el protocolo Capture, Schema Registry y Kafka.

Event Hubs replica de forma transparente los datos de configuración, metadatos y eventos en tres zonas de disponibilidad de la región. La redundancia de zona proporciona conmutación automática por error sin requerir intervención de su parte. Todos los componentes de Event Hubs, incluidos el proceso, las redes y el almacenamiento, se replican entre zonas. Event Hubs tiene suficientes reservas de capacidad para controlar instantáneamente la pérdida completa de una zona. Incluso si una zona de disponibilidad completa deja de estar disponible, Event Hubs sigue funcionando sin pérdida de datos ni interrupción en las aplicaciones de streaming.



Soporte para regiones

Los espacios de nombres de Event Hubs con redundancia de zona se pueden implementar en cualquier región de Azure que admita zonas de disponibilidad.

Requisitos

- Los niveles Estándar y Premium admiten zonas de disponibilidad sin ninguna configuración adicional necesaria.
- En el nivel dedicado, las zonas de disponibilidad requieren un mínimo de tres CUs.

Cost

La redundancia de zona en Event Hubs no agrega ningún costo adicional.

Configurar soporte de zonas de disponibilidad

Los espacios de nombres de Event Hubs admiten automáticamente la redundancia de zona cuando se implementan en regiones admitidas. No es necesario realizar ninguna otra configuración.

Comportamiento cuando todas las zonas están en buen estado

Cuando los espacios de nombres de Event Hubs usan redundancia de zona y todas las zonas de disponibilidad funcionan normalmente, espere el siguiente comportamiento:

- **Enrutamiento de tráfico entre zonas:** Event Hubs funciona en un modelo activo-activo donde la infraestructura de tres zonas de disponibilidad procesa simultáneamente eventos entrantes.
- **Replicación de datos entre zonas:** Event Hubs usa la replicación sincrónica entre zonas de disponibilidad. Cuando un productor de eventos envía un evento, Event Hubs lo escribe en réplicas en varias zonas antes de confirmar al cliente que la operación de escritura se ha completado. Este enfoque garantiza una pérdida de datos cero, incluso si una zona completa deja de estar disponible. El enfoque de replicación sincrónica proporciona garantías de coherencia sólidas al tiempo que mantiene una baja latencia a través de protocolos de replicación optimizados.

Comportamiento durante un fallo de zona

Cuando los espacios de nombres de Event Hubs utilizan redundancia de zona y ocurre una interrupción en una zona de disponibilidad, se debe esperar el siguiente comportamiento:

- **Detección y respuesta:** Event Hubs es responsable de detectar automáticamente un error en una zona de disponibilidad. No es necesario iniciar una conmutación por error de zona.

- **Notificación:** Microsoft no le notifica automáticamente cuando una zona está inactiva. Sin embargo, puede usar Azure Service Health para comprender el estado general del servicio, incluidos los errores de zona, y puede configurar alertas de Service Health para notificarle problemas.
- **Solicitudes activas:** Durante un error de zona, Event Hubs podría quitar las solicitudes activas. Si los clientes controlan correctamente los errores transitorios mediante el reintento después de un breve período de tiempo, normalmente evitan un impacto significativo.
- **Pérdida de datos esperada:** No se produce ninguna pérdida de datos durante un error de zona porque Event Hubs replica de forma sincrónica eventos entre zonas antes de la confirmación.
- **Tiempo de inactividad esperado:** Un error de zona puede provocar unos segundos de tiempo de inactividad. Si los clientes controlan correctamente los errores transitorios mediante el reintento después de un breve período de tiempo, normalmente evitan un impacto significativo.
- **Reenrutamiento del tráfico:** Event Hubs detecta la pérdida de la zona y redirige automáticamente las nuevas solicitudes a otra réplica en una de las zonas de disponibilidad correctas.

  Los SDK de cliente de Event Hubs normalmente controlan la administración de conexiones y la lógica de reintento de forma transparente.

Recuperación de zona

Cuando se recupera una zona de disponibilidad, Event Hubs reintegra automáticamente la zona a la topología activa del servicio. La zona recuperada comienza a aceptar nuevas conexiones y procesar eventos junto con las demás zonas. Los datos que se han replicado en zonas supervivientes durante la interrupción permanecen intactos y la replicación sincrónica normal se reanuda en todas las zonas. No es necesario tomar medidas para la recuperación y la reintegración de zonas.

Prueba de fallos de zona

Event Hubs administra el enrutamiento del tráfico, la conmutación por error y la recuperación de zonas en caso de fallos de zona, por lo que no es necesario validar los procesos en caso de fallos de zona de disponibilidad ni proporcionar más intervención.

Resistencia a errores en toda la región

Event Hubs proporciona dos tipos de compatibilidad con varias regiones:

- La replicación geográfica (niveles Premium y Dedicado) proporciona replicación activa-activa de metadatos y datos de eventos entre una región primaria y una o varias regiones secundarias. Use la replicación geográfica para la mayoría de las aplicaciones que necesitan permanecer resistentes a las interrupciones de la región y tener una tolerancia baja para la pérdida de datos de eventos.

- La recuperación ante desastres geográfica de metadatos (nivel Estándar y superior) proporciona la replicación activa-pasiva de la configuración y los metadatos entre una región primaria y secundaria, pero no replica los datos de eventos. Use la recuperación ante desastres geográficas para las aplicaciones que pueden tolerar la pérdida de datos de eventos durante un escenario de desastre y que necesite reanudar rápidamente las operaciones en otra región.

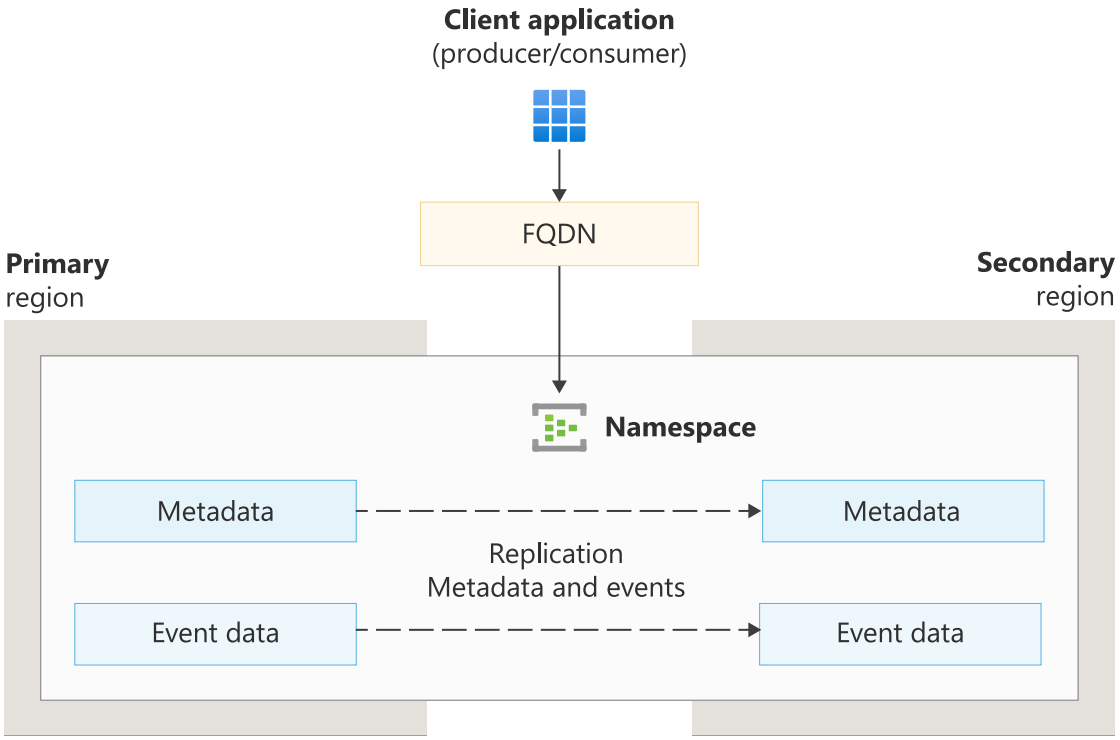
Tanto la replicación geográfica como la recuperación geográfica de desastres de metadatos requieren que usted inicie manualmente la conmutación por error o la promoción de una región secundaria para que se convierta en la nueva región primaria. Microsoft no realiza automáticamente la conmutación por error ni la promoción, incluso si la región primaria está inactiva.

Replicación geográfica

Los niveles Premium y Dedicado admiten la replicación geográfica. Esta característica replica ambos metadatos (como entidades, configuración y propiedades) y datos (como cargas de eventos) para el espacio de nombres. Tú configuras el enfoque de replicación para la configuración del espacio de nombres y los datos de eventos. Esta característica garantiza que los eventos permanecen disponibles en otra región y le permite cambiar a la región secundaria cuando sea necesario. También replica los metadatos y los datos del Registro de esquemas.

Use la replicación geográfica para escenarios que requieran resistencia a interrupciones de regiones y que tengan una tolerancia baja para la pérdida de datos de eventos.

El espacio de nombres se extiende esencialmente entre regiones. Una región actúa como principal y las demás regiones sirven como secundarias. La suscripción de Azure muestra un único espacio de nombres, independientemente de cuántas regiones secundarias configure para la replicación geográfica.



En cualquier momento, puede *promover* una región secundaria a una región primaria. Al promover una región secundaria, Event Hubs redirige el nombre de dominio completo (FQDN) del espacio de nombres a la región secundaria seleccionada y degrada la región primaria anterior a una región secundaria. Decide si desea realizar una *promoción planeada*, lo que significa que espera a que se complete la replicación de datos o una *promoción forzada*, lo que podría provocar la pérdida de datos.

> **Nota:**
>
> La replicación geográfica de Event Hubs usa el término *promoción* porque representa mejor el proceso de promoción de una región secundaria a una región primaria (y, posteriormente, degradación de una región primaria a una región secundaria). También podría ver el término *conmutación por error* utilizado para describir el proceso general.

En esta sección se resumen aspectos importantes de la replicación geográfica. Revise la documentación completa para comprender exactamente cómo funciona. Para más información, consulte Replicación geográfica de Event Hubs.

Soporte para regiones

Puede elegir cualquier región de Azure que admita Event Hubs como región primaria o regiones secundarias. No es necesario usar regiones emparejadas de Azure, por lo que puede elegir regiones secundarias en función de los requisitos de latencia, cumplimiento o residencia de datos.

Requisitos

Para habilitar la replicación geográfica, el espacio de nombres debe usar el nivel Premium o Dedicado.

Consideraciones

Al habilitar la replicación geográfica, tenga en cuenta los siguientes factores:

- **Formato de punto de comprobación:** El formato de los puntos de control cambia. Para más información, consulte Replicación geográfica: Consumo de datos.
- **Puntos de conexión privados:** Si usa puntos de conexión privados para conectarse al espacio de nombres, también debe configurar las redes en las regiones primarias y secundarias. Para obtener más información, consulte Puntos de conexión privados.

Cost

Para comprender cómo funcionan los precios para la replicación geográfica, consulte Precios.

Configuración de la compatibilidad con varias regiones

- **Habilite la replicación geográfica en un espacio de nombres nuevo o existente.** Para configurar la replicación activa-activa para un espacio de nombres recién creado, consulte Habilitación de la replicación geográfica en un nuevo espacio de nombres. Para configurar la replicación activa-activa en un espacio de nombres existente, consulte Habilitación de la replicación geográfica en un espacio de nombres existente.
- **Cambie el enfoque de replicación.** Para cambiar entre modos de replicación sincrónicos y asincrónicos, consulte Cambiar modo de replicación.
- **Deshabilitar la replicación geográfica.** Para deshabilitar la replicación geográfica en una región secundaria, consulte Eliminación de una región secundaria.

Comportamiento cuando todas las regiones están en buen estado

En esta sección se describe qué esperar cuando se configura un espacio de nombres de Event Hubs para la replicación geográfica y la región primaria está operativa.

- **Enrutamiento del tráfico entre regiones:** Las aplicaciones cliente se conectan a través del FQDN de su espacio de nombres, y su tráfico se dirige a la región primaria.

Solo la región primaria procesa activamente los eventos de los clientes durante las operaciones normales. La región secundaria recibe eventos replicados, pero de lo contrario permanece pasiva en modo de espera.

- **Replicación de datos entre regiones:** El comportamiento de replicación de datos entre las regiones principal y secundaria depende de si configura el emparejamiento de replicación para usar la replicación sincrónica o asincrónica.

  - *Síncrono:* Los eventos se replican en la región secundaria antes de que se complete la operación de escritura.

    Este modo proporciona la mayor seguridad de que los datos del evento están seguros, ya que deben confirmarse tanto en la región primaria como en la secundaria. Sin embargo, la replicación sincrónica aumenta considerablemente la latencia de escritura para los eventos entrantes. También requiere que la región secundaria esté disponible para aceptar la operación de escritura, por lo que una interrupción en cualquier región secundaria hace que se produzca un error en la operación de escritura.
    - *Asíncrono:* Los eventos se escriben en la región primaria y, a continuación, se completa la operación de escritura. Poco tiempo después, replica los eventos en la región secundaria.

    Este modo proporciona un mayor rendimiento de escritura que la replicación sincrónica porque no hay ninguna latencia de replicación entre regiones durante las operaciones de escritura. Además, el modo de replicación asincrónica puede tolerar la pérdida de una región secundaria, a la vez que permite operaciones de escritura en la región primaria. Sin embargo, si la región primaria tiene una interrupción, es posible que los datos que aún no se hayan replicado en la región secundaria no estén disponibles o se pierdan.

    Al configurar la replicación asincrónica, establezca el retardo máximo aceptable que puede tener la replicación. En cualquier momento, puede comprobar el retraso de replicación actual mediante métricas de Azure Monitor.

    Si el retraso de replicación asincrónico aumenta más allá del máximo especificado, la región primaria comienza a limitar las solicitudes entrantes para que la replicación pueda alcanzar el nivel deseado. Para evitar esta situación, es importante seleccionar regiones secundarias que no estén demasiado distantes geográficamente y asegurarse de que la capacidad sea suficiente para el rendimiento.

    Para obtener más información, consulte Modos de replicación.

Comportamiento durante una falla de región

En esta sección se describe qué esperar cuando se configura un espacio de nombres de Event Hubs para la replicación geográfica y se produce una interrupción en la región principal o secundaria.

- **Detección y respuesta:** Usted es responsable de decidir cuándo promover la región secundaria del espacio de nombres para convertirse en la nueva región primaria. Microsoft no toma esta decisión ni inicia el proceso por ti, incluso si hay una interrupción de la región. Para obtener más información sobre cómo promover una región secundaria a la nueva principal, consulte Promover secundaria.

  Al promover una región secundaria, elija si desea realizar una *promoción planeada* o una *promoción forzada*. Una promoción planeada espera a que la región secundaria se ponga al día antes de aceptar nuevo tráfico. Este enfoque elimina la pérdida de datos, pero presenta tiempo de inactividad.

  Durante una interrupción en la región principal, típicamente usted necesita realizar una promoción forzada. Si la región primaria está disponible y activa una promoción por otro motivo específico, podría elegir una promoción planeada.

- **Notificación:** Microsoft no le notifica automáticamente cuando una región está inactiva. Sin embargo, puede usar Azure Service Health para comprender el estado general del servicio, incluidos los errores de región, y puede configurar alertas de Service Health para notificarle problemas.

  Use esa información y otras métricas para decidir cuándo promover una región secundaria a una región primaria.

- **Solicitudes activas:** El comportamiento depende de si la interrupción de la región se produce en la región primaria o en una región secundaria:

  - *Interrupción de la región primaria:* Si la región primaria no está disponible, se finalizan todas las solicitudes activas. Las aplicaciones cliente deben reintentar las operaciones una vez completada la promoción.

  - *Interrupción de la región secundaria:* Una interrupción en la región secundaria puede provocar problemas para las solicitudes activas en las situaciones siguientes:

    - Si usa el modo de replicación sincrónica, la región primaria no puede completar las operaciones de escritura si alguna región secundaria no está disponible.

    - Si usa el modo de replicación asincrónica, el espacio de nombres limita su capacidad y no acepta nuevos eventos después de que el retraso de la replicación alcance el máximo que ha configurado.

Para seguir usando el espacio de nombres en la región primaria, quite el espacio de nombres secundario de la configuración de replicación geográfica.

- **Pérdida de datos esperada:** La cantidad de pérdida de datos depende del tipo de promoción que realice (planeado o forzado) y del modo de replicación (sincrónico o asincrónico):

  - *Promoción planeada:* No se espera ninguna pérdida de datos. Sin embargo, durante una interrupción de la región, es posible que una promoción planeada no sea posible porque requiere que todas las regiones primarias y secundarias estén disponibles.

  - *Promoción forzada, replicación sincrónica:* No se espera ninguna pérdida de datos.

  - *Promoción forzada, replicación asincrónica:* Es posible que experimente alguna pérdida de datos para eventos recientes que no se replican en la región secundaria. La cantidad depende del retraso de replicación. Para comprobar el retraso de replicación actual, use las métricas de Azure Monitor.

  Si realiza una promoción forzada, no puede recuperar los datos perdidos, incluso después de que la región primaria esté disponible.

- **Tiempo de inactividad esperado:** La cantidad de tiempo de inactividad esperado depende de si realiza una promoción planeada o forzada:

  - *Promoción planeada:* El primer paso de una promoción planeada replica los datos en la región secundaria. Ese proceso suele completarse rápidamente, pero en algunas situaciones, puede tardar hasta la longitud del retraso de replicación. Una vez completada la replicación, el proceso de promoción suele tardar entre 5 y 10 minutos. A veces, los servidores del sistema de nombres de dominio (DNS) pueden tardar más tiempo en actualizar las entradas y replicar completamente sus registros en los clientes.

    La región primaria no admite operaciones de escritura durante el proceso completo de promoción.

    Es posible que esta opción no sea posible durante una interrupción de la región porque requiere que todas las regiones primarias y secundarias estén disponibles.

  - *Promoción forzada:* Durante una promoción forzada, Event Hubs no espera a que se complete la replicación de datos e inicia la promoción inmediatamente. El proceso de promoción suele tardar entre 5 y 10 minutos. A veces, las entradas DNS pueden tardar más tiempo en replicarse y actualizarse por completo en todos los clientes.

    La región primaria no admite operaciones de escritura durante el proceso completo de promoción.

- **Reenrutamiento del tráfico:** Una vez completada la promoción, el FQDN del espacio de nombres apunta a la nueva región primaria. Pero este redireccionamiento depende de la rapidez con la que se actualizan los registros DNS de los clientes, incluido para que sus servidores DNS respeten el período de vida (TTL) de los registros DNS del espacio de nombres.

  En algunas situaciones, debe configurar las aplicaciones de consumo para que se comporten de forma coherente después de que se produzca la promoción de la región. Para más información, consulte Replicación geográfica: Consumo de datos.

Recuperación de regiones

Después de que la región primaria original se recupere, si desea devolver el namespace a su región primaria original, siga el mismo proceso de promoción de la región.

Si realizó una promoción forzada durante la interrupción de la región, no podrá recuperar los datos perdidos, incluso después de que la región primaria esté disponible.

Prueba de fallos de región

Para probar la replicación geográfica, promueva temporalmente la región secundaria a principal y compruebe que las aplicaciones cliente pueden cambiar entre regiones con una interrupción mínima.

Supervise la duración de la promoción y compruebe que los runbooks y la automatización funcionan correctamente. Después de realizar las pruebas, puede revertir a la configuración original.

Comprenda el posible tiempo de inactividad y pérdida de datos que puede experimentar durante y después del proceso de promoción. Pruebe la replicación geográfica en un entorno que no sea de producción que refleje la configuración del espacio de nombres de producción.

Recuperación ante desastres geográfica de metadatos

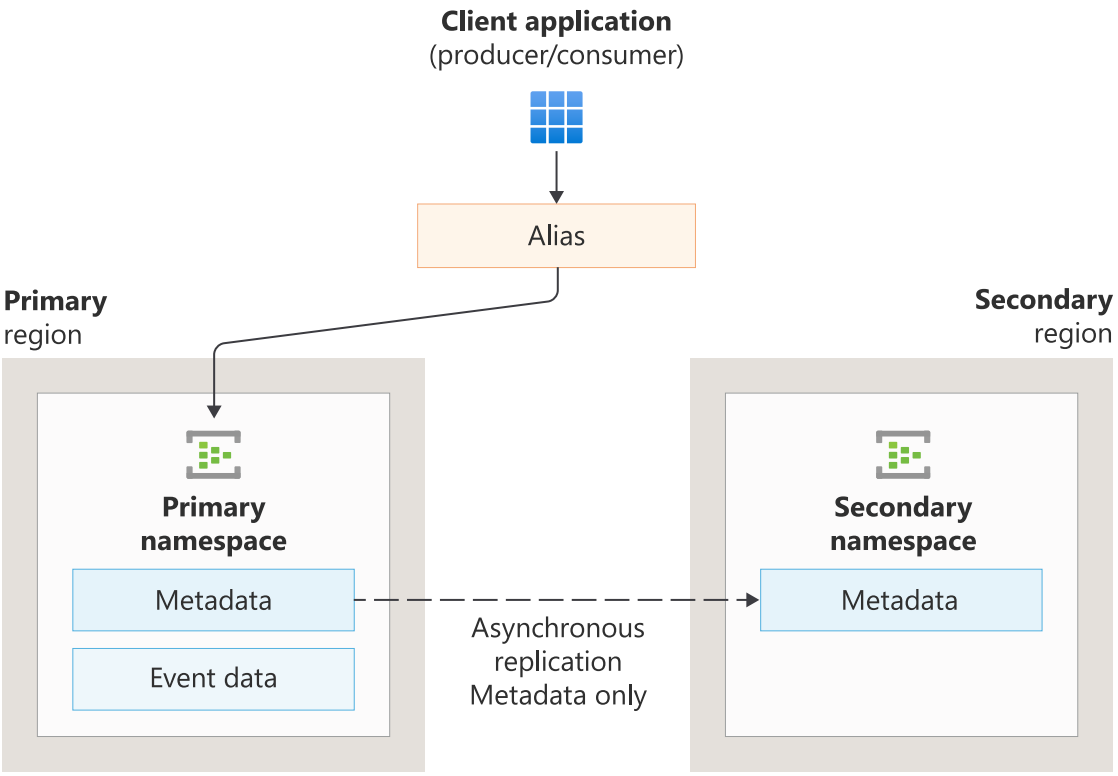
El nivel Estándar y versiones posteriores admiten la recuperación ante desastres geográfica de metadatos. Esta característica mejora la recuperación de escenarios de desastres, incluida la pérdida grave de una región. La recuperación ante desastres geográfica solo replica la configuración y los metadatos del espacio de nombres. Sin embargo, no replica los datos de eventos. Para admitir la recuperación ante desastres, esta característica garantiza que un espacio de nombres de otra región esté preconfigurado y listo para aceptar inmediatamente eventos de los clientes. La recuperación ante desastres geográfica sirve como una solución de

recuperación unidireccional y no admite la conmutación por recuperación en la región primaria anterior.

La recuperación ante desastres geográfica de metadatos funciona mejor para las aplicaciones que no necesitan mantener todos los eventos y pueden tolerar cierta pérdida de datos durante un escenario de desastre. Por ejemplo, si sus eventos representan lecturas de sensores que luego usted agrega, podría decidir que puede permitirse perder algunos eventos de una región fallida si puede reanudar rápidamente el procesamiento de nuevos eventos en otra región.

> **Importante**
>
> La recuperación ante desastres geográfica permite la continuidad de las operaciones que tienen la misma configuración, pero **no replica los datos de eventos**. Si necesita replicar datos de eventos, considere la posibilidad de usar la replicación geográfica.

Al configurar la recuperación ante desastres geográfica de metadatos, se crea un *alias* al que se conectan las aplicaciones cliente. El alias es un FQDN que dirige todo el tráfico al espacio de nombres principal de forma predeterminada.



Si se produce un error en la región primaria u otro tipo de desastre, puede iniciar manualmente una conmutación por error unidireccional desde la región primaria a la región secundaria en cualquier momento. La conmutación por error se completa casi al instante. Durante el proceso de conmutación por error, el alias de recuperación ante desastres geográficos se redirige al espacio de nombres secundario.

En esta sección se resumen aspectos importantes de la recuperación ante desastres geográfica. Revise la documentación completa para comprender exactamente cómo funciona. Para más información, consulte Recuperación ante desastres geográfica de Event Hubs.

Soporte para regiones

Puede seleccionar cualquier región de Azure que admita Event Hubs como espacio de nombres principal o secundario. No es necesario usar regiones emparejadas de Azure, por lo que puede elegir regiones secundarias en función de los requisitos de latencia, cumplimiento o residencia de datos.

Requisitos

- **Nivel de espacio de nombres principal:** El espacio de nombres principal debe estar en el nivel Estándar o superior para usar la recuperación ante desastres geográfica de metadatos.
- **Nivel de espacio de nombres secundario:** La recuperación ante desastres geográfica de metadatos admite combinaciones específicas de niveles para los espacios de nombres principal y secundario. Para obtener más información, vea Pares de espacios de nombres admitidos.

Consideraciones

- **Asignaciones de roles:** las asignaciones de control de acceso basado en roles (RBAC) de Microsoft Entra a entidades en el espacio de nombres principal no se replican en el espacio de nombres secundario. Cree asignaciones de roles manualmente en el espacio de nombres secundario para proteger el acceso a estas.
- **Registro de Esquema:** Los metadatos del Registro de Esquema se replican al usar la recuperación ante desastres geográfica de metadatos, pero los esquemas inscritos en el registro de esquema no se replican.
- **Diseño de aplicaciones:** La recuperación ante desastres geográfica requiere consideraciones específicas al diseñar las aplicaciones cliente. Para más información, consulte Consideraciones.
- **Puntos de conexión privados:** Si usa puntos de conexión privados para conectarse al espacio de nombres, configure las redes en la región primaria y secundaria. Para obtener más información, consulte Puntos de conexión privados.

Cost

Al habilitar la recuperación ante desastres geográficos de metadatos, usted paga por los espacios de nombres principal y secundario.

Configuración de la compatibilidad con varias regiones

- **Cree un emparejamiento de recuperación ante desastres geográficos para metadatos.** Para configurar la recuperación ante desastres entre los espacios de nombres principal y secundario, consulte Configuración y proceso de conmutación por error.

- **Deshabilite la recuperación ante desastres geográfica de metadatos.** Para interrumpir el emparejamiento entre espacios de nombres, consulte Configuración y flujo de conmutación por error.

Planeamiento y administración de capacidad

Cuando planee implementaciones de varias regiones, asegúrese de que ambas regiones tienen capacidad suficiente para controlar la carga completa si se produce un error en una región. La región secundaria permanece pasiva durante las operaciones normales, pero debe controlar inmediatamente el tráfico después de la conmutación por error. Planee cómo escalar la capacidad del espacio de nombres secundario para que pueda recibir tráfico de producción sin demora. Si puede tolerar un tiempo de inactividad adicional durante el proceso de conmutación por error, puede optar por escalar la capacidad del espacio de nombres secundario durante o después de la conmutación por error. Para reducir el tiempo de inactividad, aprovisione la capacidad con antelación en el espacio de nombres secundario para que permanezca lista para recibir la carga de producción.

Comportamiento cuando todas las regiones están en buen estado

En esta sección se describe qué esperar cuando se configura un espacio de nombres de Event Hubs para la recuperación ante desastres geográfica y la región primaria está operativa.

- Enrutamiento del tráfico entre regiones: Las aplicaciones cliente se conectan a través del alias de recuperación ante desastres geográfico de su espacio de nombres, y su tráfico se enruta al espacio de nombres principal en la región primaria.

  Solo el espacio de nombres principal procesa activamente eventos de los clientes durante las operaciones normales. El espacio de nombres secundario permanece pasivo en modo de espera y se produce un error en las solicitudes de acceso a los datos.

- **Replicación de datos entre regiones:** Solo los metadatos de configuración se replican entre los espacios de nombres. La replicación de la configuración se produce de forma continua y asincrónica.

  Todos los datos de eventos permanecen solo en el espacio de nombres principal y no se replican en el espacio de nombres secundario.

Comportamiento durante una falla de región

En esta sección se describe qué esperar cuando se configura un espacio de nombres de Event Hubs para la recuperación ante desastres geográficos y se produce una interrupción en la región principal.

- **Detección y respuesta:** usted es responsable de monitorear la salud de la región e iniciar la conmutación por error manualmente. Microsoft no realiza la conmutación por error ni promueve automáticamente una región secundaria, incluso si la región primaria está inactiva.

  Para obtener más información sobre cómo iniciar la conmutación por error, consulte Conmutación por error manual.

  La conmutación por error es una operación unidireccional, por lo que deberá restablecer la pareja de recuperación ante desastres geográficos más adelante. Para obtener más información, consulte Recuperación de regiones.

- **Notificación:** Microsoft no le notifica automáticamente cuando una región está inactiva. Sin embargo, puede usar Azure Service Health para comprender el estado general del servicio, incluidos los errores de región, y puede configurar alertas de Service Health para notificarle problemas.

  Use esa información y otras métricas para decidir cuándo conmutar a la región secundaria.

- **Solicitudes activas:** Las solicitudes activas terminan cuando se inicia la conmutación por error. Las aplicaciones cliente deben reintentar las operaciones una vez completada la conmutación por error.

- **Pérdida de datos esperada:**

  - *Metadatos:* Normalmente, la configuración y los metadatos se replican en el espacio de nombres secundario. Pero la replicación de metadatos se produce de forma asincrónica, por lo que es posible que los cambios recientes no se repliquen, especialmente los cambios complejos. Compruebe la configuración del espacio de nombres secundario antes de que los clientes accedan a él.

- *Datos de eventos:* Los datos del evento no se replican entre regiones. Si la región primaria deja de funcionar, los eventos del espacio de nombres principal no estarán disponibles.

  Los eventos no se pierden permanentemente a menos que un desastre catastrófico cause una pérdida total de la región primaria. Si la región se recupera, puede recuperar eventos del espacio de nombres principal más adelante.

- **Tiempo de inactividad esperado:** Normalmente, la conmutación ocurre en un plazo de 5 a 10 minutos. Los clientes pueden tardar más tiempo en replicar y actualizar completamente las entradas DNS.

- **Reenrutamiento del tráfico:** Los clientes que usan el alias de recuperación ante desastres geográfica para conectarse al espacio de nombres se redirigen automáticamente al espacio de nombres secundario después de la conmutación por error. Pero este redireccionamiento depende de los servidores DNS que respetan el TTL de los registros DNS del espacio de nombres y los clientes que reciben esos registros DNS actualizados.

Recuperación de regiones

Una vez recuperada la región primaria original, debe reestablecer manualmente el emparejamiento y, opcionalmente, realizar una conmutación de retorno. Cree un nuevo emparejamiento de recuperación ante desastres geográfica con la región recuperada como secundaria y, a continuación, realice otra conmutación por error si desea volver a la región original. Este proceso implica la posible pérdida de datos de eventos enviados al servidor principal temporal.

Si el desastre provoca la pérdida de todas las zonas de la región primaria, los datos podrían ser irrecuperables. En otros escenarios, los datos del evento permanecen en el espacio de nombres principal de antes de que se pueda recuperar la conmutación por error. Puede obtener eventos históricos del espacio de nombres principal anterior después de restaurar el acceso. Usted es responsable de configurar las aplicaciones para recibir y procesar esos eventos. Microsoft no los restaura automáticamente en tu región secundaria.

Prueba de fallos de región

Para probar los procesos de respuesta y recuperación ante desastres, realice una conmutación por error planeada durante una ventana de mantenimiento. Inicie la conmutación por error desde el espacio de nombres principal al espacio de nombres secundario y compruebe que las aplicaciones pueden conectarse y procesar eventos desde el nuevo espacio de nombres principal.

Supervise la duración de la conmutación por error y compruebe que los runbooks y la automatización funcionan correctamente. Después de realizar las pruebas, puede revertir a la configuración original.

Comprenda el posible tiempo de inactividad y la pérdida de datos que puede experimentar durante y después del proceso de conmutación por error. Pruebe la replicación geográfica en un entorno que no sea de producción que refleje la configuración del espacio de nombres de producción.

Soluciones personalizadas de varias regiones para la resistencia

La replicación geográfica y la recuperación ante desastres geográficos de metadatos proporcionan resiliencia frente a interrupciones regionales y otros problemas, y son compatibles con la mayoría de las cargas de trabajo. Algunos niveles de Event Hubs no admiten estas características, o es posible que necesite la replicación personalizada o necesite mantener varias regiones activas simultáneamente.

Varios patrones de diseño pueden lograr diferentes tipos de compatibilidad con varias regiones en Event Hubs. Muchos de los patrones requieren la implementación de varios espacios de nombres y el uso de servicios como Azure Functions para replicar eventos entre ellos. Para obtener más información, consulte Federación multisitio y de varias regiones.

Copias de seguridad y restauración

Event Hubs no está diseñado como una ubicación de almacenamiento a largo plazo para los datos. Normalmente, los datos se almacenan en un centro de eventos durante un breve período de tiempo y, a continuación, se procesan o se conservan en otro sistema de almacenamiento de datos. Puede configurar el período de retención de datos para su centro de eventos según sus requisitos y el nivel que utiliza su espacio de nombres. Para más información, consulte Retención de eventos.

Si necesita conservar una copia de los eventos, considere la posibilidad de usar Event Hubs Capture, que guarda copias de eventos en una cuenta de Azure Blob Storage.

Acuerdo de nivel de servicio

El contrato de nivel de servicio (SLA) para los servicios de Azure describe la disponibilidad esperada de cada servicio y las condiciones que la solución deberá cumplir para lograr esa expectativa de disponibilidad. Para obtener más información, consulte Acuerdos de Nivel de Servicio para servicios en línea.

El SLA de disponibilidad del namespace es superior cuando se utilizan los niveles Premium o Dedicado.

Contenido relacionado

- Documentación de Event Hubs
- Introducción a la confiabilidad de Azure

Last updated on 22/01/2026

Confiabilidad de Azure Functions

Azure Functions es un servicio de proceso controlado por eventos que ejecuta pequeños bloques de código o *funciones* sin aprovisionamiento o administración de infraestructuras explícitas. Las funciones pueden responder a eventos como solicitudes HTTP, temporizadores, mensajes de cola y cambios en otros servicios de Azure. Esta funcionalidad hace que Functions sea adecuado para el procesamiento de datos, la integración del sistema y las tareas en segundo plano.

Cuando se usa Azure, la confiabilidad es una responsabilidad compartida. Microsoft proporciona una variedad de funcionalidades para admitir resistencia y recuperación. Es responsable de comprender cómo funcionan esas funcionalidades dentro de todos los servicios que usa y de seleccionar las funcionalidades que necesita para cumplir los objetivos empresariales y los objetivos de tiempo de actividad.

En este artículo se describe cómo hacer que Functions sea resistente a varias posibles interrupciones y problemas, incluidos errores transitorios, errores de zona de disponibilidad y errores en toda la región. También resalta información clave sobre el acuerdo de nivel de servicio (SLA) de Functions.

Recomendaciones de implementación de producción

Azure Well-Architected Framework proporciona recomendaciones sobre confiabilidad, seguridad, costo, operaciones y rendimiento. Para obtener información sobre cómo estas áreas influyen entre sí y contribuyen a una solución confiable de Functions, consulte Procedimientos recomendados de arquitectura para Functions.

Introducción a la arquitectura de confiabilidad

Al implementar Functions, es importante familiarizarse con estos conceptos:

- **Planes de hospedaje:** Los planes representan el entorno de hospedaje de las aplicaciones de funciones. El plan determina los recursos de proceso disponibles, el modelo de precios y el comportamiento de escalado.
- **Cuentas de almacenamiento:** Al crear una aplicación de funciones, debe especificar una cuenta de almacenamiento de host. La cuenta de almacenamiento gestiona aspectos de las operaciones internas de la aplicación de funciones, incluidos el almacenamiento de código

de funciones, el registro y la gestión de concurrencia (como en el caso de las concesiones de blobs para tipos específicos de desencadenadores).

También puede usar una cuenta de almacenamiento para la implementación. Esta cuenta de almacenamiento puede ser la misma que la cuenta de almacenamiento de host o una cuenta de almacenamiento diferente.

> ⓘ **Importante**
>
> Las cuentas de almacenamiento son una parte fundamental de la arquitectura de confiabilidad de Functions. Configúrelos para satisfacer los requisitos de resistencia de la aplicación de funciones.

- **Desencadenadores y enlaces**: Los desencadenadores y enlaces permiten que la función responda a eventos, reciba datos de otros servicios y escriba datos en otros servicios.
- **Durable Functions**: Durable Functions es una característica de Functions. Proporciona funciones con estado, como orquestaciones de larga duración y entidades con estado.

  Cuando se usan funciones duraderas, se configura un proveedor de almacenamiento que almacena el estado. Evalúe las características de confiabilidad del almacén de estado que elija y configúrela para satisfacer los requisitos de resistencia.

Resistencia a errores transitorios

Los errores transitorios son errores breves e intermitentes en los componentes. Se producen con frecuencia en un entorno distribuido como la nube y son una parte normal de las operaciones. Los errores transitorios se corrigen después de un breve período de tiempo. Es importante que las aplicaciones puedan controlar errores transitorios, normalmente mediante el reintento de solicitudes afectadas.

Todas las aplicaciones hospedadas en la nube deben seguir las instrucciones de control de errores transitorios de Azure cuando se comunican con cualquier API, bases de datos y otros componentes hospedados en la nube. Para obtener más información, consulte Recomendaciones para controlar errores transitorios.

Tenga en cuenta las siguientes recomendaciones para controlar errores transitorios en las aplicaciones de funciones:

- **Desencadenadores y enlaces:** La plataforma Functions incluye el control de errores transitorios integrado para desencadenadores y enlaces. Cuando se produce un error transitorio mientras se desencadena un desencadenador compatible o se leen o escriben datos de enlace admitidos, la plataforma puede reintentar automáticamente la operación. Este comportamiento de reintento integrado garantiza que los problemas de conectividad temporal o las interrupciones del servicio no impidan que la función se ejecute. Para obtener más información, consulte Control de errores y reintentos de Functions.

  Esta protección solo cubre errores transitorios. La plataforma no reintenta los errores persistentes, como un cadena de conexión mal configurado o un recurso eliminado.

  La plataforma trata los errores persistentes y los errores transitorios repetidos como errores. Configure el registro para capturar información sobre los errores de ejecución de funciones. Para obtener más información, consulte Configuración de la supervisión de Funciones.

- **Tu código de función:** En tu código de función, eres responsable de controlar errores transitorios cuando tu función llame a servicios externos. Implemente la lógica de reintento, los tiempos de espera y los patrones de cortacircuito según corresponda para las llamadas de servicio externas realizadas en su código de función. Diseñe las funciones para que sean idempotentes siempre que sea posible, de modo que los reintentos no creen efectos secundarios repetidos.

- **Clientes:** Las aplicaciones cliente que se conectan a funciones de forma sincrónica, como a través de una conexión HTTP, deben ser resistentes a errores transitorios.

Resistencia a errores de zona de disponibilidad

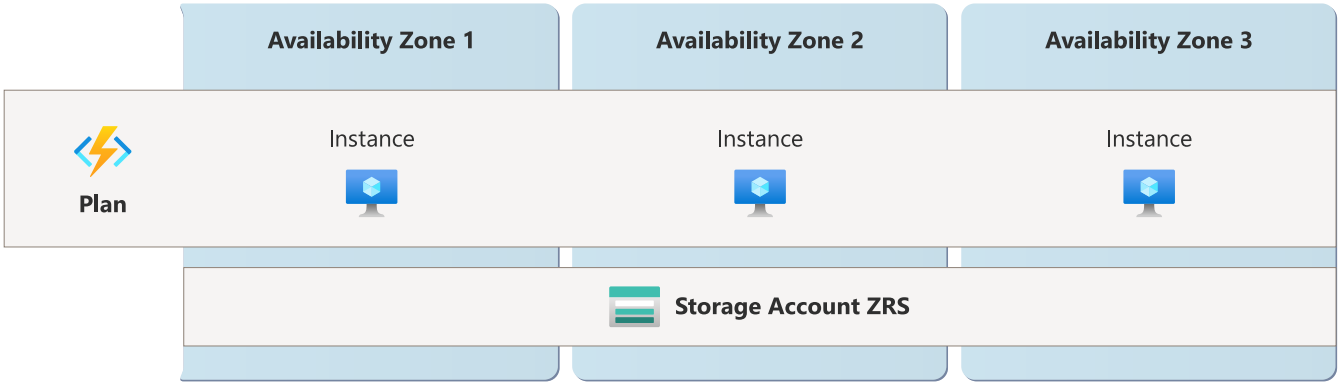
Las zonas de disponibilidad son grupos físicamente independientes de centros de datos dentro de una región de Azure. Cuando una zona falla, los servicios pueden transferirse a una de las zonas restantes.

Los planes Flex Consumption admiten implementaciones con redundancia de zona.

Cuando se habilita la redundancia de zona, la plataforma distribuye automáticamente las instancias del plan en todas las zonas de disponibilidad de la región seleccionada. Si alguna zona de disponibilidad de la región tiene un problema, las funciones continúan ejecutándose en instancias de zonas saludables.

Debe habilitar el almacenamiento con redundancia de zona (ZRS) en la cuenta de almacenamiento del host, para garantizar que también sea resistente a las interrupciones de

zona.



El plan es *no zonal* o *regional* cuando no se habilita la redundancia de zona. La plataforma puede ubicar instancias de plan en cualquier zona de disponibilidad de la región o en la misma zona. Las instancias del plan no son tolerantes a fallos en las zonas de disponibilidad. Es posible que su plan experimente tiempo de inactividad durante una interrupción en cualquier zona de la región.

Requisitos

- **Compatibilidad con regiones:** Puede implementar planes de consumo flexible con redundancia de zona en un conjunto específico de regiones. Puede recuperar la lista actual de regiones admitidas mediante la CLI de Azure. Para más información, consulte Visualización de regiones que admiten zonas de disponibilidad.

- **Cuenta de almacenamiento de host:** Debe configurar la cuenta de almacenamiento de host predeterminada de la aplicación de funciones para usar ZRS. Si usa una cuenta de almacenamiento de host que no está configurada para ZRS, es posible que la aplicación se comporte inesperadamente durante una interrupción de zona.

- **Cuenta de almacenamiento de contenedor de implementación:** Si usa una cuenta de almacenamiento independiente para el contenedor de implementación de la aplicación, actualícela también para que sea redundante de zona.

Consideraciones

Comportamientos no en tiempo de ejecución: La redundancia de zona solo garantiza un tiempo de actividad continuado para las aplicaciones implementadas. Una interrupción de zona de disponibilidad puede afectar a aspectos de Functions, aunque la aplicación siga sirviendo tráfico. Estos comportamientos incluyen el escalado del plan, la creación de aplicaciones, la configuración de la aplicación y la publicación de aplicaciones.

Distribución de instancias a través de zonas

Al configurar las aplicaciones del plan de consumo Flex como redundantes por zona, la plataforma distribuye automáticamente las instancias del plan en varias zonas de la región seleccionada, aplicando reglas diferentes para instancias siempre activas en comparación con las instancias bajo demanda.

- **Las instancias siempre disponibles** se distribuyen entre al menos dos zonas mediante la distribución circular.

  Para garantizar la resistencia de zona, la plataforma mantiene automáticamente al menos dos instancias listas para siempre para cada función o grupo de escalado por función, independientemente de la configuración siempre lista para la aplicación. Las instancias que crea la plataforma son administradas por la plataforma, facturadas como instancias listas para siempre y no cambian las opciones de configuración siempre preparadas.

- **Las instancias a petición** se generan a partir de volúmenes fuente de eventos cuando la aplicación se escala más allá del número de instancias siempre listas. Las instancias a petición se distribuyen entre zonas de disponibilidad de la mejor manera posible. La plataforma prioriza un escalado horizontal más rápido en lugar de una distribución uniforme entre zonas. La plataforma intenta equilibrar la distribución a lo largo del tiempo.

Costo

No se incurre en ningún costo adicional al habilitar la redundancia de zona. Los precios de un plan con redundancia de zona son los mismos que los de un plan de zona única. Sin embargo, habilitar la redundancia de zona afecta al número mínimo de instancias del plan.

Cuando se habilitan zonas de disponibilidad en una aplicación con una configuración de instancia del tipo siempre lista, y hay menos de dos instancias para cada función o grupo de escalado por función, la plataforma crea automáticamente dos instancias del tipo siempre lista para cada función o grupo de escalado por función. Estas nuevas instancias también se facturan como instancias siempre listas.

Para más información sobre los precios completos, consulte Precios de Functions.

Configurar soporte de zonas de disponibilidad

- **Cree un nuevo plan de Funciones con redundancia de zona.** Puede habilitar la redundancia de zona al crear un nuevo plan. Para obtener más información, consulte Creación de una

aplicación de funciones con redundancia de zona.

- **Habilite la redundancia de zona en un plan existente.** Puede activar o desactivar las zonas de disponibilidad para los planes de Elastic Premium existentes. Los planes elastic Premium tienen un comportamiento de capacidad específico que difiere de los planes dedicados (App Service) y requiere pasos de configuración adicionales. Para ver los pasos detallados, consulte Habilitación de la redundancia de zona en un plan existente.

Planeamiento y administración de capacidad

Las aplicaciones de funciones con redundancia de zona siguen ejecutándose incluso cuando las zonas de la región experimentan una interrupción.

Durante una interrupción en la zona, Functions detecta instancias perdidas e intenta automáticamente buscar o crear instancias de reemplazo en las zonas saludables. La plataforma realiza este proceso en base al mejor esfuerzo y no garantiza el éxito. Si la carga de trabajo requiere un número específico de instancias para mantener el nivel de servicio esperado, considere *la posibilidad de sobreaprovisionar* el número de instancias siempre listas. El sobreaprovisionamiento permite a la solución tolerar la pérdida de capacidad y seguir funcionando sin reducir el rendimiento. Para más información, consulte Administración de la capacidad mediante el sobreaprovisionamiento.

Comportamiento cuando todas las zonas están en buen estado

En esta sección se describe qué esperar cuando un plan es con redundancia de zona, la cuenta de almacenamiento de host usa ZRS y todas las zonas de disponibilidad están operativas.

- **Operación entre zonas:** Al configurar la redundancia de zona en Functions, las solicitudes se distribuyen automáticamente entre las instancias de cada zona de disponibilidad. Una solicitud puede ir a cualquier instancia de cualquier zona de disponibilidad.

- **Replicación de datos entre zonas:** Functions es un servicio de proceso sin estado, por lo que no hay datos para replicar entre zonas. La plataforma replica la configuración entre zonas automáticamente.

  Si la cuenta de almacenamiento de host usa ZRS, Azure Storage replica sincrónicamente sus datos en varias zonas de disponibilidad.

  Para funciones duraderas, revise el proveedor de almacenamiento para obtener información sobre cómo replica datos entre zonas.

Comportamiento durante un fallo de zona

En esta sección se describe qué esperar cuando un plan tiene redundancia de zona, la cuenta de almacenamiento del host usa ZRS y hay una interrupción en una zona de disponibilidad.

- **Detección y respuesta:** La plataforma Functions es responsable de detectar un error en una zona de disponibilidad. No es necesario iniciar una conmutación por error de zona.
- **Notificación**: Microsoft no le notifica automáticamente cuando una zona está inactiva. Sin embargo, puede usar Azure Resource Health para supervisar el estado de un recurso individual y puede configurar alertas de Resource Health para notificarle problemas. También puede usar Azure Service Health para comprender el estado general del servicio, incluidos los errores de zona, y puede configurar alertas de Service Health para notificarle problemas.
- **Solicitudes activas:** Cuando una zona de disponibilidad no está disponible, las solicitudes en curso que se conectan a una instancia de la zona de disponibilidad errónea finalizan y se deben reintentar. Asegúrese de que las aplicaciones están preparadas siguiendo las instrucciones de control de errores transitorios.
- **Pérdida de datos esperada:** No se espera que los errores de zona causen pérdida de datos porque Functions es un servicio sin estado.

  Si la cuenta de almacenamiento de host utiliza ZRS, el almacenamiento garantiza que no se pierdan datos debido a un fallo de zona.

  Para las funciones duraderas, compruebe su proveedor de almacenamiento para saber si la pérdida de datos sea posible durante un fallo de zona.
- **Tiempo de inactividad esperado:** Durante las interrupciones de zona, las conexiones pueden experimentar interrupciones breves que suelen durar unos segundos a medida que se redistribuye el tráfico. Asegúrese de que las aplicaciones están preparadas siguiendo las instrucciones de control de errores transitorios.
- **Reenrutamiento del tráfico:** Functions detecta las instancias perdidas de esa zona e intenta buscar nuevas instancias de reemplazo. Una vez que Functions encuentra reemplazos, distribuye el tráfico entre las nuevas instancias según sea necesario.

  > ⓘ **Importante**

> Azure no garantiza que las solicitudes de más instancias se realicen correctamente en un escenario de zona descendente. La plataforma intenta reponer instancias perdidas de forma óptima. Si necesita capacidad garantizada durante un error de zona de disponibilidad, cree y configure los planes para tener en cuenta la pérdida de zona mediante el sobreaprovisionamiento de la capacidad.

- **Comportamientos no en tiempo de ejecución:** Las aplicaciones de un plan de aplicación de funciones con redundancia de zona siguen ejecutándose y atendiendo el tráfico incluso si una zona de disponibilidad experimenta una interrupción. Sin embargo, los comportamientos que no son de ejecución pueden verse afectados durante una interrupción de la zona de disponibilidad. Estos comportamientos incluyen el escalado de aplicaciones de funciones, la creación de aplicaciones, la configuración de la aplicación y la publicación de aplicaciones.

Recuperación de zona

Cuando se recupera la zona de disponibilidad, Functions restaura automáticamente las instancias de la zona de disponibilidad, quita las instancias temporales creadas en las otras zonas de disponibilidad y vuelve a enrutar el tráfico entre las instancias como normal.

Prueba de fallos de zona

La plataforma Functions administra el enrutamiento del tráfico, la conmutación por error y la recuperación de zona para los recursos con redundancia de zona. No es necesario iniciar nada. Azure administra completamente esta característica, por lo que no es necesario validar los procesos de error de zona de disponibilidad.

Resistencia a errores en toda la región

Functions es un servicio de una sola región. Si la región deja de estar disponible, el recurso de Functions tampoco está disponible.

Soluciones personalizadas de varias regiones para la resistencia

Para evitar interrupciones en su servicio durante caídas regionales, puede implementar de forma redundante las mismas funciones en aplicaciones de funciones en varias regiones.

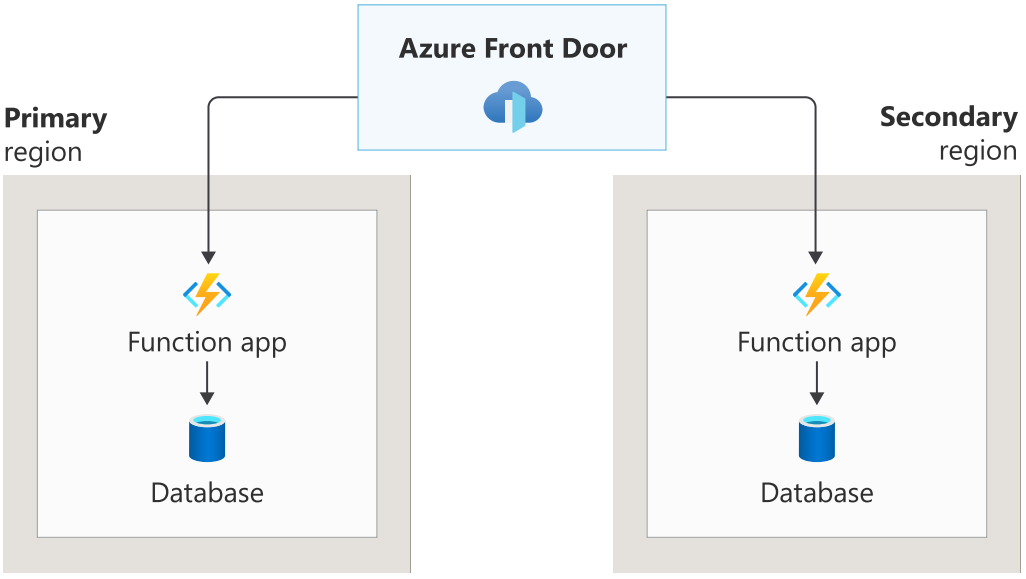
Usted es responsable de:

- Implementación de aplicaciones de funciones en varias regiones.
- Administración de la distribución del tráfico entre regiones.
- Implementación de mecanismos de conmutación por error.
- Garantizar la coherencia de los datos entre regiones (si procede).
- Supervisión y administración de implementaciones entre regiones.

Al ejecutar el mismo código de función en varias regiones, tenga en cuenta los patrones *activo-activo* y *activo-pasivo* . En las secciones siguientes se presentan estos patrones, pero no se proporcionan instrucciones detalladas ni pasos de configuración.

Patrón activo-activo para funciones de desencadenador HTTP

En un patrón activo-activo, las funciones en las dos regiones ejecutan activamente y procesan eventos, ya sea de forma duplicada o de manera alternada. Utilice un patrón activo-activo en combinación con Azure Front Door para sus funciones críticas desencadenadas por HTTP, que pueden enrutar y distribuir solicitudes HTTP entre funciones que se están ejecutando en varias regiones. Azure Front Door también puede comprobar periódicamente el estado de cada punto de conexión. Si un servicio de una región deja de responder a las comprobaciones de estado, Azure Front Door lo quita de la rotación y solo reenvía el tráfico a los servicios saludables restantes.



Patrón activo-pasivo para funciones de desencadenador que no son HTTP

En el caso de funciones dirigidas por eventos que no se activan mediante HTTP (como activadores de Azure Service Bus y Azure Event Hubs), use un patrón activo-pasivo. En un patrón

activo-pasivo, las instancias de función se ejecutan en la región que recibe eventos, mientras que las instancias de la región secundaria permanecen inactivas. Este patrón garantiza que solo una función procese cada mensaje, lo que ayuda a mantener la coherencia de los datos. Permite conmutar por error a la región secundaria durante un desastre, como una interrupción de la región.

Considere la conmutación por error de la app de funciones junto con los comportamientos de conmutación por error de otros servicios que use, como:

- Service Bus, replicación geográfica y recuperación ante desastres
- Replicación geográfica de Event Hubs y recuperación ante desastres geográfica

Considere una topología de ejemplo que usa un activador de Event Hubs, donde el espacio de nombres de Event Hubs está configurado para la recuperación ante desastres geográficos. En este caso, el patrón activo-pasivo requiere los siguientes componentes:

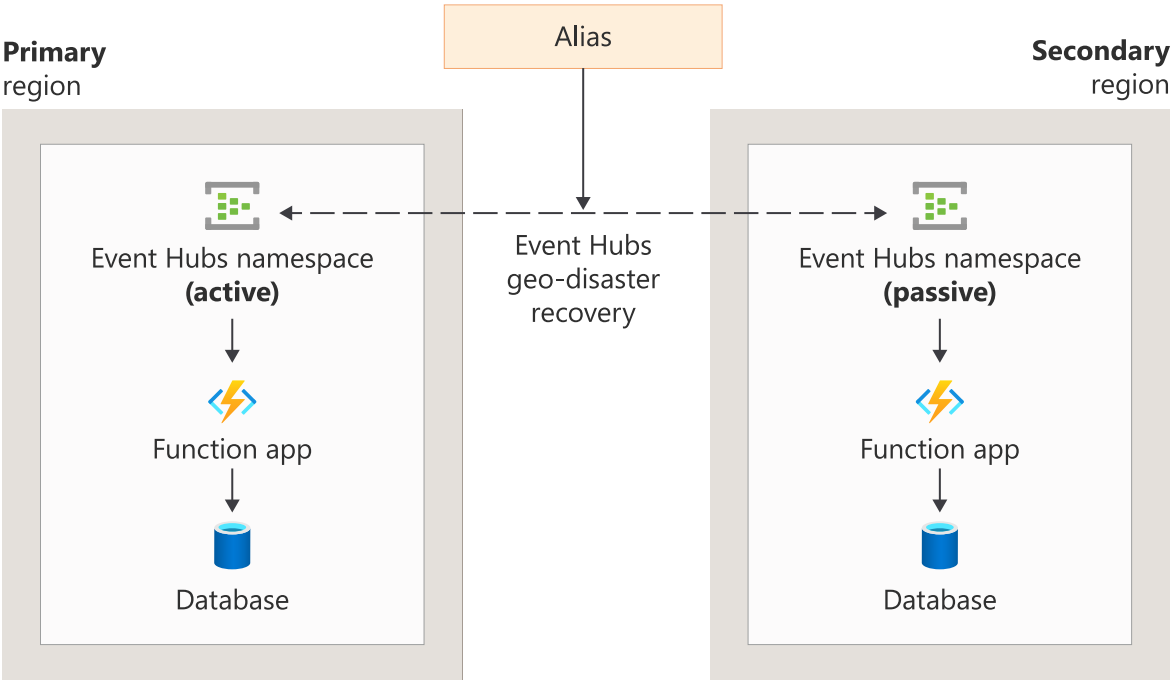
- Espacios de nombres de Event Hubs implementados tanto en una región primaria como en una secundaria.

- Se habilitó la recuperación ante desastres geográfica para vincular los centros de eventos principales y secundarios. Esta configuración también crea un *alias* que puede usar para conectarse al espacio de nombres de Event Hubs y cambiar el alias de la base de datos principal a la secundaria sin cambios en la información de conexión.

- Aplicaciones de funciones desplegadas tanto en la región primaria como en la secundaria. La aplicación de la región secundaria permanece inactiva porque no recibe mensajes.

- Los desencadenadores de cada aplicación de funciones utilizan la cadena de conexión *direct* (noalias) para su espacio de nombres correspondiente de Event Hubs.

- Los publicadores del espacio de nombres de Event Hubs publican al alias de la cadena de conexión.



Antes de la conmutación por error, los publicadores que envían eventos al alias compartido dirigen el tráfico hacia el centro de eventos principal. La aplicación de funciones principal escucha exclusivamente al hub de eventos principal. La aplicación de funciones secundaria permanece pasiva e inactiva.

Cuando se inicia el failover, los publicadores que envían eventos al alias compartido son dirigidos al centro de eventos secundario. La aplicación de funciones secundaria se activa y se desencadena automáticamente. El centro de eventos puede controlar todo el proceso de conmutación por error y cada aplicación de funciones solo se ejecuta cuando su centro de eventos correspondiente está activo.

Durable Functions

Para la recuperación ante desastres en varias regiones para funciones duraderas, consulte Disaster recovery and geo-distribution in Azure durable functions.

Resistencia al mantenimiento del servicio

Functions realiza actualizaciones de servicio periódicas y otras tareas de mantenimiento.

- **Resistencia de errores transitorios:** Durante el mantenimiento del servicio, las instancias que ejecutan la aplicación de funciones pueden reiniciarse o experimentar interrupciones temporales. Asegúrese de que las aplicaciones cliente que interactúan con la aplicación de funciones son resistentes a errores transitorios.

- **Habilite la redundancia de zona:** Al habilitar la redundancia de zona en el plan, también se mejora la resistencia durante las actualizaciones de la plataforma. Implementar varias instancias en tu plan y habilitar la redundancia de zonas proporciona una capa adicional de resiliencia si una instancia o una zona se vuelven defectuosas durante un proceso de actualización.

Resistencia a las implementaciones de aplicaciones

Las implementaciones de aplicaciones presentan el riesgo de problemas en un entorno de producción. Prepárese para revertir una actualización si provoca problemas. Controlar cómo se implementan las actualizaciones para reducir la interrupción de los reinicios de la aplicación.

Los planes de consumo flexible admiten estrategias de actualización del sitio, que proporcionan varias maneras de implementar las actualizaciones de la aplicación. Estas estrategias incluyen actualizaciones graduales para implementaciones sin tiempo de inactividad.

Acuerdo de nivel de servicio

El acuerdo de nivel de servicio (SLA) para Azure servicios describe la disponibilidad esperada de cada servicio y las condiciones que la solución debe cumplir para lograr esa expectativa de disponibilidad. Para obtener más información, consulte Acuerdos de Nivel de Servicio para servicios en línea.

Functions proporciona acuerdos de nivel de servicio de disponibilidad distintos para el plan de consumo y para otros tipos de plan.

Contenido relacionado

- Recuperación ante desastres y distribución geográfica en funciones duraderas de Azure
- Creación de una instancia de Azure Front Door
- Consideraciones sobre la conmutación por error de Event Hubs
- Estrategias de arquitectura para usar regiones y zonas de disponibilidad
- Confiabilidad en Azure

Last updated on 28/04/2026

Reliability in Azure Logic Apps

Azure Logic Apps helps you more easily integrate and orchestrate data between apps, cloud services, and on-premises systems by reducing how much code that you have to write.

When you use Azure, reliability is a shared responsibility. Microsoft provides a range of capabilities to support resiliency and recovery. You're responsible for understanding how those capabilities work within all of the services you use, and selecting the capabilities you need to meet your business objectives and uptime goals.

This article describes how to make logic app workflows resilient to a variety of potential outages and problems, including transient faults, availability zone outages, and region outages. It also highlights key information about the Azure Logic Apps service-level agreement (SLA).

Production deployment recommendations for reliability

For enterprise and secure workflows that have isolation or network security requirements, we recommend that you create and run Standard workflows in single-tenant Azure Logic Apps, rather than Consumption workflows in multitenant Azure Logic Apps. For more information, see Create and deploy to different environments.

Reliability architecture overview

This section describes some of the important aspects of how the service works that are most relevant from a reliability perspective. The section introduces the logical architecture, which includes some of the resources and features that you deploy and use. It also discusses the physical architecture, which provides details on how the service works under the covers.

Logical architecture

The primary resource that you deploy is a *logic app*. Consumption logic apps have only one workflow, while Standard logic apps can have more than one workflow. Most workflows use one or more connections to access other apps, services, and systems.

If you access data in on-premises systems, you might deploy an on-premises data gateway. Each gateway resource represents a separate data gateway installation on a local computer. You can configure an on-premises data gateway for high availability by using multiple computers. For more information, see High availability support.

When you use Azure Logic Apps for business-to-business (B2B) enterprise integration scenarios, you might deploy integration accounts where you define and store the artifacts that logic app workflows use.

Physical architecture

For Consumption logic apps, Azure Logic Apps automatically manages the compute infrastructure, state storage, and other resources. You don't need to configure or manage any virtual machines (VMs). Consumption logic apps share compute infrastructure between many customers.

Resilience to transient faults

Transient faults are short, intermittent failures in components. They occur frequently in a distributed environment like the cloud, and they're a normal part of operations. Transient faults correct themselves after a short period of time. It's important that your applications can handle transient faults, usually by retrying affected requests.

All cloud-hosted applications should follow the Azure transient fault handling guidance when they communicate with any cloud-hosted APIs, databases, and other components. For more information, see Recommendations for handling transient faults.

In Azure Logic Apps, many workflow triggers and actions automatically support *retry policies*, which automatically retry requests that fail because of transient faults. For more information about how to change or turn off retry policies, see Handle errors and exceptions in Azure Logic Apps.

If an action fails, you can customize the behavior of subsequent actions. You can also create *scopes* to group related actions that might fail or succeed together.

For more information, see Handle errors and exceptions in Azure Logic Apps.

Resilience to availability zone failures

Availability zones are physically separate groups of datacenters within an Azure region. When one zone fails, services can fail over to one of the remaining zones.

Azure Logic Apps supports *zone redundancy*, which spreads compute resources and state across multiple availability zones. When you distribute logic app workload resources across availability zones, you improve resiliency and reliability for your production logic app workloads.

New and existing Consumption logic app workflows in multitenant Azure Logic Apps are automatically zone-redundant.

Requirements

- **Region support**: Consumption logic apps deployed in any region that supports availability zones are automatically zone redundant. Japan West is the exception. It doesn't support zone-redundant logic apps because some dependent services don't support zone redundancy.

Considerations

- **Connectors**: Built-in connectors are automatically zone redundant when your logic app is zone redundant.
- **Integration accounts**: Premium SKU integration accounts are zone redundant by default.

Cost

No extra cost applies to use zone redundancy. All new and existing Consumption logic apps are zone-redundant automatically.

Configure availability zone support

Consumption logic app workflows automatically support zone redundancy, so no configuration is required.

Behavior when all zones are healthy

This section describes what to expect when logic app resources are configured for zone redundancy and all availability zones are operational.

- **Traffic routing between zones**: During normal operations, workflow invocations can use compute resources from any availability zone in the region.
- **Data replication between zones**: For stateful workflows, workflow state is synchronously replicated between availability zones by using zone-redundant storage (ZRS).

Behavior during a zone failure

This section describes what to expect when an availability zone outage occurs and logic app resources are configured for zone redundancy.

- **Detection and response**: Azure Logic Apps is responsible for detecting a failure in an availability zone. You don't need to do anything to initiate a zone failover.
- **Notification**: Microsoft doesn't automatically notify you when a zone is down. However, you can use Azure Service Health to understand the overall health of the service, including any zone failures, and you can set up Service Health alerts to notify you of problems.
- **Active requests**: If an availability zone becomes unavailable, Azure Logic Apps terminates any in-progress workflow executions that run on a VM in the faulty availability zone. The platform automatically resumes the workflow on another VM in a different availability zone. Because of this behavior, active workflows might experience some transient faults or higher latency when new VMs are added to the remaining availability zones.
- **Expected downtime**: No downtime is expected in Azure Logic Apps. But if dependencies exist on other services that experience downtime, your logic app might also be affected.
- **Expected data loss**: No data loss is expected.
- **Traffic rerouting**: Incoming traffic is automatically distributed to infrastructure in healthy zones.

Zone recovery

When the availability zone recovers, Azure Logic Apps automatically restores instances in the availability zone, removes any temporary instances created in the other availability zones, and reroutes traffic between your instances as normal.

Test for zone failures

Azure Logic Apps manages traffic routing, failover, and failback for zone-redundant logic app resources. You don't need to initiate anything. This feature is fully managed, so you don't need to validate availability zone failure processes.

Resilience to region-wide failures

Each logic app is deployed into a single Azure region. If the region becomes unavailable, your logic app is also unavailable.

Custom multi-region solutions for resiliency

For higher resiliency, you can deploy a standby or backup logic app in a secondary region and fail over to that other region if the primary region is unavailable. To set up this capability, complete the following tasks:

- Deploy your logic app in both primary and secondary regions.
- Reconfigure connections to resources as needed.
- Configure load balancing and failover policies.
- Plan to monitor the primary instance health and initiate failover.

For more information about multi-region deployments for your logic app workflows, see the following articles:

- Multi-region deployments in Azure Logic Apps
- Set up cross-region disaster recovery for integration accounts in Azure Logic Apps
- Create replication tasks for Azure resources by using Azure Logic Apps

Service-level agreement

The service-level agreement (SLA) for Azure services describes the expected availability of each service and the conditions that your solution must meet to achieve that availability expectation. For more information, see SLAs for online services .

Related content

- Reliability in Azure
- Handle errors and exceptions in Azure Logic Apps

ⓘ **Nota:** El autor ha creado este artículo con ayuda de inteligencia artificial. Más información

---

Last updated on 22/01/2026

Confiabilidad en Azure Service Bus

Azure Service Bus es un servicio de agente de mensajes empresariales totalmente administrado que proporciona funcionalidades de mensajería asincrónica confiables para desacoplar aplicaciones y servicios. Service Bus admite colas para la comunicación de punto a punto y temas con suscripciones para patrones de mensajería de publicación y suscripción. El servicio proporciona características de confiabilidad integradas, incluida la durabilidad del mensaje, garantías de entrega al menos una vez y colas de mensajes fallidos para controlar el procesamiento de mensajes erróneos.

Cuando se usa Azure, la confiabilidad es una responsabilidad compartida. Microsoft proporciona una variedad de funcionalidades para admitir resistencia y recuperación. Es responsable de comprender cómo funcionan esas funcionalidades dentro de todos los servicios que usa y de seleccionar las funcionalidades que necesita para cumplir los objetivos empresariales y los objetivos de tiempo de actividad.

En este artículo se describe cómo hacer que Service Bus sea resistente a diferentes posibles interrupciones y problemas, incluidos errores transitorios, interrupciones de zona de disponibilidad y interrupciones de región. También resalta cierta información clave sobre el acuerdo de nivel de servicio (SLA) de Service Bus.

Recomendaciones de implementación de producción

Azure Well-Architected Framework proporciona recomendaciones sobre confiabilidad, rendimiento, seguridad, costo y operaciones. Para obtener información sobre cómo estas áreas influyen entre sí y contribuyen a una solución confiable de Azure Service Bus, consulte Procedimientos recomendados de arquitectura para Service Bus.

Introducción a la arquitectura de confiabilidad

En esta sección se describen algunos de los aspectos importantes de cómo funciona el servicio que es más relevante desde una perspectiva de confiabilidad. En la sección se presenta la arquitectura lógica, que incluye algunos de los recursos y características que se implementan y usan. También se describe la arquitectura física, que proporciona detalles sobre cómo funciona el servicio en segundo plano.

Arquitectura lógica

Un *espacio de nombres* actúa como contenedor de administración para Service Bus y se puede configurar para usar el nivel Básico, Estándar o Premium. Para configurar el servicio a nivel de espacio de nombres, asigne capacidad, configure la seguridad de red y habilite la replicación geográfica y la recuperación ante desastres geográficos.

Dentro de un espacio de nombres, se implementan *colas* y *temas*, que son entidades de mensajería que tienen una semántica diferente. Para más información, consulte Colas, temas y suscripciones de Service Bus.

Puede configurar opcionalmente *particiones* en el espacio de nombres, las cuales distribuyen colas y tópicos en varios agentes de mensajes y almacenes de mensajería. Un espacio de nombres puede usar varias particiones para controlar el procesamiento paralelo y el escalado horizontal. Service Bus solo garantiza la ordenación dentro de una sola partición. La creación de particiones desempeña un papel clave en el diseño de confiabilidad de la aplicación. Al diseñar la aplicación, realice un equilibrio entre maximizar la disponibilidad y la coherencia. Para el nivel Premium, habilite la creación de particiones en el espacio de nombres. Para los espacios de nombres de nivel Básico y Estándar, configure particiones en la entidad y opcionalmente cuando envíe mensajes.

Puede usar Service Bus y sus enfoques de diseño asincrónicos para aumentar la disponibilidad de las aplicaciones. Para más información, consulte Patrones de mensajería asincrónica y alta disponibilidad.

Arquitectura física

Service Bus proporciona los recursos de proceso y almacenamiento subyacentes. Para cada espacio de nombres, varios *agentes de mensajes* procesan los mensajes y varios *almacenes de mensajería* almacenan los mensajes. Cada almacén de mensajería tiene tres copias: una copia principal y dos copias secundarias. Service Bus mantiene las tres copias sincronizadas para las operaciones de datos y administración. Si se produce un error en la copia principal, Service Bus promueve una copia secundaria a principal sin tiempo de inactividad para los clientes.

En el caso de los espacios de nombres que usan el nivel Básico o Estándar, Service Bus proporciona redundancia a través de una infraestructura multiinquilino compartida que replica automáticamente los mensajes entre zonas de disponibilidad donde estén disponibles. El servicio mantiene varios almacenes de mensajería y mantiene todas las copias sincronizadas para las operaciones de administración y datos.

En el caso de los espacios de nombres de nivel Premium, Service Bus proporciona unidades de mensajería dedicadas (RU) que cada una tiene recursos dedicados de CPU y memoria. Los espacios de nombres de nivel Premium se pueden escalar automáticamente en función de las

demandas de carga de trabajo. Para obtener más información, consulte Actualización automática de las unidades de mensajería de un espacio de nombres de Service Bus.

La infraestructura de Service Bus abarca varias máquinas físicas y bastidores que se distribuyen entre dominios de error, lo que reduce el riesgo de fallos catastróficos que afectan al espacio de nombres. En las regiones que tienen zonas de disponibilidad, la infraestructura se extiende entre centros de datos físicos independientes. El servicio implementa mecanismos transparentes de detección de errores y conmutación por error para que siga funcionando dentro de los niveles de servicio garantizados y normalmente sin interrupciones perceptibles cuando se produzcan estos errores.

Resistencia a errores transitorios

Los errores transitorios son errores breves e intermitentes en los componentes. Se producen con frecuencia en un entorno distribuido como la nube y son una parte normal de las operaciones. Los errores transitorios se corrigen después de un breve período de tiempo. Es importante que las aplicaciones puedan controlar errores transitorios, normalmente mediante el reintento de solicitudes afectadas.

Todas las aplicaciones hospedadas en la nube deben seguir las instrucciones de control de errores transitorios de Azure cuando se comunican con cualquier API, bases de datos y otros componentes hospedados en la nube. Para obtener más información, consulte Recomendaciones para controlar errores transitorios.

El SDK de Service Bus incluye lógica de reintento automático con retroceso exponencial para las operaciones que producen errores debido a condiciones transitorias, como tiempos de espera de red o falta de disponibilidad temporal del servicio. Cuando las aplicaciones experimentan desconectaciones transitorias de Service Bus, el SDK intenta volver a conectarse automáticamente mediante la directiva de reintento configurada.

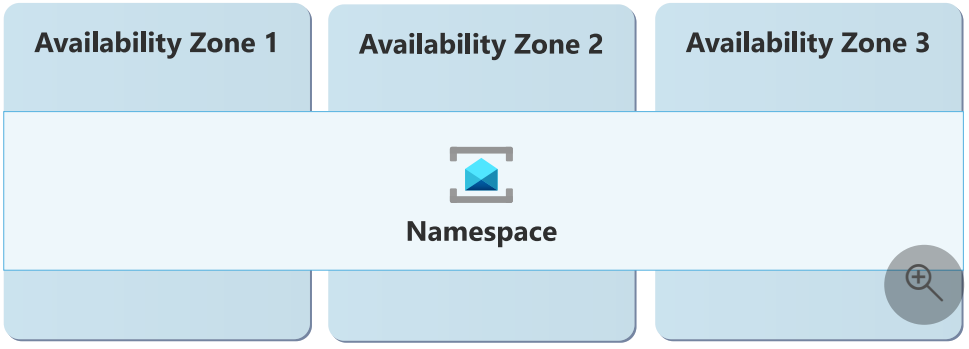
Para optimizar el control de errores transitorios en las aplicaciones, use el SDK de Service Bus más reciente, que incluye las características de administración de conexiones y lógica de reintento más actuales. Para más información, consulte Biblioteca cliente de Service Bus para .NET.

Resistencia a errores de zona de disponibilidad

Las zonas de disponibilidad son grupos físicamente independientes de centros de datos dentro de una región de Azure. Cuando una zona falla, los servicios pueden transferirse a una de las zonas restantes.

Service Bus admite implementaciones con redundancia de zona en todos los niveles de servicio. Al crear un espacio de nombres de Service Bus en una región admitida, la redundancia de zona se habilita automáticamente sin costo adicional. El modelo de implementación con redundancia de zona se aplica a todas las características de Service Bus, incluidas las particiones y las sesiones.

Service Bus replica de forma transparente los datos de configuración, metadatos y mensajes en varias zonas de disponibilidad de la región. La redundancia de zona proporciona una conmutación automática por error sin su intervención. Todos los componentes de Service Bus, incluidos los procesos, las redes y el almacenamiento, se replican entre zonas. Service Bus tiene suficientes reservas de capacidad para controlar instantáneamente la pérdida completa de una zona. Sigue funcionando sin pérdida de datos ni interrupción en las aplicaciones de mensajería, incluso si una zona de disponibilidad completa deja de estar disponible.



Requisitos

- **Compatibilidad con regiones:** Puede implementar espacios de nombres de Service Bus con redundancia de zona en regiones de Azure que admiten zonas de disponibilidad. Service Bus habilita automáticamente la compatibilidad con la zona de disponibilidad al crear un espacio de nombres en una región admitida, sin que se requiera ninguna configuración adicional.

- **Niveles:** Todos los niveles de Service Bus (Básico, Estándar y Premium) admiten zonas de disponibilidad sin requisitos adicionales.

Consideraciones

Los espacios de nombres de Service Bus incluyen una propiedad `zoneRedundant`. Anteriormente, esta propiedad era necesaria para habilitar zonas de disponibilidad, pero este requisito ha cambiado y la `zoneRedundant` propiedad está en desuso. Esta propiedad puede aparecer como `false` incluso cuando está habilitada la redundancia de zona. Todos los espacios de nombres en regiones que tienen zonas de disponibilidad son redundantes a nivel de zona.

Cost

La redundancia de zona en Service Bus no agrega ningún costo adicional.

Configurar soporte de zonas de disponibilidad

Los espacios de nombres de Service Bus admiten automáticamente la redundancia de zona al implementarlos en regiones admitidas. No es necesario realizar ninguna otra configuración.

Comportamiento cuando todas las zonas están en buen estado

En esta sección se describe qué esperar cuando los espacios de nombres de Service Bus están configurados para la redundancia de zona y todas las zonas de disponibilidad están operativas.

- **Enrutamiento de tráfico entre zonas:** Service Bus usa un modelo activo-activo en el que los mensajes se distribuyen entre varias zonas de disponibilidad. Las conexiones de cliente se equilibran automáticamente entre zonas y el servicio enruta las operaciones a la infraestructura de mensajería disponible independientemente de la zona.

- **Replicación de datos entre zonas:** Service Bus usa la replicación sincrónica entre zonas de disponibilidad, incluidos los metadatos y los datos de mensajes. Varias copias del almacén de mensajería deben confirmar las operaciones de escritura antes de que el servicio las considere completadas, lo que garantiza la coherencia de los datos entre zonas durante las operaciones normales.

Comportamiento durante un fallo de zona

En esta sección se describe qué esperar cuando los espacios de nombres de Service Bus están configurados para redundancia zonal y ocurre una interrupción en una zona de disponibilidad.

- **Detección y respuesta:** Microsoft detecta automáticamente errores de zona e inicia la conmutación de emergencia a zonas saludables. No se requiere ninguna acción del cliente para la conmutación por error de nivel de zona.

- **Notificación:** Microsoft no le notifica automáticamente cuando una zona está inactiva. Sin embargo, puede usar Azure Service Health para comprender el estado general del servicio, incluidos los errores de zona, y puede configurar alertas de Service Health para notificarle problemas.

- **Solicitudes activas:** Durante un error de zona, Service Bus podría quitar las solicitudes activas. Si los clientes controlan correctamente los errores transitorios mediante el

reintento después de un breve período de tiempo, normalmente evitan un impacto significativo.

- **Pérdida de datos esperada:** No se produce ninguna pérdida de datos durante un error de zona porque Service Bus replica de forma sincrónica mensajes entre zonas antes de la confirmación.
- **Tiempo de inactividad esperado:** Un error de zona puede provocar unos segundos de tiempo de inactividad. Si los clientes controlan correctamente los errores transitorios mediante el reintento después de un breve período de tiempo, normalmente evitan un impacto significativo.
- **Reenrutamiento del tráfico:** Service Bus detecta la pérdida de la zona y redirige automáticamente las nuevas solicitudes a otra réplica en una de las zonas de disponibilidad correctas.

  El SDK de Service Bus normalmente controla la administración de conexiones y la lógica de reintento de forma transparente.

Recuperación de zona

Cuando se recupera una zona de disponibilidad, Service Bus reintegra automáticamente la zona en la topología activa del servicio. A continuación, la zona recuperada acepta nuevas conexiones y procesa mensajes junto con las demás zonas. Los datos que se replican en zonas supervivientes durante la interrupción permanecen intactos y la replicación sincrónica normal se reanuda en todas las zonas. No es necesario tomar medidas para la recuperación o la reintegración de zonas.

Prueba de fallos de zona

Service Bus administra el enrutamiento del tráfico, la conmutación por error y la recuperación de zona para los fallos de zona, por lo que no es necesario validar los procesos de fallo de zona de disponibilidad ni proporcionar datos adicionales.

Resistencia a errores en toda la región

Service Bus ofrece dos tipos de compatibilidad con varias regiones, y ambos requieren espacios de nombres de nivel Premium.

- La replicación geográfica proporciona replicación activa-pasiva de metadatos y datos de mensajes entre una región primaria y una región secundaria. Use Geo-Replication para la

mayoría de las aplicaciones que deben permanecer resistentes a las interrupciones de regiones y tener baja tolerancia a la pérdida de datos de mensajes.

- Metadata Geo-Disaster Recovery proporciona la replicación activa-pasiva de la configuración y los metadatos entre una región primaria y secundaria, pero no replica los datos del mensaje. Considere usar Recuperación ante Desastres Geográficos para aplicaciones que controlan su propia replicación de datos o no necesitan replicación de datos.

Tanto la replicación geográfica como la recuperación ante desastres geográfica de metadatos requieren que inicie manualmente la conmutación por error o la promoción de una región secundaria para que se convierta en la nueva región primaria. Microsoft no inicia automáticamente la conmutación por error ni la promoción, incluso si la región principal deja de funcionar.

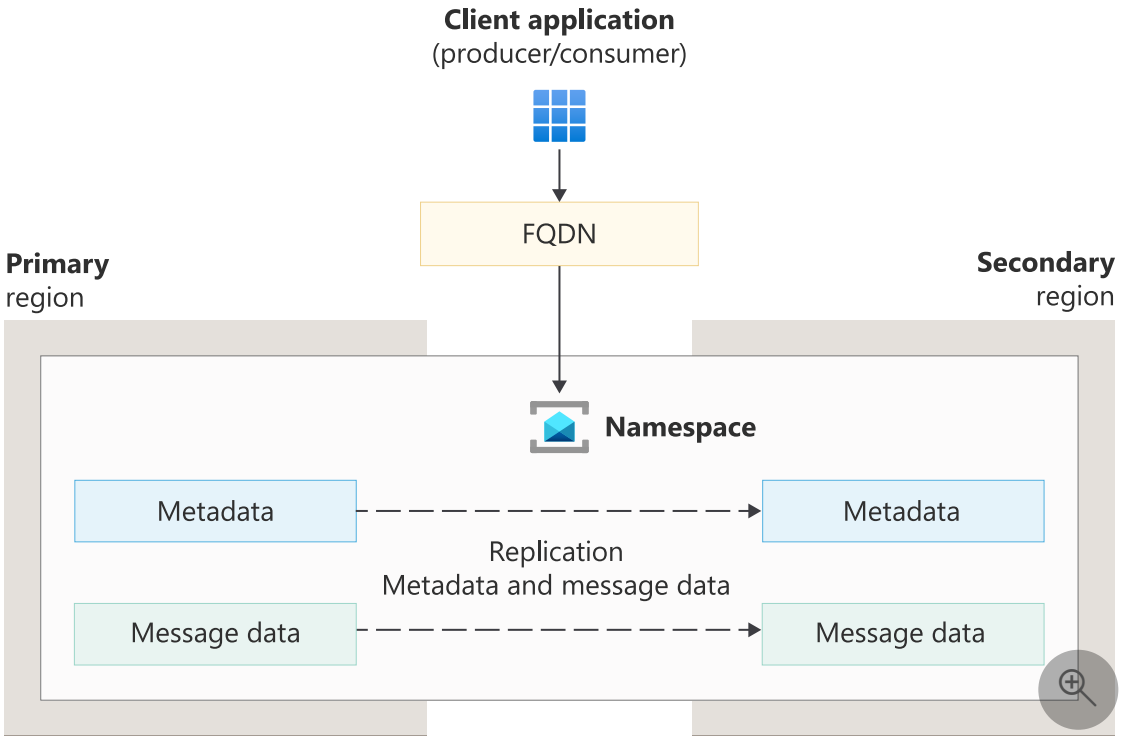
Los espacios de nombres de los niveles Básico y Estándar no incluyen características nativas de varias regiones, pero puede implementar patrones de replicación de nivel de aplicación mediante varios espacios de nombres entre regiones. Para obtener más información, consulte Soluciones personalizadas de varias regiones para lograr resistencia.

Geo-Replication

El nivel Premium admite la replicación geográfica. Esta característica replica metadatos, como entidades, configuración y propiedades, para el espacio de nombres. También replica datos, como los mensajes de las colas y los temas, junto con las propiedades y el estado del mensaje. Configure el enfoque de replicación para la configuración y los datos del espacio de nombres. Esta característica garantiza que los mensajes permanecen disponibles en otra región y le permite cambiar a la región secundaria cuando sea necesario.

Use Geo-Replication para escenarios que requieran resistencia durante interrupciones de regiones y que tengan una tolerancia baja a la pérdida de datos de mensajes.

El espacio de nombres se extiende esencialmente entre regiones. Una región actúa como principal y las demás regiones sirven como secundaria. La suscripción de Azure muestra un único espacio de nombres.



En cualquier momento, puede *promover* la región secundaria a una región primaria. Al promover la región secundaria, Service Bus redirige el nombre de dominio completamente calificado (FQDN) del espacio de nombres a la región secundaria seleccionada y convierte la región primaria anterior en una región secundaria. Decide si desea realizar una *promoción planeada*, lo que significa que espera a que se complete la replicación de datos o una *promoción forzada*, lo que podría provocar la pérdida de datos.

> ⓘ **Nota:**
>
> Service Bus Geo-Replication usa el término *promoción* porque representa mejor el proceso de promoción de una región secundaria a una región primaria (y, posteriormente, degradación de una región primaria a una región secundaria). El término *conmutación por error* también se emplea para describir este proceso general.

En esta sección se resumen aspectos importantes de la replicación geográfica. Revise la documentación completa para obtener información sobre cómo funciona. Para más información, consulte Replicación geográfica de Service Bus.

Requisitos

- **Compatibilidad con regiones:** Puede elegir cualquier región de Azure que admita Service Bus como región primaria o secundaria. No es necesario usar regiones emparejadas de Azure, por lo que seleccione regiones secundarias en función de los requisitos de latencia, cumplimiento o residencia de datos.

- **Nivel:** Para habilitar la replicación geográfica, el espacio de nombres debe usar el nivel Premium.
- **Recuperación ante Desastres Geográficos de Metadatos:** No se puede configurar un espacio de nombres para usar tanto la Geo-Replicación como la Recuperación ante Desastres Geográficos.

Consideraciones

- **Limitaciones de características:** Al habilitar la replicación geográfica, se aplican algunas restricciones. Para más información, consulte Replicación geográfica de Service Bus.
- **Puntos de conexión privados:** Si usa puntos de conexión privados para conectarse al espacio de nombres, también debe configurar las redes en las regiones primarias y secundarias. Para obtener más información, consulte Puntos de conexión privados.

Cost

Para obtener información sobre cómo funcionan los precios para la replicación geográfica, consulte Precios.

Configuración de la compatibilidad con varias regiones

- **Habilite Geo-Replication en un nuevo espacio de nombres.** Para habilitar Geo-Replication en un espacio de nombres durante la creación, consulte Set up Geo-Replication.
- **Migración de metadatos de Recuperación ante desastres geográfica a replicación geográfica.**Cambie del uso de metadatos de Recuperación ante desastres geográfica a replicación geográfica.
- **Cambie el enfoque de replicación.** Para cambiar entre modos de replicación sincrónicos y asincrónicos, consulte Cambiar el modo de replicación.
- **Desactive la replicación geográfica.** Para desactivar Geo-Replication para una región secundaria, consulte Eliminar región secundaria.

Comportamiento cuando todas las regiones están en buen estado

En esta sección se describe qué esperar cuando se configura un espacio de nombres de Service Bus para Geo-Replication y la región primaria está operativa.

- **Enrutamiento del tráfico entre regiones:** Las aplicaciones cliente se conectan a través del FQDN de su espacio de nombres, y su tráfico se dirige a la región primaria.

  Solo la región primaria procesa activamente los mensajes de los clientes durante las operaciones normales. La región secundaria recibe mensajes replicados, pero por lo demás permanece en modo de espera.

- **Replicación de datos entre regiones:** El comportamiento de replicación de datos entre la región principal y secundaria depende de si el emparejamiento de replicación usa la replicación sincrónica o asincrónica.

  - *Síncrono:* Los mensajes se replican en la región secundaria antes de que se complete la operación de escritura.

    Este modo proporciona la mayor garantía de que los datos del mensaje son seguros porque se deben confirmar en la región primaria y secundaria. Pero la replicación sincrónica aumenta considerablemente la latencia de escritura para los mensajes entrantes. También requiere que la región secundaria esté disponible para aceptar la operación de escritura, por lo que una interrupción en la región secundaria hace que se produzca un error en la operación de escritura.

  - *Asincrónico:* El servicio escribe mensajes en la región primaria y, a continuación, completa la operación de escritura. Poco tiempo después, replica los mensajes en la región secundaria.

    Este modo proporciona un mayor rendimiento de escritura que la replicación sincrónica porque no hay ninguna latencia de replicación entre regiones durante las operaciones de escritura. También puede tolerar la pérdida de la región secundaria, a la vez que permite operaciones de escritura en la región primaria. Pero si se produce una interrupción en la región primaria, es posible que los datos que no se repliquen en la región secundaria no estén disponibles o se pierdan.

    Al configurar la replicación asincrónica, se establece el tiempo de retraso máximo aceptable para la replicación. En cualquier momento, puede comprobar el retraso de replicación actual mediante métricas de Azure Monitor.

    Si el retraso de replicación asincrónica aumenta más allá del máximo que especifique, la región primaria comienza a limitar las peticiones entrantes para que la replicación sea capaz de alcanzar el nivel esperado. Para evitar esta situación, seleccione regiones secundarias que no estén demasiado distantes geográficamente y asegúrese de que la capacidad sea suficiente para el rendimiento.

    Algunos tipos de metadatos se replican de forma sincrónica incluso cuando se selecciona el modo de replicación asincrónica.

Para obtener más información, consulte Modos de replicación.

Comportamiento durante una interrupción regional

En esta sección se describe qué esperar cuando se configura un espacio de nombres de Service Bus para la replicación geográfica y hay una interrupción en la región principal o secundaria.

- **Detección y respuesta:** Usted es responsable de decidir cuándo promover la región secundaria del espacio de nombres para convertirse en la nueva región primaria. Microsoft no toma esta decisión ni inicia el proceso por ti, incluso si hay una interrupción de la región. Para conocer los criterios que se deben tener en cuenta al decidir si se debe realizar una conmutación por error, consulte Escenarios recomendados para desencadenar la promoción.

  Para obtener más información sobre cómo promover una región secundaria a la nueva principal, consulte Flujo de promoción.

  Al promover una región secundaria, elija entre una *promoción planeada* o una *promoción forzada*. Una promoción planeada espera a que la región secundaria se ponga al día antes de aceptar tráfico nuevo. Este enfoque evita la pérdida de datos, pero produce tiempo de inactividad.

  Durante una interrupción en la región primaria, generalmente se necesita realizar una promoción forzada. Si la región primaria está disponible y desencadena una promoción por otro motivo, puede elegir una promoción planeada en su lugar.

- **Notificación:** Microsoft no le notifica automáticamente cuando una región está inactiva. Sin embargo, puede usar Azure Service Health para comprender el estado general del servicio, incluidos los errores de región, y puede configurar alertas de Service Health para notificarle problemas.

- **Solicitudes activas:** El comportamiento depende de si la interrupción de la región se produce en la región primaria o en la región secundaria:

  - *Interrupción de la región primaria:* Si la región primaria no está disponible, se finalizan todas las solicitudes activas. Las aplicaciones cliente deben reintentar las operaciones una vez completada la promoción.

  - *Interrupción de la región secundaria:* Una interrupción en la región secundaria puede provocar problemas para las solicitudes activas en las situaciones siguientes:

    - Si usa el modo de replicación sincrónica, la región primaria no puede completar las operaciones de escritura si alguna región secundaria no está disponible.

- Si utiliza el modo de replicación asincrónica, su espacio de nombres limita el flujo y no acepta nuevos mensajes una vez que el retraso de replicación alcanza el máximo que haya configurado.

  Para seguir usando el espacio de nombres en la región primaria, quite el espacio de nombres secundario de la configuración de Geo-Replication.

- **Pérdida de datos esperada:** La cantidad de pérdida de datos depende de si realiza una promoción planeada o forzada y si el modo de replicación es sincrónico o asincrónico:

  - *Promoción planeada:* No se espera ninguna pérdida de datos. Pero durante una interrupción de la región, es posible que no sea posible una promoción planeada porque requiere que todas las regiones primarias y secundarias estén disponibles.

  - *Promoción forzada, replicación sincrónica:* No se espera ninguna pérdida de datos.

  - *Promoción forzada, replicación asincrónica:* Es posible que experimente alguna pérdida de datos para los mensajes recientes que no se replican en la región secundaria y para los cambios de estado que no se replican. La cantidad depende del retraso de replicación. Para comprobar el retraso de replicación actual, use las métricas de Azure Monitor.

  Si realiza una promoción forzada, no puede recuperar los datos perdidos, incluso después de que la región primaria esté disponible.

- **Tiempo de inactividad esperado:** La cantidad de tiempo de inactividad esperado depende de si realiza una promoción planeada o forzada:

  - *Promoción planeada:* El primer paso de una promoción planeada replica los datos en la región secundaria. Ese proceso suele completarse rápidamente, pero en algunas situaciones, puede tardar hasta la longitud del retraso de replicación. Una vez completada la replicación, el proceso de promoción suele tardar entre 5 y 10 minutos. A veces, los servidores del sistema de nombres de dominio (DNS) pueden tardar más tiempo en actualizar las entradas y replicar completamente sus registros en los clientes.

    La región primaria no admite operaciones de escritura durante el proceso completo de promoción.

    Es posible que esta opción no sea posible durante una interrupción de la región porque requiere que todas las regiones primarias y secundarias estén disponibles.

  - *Promoción forzada:* Durante una promoción forzada, Service Bus no espera a que la replicación de datos se complete e inicie la promoción inmediatamente. El proceso de

promoción suele tardar entre 5 y 10 minutos. A veces, las entradas DNS pueden tardar más tiempo en replicarse y actualizarse por completo en todos los clientes.

La región primaria no admite operaciones de escritura durante el proceso completo de promoción.

- **Reenrutamiento del tráfico:** Una vez completada la promoción, el FQDN del espacio de nombres apunta a la nueva región primaria. Pero este redireccionamiento depende de la rapidez con la que se actualizan los registros DNS de los clientes, incluido si sus servidores DNS respetan el período de vida (TTL) de los registros DNS del espacio de nombres.

Recuperación de regiones

Una vez que la región primaria original se haya recuperado, si desea devolver el espacio de nombres a esa región, siga el mismo proceso de promoción regional.

Si ha realizado una promoción forzada durante la interrupción de la región, no puede recuperar los datos perdidos, incluso después de que la región primaria esté disponible.

Prueba de fallos de región

Para probar la replicación geográfica, promueva temporalmente la región secundaria a principal y compruebe que las aplicaciones cliente pueden cambiar entre regiones con una interrupción mínima.

Supervise la duración de la promoción y compruebe que los runbooks y la automatización funcionan correctamente. Después de realizar las pruebas, puede revertir a la configuración original.

Comprenda el posible tiempo de inactividad y pérdida de datos que puede experimentar durante y después del proceso de promoción. Pruebe la Geo-Replication en un entorno no de producción que reproduzca la configuración de su espacio de nombres de producción.

Recuperación de metadatos ante desastres geográficos

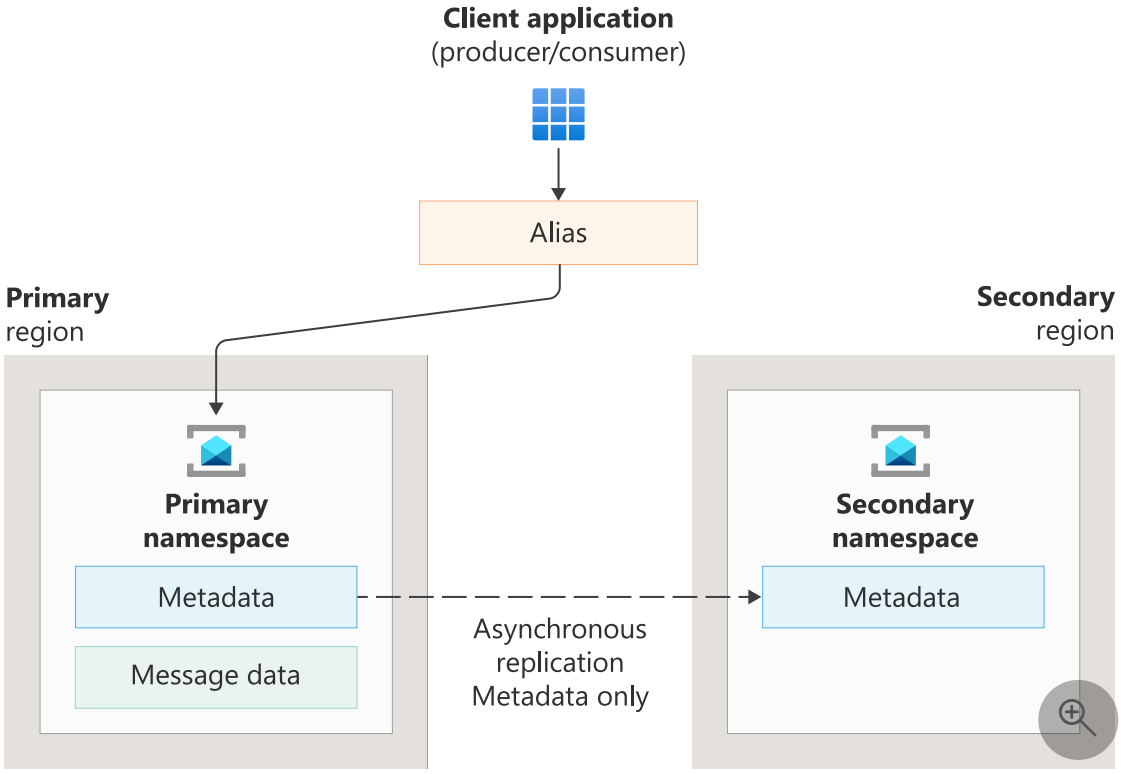
El nivel Premium admite metadatos Geo-Disaster Recovery. Esta característica mejora la recuperación de escenarios de desastres, incluida la pérdida grave de una región. Geo-Disaster Recovery solo replica la configuración y los metadatos del espacio de nombres, y no replica los datos del mensaje. Para admitir la recuperación ante desastres, esta característica garantiza que un espacio de nombres de otra región esté preconfigurado y listo para aceptar inmediatamente mensajes de los clientes. La recuperación ante desastres geográfica sirve

como una solución de recuperación unidireccional y no admite la conmutación por recuperación en la región primaria anterior.

La recuperación de metadatos Geo-Disaster funciona mejor para las aplicaciones que no necesitan mantener todos los mensajes y pueden tolerar cierta pérdida de datos durante un escenario de desastre. La recuperación de metadatos Geo-Disaster Recovery también puede ser adecuada para aplicaciones que replican los datos por sí mismas o que no necesitan replicación de datos en absoluto. Por ejemplo, cuando los mensajes contienen imágenes grandes que posteriormente se convierten en miniaturas, la pérdida de algunos mensajes de una región con errores puede ser aceptable si puede reanudar rápidamente el procesamiento de mensajes nuevos en otra región y reconstruir los mensajes perdidos más adelante.

> **Importante**
>
> Recuperación ante Desastres Geográficos permite la continuidad de las operaciones que tienen la misma configuración, pero **no duplica los datos de los mensajes**. Si debe replicar datos de mensajes, considere la posibilidad de usar la replicación geográfica.

Al configurar metadatos Geo-Disaster Recovery, se crea un *alias* al que se conectan las aplicaciones cliente. El alias es un FQDN que dirige todo el tráfico al espacio de nombres principal de forma predeterminada.



Si se produce un error en la región primaria u otro tipo de desastre, puede iniciar manualmente una conmutación por error unidireccional desde la región primaria a la región

secundaria en cualquier momento. Puede realizar una *conmutación por error segura*, que espera a que se complete la replicación antes de cambiar a la secundaria. Es posible que esta opción no esté disponible durante una interrupción de la región. Después de que se inicie una conmutación por error, se completa casi al instante. Durante el proceso de conmutación por error, el alias de recuperación ante desastres geográfica se redirige al espacio de nombres secundario.

En esta sección se resumen los aspectos importantes de Geo-Disaster Recovery. Revise la documentación completa para obtener información sobre cómo funciona. Para obtener más información, consulte Service Bus Geo-Disaster Recovery.

Requisitos

- **Compatibilidad con regiones:** Puede elegir cualquier región de Azure que admita Service Bus como espacio de nombres principal o secundario. No es necesario usar regiones emparejadas de Azure, por lo que seleccione regiones secundarias en función de los requisitos de latencia, cumplimiento o residencia de datos.
- **Nivel:** Para habilitar los metadatos Geo-Disaster Recovery, ambos espacios de nombres deben usar el nivel Premium.
- **Particionamiento:** No es posible asociar un espacio de nombres particionado con un espacio de nombres no particionado.
- **Recuperación ante Desastres Geográficos de Metadatos:** No se puede configurar un espacio de nombres para usar tanto la Geo-Replicación como la Recuperación ante Desastres Geográficos.

Consideraciones

- **Limitaciones de características:** Al habilitar Geo-Disaster Recuperación, se aplican algunas restricciones. Para obtener más información, vea Puntos importantes que se deben tener en cuenta y Consideraciones.
- **Asignaciones de roles:** las asignaciones de control de acceso basado en roles (RBAC) de Microsoft Entra a entidades en el espacio de nombres principal no se replican en el espacio de nombres secundario. Cree asignaciones de roles manualmente en el espacio de nombres secundario para proteger el acceso a esas entidades.
- **Diseño de aplicaciones:** La Recuperación ante desastres geográficos requiere consideraciones específicas al diseñar sus aplicaciones cliente. Para más información, consulte Consideraciones.

- **Puntos de conexión privados:** Si usa puntos de conexión privados para conectarse al espacio de nombres, configure las redes en la región primaria y secundaria. Para obtener más información, consulte Puntos de conexión privados.
- **Espacios de nombres migrados de Estándar a Premium:** Si tu espacio de nombres está en el nivel Estándar y lo migras al nivel Premium, necesitas gestionar el alias de una manera diferente. Para obtener más información, consulte Service Bus de Standard a Premium.

Cost

Al habilitar la recuperación ante desastres geográfica de metadatos, usted paga por los espacios de nombres principal y secundario.

Configuración de la compatibilidad con varias regiones

- **Cree un emparejamiento de recuperación de desastres geográfico de metadatos.** Para configurar la recuperación ante desastres entre los espacios de nombres principal y secundario, consulte Configuración y proceso de conmutación por error.
- **Desactive los metadatos de Recuperación de Geo-Desastres.** Para interrumpir el emparejamiento entre espacios de nombres, consulte Configuración y flujo de conmutación por error.

Planeamiento y administración de capacidad

Cuando planee implementaciones de varias regiones, asegúrese de que ambas regiones tienen capacidad suficiente para controlar la carga completa si se produce un error en una región. La región secundaria permanece pasiva durante las operaciones normales, pero debe controlar inmediatamente el tráfico después de la conmutación por error. Planee cómo escalar la capacidad del espacio de nombres secundario para que pueda recibir tráfico de producción sin demora. Si puede tolerar un tiempo de inactividad adicional durante el proceso de conmutación por error, puede escalar la capacidad del espacio de nombres secundario durante o después de la conmutación por error. Para reducir el tiempo de inactividad, aprovisione la capacidad con antelación en el espacio de nombres secundario para que permanezca lista para recibir la carga de producción.

Comportamiento cuando todas las regiones están en buen estado

En esta sección se describe qué esperar cuando se configura un espacio de nombres de Service Bus para Geo-Disaster Recovery y la región primaria está operativa.

- **Enrutamiento de tráfico entre regiones:** Las aplicaciones de cliente se conectan a través del alias de Recuperación ante Desastres Geo para el espacio de nombres, y su tráfico se dirige al espacio de nombres primario en la región primaria.

  Solo el espacio de nombres principal procesa activamente los mensajes de los clientes durante las operaciones normales. El espacio de nombres secundario permanece en modo de espera y se produce un error en las solicitudes de acceso a los datos.

- **Replicación de datos entre regiones:** Solo los metadatos de configuración se replican entre los espacios de nombres. La replicación de la configuración se produce de forma continua y asincrónica.

  Todos los datos del mensaje permanecen solo en el espacio de nombres principal y no se replican en el espacio de nombres secundario.

Comportamiento durante una interrupción regional

En esta sección se describe qué esperar cuando se configura un espacio de nombres de Service Bus para Geo-Disaster Recovery y se produce una interrupción en la región primaria.

- **Detección y respuesta:** usted es responsable de monitorear la salud de la región e iniciar la conmutación por error manualmente. Microsoft no inicia la conmutación por error ni promueve automáticamente una región secundaria, incluso si la región primaria está inactiva.

  Para obtener más información sobre cómo iniciar la conmutación por error, consulte el Flujo de conmutación por error.

  Al iniciar una conmutación por error, elija si desea realizar una *conmutación por error segura* o una *conmutación por error estándar*, como una conmutación por error forzada o manual. Una conmutación por error segura espera a que la replicación se complete en la región secundaria antes de que se inicie la conmutación por error. Este enfoque reduce la pérdida de metadatos, pero presenta tiempo de inactividad. La conmutación por error segura requiere que los espacios de nombres estén en la misma suscripción de Azure.

  Durante una interrupción en la región primaria, normalmente debe realizar una conmutación por error forzada. Si la región primaria está disponible y desencadena una conmutación por error por otro motivo, puede elegir una conmutación por error planeada.

  La conmutación por error es una operación unidireccional, por lo que es necesario reestablecer el emparejamiento de Geo-Disaster Recovery más adelante. Para obtener más información, consulte Recuperación de regiones.

- **Notificación:** Microsoft no le notifica automáticamente cuando una región está inactiva. Sin embargo, puede usar Azure Service Health para comprender el estado general del servicio, incluidos los errores de región, y puede configurar alertas de Service Health para notificarle problemas.

- **Solicitudes activas:** Las solicitudes activas terminan cuando se inicia la conmutación por error. Las aplicaciones cliente deben reintentar las operaciones una vez completada la conmutación por error.

- **Pérdida de datos esperada:**

  - *Metadatos:* Normalmente, la configuración y los metadatos se replican en el espacio de nombres secundario. La replicación de metadatos se produce de forma asincrónica, por lo que es posible que los cambios recientes no se repliquen, especialmente los cambios complejos. Compruebe la configuración del espacio de nombres secundario antes de que los clientes accedan a él.

  - *Datos del mensaje:* Los datos del mensaje no se replican entre regiones. Si la región primaria deja de funcionar, los mensajes del espacio de nombres principal no estarán disponibles.

    Los mensajes no se pierden permanentemente a menos que un desastre catastrófico cause una pérdida total de la región primaria. Si la región se recupera, puede recuperar mensajes del espacio de nombres principal más adelante.

- **Tiempo de inactividad esperado:** Normalmente, la conmutación ocurre en un plazo de 5 a 10 minutos. Los clientes pueden tardar más tiempo en replicar y actualizar completamente las entradas DNS.

- **Reenrutamiento del tráfico:** Los clientes que usan el alias de recuperación de Geo-Disaster para conectarse al espacio de nombres se redirigen automáticamente al espacio de nombres secundario después de la conmutación por error. Pero este redireccionamiento depende de los servidores DNS que respetan el TTL de los registros DNS del espacio de nombres y los clientes que reciben esos registros DNS actualizados.

Recuperación de regiones

Una vez que la región primaria original se recupere, debe reestablecer manualmente el emparejamiento y, opcionalmente, realizar una conmutación por recuperación. Cree un nuevo emparejamiento de recuperación de Geo-Disaster con la región recuperada como secundaria y luego realice otra conmutación por error si desea regresar a la región original. Este proceso implica la posible pérdida de datos de los mensajes enviados al espacio de nombres principal temporal.

Si el desastre provoca la pérdida de todas las zonas de la región primaria, los datos podrían ser irrecuperables. En otros escenarios, los datos del mensaje que permanecen en el espacio de nombres principal antes de la conmutación por error son recuperables. Puede obtener mensajes históricos del espacio de nombres principal anterior después de restaurar el acceso. Usted es responsable de configurar las aplicaciones para recibir y procesar esos mensajes. Microsoft no los restaura automáticamente en tu región secundaria.

Prueba de fallos de región

Para probar los procesos de respuesta y recuperación ante desastres, realice una conmutación por error planeada durante una ventana de mantenimiento. Inicie la conmutación por error desde el espacio de nombres principal al espacio de nombres secundario y compruebe que las aplicaciones pueden conectarse y procesar mensajes desde el nuevo espacio de nombres principal.

Supervise la duración de la conmutación por error y compruebe que los runbooks y la automatización funcionan correctamente. Después de realizar las pruebas, puede revertir a la configuración original.

Comprenda el posible tiempo de inactividad y la pérdida de datos que puede experimentar durante y después del proceso de conmutación por error. Pruebe la recuperación ante desastres geográfica de metadatos en un entorno de preproducción que refleje la configuración de su espacio de nombres de producción.

Soluciones personalizadas de varias regiones para la resistencia

La Geo-Replicación y la Recuperación ante Desastres mediante Metadata proporcionan resiliencia frente a interrupciones regionales y otros problemas, y son adecuadas para la mayoría de las cargas de trabajo. Es posible que estas funcionalidades no se adapten a sus necesidades en las situaciones siguientes:

- Tiene requisitos para la replicación personalizada o para mantener varias regiones activas simultáneamente.
- Actualmente usas un nivel de Service Bus que no admite estas características.

Hay varios patrones de diseño que proporcionan distintos tipos de compatibilidad con varias regiones en Service Bus. Muchos de estos patrones requieren que implemente varios espacios de nombres y configure la aplicación para usarlos correctamente. Para obtener más información, consulte los siguientes recursos:

- Proteger aplicaciones de Service Bus contra interrupciones y desastres
- Replicación de mensajes y federación entre regiones

Resistencia al mantenimiento del servicio

Service Bus realiza un mantenimiento normal. Durante el mantenimiento planeado, los espacios de nombres se mueven a un nodo redundante que contiene las actualizaciones más recientes. Durante el traslado, el SDK del cliente se desconecta y, a continuación, se reconecta automáticamente al espacio de nombres. Las actualizaciones normalmente se completan en un plazo de 30 segundos. Es importante que las aplicaciones estén preparadas para las desconexiones de red transitorias que pueden producirse durante períodos de mantenimiento.

Para más información, consulte Eventos de mantenimiento de Azure para Service Bus.

Copias de seguridad y restauración

Service Bus no está diseñado para el almacenamiento de datos a largo plazo. Los datos normalmente se almacenan en un tema o cola durante un breve período de tiempo. Después, se procesa o se conserva en otro sistema de almacenamiento de datos y, a continuación, se elimina. Debido a este diseño, Service Bus mantiene automáticamente réplicas de datos de mensajes, pero no proporciona ninguna funcionalidad de copia de seguridad y restauración para esos datos.

En escenarios que requieren retención de mensajes a largo plazo, considere la posibilidad de implementar el archivado de nivel de aplicación en Azure Storage u otros servicios de almacenamiento duraderos.

Acuerdo de nivel de servicio

El contrato de nivel de servicio (SLA) para los servicios de Azure describe la disponibilidad esperada de cada servicio y las condiciones que la solución deberá cumplir para lograr esa expectativa de disponibilidad. Para obtener más información, consulte Acuerdos de Nivel de Servicio para servicios en línea.

Service Bus proporciona un SLA para todos los espacios de nombres. El SLA de disponibilidad es mayor cuando el espacio de nombres cumple los siguientes criterios:

- Usa el nivel Premium.
- Se encuentra en una región que tiene zonas de disponibilidad.
- Usa particionamiento.

Contenido relacionado

- Introducción a Service Bus
- Recuperación ante desastres geográficos de Service Bus
- Replicación geográfica de Service Bus

Last updated on 27/01/2026

Confiabilidad en Azure Web PubSub Service

Azure Web PubSub Service es un servicio de mensajería totalmente administrado en tiempo real que permite la comunicación bidireccional entre servidores y clientes mediante el protocolo WebSocket. Un único recurso de Web PubSub puede escalar a un millón de conexiones WebSocket simultáneas. El servicio admite varios patrones de mensajería, como difusión de servidor a cliente, mensajería a grupos con nombre, pub/sub de cliente a cliente y streaming de tokens de IA.

Cuando se usa Azure, la confiabilidad es una responsabilidad compartida. Microsoft proporciona una variedad de capacidades para apoyar la resiliencia y la recuperación. Es responsable de comprender cómo funcionan esas funcionalidades dentro de todos los servicios que usa y de seleccionar las funcionalidades que necesita para cumplir los objetivos empresariales y los objetivos de tiempo de actividad.

En este artículo se describe cómo hacer que el servicio Azure Web PubSub sea resistente a una variedad de posibles interrupciones y problemas, incluidos errores transitorios, errores de zona de disponibilidad y errores en toda la región. También describe cómo el servicio controla el mantenimiento y resalta la información clave sobre el contrato de nivel de servicio (SLA) de Azure Web PubSub servicio.

Recomendaciones de implementación de producción para la confiabilidad

Para cargas de trabajo de producción, siga estas recomendaciones:

- ✓ Use el nivel Premium. El nivel Premium es resistente a los errores de zona de disponibilidad en regiones admitidas y permite configurar la replicación geográfica.
- ✓ Use el SDK de cliente de Azure Web PubSub al compilar aplicaciones cliente o siga las instrucciones de control de errores transitorios mediante la reconexión segura. Las conmutaciones por error de zona, las conmutaciones por error de región y los errores transitorios interrumpen las conexiones activas.
- ✓ Habilite la replicación geográfica para protegerse frente a errores en toda la región. Dimensione cada réplica con suficientes unidades para manejar toda la carga de tráfico esperada durante un evento de failover.

Introducción a la arquitectura de confiabilidad

En esta sección se describen algunos de los aspectos importantes de cómo funciona el servicio que es más relevante desde una perspectiva de confiabilidad. En la sección se presenta la arquitectura lógica, que incluye algunos de los recursos y características que se implementan y usan. También se describe la arquitectura física, que proporciona detalles sobre cómo funciona el servicio en segundo plano.

Arquitectura lógica

El recurso que crea es un *recurso de Web PubSub*. Configure un recurso con una serie de *unidades*, que representa la capacidad del recurso, incluido el número máximo de conexiones simultáneas. Para obtener más información, consulte Performance guide for Azure Web PubSub Service.

Un recurso de Web PubSub tiene un punto de conexión único globalmente similar a `contoso.webpubsub.azure.com`. Los clientes establecen conexiones de WebSocket a este punto de conexión. Los servidores de aplicaciones se conectan al mismo punto de conexión para enviar mensajes y recibir eventos de los clientes.

Para obtener más información, consulte Azure Web PubSub service internals.

Arquitectura física

Azure Web PubSub Service administra el estado de conexión de WebSocket y el enrutamiento de mensajes en un conjunto de recursos de proceso. Microsoft administra la infraestructura subyacente. No ve ni interactúa directamente con máquinas virtuales individuales que usa el servicio u otros componentes de infraestructura.

Resistencia a errores transitorios

Los errores transitorios son errores breves e intermitentes en los componentes. Se producen con frecuencia en un entorno distribuido como la nube y son una parte normal de las operaciones. Los errores transitorios se corrigen después de un breve período de tiempo. Es importante que las aplicaciones puedan controlar errores transitorios, normalmente mediante el reintento de solicitudes afectadas.

Todas las aplicaciones hospedadas en la nube deben seguir las instrucciones de control de errores transitorios de Azure cuando se comunican con cualquier API, bases de datos y otros componentes hospedados en la nube. Para obtener más información, consulte Recomendaciones para controlar errores transitorios.

WebSocket es un protocolo de conexión de larga duración. Los eventos de red transitorios, los reinicios de la infraestructura de back-end y las operaciones de mantenimiento del servicio pueden quitar una conexión activa. Una reconexión básica restaura la conexión, pero sin lógica adicional, el cliente pierde los mensajes que estaban en curso o en cola durante la interrupción.

Azure Web PubSub Service soluciona este problema a través de *subprotocolos fiables* que operan sobre la conexión WebSocket sin procesar. Los subprotocolos realizan un seguimiento de la secuencia de mensajes y el estado de conexión para que, cuando una conexión cae, el cliente renegocia con el servicio y se reanuda desde donde se dejó.

Normalmente, no hay pérdida de mensajes después de que una conexión se corte y vuelva a establecerse. Sin embargo, hay algunas situaciones en las que puede producirse una pérdida de mensajes. Por ejemplo, si el cliente se desconecta durante más de un minuto y, a continuación, se vuelve a conectar con el mismo identificador de conexión, la operación de reconexión muestra un estado de error para indicar que puede producirse una pérdida de mensajes.

Para aprovechar los subprotocolos confiables, siga estas recomendaciones:

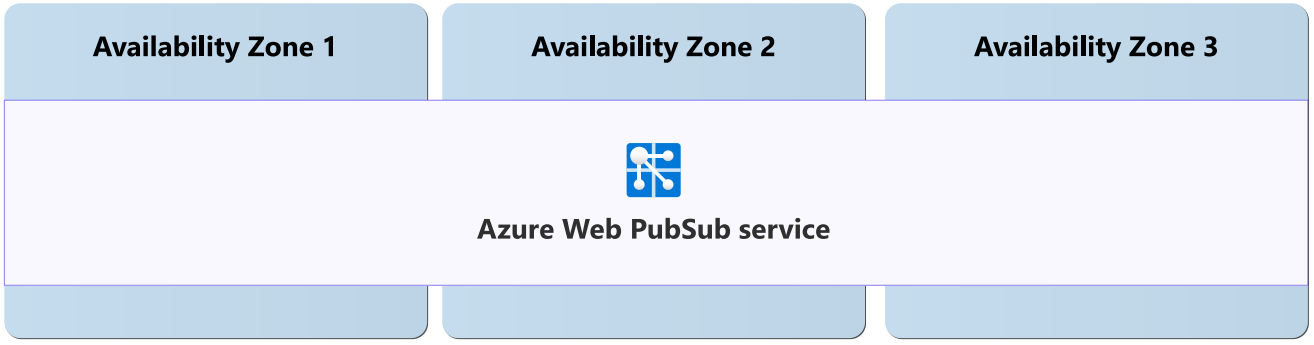
- **Siempre que sea posible, use un SDK de cliente de Azure Web PubSub.** El SDK implementa automáticamente el subprotocolo confiable. No se requiere configuración adicional. Para obtener más información, consulte:
  - SDK del lado cliente de Web PubSub para JavaScript
  - Azure Web PubSub biblioteca cliente para .NET
  - Azure Web PubSub biblioteca cliente para Python
  - Biblioteca cliente de Azure WebPubSub para Java
- **Si no puede usar el SDK, implemente uno de los subprotocolos confiables** directamente en el código de cliente de WebSocket. Para obtener instrucciones completas sobre la especificación y la implementación, consulte Creación de clientes de WebSocket confiables.

Resistencia a errores de zona de disponibilidad

Las zonas de disponibilidad son grupos físicamente independientes de centros de datos dentro de una región de Azure. Cuando una zona falla, los servicios pueden transferirse a una de las zonas restantes.

Azure Web PubSub Service admite implementaciones con redundancia de zona cuando se usa el nivel Premium. Al crear o actualizar un recurso de Web PubSub de nivel Premium en una región que admita zonas de disponibilidad, la redundancia de zona se habilita automáticamente. El

servicio distribuye su infraestructura en varias zonas de disponibilidad de la región. Si se produce un error en una zona, el servicio enruta el tráfico a la infraestructura en la zona en buen estado.



Requisitos

- **Compatibilidad con regiones:** La redundancia de zona se admite en la mayoría de las regiones en las que se aplican ambas condiciones:
  - Servicio de Azure Web PubSub está disponible. Para obtener una lista de las regiones en las que el servicio está disponible, consulte Disponibilidad del producto por región.
  - La región admite zonas de disponibilidad. Para obtener una lista de regiones con zonas de disponibilidad, consulte Azure lista de regiones.

  Sin embargo, El oeste de Japón no admite actualmente redundancia de zona para Azure Web PubSub.

- **Nivel:** La redundancia de zona está disponible en el nivel Premium.

Costo

La redundancia de zona no implica un costo adicional, y pagas la tarifa estándar del nivel Premium. Para obtener más información, consulte precios de servicio Azure Web PubSub.

Configurar soporte de zonas de disponibilidad

La redundancia de zona no requiere ninguna configuración más allá de seleccionar el nivel Premium. Se habilita automáticamente en ambos casos:

- **Cree un nuevo recurso de Web PubSub con redundancia de zona.** Seleccione una SKU de nivel Premium al crear el recurso. Para obtener más información, consulte Crear un recurso de Azure Web PubSub.

- **Actualice un recurso existente al nivel Premium.** La redundancia de zona se habilita automáticamente al actualizar un recurso existente a una SKU de nivel Premium. La

actualización de Estándar a Premium no provoca tiempo de inactividad del servicio. Para obtener más información, consulte Escalar una instancia de servicio de Azure Web PubSub.

Comportamiento cuando todas las zonas están en buen estado

En esta sección se describe qué esperar al configurar un recurso de Azure Web PubSub para la redundancia de zona y todas las zonas de disponibilidad están operativas.

- **Operación entre zonas:** Azure Web PubSub Service gestiona automáticamente cómo se distribuyen las conexiones y las operaciones entre zonas de disponibilidad. La infraestructura de varias zonas procesa el tráfico en un modelo activo-activo. No es necesario configurar nada para aprovechar este comportamiento. El servicio enruta los mensajes entre instancias entre zonas automáticamente, por lo que un mensaje enviado por un cliente de una zona se entrega a los clientes conectados en cualquier otra zona.

- **Replicación de datos entre zonas:** Azure Web PubSub Service no conserva datos de clientes. El servicio mantiene los metadatos de sesión, como el estado de conexión y la información de secuencia de mensajes para las conexiones activas. Estos metadatos se replican sincrónicamente entre zonas de disponibilidad.

Comportamiento durante un fallo de zona

En esta sección se describe qué esperar al configurar un recurso de Azure Web PubSub para la redundancia de zona y hay una interrupción en una de las zonas de disponibilidad.

- **Detección y respuesta:** La plataforma del servicio Azure Web PubSub es responsable de detectar un error en una zona de disponibilidad. No es necesario realizar ninguna acción para iniciar una conmutación por error de zona.

- **Notificación**: Microsoft no le notifica automáticamente cuando una zona está inactiva. Sin embargo, puede usar Azure Resource Health para supervisar el estado de un recurso individual y puede configurar alertas de Resource Health para notificarle problemas. También puede usar Azure Service Health para comprender el estado general del servicio, incluidos los errores de zona, y puede configurar alertas de Service Health para notificarle problemas.

- **Solicitudes activas:** Durante un fallo de zona, las conexiones activas de WebSocket a la infraestructura de la zona afectada se interrumpen. Si los clientes controlan los errores transitorios correctamente, por ejemplo, volviendo a conectarse después de un breve período de tiempo, normalmente evitan un impacto significativo.

- **Expected data loss:** Azure Web PubSub Service no conserva los mensajes, por lo que no se espera que un error de zona cause pérdida de datos en el servicio Azure Web PubSub. Sin embargo, las conexiones activas se interrumpen durante un evento de caída de zona, por lo que es posible que se pierdan los datos que se están transmitiendo activamente.

  Si los publicadores usan un SDK de cliente de Azure Web PubSub o implementan los subprotocolos confiables, el servicio los reconoce después de que el servicio los reciba. Cuando se confirma un mensaje, se replica en todas las zonas de disponibilidad, por lo que el error de zona del publicador no hace que se pierda el mensaje. Sin embargo, si un suscriptor no recibe el mensaje antes de quitarlo, es posible que no reciba el mensaje.

- **Tiempo de inactividad esperado:** La reconexión de conexiones activas eliminadas suele tardar unos segundos. Los clientes que implementan la lógica de reconexión experimentan una interrupción mínima.

- **Redistribution:** Azure Web PubSub Service detecta la pérdida de la zona y redistribuye automáticamente el tráfico entre las zonas correctas. No tiene que realizar ninguna acción.

Recuperación de zona

Cuando se recupera una zona de disponibilidad, Azure Web PubSub Service lo vuelve a reinsertar automáticamente en la topología de servicio activa. No es necesario realizar ninguna acción para la recuperación de zona.

Una vez recuperada una zona, es posible que se dirijan nuevas conexiones a la infraestructura de la zona recuperada. Las conexiones existentes no se moverán ni se reequilibrarán en la zona recuperada, pero se volverán a equilibrar gradualmente a medida que las conexiones existentes bajen y se vuelvan a conectar con el tiempo. El desequilibrio de conexiones entre zonas no tiene ningún impacto en la carga de trabajo.

Prueba de fallos de zona

Azure Web PubSub Service administra automáticamente el enrutamiento del tráfico, la conmutación por error y la recuperación de zona para los recursos de nivel Premium con redundancia de zona. No es necesario iniciar nada. Dado que la redundancia de zona está totalmente administrada, no es necesario validar los procesos de fallo de zona de disponibilidad.

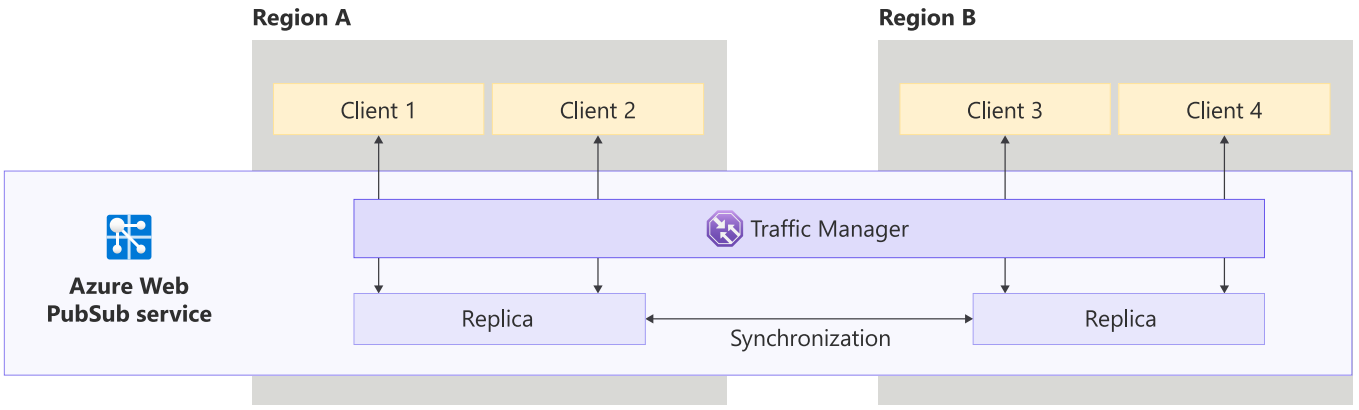
Resistencia a errores en toda la región

Azure Web PubSub Service es un servicio de una sola región. Si la región deja de estar disponible, el recurso de Web PubSub tampoco está disponible.

Para proteger la aplicación frente a un error en toda la región, puede usar la *replicación geográfica*, que está disponible en el nivel Premium. Como alternativa, puede crear una solución de varias regiones personalizada mediante la implementación de varios recursos de Web PubSub en diferentes regiones.

Replicación geográfica

La replicación geográfica permite agregar *replicas* del recurso de Web PubSub en otras regiones de Azure. Todas las réplicas comparten un único punto de conexión (`contoso.webpubsub.azure.com`). Detrás de este punto de conexión, Azure Traffic Manager utiliza el enrutamiento basado en DNS para dirigir a cada cliente a la réplica regional más cercana que esté en buen estado. Si se produce un error en una región, el Traffic Manager detecta el error a través de comprobaciones de estado y deja de dirigir a los clientes a esa réplica. Las nuevas conexiones de cliente se enrutan automáticamente a la réplica saludable más cercana.



La región en la que creó el recurso Web PubSub en se denomina *región primaria* y su réplica es la *réplica principal*. El plano de control del recurso principal administra la configuración del recurso de Web PubSub.

Requisitos

- **Region support:** Puede agregar réplicas en cualquier región en la que Azure Web PubSub Service esté disponible.
- **Nivel:** Debe usar el nivel Premium para habilitar la replicación geográfica.
- **Límite de réplica:** Cada recurso principal de Web PubSub admite hasta ocho réplicas.

Consideraciones

- **Herencia de configuración:** Las réplicas heredan la mayoría de las opciones de configuración del recurso principal. Algunas opciones deben configurarse por separado en cada réplica. Para obtener la lista completa de opciones de configuración que no se heredan, consulte Geo-replication en Azure Web PubSub.
- **Cambios de configuración:** El plano de control principal, en la región primaria, procesa los cambios de configuración en el recurso de Web PubSub. Si el plano de control principal no está disponible, no podrá actualizar la configuración del recurso, aunque las réplicas existentes seguirán procesando el tráfico de datos sin interrupción.

Costo

Cada réplica se factura por separado en función de su propio recuento de unidades y del volumen de mensajes salientes. Si un mensaje se transfiere entre réplicas y, a continuación, se entrega a un cliente o servidor de otra región, se factura como mensaje saliente. Para obtener más información, consulte precios de servicio Azure Web PubSub.

Configuración de la replicación geográfica

Para agregar o quitar una réplica a un recurso de Web PubSub, consulte Geo-replication en Azure Web PubSub.

Planeamiento y administración de capacidad

Cada réplica controla el tráfico de forma independiente. Durante una conmutación por error regional, los clientes de la región que ha fallado se reconectan a la réplica saludable más cercana. Para asegurarse de que las réplicas supervivientes tienen suficiente capacidad para absorber esta carga adicional, configure cada réplica con unidades que puedan controlar el tráfico total esperado de la carga de trabajo, no solo la parte que normalmente sirve.

Como alternativa, habilite el escalado automático en cada réplica para que las unidades se amplíen automáticamente en respuesta a una mayor carga. El escalado automático sigue funcionando cuando una réplica secundaria no está disponible, pero el escalado automático no funciona si el plano de control principal no está disponible. Para obtener más información sobre el escalado automático, consulte Unidades de escalado automático de un servicio de Azure Web PubSub.

Para obtener instrucciones generales sobre el sobreaprovisionamiento como estrategia, consulte Administración de la capacidad mediante el sobreaprovisionamiento.

Comportamiento cuando todas las regiones están en buen estado

En esta sección se describe qué esperar al configurar Azure Web PubSub Service para la replicación geográfica y todas las regiones están operativas.

- **Operación entre regiones:** Azure Traffic Manager enruta cada cliente a la réplica regional saludable más cercana. Los clientes de diferentes áreas geográficas pueden conectarse a diferentes réplicas. El servicio Web PubSub sincroniza los mensajes entre réplicas para que los clientes conectados a cualquier réplica puedan comunicarse entre sí.

- **Replicación de datos entre regiones:** Cuando se envía un mensaje a una réplica, el servicio transfiere sincrónicamente ese mensaje a otras réplicas para que los clientes conectados a otro lugar puedan recibirlo. La sobrecarga de sincronización es mínima para los patrones de mensajería más comunes, como la difusión a grupos grandes o la mensajería de una sola conexión. La mensajería a grupos pequeños (menos de 10 miembros) puede producir una sobrecarga de sincronización ligeramente mayor.

  Azure Web PubSub Servicio no conserva los mensajes; solo la entrega activa se sincroniza entre réplicas.

Comportamiento durante una falla de región

En esta sección se describe qué esperar al configurar Azure Web PubSub Service para la replicación geográfica y se produce una interrupción en una de las regiones de réplica.

- **Detección y respuesta:** El servicio Web PubSub es responsable de detectar un error en una región y redirigir automáticamente el tráfico entrante a una réplica en una de las demás regiones que configure.

- **Notificación**: Microsoft no le notifica automáticamente cuando una región está inactiva. Sin embargo:

  - Puede usar Azure Resource Health para supervisar el estado de un recurso individual y puede configurar alertas de Resource Health para notificarle problemas.

  - Puede usar Azure Service Health para comprender el estado general del servicio, incluidos los errores de región, y puede configurar alertas de Service Health para notificarle los problemas.

- **Solicitudes activas:** Se quitan las conexiones WebSocket activas a la réplica de la región con errores. Los clientes deben volver a conectarse después de que la réplica conmute por error.

- **Pérdida de datos esperada:** Azure Web PubSub Service no almacena los mensajes. Es posible que se pierdan los mensajes que estaban en tránsito a los clientes en la región que falló en el momento del fallo. No se espera ninguna pérdida de datos persistente porque el servicio no almacena los datos del cliente.

- **Tiempo de inactividad esperado:** Azure Traffic Manager realiza verificaciones de salud en cada réplica. Cuando una interrupción en una región provoca que una réplica falle en su verificación de estado, Traffic Manager elimina el endpoint de esa réplica de los resultados de la resolución DNS. Después de quitar el punto de conexión, deben transcurrir 90 segundos del TTL DNS antes de que los clientes vean los registros DNS actualizados. En total, la transición suele tardar unos minutos. Los clientes bien diseñados que implementan la lógica de reconexión pueden reanudar el funcionamiento normal después de volver a conectarse a la réplica sana.

  Si el plano de control principal no está disponible, no puede realizar ningún cambio en la configuración del recurso de Web PubSub ni en sus réplicas. Sin embargo, las conexiones de WebSocket siguen funcionando en réplicas saludables.

- **Redistribution:** Azure Traffic Manager dirige la solicitud entrante a las réplicas saludables. Sin embargo, si un cliente intenta volver a conectarse antes de que Azure Traffic Manager haya detectado el fallo de conmutación de réplica y se hayan propagado las entradas DNS actualizadas al cliente, el intento de reconexión de un cliente podría continuar apuntando a la región no disponible, lo que provocaría un fallo.

  Una vez que la actualización de DNS se propaga, los clientes que se reconectan se enrutan automáticamente a la réplica saludable más cercana.

Recuperación de regiones

Cuando se recupera la región fallida, la verificación de salud del Traffic Manager detecta la réplica restaurada y vuelve a incluir su punto de conexión en la resolución DNS. Los clientes conectados actualmente a otras réplicas no se ven afectados y permanecen conectados hasta que se desconectan. Las nuevas conexiones se enrutan a la réplica de la región recuperada cuando es la réplica más cercana y en funcionamiento.

Prueba de fallos de región

Para simular un fallo de región y con el fin de probar el comportamiento de reconexión de su aplicación cliente, puede deshabilitar el punto de conexión de una réplica. Esta acción hace que Traffic Manager deje de enrutar el tráfico a esa réplica, lo que le permite observar cómo se

comportan los clientes cuando la réplica a la que se conectan deja de estar disponible. Para obtener pasos detallados, consulte Deshabilitar o habilitar el punto de conexión de réplica.

Soluciones de varias regiones personalizadas para la resistencia

Si necesita resistencia entre regiones, pero no usa la replicación geográfica, puede implementar y administrar recursos independientes de Web PubSub en varias regiones e implementar su propia lógica de conmutación por error en el servidor de aplicaciones. Este enfoque es más complejo que la replicación geográfica y no admite la tolerancia a fallos sin tiempo de inactividad para la conexión de cliente a cliente. Para una visión general detallada de la arquitectura, patrones de conmutación por error y guías para la realización de pruebas, consulte Resiliencia y recuperación ante desastres en Azure Web PubSub Service.

Copias de seguridad y restauración

Azure Web PubSub Service es un servicio de mensajería sin estado. No conserva los mensajes del cliente y no tiene capacidad de copia de seguridad ni restauración.

Para proteger la configuración de sus recursos, defina sus recursos de Web PubSub mediante la infraestructura como código (como Bicep o plantillas de ARM) y almacene esas definiciones en el control de versiones. Si necesita volver a crear un recurso, vuelva a implementarlo desde la configuración almacenada.

Resistencia al mantenimiento del servicio

Microsoft aplica periódicamente actualizaciones de servicio y realiza otro mantenimiento. La plataforma Azure controla estas actividades automáticamente, lo que garantiza que el mantenimiento sea transparente y sin problemas. No se espera ningún tiempo de inactividad durante los eventos de mantenimiento a menos que se le haya informado a través del mantenimiento planeado de Azure Service Health.

Acuerdo de nivel de servicio

El acuerdo de nivel de servicio (SLA) para Azure servicios describe la disponibilidad esperada de cada servicio y las condiciones que la solución debe cumplir para lograr esa expectativa de disponibilidad. Para obtener más información, consulte Acuerdos de Nivel de Servicio para servicios en línea.

Contenido relacionado

- ¿Qué es Azure Web PubSub servicio?
- Replicación geográfica en Azure Web PubSub
- Resiliencia y recuperación ante desastres en Azure Web PubSub Service
- Escalar una instancia del Servicio Web PubSub de Azure
- Escalar automáticamente las unidades de un servicio de Azure Web PubSub
- Confiabilidad en Azure

ⓘ **Nota:** El autor ha creado este artículo con ayuda de inteligencia artificial. Más información

---

Last updated on 30/04/2026

Confiabilidad en el Registro de dispositivos de Azure

Azure Registro de dispositivos almacena información sobre los recursos y dispositivos en la nube. Device Registry representa los recursos como recursos de Azure en la nube en un único registro. El registro único es una fuente de verdad para los metadatos del dispositivo y del recurso y las funcionalidades de administración de recursos. El Registro de dispositivos se puede usar junto con Operaciones de IoT de Azure.

Al usar Azure, relibilidad es una responsabilidad compartida. Microsoft proporciona una variedad de funcionalidades para admitir resistencia y recuperación. Es responsable de comprender cómo funcionan esas funcionalidades dentro de todos los servicios que usa y de seleccionar las funcionalidades que necesita para cumplir los objetivos empresariales y los objetivos de tiempo de actividad.

En este artículo se describe cómo hacer que el Registro de dispositivos sea resistente a una variedad de posibles interrupciones y problemas, como errores transitorios, interrupciones de zona de disponibilidad y interrupciones en regiones.

Introducción a la arquitectura de confiabilidad

Device Registry representa los recursos industriales y los dispositivos IoT como recursos Azure Resource Manager. Puede usar herramientas y funcionalidades de Resource Manager estándar para administrar los dispositivos, incluidos Azure Policy para la aplicación de directivas, Azure Resource Graph para consultar dispositivos y Bicep para definir los dispositivos como código. Para obtener más información sobre Azure Registro de dispositivos y su arquitectura, consulte Integración con Azure Registro de dispositivos (versión preliminar): Azure IoT Hub.

> **Nota:**
>
> Operaciones de IoT de Azure incluye otros componentes más allá del Registro de dispositivos. Para obtener información detallada sobre las características de alta disponibilidad y pérdida de datos cero de los componentes de Operaciones de IoT de Azure, consulte Operaciones de IoT de Azure preguntas más frecuentes.

Resistencia a errores transitorios

Los errores transitorios son errores breves e intermitentes en los componentes. Se producen con frecuencia en un entorno distribuido como la nube y son una parte normal de las operaciones. Los errores transitorios se corrigen después de un breve período de tiempo. Es importante que las aplicaciones puedan controlar errores transitorios, normalmente mediante el reintento de solicitudes afectadas.

Todas las aplicaciones hospedadas en la nube deben seguir las instrucciones de control de errores transitorios Azure cuando se comunican con cualquier API, bases de datos y otros componentes hospedados en la nube. Para obtener más información, consulte Recomendaciones para controlar errores transitorios.

Los clientes interactúan con el Registro de dispositivos mediante Azure Resource Manager. Normalmente, usa el portal de Azure, CLI de Azure o SDK de Azure para interactuar con los recursos del Registro de dispositivos y estas herramientas proporcionan control automático de errores transitorios. Si usa las API de Resource Manager directamente, asegúrese de controlar errores transitorios.

Resistencia a errores de zona de disponibilidad

Availability zones son grupos de centros de datos físicamente independientes dentro de una región de Azure. Cuando una zona falla, los servicios pueden transferirse a una de las zonas restantes.

Azure Device Registry tiene redundancia de zona de forma predeterminada, lo que significa que replica automáticamente sus datos en varias zonas de disponibilidad. Esta configuración mejora la resistencia del servicio al proporcionar alta disponibilidad. Si se produce un error en una zona, el servicio puede seguir funcionando desde otra zona sin interrupción.

Microsoft administra la configuración y la puesta en marcha de la redundancia de zona en el Registro de dispositivos de Azure. No es necesario realizar más configuraciones para habilitar esta redundancia de zona. Microsoft garantiza que el servicio esté configurado para proporcionar el mayor nivel de disponibilidad y confiabilidad.

Requisitos

Compatibilidad con regiones: La siguiente lista de regiones admite zonas de disponibilidad en el Registro de dispositivos:

Expandir tabla

Americas	Europa
East US	Centro-oeste de Alemania
Este de EE. UU. 2	Norte de Europa
West US	West Europe
Oeste de EE. UU. 2	
Oeste de EE. UU. 3	

Cost

No hay ningún costo adicional para usar la redundancia de zona para el Registro de dispositivos.

Configurar soporte de zonas de disponibilidad

Al crear un recurso de Registro de Dispositivos en Operaciones de IoT de Azure, es automáticamente zona-redundante por defecto. No es necesario realizar más configuraciones.

Comportamiento cuando todas las zonas están en buen estado

En la siguiente información se describe lo que sucede cuando tiene un registro de dispositivos con redundancia de zona y todas las zonas de disponibilidad están operativas:

- **Enrutamiento de tráfico entre zonas:** El sistema distribuye automáticamente las solicitudes en cada zona de disponibilidad. Una solicitud podría usar la infraestructura del Registro de dispositivos en cualquier zona de disponibilidad.
- **Replicación de datos entre zonas:** Los datos del dispositivo se replican sincrónicamente entre zonas de disponibilidad.

Comportamiento durante un fallo de zona

En la siguiente información se describe lo que sucede cuando tiene un registro de dispositivos con redundancia de zona y una zona de disponibilidad experimenta una interrupción.

- **Detección y respuesta:** Dado que el Registro de dispositivos detecta y responde automáticamente a errores en una zona de disponibilidad, no es necesario hacer nada para iniciar una conmutación por error de zona de disponibilidad.

- **Notificación:** Microsoft no le notifica automáticamente cuando una zona está inactiva. Sin embargo, puede usar Azure Service Health para comprender el estado general del servicio, incluidos los errores de zona, y puede configurar alertas de Service Health para notificarle problemas.
- **Solicitudes activas:** Es posible que se quiten algunas solicitudes activas, por lo que es posible que deba reintentarse de la misma manera que otros errores transitorios. Para asegurarse de que la aplicación es resistente a los errores transitorios, consulte Resistencia a errores transitorios.
- **Pérdida de datos esperada:** No se espera que un error de zona cause ninguna pérdida de datos.
- **Tiempo de inactividad esperado:** No se espera que un error de zona provoque tiempo de inactividad en los recursos.
- **Redistribución:** El Registro de dispositivos enruta automáticamente nuevas solicitudes a la infraestructura en una zona correcta.

Recuperación de zona

Cuando se recupera la zona de disponibilidad, el Registro de dispositivos restaura automáticamente las operaciones en la zona de disponibilidad.

Prueba de fallos de zona

La plataforma Device Registry gestiona el enrutamiento del tráfico, la conmutación por error y recuperación entre zonas de disponibilidad. No es necesario iniciar nada. Dado que esta característica está totalmente administrada, no es necesario validar los procesos de error de zona de disponibilidad.

Resistencia a errores en toda la región

Device Registry es un servicio de una sola región. Si la región deja de estar disponible, los recursos del Registro de dispositivos tampoco están disponibles. Sin embargo, los datos de su registro se replican en la región emparejada de Azure para permitir la conmutación automática por error en algunas situaciones.

Conmutación por error gestionada por Microsoft a una región emparejada

Los datos de su registro se replican automáticamente a la región de Azure emparejada. En caso de una interrupción prolongada de la región, Microsoft podría optar por conmutar a la región emparejada. Si esto sucede, el registro sigue estando disponible en la región emparejada.

Requisitos

Compatibilidad con regiones: La replicación predeterminada y la conmutación por error se admiten en todas las regiones en las que el Registro de dispositivos está disponible, ya que todas estas regiones están emparejadas.

Cost

No hay costos adicionales para la replicación de datos entre regiones ni para la conmutación por error.

Configuración de la replicación y preparación de la conmutación por error

De manera predeterminada, la replicación de datos entre regiones se configura automáticamente al crear recursos del Registro de dispositivos en una región con un par. Este proceso es una opción predeterminada y no requiere intervención de usted.

Comportamiento cuando todas las regiones están en buen estado

En esta sección se describe lo que se debe esperar cuando se configura un registro de dispositivos para la replicación y la conmutación por error entre regiones, y cuando la región primaria está operativa.

- **Replicación de datos entre regiones:** Los datos se replican automáticamente en la región emparejada. La replicación se produce de forma asincrónica, lo que significa que se espera cierta pérdida de datos si se produce una conmutación por error.
- **Enrutamiento del tráfico entre regiones:** En las operaciones normales, el tráfico solo fluye a la región primaria.

Comportamiento durante una falla de región

En esta sección se describe qué esperar cuando se configura un registro de dispositivos para la replicación y la conmutación por error entre regiones y se produce una falla en la región primaria.

- **Detección y respuesta:** Microsoft puede decidir realizar una conmutación por error si se pierde la región primaria. Este proceso puede tardar varias horas después de la pérdida de la región primaria o incluso más tiempo en algunos escenarios. Es posible que la conmutación por error de los recursos del Registro de dispositivos no se produzca al mismo tiempo que otros servicios de Azure.

- **Notificación:** Los eventos de fallo de región se pueden supervisar a través de Azure Service Health. Configure alertas para recibir notificaciones de problemas de nivel de región.

- **Solicitudes activas:** Es probable que se pierdan todas las solicitudes que la región primaria está procesando durante una conmutación por error. Los clientes deben reintentar solicitudes una vez completada la conmutación por error.

- **Pérdida de datos esperada:** Los datos se replican de forma asincrónica en la región emparejada. Por lo tanto, se anticipa cierta pérdida de datos tras el failover. Puede esperar menos de 15 minutos de pérdida de datos tras una conmutación por error en la región.

- **Tiempo de inactividad esperado:** Se espera aproximadamente 24 horas de tiempo de inactividad desde el momento en que se pierde la región cuando el recurso está disponible en la región emparejada.

- **Reenrutamiento del tráfico:** Durante el proceso de conmutación por error, el Registro de dispositivos actualiza los registros DNS para que apunten a la región emparejada. Todas las solicitudes posteriores se envían a la región emparejada.

  Una vez completada la operación de conmutación por error del registro, se espera que todas las operaciones del dispositivo y las aplicaciones de back-end continúen funcionando sin necesidad de intervención manual.

Recuperación de regiones

Cuando se recupera la región primaria, el Registro de dispositivos restaura automáticamente las operaciones en la región.

Prueba de fallos de región

La plataforma Device Registry administra el enrutamiento del tráfico, la conmutación por error y la conmutación por recuperación entre regiones emparejadas. No es necesario iniciar nada. Dado que esta característica está totalmente administrada, no es necesario validar los procesos de error de región emparejada.

Copias de seguridad y restauración

Para proteger la configuración de recursos, defina los recursos del Registro de dispositivos mediante la infraestructura como código (como Bicep o plantillas de ARM) y almacene esas definiciones en el control de código fuente. Si necesita volver a crear un recurso, vuelva a implementarlo desde la configuración almacenada.

Resistencia al mantenimiento del servicio

Microsoft aplica periódicamente actualizaciones de servicio y realiza otro mantenimiento. La plataforma Azure controla estas actividades automáticamente, lo que garantiza que el mantenimiento sea transparente y sin problemas. No se espera ningún tiempo de inactividad durante los eventos de mantenimiento a menos que se le haya informado a través del mantenimiento planeado de Azure Service Health.

Contenido relacionado

- ¿Qué es Operaciones de IoT de Azure?
- Integración con Azure Device Registry (vista previa) - Azure IoT Hub
- Confiabilidad en Azure

ⓘ**Nota:** El autor ha creado este artículo con ayuda de inteligencia artificial. Más información

---

Last updated on 19/08/2026

Confiabilidad en Azure IoT Hub

Azure IoT Hub es un servicio administrado hospedado en la nube que actúa como un centro de mensajes central para la comunicación entre una aplicación de IoT y sus dispositivos conectados.

Al usar Azure, relibilidad es una responsabilidad compartida. Microsoft proporciona una variedad de capacidades para admitir resiliencia y recuperación. Es responsable de comprender cómo funcionan esas funcionalidades dentro de todos los servicios que usa y de seleccionar las funcionalidades que necesita para cumplir los objetivos empresariales y los objetivos de tiempo de actividad.

En este artículo se describe cómo hacer que IoT Hub sean resistentes a una variedad de posibles interrupciones y problemas, incluidos errores transitorios, interrupciones de zona de disponibilidad y interrupciones de regiones. También se describe cómo puede usar copias de seguridad para recuperarse de otros tipos de problemas y se resalta cierta información clave sobre el contrato de nivel de servicio (SLA) de IoT Hub.

Recomendaciones de implementación de producción para la confiabilidad

En el caso de las cargas de trabajo de producción, se recomienda seguir estas recomendaciones:

- ✓ Implemente el centro de IoT en una región que admita redundancia de zona para los componentes de proceso y de datos. Para obtener más información, consulte los requisitos.
- ✓ Implemente los patrones de reintento adecuados en todos los dispositivos y aplicaciones que se comunican con IoT Hub.
- ✓ Diseñe la lógica de reconexión de dispositivos para controlar errores transitorios y conmutaciones por error de servicio. Para más información, consulte Administración de reconexión de dispositivos para crear aplicaciones resistentes.
- ✓ Para implementaciones a mayor escala, siga estas recomendaciones:
  - Implemente patrones de reintentos exponenciales y basados en vibración en dispositivos y aplicaciones al volver a conectarse a IoT Hub, lo que ayuda a evitar las tormentas de reconexión durante las conmutaciones por error del lado del servicio o las interrupciones de red.

- Comprenda IoT Hub límites y cuotas de servicio y planee cómo las controla la solución. Al considerar desde el principio el comportamiento de limitación del lado del servicio, los límites de conexión y las unidades de rendimiento, puede diseñar su solución para permitir una escalabilidad predecible y evitar la necesidad de reestructuración arquitectónica a medida que su flota crece. Para obtener orientación adicional, consulte cuotas y escalado de IoT Hub.
- Utilice Azure IoT Hub junto con Azure IoT Hub Device Provisioning Service (DPS). DPS permite la incorporación segura, sin intervención táctil y la asignación de dispositivos en uno o varios concentradores. Incluso si no prevé tener una gran flota, incorporando DPS desde el principio, los flujos de trabajo de fabricación e incorporación de dispositivos se pueden escalar sin necesidad de cambios de firmware o infraestructura más adelante. Para más información, consulte Implementación de soluciones de IoT a escala con DPS.
- Considere la posibilidad de usar directivas de asignación de DPS para distribuir dispositivos en varias instancias de IoT Hub para mejorar la disponibilidad y la resistencia de las regiones. Este enfoque permite el escalado horizontal de la capacidad de ingesta y admite el crecimiento futuro de la flota sin necesidad de volver a aprovisionar el dispositivo.

Introducción a la arquitectura de confiabilidad

Al crear un centro de IoT, se implementa un recurso de *IoT Hub* que incluye toda la funcionalidad necesaria para administrar y comunicarse con los dispositivos. Los componentes principales de un centro de IoT incluyen:

- **Registro de identidad del dispositivo:** Una base de datos que almacena información sobre los dispositivos y módulos que se pueden conectar al centro de IoT. Cada dispositivo debe tener una entrada en el registro de identidad para poder conectarse. Para obtener más información, consulte Descripción del registro de identidades en IoT Hub.

- **Componentes de mensajería:** IoT Hub controla la mensajería bidireccional entre dispositivos y sus aplicaciones de back-end, incluidos los datos de telemetría de dispositivo a nube, comandos de nube a dispositivo e invocaciones directas de métodos.

- **Dispositivos gemelos:** Documentos JSON que almacenan información de estado del dispositivo, incluidos metadatos, configuraciones y condiciones. Los dispositivos y las aplicaciones de back-end pueden leer y actualizar dispositivos gemelos.

Con fines de confiabilidad, los componentes de IoT Hub se clasifican en dos grupos:

- **Componentes de proceso:** Administrar conexiones de dispositivos, autenticar solicitudes, enrutar mensajes y procesar invocaciones de método directo. Estos componentes determinan la capacidad de IoT Hub de aceptar y procesar solicitudes.
- **Componentes de datos:** Almacene el registro de identidad del dispositivo, los dispositivos gemelos y los mensajes del dispositivo a la nube. Estos componentes determinan la disponibilidad y la durabilidad de los datos.

Esta distinción es importante porque las distintas regiones admiten distintos tipos de redundancia para estos componentes.

Integración con otros servicios

Puede integrar IoT Hub con el Registro de Dispositivos de Azure. Si usa este enfoque, revise Fiabilidad en el Registro de Dispositivos de Azure para comprender la fiabilidad en ambos servicios.

Si usa IoT Hub Device Provisioning Service (DPS) para el aprovisionamiento de dispositivos, la confiabilidad de la solución también depende de DPS. Para obtener más información, consulte Alta disponibilidad y recuperación ante desastres del servicio de aprovisionamiento de dispositivos de IoT Hub.

Resistencia a errores transitorios

Los errores transitorios son errores breves e intermitentes en los componentes. Se producen con frecuencia en un entorno distribuido como la nube y son una parte normal de las operaciones. Los errores transitorios se corrigen después de un breve período de tiempo. Es importante que las aplicaciones puedan controlar errores transitorios, normalmente mediante el reintento de solicitudes afectadas.

Todas las aplicaciones hospedadas en la nube deben seguir las instrucciones de control de errores transitorios Azure cuando se comunican con cualquier API, bases de datos y otros componentes hospedados en la nube. Para obtener más información, consulte Recomendaciones para controlar errores transitorios.

IoT Hub proporciona una garantía de tiempo de actividad razonablemente alta, pero los errores transitorios pueden producirse en cualquier plataforma informática distribuida. Para

controlar los errores transitorios, cree los patrones de reintento adecuados en los componentes que interactúan con las aplicaciones en la nube.

En el caso de implementaciones a gran escala, es una buena práctica usar el jitter de reintentos aleatorio. Jitter ayuda a distribuir la carga entre particiones del hub y reduce la probabilidad de limitación durante los eventos de reconexión a gran escala.

Resistencia a errores de zona de disponibilidad

Availability zones son grupos de centros de datos físicamente independientes dentro de una región de Azure. Cuando una zona falla, los servicios pueden transferirse a una de las zonas restantes.

IoT Hub admite dos tipos distintos de compatibilidad con zonas de disponibilidad:

- *Redundancia de zona para los datos*, que replica automáticamente los datos entre varias zonas de disponibilidad para los componentes de almacenamiento subyacentes que almacenan el registro de identidad del dispositivo y los mensajes del dispositivo a la nube.
- *Redundancia de zona para cómputo*, que proporciona resiliencia en los componentes responsables de administrar los dispositivos y el enrutamiento de mensajes.

Cuando una región admite ambos tipos de datos redundantes por zona, los datos de servicio críticos, incluido el Registro de Identidad del Dispositivo, se replican sincrónicamente entre las zonas de disponibilidad dentro de la región. Como resultado, en caso de error de una sola zona, no se espera ninguna pérdida de datos y el servicio vuelve a enrutar automáticamente la conectividad del dispositivo y el tráfico de mensajes a una zona correcta. Aunque las solicitudes en curso pueden verse afectadas transitoriamente durante el evento de conmutación por error, la continuidad del servicio se mantiene y la lógica de reintento del dispositivo normalmente garantiza la recuperación.

Requisitos

Compatibilidad con regiones: El tipo de soporte de zona de disponibilidad para IoT Hub depende de la región en la que se implemente.

Expandir tabla

Región	Redundancia de zona para datos	Redundancia de zona para procesos
Australia East	✓	✓
Sur de Brasil	✓	✓
Canada Central	✓	✓
Centro de la India	X	✓
Central US	✓	✓
East US	X	✓
Centro de Francia	✓	✓
Centro-oeste de Alemania	✓	✓
Japón Oriental	✓	✓
Centro de Corea del Sur	X	✓
Norte de Europa	✓	✓
Norway East	X	✓
Qatar Central	X	✓
Centro-sur de EE. UU.	X	✓
Sudeste Asiático	✓	✓
UK South	✓	✓
West Europe	X	✓
Oeste de EE. UU. 2	✓	✓
Oeste de EE. UU. 3	X	✓

Los centros de IoT que cree en regiones que no están en esta lista no son resistentes a las interrupciones de zona.

Cost

No hay ningún costo adicional para la redundancia de zona con IoT Hub.

Configurar soporte de zonas de disponibilidad

Los recursos de IoT Hub admiten automáticamente la redundancia de zona cuando se implementan en regiones admitidas. No es necesario realizar ninguna otra configuración.

Comportamiento cuando todas las zonas están en buen estado

En esta sección se describe qué esperar cuando los recursos de IoT Hub se configuran para redundancia zonal y todas las zonas de disponibilidad están operativas.

- **Replicación de datos entre zonas:** Cuando el centro de IoT se implementa en una región que admite la redundancia de zona para los datos, los cambios en los datos se replican entre zonas de disponibilidad automáticamente. La replicación se produce sincrónicamente.

- **Enrutamiento de tráfico entre zonas:** Cuando el centro de IoT se implementa en una región que admite redundancia de zona para componentes de proceso, las solicitudes se enrutan a una instancia principal del servicio en una de las zonas de disponibilidad. Azure selecciona automáticamente la instancia activa y la zona.

Comportamiento durante un fallo de zona

En esta sección se describe qué esperar cuando los recursos de IoT Hub están configurados para la redundancia de zona y ocurre una interrupción en la zona de disponibilidad.

- **Detección y respuesta:** El servicio IoT Hub es responsable de detectar un fallo en una zona de disponibilidad. No es necesario hacer nada para iniciar una conmutación por error de la zona.

- **Notification**: Microsoft no le notifica automáticamente cuando una zona está inactiva. Sin embargo, puede usar Azure Resource Health para supervisar el estado de un recurso individual y puede configurar Resource Health alertas para notificarle problemas. También puede usar Azure Service Health para comprender el estado general del servicio, incluidos los errores de zona, y puede configurar alertas de Service Health para notificarle problemas.

- **Solicitudes activas:** Durante un fallo de zona, las solicitudes activas podrían ser eliminadas. Los clientes y dispositivos deben tener una lógica de reintento suficiente implementada para controlar los errores transitorios.

- **Pérdida de datos esperada:** Cuando el centro de IoT se implementa en una región que admite redundancia de zona para los datos, un error de zona no debe provocar ninguna pérdida de datos.
- **Tiempo de inactividad esperado:** Cuando IoT Hub se implementa en una región que admite redundancia de zona para componentes informáticos y de datos, un error de zona no debe provocar tiempo de inactividad en los recursos de IoT Hub.
- **Redirección del tráfico:** Cuando IoT Hub se implementa en una región que admite redundancia de zona para componentes de computación, IoT Hub detecta la pérdida de la zona. A continuación, las nuevas solicitudes se redirigen automáticamente a una nueva instancia principal del servicio en una de las zonas de disponibilidad correctas.

Recuperación de zona

Cuando se recupera la zona de disponibilidad, IoT Hub restaura automáticamente las instancias de la zona de disponibilidad y vuelve a enrutar el tráfico entre las instancias como normal.

Prueba de fallos de zona

Dado que IoT Hub administra completamente el enrutamiento del tráfico, la conmutación por error y la recuperación para fallos de zona de disponibilidad, no es necesario validar los fallos de zona de disponibilidad ni proporcionar ningún dato adicional.

Resistencia a errores en toda la región

IoT Hub es un servicio de una sola región. Si la región deja de estar disponible, los recursos de IoT Hub tampoco están disponibles. Aunque IoT Hub admite la replicación de datos asincrónica en una región de Azure emparejada con fines de recuperación ante desastres, no hay ninguna conmutación por error integrada entre regiones para la conectividad del dispositivo.

Si los recursos están en una región *no emparejada*, Microsoft no replica la configuración y los datos entre regiones y no existe conmutación por error integrada entre regiones. Sin embargo, puede implementar recursos independientes en varias regiones. En este escenario, es responsabilidad suya administrar la replicación, la distribución del tráfico y la conmutación por error.

Si el centro de IoT está en una región no emparejada, o si el comportamiento predeterminado de replicación y conmutación por error no satisface sus necesidades, diseñe e implemente una

estrategia de conmutación por error personalizada de varias regiones, incluidos los pasos siguientes:

- Aprovisionamiento de un IoT Hub secundario en una región de Azure diferente.
- Implementar un mecanismo de redirección de puntos de conexión para dirigir los dispositivos a la región alternativa cuando sea necesario. Por ejemplo, puede aprovisionar cada dispositivo en ambos centros con antelación y configurar ambas cadenas de conexión en los dispositivos para que puedan cambiar entre centros cuando sea necesario.

Conmutación por error gestionada por Microsoft a una región emparejada

Si los recursos están en una región emparejada, los datos del centro de IoT se replican en la región emparejada. Este enfoque está diseñado para facilitar la recuperación ante desastres. Si hay un fallo en la región principal de tu IoT Hub, no es necesario realizar ninguna acción para que la conectividad de los dispositivos continúe en la región emparejada.

Tipos de conmutación por error

El centro de IoT puede conmutar por error a la región emparejada en los escenarios siguientes:

- *Conmutación por error iniciada por el cliente:* Puede desencadenar la conmutación por error manual en la región emparejada usted mismo, independientemente de si la región está experimentando tiempo de inactividad o no. Puede usar este enfoque para realizar conmutaciones por error planeadas y simulacros.

- *Conmutación por error iniciada por Microsoft:* Si se pierde una región, Microsoft puede iniciar una conmutación por error de los centros de IoT a la región emparejada. Sin embargo, es poco probable que Microsoft inicie la conmutación por error, excepto si ocurre un retraso significativo y en la medida de lo posible. La conmutación por error de los recursos de IoT Hub puede producirse en un momento diferente a la conmutación por error de otros servicios de Azure. Este proceso es una opción predeterminada y no requiere intervención de usted.

Requisitos

- **Compatibilidad con regiones:** La replicación predeterminada y la conmutación por error solo se admiten en regiones emparejadas.

- **Tier:** Las opciones de replicación y conmutación por error de la región emparejada están disponibles para todos los niveles de IoT Hub.

Consideraciones

No utilice la conmutación por error iniciada por el cliente para migrar permanentemente el concentrador entre las regiones emparejadas de Azure. Por lo general, los dispositivos se encuentran cerca de la región primaria del centro. Si mueve el centro a la región secundaria, la latencia aumenta para las operaciones entre los dispositivos y el centro de IoT en la ubicación secundaria.

Cost

En el caso de los centros emparejados, no hay ningún costo adicional para la conmutación por error o la replicación de datos entre regiones.

Si el hub de IoT se encuentra en una región no emparejada, consulte Soluciones personalizadas de varias regiones para obtener resiliencia para posible información sobre los costos.

Configuración de la replicación y preparación de la conmutación por error

De forma predeterminada, la replicación de datos entre regiones se configura automáticamente al crear un centro de IoT en una región emparejada. Este proceso es una opción predeterminada y no requiere intervención de usted. En regiones, excepto el Sur de Brasil y sudeste de Asia (Singapur), los datos del cliente no se almacenan ni procesan fuera de la geografía en la que se implementa la instancia de servicio.

Si el centro de IoT se encuentra en las regiones Sur de Brasil o Sudeste de Asia (Singapur), puede deshabilitar la replicación de datos y no participar en la conmutación por error. Para obtener más información, consulte Deshabilitación de la recuperación ante desastres (DR).

Si el centro de IoT está en una región no emparejada, debe planear su propio enfoque de replicación y conmutación por error entre regiones. Para obtener más información, consulte Soluciones personalizadas de varias regiones para lograr resistencia.

Comportamiento cuando todas las regiones están en buen estado

En esta sección se describe qué esperar cuando se configura un centro de IoT para la replicación y la conmutación por error entre regiones, y la región primaria está operativa.

- **Replicación de datos entre regiones:** Los datos, incluido el Registro de identidad de dispositivo, se replican automáticamente en la región emparejada. La replicación se produce de forma asincrónica, lo que significa que se espera cierta pérdida de datos si se produce una conmutación por error. No hay replicación de datos entre regiones para centros de IoT en regiones no emparejadas.

- **Enrutamiento del tráfico entre regiones:** En las operaciones normales, el tráfico solo fluye a la región primaria.

Comportamiento durante una falla de región

En esta sección se proporciona una descripción de qué se puede esperar cuando un centro IoT está configurado para la replicación y la conmutación por error interregional y hay una interrupción en la región primaria.

- **Detección y respuesta:** La responsabilidad de detectar y responder a una interrupción en la región primaria puede variar.

  - *Conmutación por error iniciada por el cliente:* es responsable de detectar una pérdida de región y decidir cuándo conmutar por error. Para obtener más información sobre cómo realizar la conmutación por error iniciada por el cliente entre regiones emparejadas, consulte Realización de una conmutación por error manual para un centro de IoT.

    Hay límites sobre la frecuencia con la que puede realizar la conmutación por error o la conmutación por recuperación iniciadas por el cliente:

    - A los usuarios se les permite realizar dos operaciones correctas de conmutación por error y dos operaciones correctas de conmutación por recuperación al día.

    - No se permite realizar operaciones consecutivas de conmutación por error o conmutación por recuperación. Debe esperar una hora entre estas operaciones.

  - *Conmutación por error iniciada por Microsoft:* Microsoft puede decidir realizar una conmutación por error si se pierde la región primaria. Este proceso puede tardar varias horas después de la pérdida de la región primaria o incluso más tiempo en algunos escenarios. Puede que la conmutación por error de los recursos de IoT Hub no se produzca al mismo tiempo que otros servicios de Azure.

- **Solicitudes activas:** Es probable que se pierdan todas las solicitudes que la región primaria está procesando durante una conmutación por error. Los clientes deben

reintentar solicitudes una vez completada la conmutación por error.

- **Pérdida de datos esperada:** En el caso de las regiones emparejadas, los datos se replican de forma asincrónica en la región emparejada. Por lo tanto, se anticipa cierta pérdida de datos tras el failover. Este proceso se aplica tanto a las conmutaciones por error administradas por Microsoft como a las administradas por el cliente. En la tabla siguiente se describe el *objetivo de punto de recuperación (RPO)* o la pérdida de datos esperada, de cada tipo de datos que almacenan los centros de IoT.

Expandir tabla

Tipo de dato	RPO
Registro de identidad	Pérdida de datos de 0 a 5 minutos
Datos del dispositivo gemelo	Pérdida de datos de 0 a 5 minutos
Mensajes de nube a dispositivo [1]	Pérdida de datos de 0 a 5 minutos
Trabajos principales [1] y de dispositivo	Pérdida de datos de 0 a 5 minutos
Mensajes de dispositivo a nube	Todos los mensajes no leídos se pierden
Mensajes de comentarios de la nube al dispositivo	Todos los mensajes no leídos se pierden

[1]Los mensajes de la nube al dispositivo y los trabajos principales no se recuperan como parte de la conmutación por error iniciada por el cliente.

- **Tiempo de inactividad esperado:** El tiempo de inactividad esperado durante la conmutación por error de la región depende del tipo de conmutación por error.
  - *Conmutación por error iniciada por el cliente:* Se espera aproximadamente de 10 minutos a 2 horas de tiempo de inactividad desde el momento en que se pierde la región hasta cuando el recurso esté operativo en la región emparejada. El número de dispositivos registrados en la instancia de centro de IoT que se conmuta por error afecta al tiempo de recuperación. Puede esperar que el tiempo de recuperación de un centro que hospede aproximadamente 100 000 dispositivos sea de aproximadamente 15 minutos.
  - *Conmutación por error iniciada por Microsoft:* Espere aproximadamente de 2 a 26 horas de tiempo de inactividad desde el momento en que se pierde la región hasta que el recurso esté disponible en la región emparejada.

La gran cantidad de tiempo de recuperación se debe a que Microsoft debe realizar la operación de conmutación por error en nombre de todos los clientes afectados de esa región. En el caso de los sistemas críticos, debe usar la conmutación por error iniciada por el cliente para reducir el tiempo de inactividad. Sin embargo, si ejecuta una solución de IoT menos crítica que puede mantener un tiempo de inactividad de aproximadamente un día, puede depender de la opción iniciada por Microsoft para satisfacer los objetivos generales de recuperación ante desastres de la solución de IoT.

- Para ambos tipos de conmutación por error, el nombre de dominio completo de la instancia de IoT Hub sigue siendo el mismo después de la conmutación por error, lo que significa que la cadena de conexión también sigue siendo la misma. Sin embargo, dado que cambia la dirección IP subyacente, los clientes deben esperar a que los registros del sistema de nombres de dominio (DNS) se actualicen antes de acceder al centro de IoT después de la conmutación por error.

  > **ⓘ Importante**
  >
  > Los SDK de IoT Hub no almacenan en caché la dirección IP del IoT hub. El código de usuario que interconecta con los SDK tampoco debe almacenar en caché la dirección IP del centro de IoT.

  Debido a estos factores, el tiempo que tardan las operaciones de tiempo de ejecución en volverse completamente operativas en su instancia de IoT Hub después del proceso de conmutación por error se puede expresar mediante la siguiente función:

  > Tiempo de recuperación = RTO [10 minutos a 2 horas para la conmutación por error iniciada por el cliente o de 2 a 26 horas para la conmutación por error iniciada por Microsoft] + Retraso de propagación de DNS + Tiempo que tarda la aplicación cliente en actualizar cualquier dirección IP de IoT Hub almacenada en caché

- **Redirección de tráfico:** Durante el proceso de conmutación por error, IoT Hub actualiza los registros DNS para redirigirlos a la región emparejada. Todas las solicitudes posteriores se envían a la región emparejada.

  Una vez completada la operación de conmutación por error del centro de IoT, se espera que todas las operaciones del dispositivo y las aplicaciones de back-end sigan funcionando sin necesidad de intervención manual. Esta continuidad garantiza que los mensajes del dispositivo a la nube sigan funcionando y que todo el registro de

dispositivos esté intacto. Si emite eventos a través de Azure Event Grid, se pueden consumir a través de las mismas suscripciones que configuró anteriormente, siempre que esas suscripciones de Event Grid sigan estando disponibles. No se requiere ningún control adicional para los puntos de conexión personalizados.

Se requiere una configuración posterior a la conmutación por error

En función de dónde redirija los mensajes del centro de IoT, es posible que tenga que realizar pasos adicionales después de que se complete la conmutación por error.

- **Azure Event Hubs:** El nombre y el punto de conexión compatibles con Event Hubs del extremo de eventos integrados del IoT Hub cambian tras la conmutación por error. Este cambio se produce porque el cliente de Event Hubs no tiene visibilidad de eventos de IoT Hub.

  Cuando recibe mensajes de telemetría del punto de conexión integrado mediante el cliente de Event Hubs o el host del procesador de eventos, use la cadena de conexión del centro de IoT para establecer la conexión. Este enfoque garantiza que las aplicaciones de back-end sigan funcionando sin necesidad de intervención manual después de la conmutación por error.

  Si usa el nombre y el punto de conexión compatibles con Event Hubs directamente en la aplicación, tiene que capturar el nuevo punto de conexión compatible con Event Hubs después de la conmutación por error para continuar con las operaciones. Para recuperar el punto de conexión y el nombre, puede usar el portal de Azure o el SDK de .NET:

  - *El portal de Azure:* Para obtener más información sobre cómo usar el portal para obtener el punto de conexión compatible con Event Hubs y el nombre que sea compatible con Event Hubs, consulte Conectar al punto de conexión integrado.

  - *El SDK de .NET:* Para usar la cadena de conexión del IoT hub para volver a capturar el punto de conexión compatible con Event Hubs, use el código de muestra. En este ejemplo de código se usa el cadena de conexión para obtener el nuevo punto de conexión de Event Hubs y volver a establecer la conexión. Debe tener Visual Studio instalado.

- **Azure Functions y Azure Stream Analytics:** Si usa Azure Functions o Stream Analytics para conectarse al punto de conexión de eventos integrado, debe actualizar el punto de conexión de Event Hubs al que se conecta la función o el trabajo, siguiendo el mismo proceso descrito en el punto de viñeta anterior. Después, realice una acción de **Reiniciar**

porque los desplazamientos del flujo de eventos se vuelven inválidos después de la conmutación por error.

- **Azure Storage:** Al enrutar a Azure Storage, enumere primero los blobs o archivos. Luego, itere sobre ellos para asegurarse de que todos los blobs o archivos se leen sin asumir el particionamiento. El intervalo de particiones puede cambiar potencialmente durante una conmutación por error iniciada por Microsoft o por el cliente. Puede usar la API List Blobs para enumerar los blobs o la API List Azure Data Lake Storage para listar los archivos. Para obtener más información, consulte Azure Storage como punto de conexión de enrutamiento.

Recuperación de regiones

Para conmutar por recuperación a la región primaria, puede desencadenar manualmente la acción de conmutación por error una segunda vez. Es importante recordar las restricciones sobre la frecuencia con la que puede conmutar por error.

Si la operación de conmutación por error original se ha realizado para recuperarse de una interrupción extendida en la región principal original, realice la conmutación por recuperación en la región principal después de que la región principal se recupere de la interrupción.

Prueba de fallos de región

Para simular un error durante una interrupción de la región, puede desencadenar una conmutación por error manual del centro de IoT. Sin embargo, dado que la conmutación por error regional provoca tiempo de inactividad y pérdida de datos, solo debe realizar conmutaciones por error de prueba en entornos que no sean de producción. Para obtener más información, consulte Comportamiento durante un error de región. Considere la posibilidad de configurar una instancia de prueba IoT Hub para iniciar la opción de conmutación por error planeada periódicamente. Las pruebas periódicas pueden ayudarle a crear confianza en su capacidad de restaurar y operar sus soluciones de un extremo a otro de forma eficaz cuando se produce un desastre real.

Soluciones personalizadas de varias regiones para la resistencia

Las funcionalidades de conmutación por error entre regiones de IoT Hub no son adecuadas para los escenarios siguientes:

- El centro de IoT se encuentra en una región no emparejada.

- Los objetivos de tiempo de actividad empresarial no se satisfacen por el tiempo de recuperación o la pérdida de datos que proporciona la opción de conmutación por error integrada.
- Debe conmutar por error a una región que no esté emparejada con la región principal.

Puede diseñar una solución de conmutación por error entre regiones adaptada a cada dispositivo individual. Un tratamiento completo de las topologías de implementación en soluciones de IoT está fuera del ámbito de este artículo, pero puede considerar un modelo de implementación de varias regiones. En este modelo, la instancia principal de IoT Hub y el back-end de la solución se ejecutan principalmente en una región de Azure. Un centro de IoT secundario y un back-end se implementan en otra región de Azure. Si el centro de IoT de la región primaria experimenta una interrupción o se interrumpe la conectividad de red desde el dispositivo a la región primaria, los dispositivos usan un punto de conexión de servicio secundario.

- **Tiempo de inactividad esperado:** Este enfoque tiene menos de un minuto de tiempo de inactividad, pero puede ser complejo de implementar.
- **Pérdida de datos esperada:** La cantidad de pérdida de datos depende de los almacenes de datos específicos que use y de la forma en que configure la replicación geográfica entre ellos.
- **Costar:** Este enfoque requiere que aprovisione al menos un centro de IoT adicional, lo que aumenta el costo total.

En un nivel alto, para implementar un modelo de conmutación por error regional con IoT Hub, debe tomar las siguientes medidas:

- **Una lógica secundaria de IoT Hub y enrutamiento de dispositivos:** Si se interrumpe el servicio de la región primaria, los dispositivos deben empezar a conectarse a la región secundaria. Como se conoce el estado de la mayoría de los servicios implicados, es habitual que los administradores de la solución desencadenen manualmente el proceso de conmutación por error entre regiones. La mejor manera de comunicar el nuevo punto de conexión a los dispositivos mientras mantiene el control sobre el proceso es hacer que comprueben periódicamente un servicio de conserjería para el punto de conexión activo actual. El servicio de conserjería puede ser una aplicación web que se replica y se mantiene accesible mediante técnicas de redirección de DNS, como Azure Traffic Manager.

> ⓘ **Nota:**
>
> Traffic Manager no tiene compatibilidad integrada con IoT Hub. Puede configurar puntos de conexión personalizados de Traffic Manager para cada centro de IoT. Configure el sondeo de estado del punto de conexión de Traffic Manager para usar el punto de conexión del centro de IoT.

- **Replicación del Registro de identidad:** Para poder usarse, el centro de IoT secundario debe contener todas las identidades de dispositivo que se pueden conectar a la solución. La solución debe mantener copias de seguridad con replicación geográfica de identidades de dispositivo y cargarlas en el centro de IoT secundario antes de cambiar el punto de conexión activo para los dispositivos. La funcionalidad de exportación de identidades de dispositivo de IoT Hub es útil en este contexto. Para obtener más información, consulte Descripción del registro de identidades en IoT Hub.

- **Lógica de combinación:** Cuando la región primaria vuelva a estar disponible, todos los datos y el estado creados en la región secundaria se deben volver a migrar a la región primaria. Este estado y estos datos tienen que ver principalmente con las identidades de dispositivo y los metadatos de aplicación, que deben combinarse con el centro de IoT principal y con otros almacenes específicos de la aplicación en la región primaria.

  Para simplificar este paso, use operaciones *idempotentes* . Las operaciones idempotentes minimizan los efectos secundarios de la posible distribución uniforme de eventos y de los duplicados o la entrega desordenada de eventos. Además, la lógica de la aplicación debe diseñarse para tolerar posibles incoherencias o un estado ligeramente obsoleto. Este escenario puede producirse debido al tiempo adicional que tarda el sistema en recuperarse en función de los RPO.

Copias de seguridad y restauración

El servicio IoT Hub permite operaciones de exportación masiva, que permiten exportar todo el registro de identidad de un IoT hub. Estos datos se transfieren a un contenedor de blobs de Azure Storage mediante una firma de acceso compartido. Este método permite crear copias de seguridad confiables de la información del dispositivo en un contenedor de blobs que controle. Para obtener más información, consulte Importar y exportar identidades de dispositivo IoT Hub en bloque.

También puede exportar la plantilla de Azure Resource Manager (ARM) de un IoT hub existente para crear una copia de seguridad de la configuración del IoT hub. Para obtener más información, consulte Migrar manualmente un IoT hub utilizando una plantilla ARM.

Para la mayoría de las soluciones, no debe confiar exclusivamente en copias de seguridad. En su lugar, utilice las otras capacidades descritas en esta guía para apoyar los requisitos de resiliencia. Sin embargo, las copias de seguridad protegen contra algunos riesgos que otros enfoques no. Para más información, consulte ¿Qué son la redundancia, la replicación y la copia de seguridad?.

Resistencia al mantenimiento del servicio

Microsoft aplica periódicamente actualizaciones de servicio y realiza otro mantenimiento. La plataforma Azure controla estas actividades automáticamente, lo que garantiza que el mantenimiento sea transparente y sin problemas. No se espera ningún tiempo de inactividad durante los eventos de mantenimiento a menos que se le haya notificado a través de Azure Service Health mantenimiento programado.

Acuerdo de nivel de servicio

El acuerdo de nivel de servicio (SLA) para Azure servicios describe la disponibilidad esperada de cada servicio y las condiciones que la solución debe cumplir para lograr esa expectativa de disponibilidad. Para obtener más información, vea SLAs for servicios en línea.

Contenido relacionado

- Escalabilidad, alta disponibilidad y recuperación ante desastres de soluciones de IoT
- Reliability in Azure Device Registry
- Confiabilidad en Azure

Last updated on 16/04/2026

Confiabilidad en Azure Notification Hubs

Azure Notification Hubs le ayuda a administrar las notificaciones push en varios sistemas de notificaciones de plataforma (PNS), como apple Push Notification Service (APN), Firebase Cloud Messaging (FCM) y Windows Push Notification Service (WNS).

Cuando se usa Azure, la confiabilidad es una responsabilidad compartida. Microsoft proporciona una variedad de capacidades para apoyar la resiliencia y la recuperación. Es responsable de comprender cómo funcionan esas funcionalidades dentro de todos los servicios que usa y de seleccionar las funcionalidades que necesita para cumplir los objetivos empresariales y los objetivos de tiempo de actividad.

En este artículo se describe cómo hacer que Notification Hubs sea resistente a varias posibles interrupciones y problemas, incluidos errores transitorios, errores de zona de disponibilidad, errores de toda la región y mantenimiento del servicio. También se describen las opciones de copia de seguridad y restauración y la información clave sobre el acuerdo de nivel de servicio (SLA) de Notification Hubs.

Recomendaciones de implementación de producción

Para cargas de trabajo de producción, siga estas recomendaciones:

- ✓ Use el nivel Básico o Estándar para que el espacio de nombres sea apto para el Acuerdo de Nivel de Servicio.
- ✓ Cuando sea posible, use instalaciones en lugar de registros en aplicaciones de dispositivo.
- ✓ Use los SDK proporcionados por Microsoft para interactuar con Notification Hubs.
- ✓ Habilitación de la redundancia de zona.
- ✓ Para prepararse para interrupciones en toda la región, habilite la recuperación ante desastres de metadatos en otra región de Azure. Planee cómo realizar copias de seguridad y restaurar registros e instalaciones de dispositivos.

Introducción a la arquitectura de confiabilidad

Azure Notification Hubs se organiza en torno a *espacios de nombres* y *centros de notificaciones*. Un espacio de nombres es un límite administrativo que contiene uno o varios hubs. Los

concentradores representan puntos de conexión para una aplicación. Los dispositivos se registran con esos puntos de conexión mediante *registros* o *instalaciones*, lo que permite al servicio enviar notificaciones push a los dispositivos. Para obtener más información, consulte Administración de registros.

Notification Hubs envía notificaciones push a sistemas de notificaciones de plataforma (PNS), como Apple Push Notification Service (APN) y Firebase Cloud Messaging (FCM). La entrega de notificaciones de un extremo a otro depende de la disponibilidad de Notification Hubs y del comportamiento de los proveedores de PNS de nivel inferior.

Para planear la confiabilidad, es importante distinguir entre los siguientes tipos de datos que Administra Notification Hubs:

- **Metadatos:** Espacio de nombres y configuración del hub, incluida la información de conexión y la configuración de recuperación ante desastres.
- **Datos de registro:** Registros de dispositivos e instalaciones que asignan usuarios y dispositivos a etiquetas y plantillas.

Resistencia a errores transitorios

Los errores transitorios son errores breves e intermitentes en los componentes. Se producen con frecuencia en un entorno distribuido como la nube y son una parte normal de las operaciones. Los errores transitorios se corrigen después de un breve período de tiempo. Es importante que las aplicaciones puedan controlar errores transitorios, normalmente mediante el reintento de solicitudes afectadas.

Todas las aplicaciones hospedadas en la nube deben seguir las instrucciones de control de errores transitorios de Azure cuando se comunican con cualquier API, bases de datos y otros componentes hospedados en la nube. Para más información, vea Recomendaciones para gestionar errores temporales.

Notification Hubs controla automáticamente los errores transitorios que se producen al conectarse a un PNS. Sin embargo, usted es responsable de gestionar los fallos transitorios cuando sus servicios o los dispositivos de sus usuarios interactúan con Notification Hubs. Los errores transitorios pueden producirse durante las operaciones de registro, las operaciones de envío de notificaciones y las operaciones de administración. Siga estas instrucciones:

- **Registros e instalaciones:** Las aplicaciones de los dispositivos deben reintentar las operaciones de registro e instalación que producen errores debido a errores transitorios. los SDK proporcionados por Microsoft controlan los reintentos automáticamente. Si no puedes

utilizar los SDK proporcionados, implementa una lógica de reintento con retroceso exponencial y fluctuación, y haz que las operaciones de registro sean idempotentes siempre que sea posible.

La creación o actualización de una instalación es idempotente, por lo que puedes reintentar la operación con total seguridad. Cuando sea posible, use instalaciones en lugar de registros.

- **Operaciones de envío y administración de notificaciones:** Use un SDK proporcionado por Microsoft para enviar notificaciones push y realizar operaciones de administración. Estos SDK reintentan automáticamente cuando se producen errores transitorios.

  Si no puedes utilizar los SDK proporcionados, implementa una lógica de reintento con retroceso exponencial y fluctuación, y haz que las operaciones de envío de notificaciones sean idempotentes siempre que sea posible.

Resistencia a errores de zona de disponibilidad

Las zonas de disponibilidad son grupos físicamente independientes de centros de datos dentro de una región de Azure. Cuando una zona falla, los servicios pueden transferirse a una de las zonas restantes.

En regiones que admiten zonas de disponibilidad, los espacios de nombres de Notification Hubs admiten una configuración *con redundancia de zona* . Notification Hubs habilita automáticamente la redundancia de zona para todos los espacios de nombres de algunas regiones. Cuando se habilita la redundancia de zona, Microsoft replica tanto los metadatos como los datos de registro en todas las zonas de disponibilidad de la región.

Requirements

- **Compatibilidad con regiones:**

Notification Hubs habilita automáticamente la redundancia de zona para todos los espacios de nombres de las siguientes regiones. No se puede deshabilitar la redundancia de zona en estas regiones:

Expandir tabla

Europe	Oriente Medio	Africa	Asia Pacífico
Centro de Francia	Qatar Central	Norte de Sudáfrica	Norte de China 3
Norte de Italia			Centro de Corea del Sur
Este de Noruega			
Centro de Polonia			
Centro de Suecia			
Norte de Suiza			

En otras regiones que admiten Notification Hubs y tienen zonas de disponibilidad, la redundancia de zona es opcional. Solo puedes habilitarlo al crear un espacio de nombres.

- **Compatibilidad con niveles:** Puede usar zonas de disponibilidad con todos los niveles de Notification Hubs.

Cost

La redundancia de zonas conlleva un cargo adicional al precio del nivel. Para más información, consulte Precios de Notification Hubs.

Configurar soporte de zonas de disponibilidad

- **Cree un nuevo espacio de nombres con redundancia de zona:** El proceso para crear un nuevo espacio de nombres con redundancia de zona depende de la región que use:
  - En las regiones en las que Notification Hubs habilita automáticamente la redundancia de zona, no es necesario configurarla.

> **Importante**
>
> En estas regiones, Notification Hubs siempre crea espacios de nombres con redundancia de zona habilitada, incluso si una implementación basada en código,

como un archivo Bicep o una plantilla de Azure Resource Manager, indica que la redundancia de zona está deshabilitada.

Si no desea un espacio de nombres con redundancia de zona, créelo en una región que admita redundancia de zona opcional.

  - En las regiones en las que la redundancia de zona es opcional, solo puede habilitarla cuando se crea un espacio de nombres. Para obtener información sobre cómo configurar un nuevo espacio de nombres con redundancia de zona, consulte Creación de un centro de notificaciones de Azure en el portal de Azure.

- **Hacer redundante una zona de espacio de nombres existente:** Notification Hubs no admite la migración in situ de un espacio de nombres existente a la compatibilidad con zonas de disponibilidad. Debe implementar un nuevo espacio de nombres y mover los registros a ese espacio de nombres. Siga las instrucciones de Mover recursos entre regiones de Azure, que también se aplican si implementa el nuevo espacio de nombres en la misma región.

Comportamiento cuando todas las zonas están en buen estado

En esta sección se describe qué esperar al configurar un espacio de nombres de Notification Hubs con redundancia de zona, cuando todas las zonas están operativas.

- **Operación entre zonas:** Notification Hubs distribuye y atiende automáticamente las solicitudes mediante el uso de la infraestructura en cualquier zona de la región.

- **Replicación de datos entre zonas:** Tanto los datos de registro como los metadatos se replican sincrónicamente en todas las zonas de la región especificada.

Comportamiento durante un fallo de zona

En esta sección se describe qué cabe esperar al configurar un espacio de nombres de Notification Hubs con redundancia de zona, si se produce una interrupción en una de las zonas.

- **Detección y respuesta:** Microsoft detecta fallos de zona y gestiona la conmutación por error dentro de la región. No es necesario iniciar la conmutación por error.

- **Notificación:** Microsoft no le notifica automáticamente cuando una zona está inactiva. Sin embargo, puede usar Azure Service Health para comprender el estado general del servicio,

incluidos los errores de zona, y puede configurar alertas de Service Health para notificarle problemas.

- **Solicitudes activas:** Las operaciones de administración en curso, los registros de dispositivos y las nuevas solicitudes para enviar notificaciones pueden fallar durante la conmutación por error. Las aplicaciones deben reintentar las operaciones con errores siguiendo las instrucciones de control de errores transitorios.

- **Pérdida de datos esperada:** No se espera la pérdida de datos durante una interrupción de una sola zona porque Notification Hubs replica de forma sincrónica el espacio de nombres y la configuración del centro y los datos de registro en todas las zonas de disponibilidad.

  Esta replicación no es una copia de seguridad. En el modelo de responsabilidad compartida, es responsable de realizar copias de seguridad de los datos de registro e instalación. Para obtener más información, consulte Copia de seguridad y restauración.

- **Tiempo de inactividad esperado:** Una breve interrupción del servicio es posible mientras Microsoft redirecciona el tráfico. Siga las instrucciones de control de errores transitorios para preparar las aplicaciones para estas interrupciones.

- **Redistribución:** El servicio redirige automáticamente las solicitudes a zonas correctas.

Recuperación de zona

Cuando se recupera la zona afectada, no es necesario realizar ninguna acción. Microsoft restaura y vuelve a equilibrar la infraestructura de Notification Hubs para usar la zona recuperada.

Prueba de fallos de zona

No se puede activar directamente una conmutación por error de zona de Notification Hubs. Para probar el comportamiento de la carga de trabajo, ejecute pruebas de resiliencia para reintentos, idempotencia y errores de dependencia en entornos no productivos. También puede usar Azure Chaos Studio para probar los componentes de la aplicación circundantes.

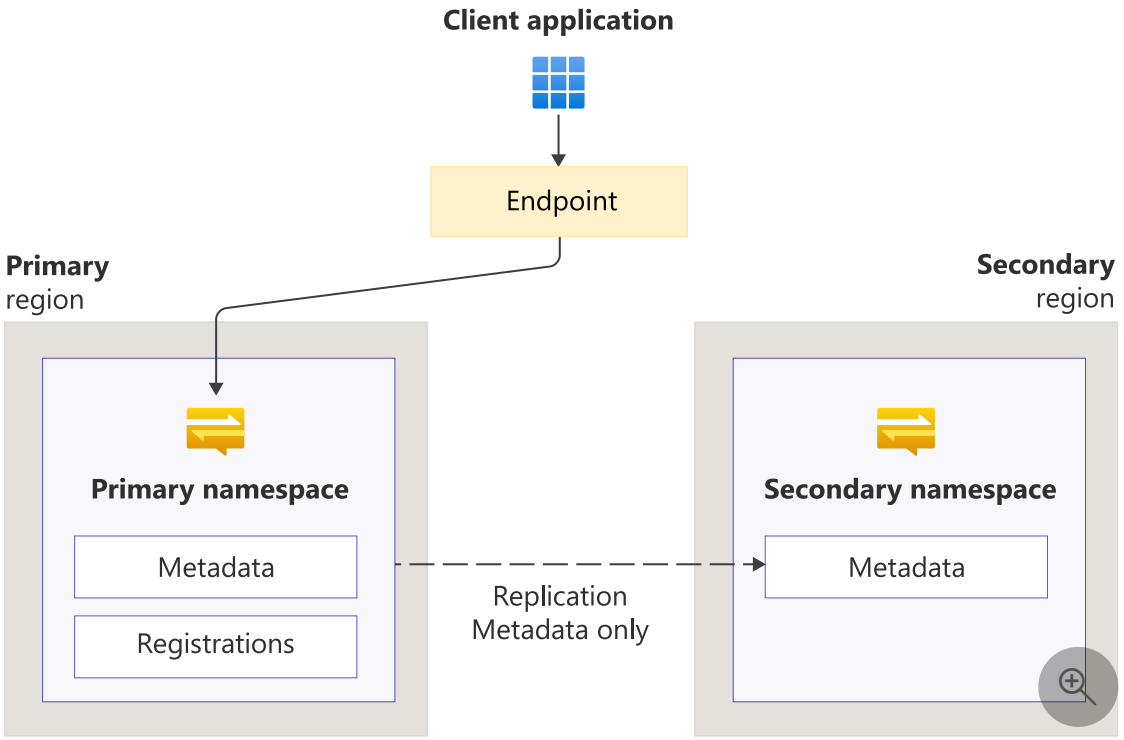
Resistencia a errores en toda la región

Notification Hubs proporciona recuperación ante desastres de metadatos mediante la replicación de los metadatos del espacio de nombres entre regiones, pero no replica los datos de registro de dispositivos. Esta funcionalidad requiere intervención manual durante una interrupción de la región e implica cierto tiempo de inactividad para el centro de notificaciones.

Si necesita reducir el tiempo de inactividad y la intervención manual durante la conmutación por error, considere la posibilidad de usar una solución de varias regiones personalizada.

recuperación geográfica ante desastres de metadatos administrados por Microsoft

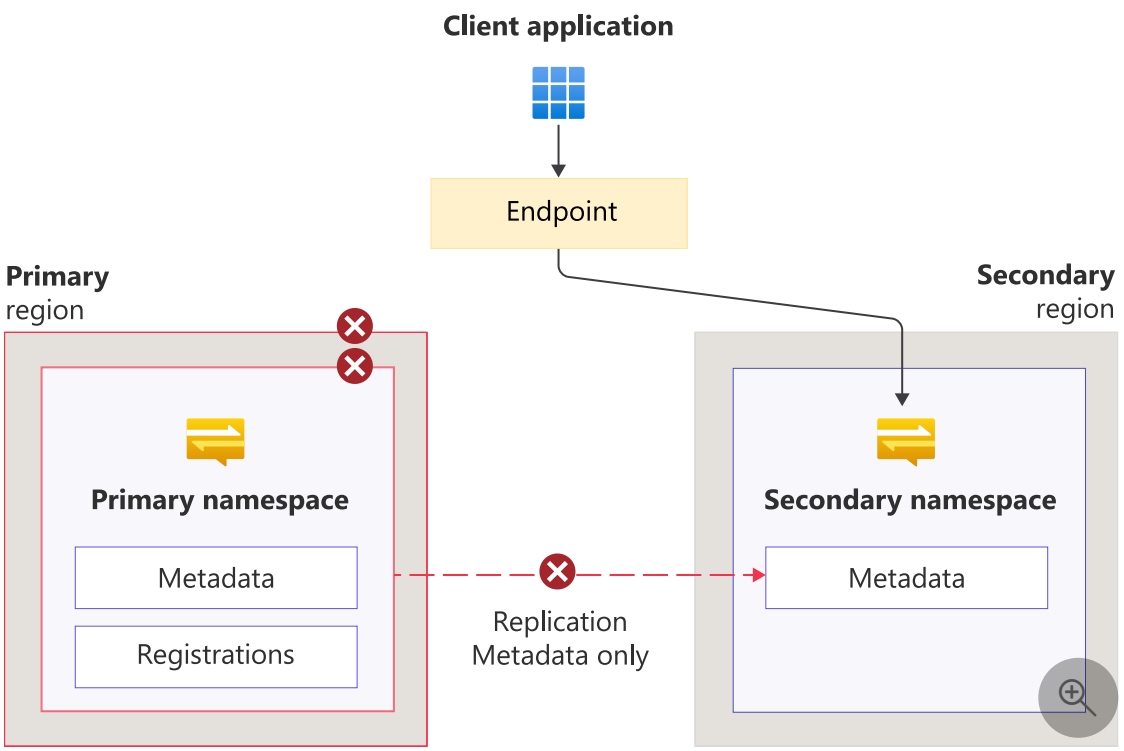
Notification Hubs admite la recuperación ante desastres de metadatos administrados por Microsoft en una región de Azure secundaria. Si la región primaria tiene una región emparejada, puede seleccionar esa región emparejada. Independientemente del estado de emparejamiento de la región primaria, también puede elegir una región secundaria en una lista de *regiones de recuperación flexibles*. A continuación, Notification Hubs replica los metadatos del espacio de nombres, como el nombre del espacio de nombres, las cadenas de conexión y otra información crítica.



> **Importante**
>
> La recuperación ante desastres geográfica de metadatos no replica los datos de registro. Si se desencadena un escenario de recuperación ante desastres, se pueden perder los datos de registro e instalación. Usted es responsable de implementar una solución para restablecer los datos de registro en su hub después de la recuperación.

Microsoft es responsable de declarar un desastre e iniciar la conmutación por error. Cuando esto sucede, Microsoft crea un nuevo espacio de nombres en la región secundaria. Dado que usa los

metadatos de la región primaria, las aplicaciones pueden conectarse a ese espacio de nombres usando el nombre del espacio de nombres existente, la cadena de conexión y los nombres del centro de eventos.



Requirements

- **Compatibilidad con regiones:** En regiones de Azure emparejadas, el espacio de nombres puede usar la región emparejada Azure como región secundaria.

  Si el espacio de nombres está en una región no emparejada o si desea replicar datos en otra región, puede seleccionar una de las siguientes regiones de recuperación flexible como región secundaria:

  Expandir tabla

Americas	Europe	Africa	Asia Pacífico
Sur de Brasil	Norte de Europa	Norte de Sudáfrica	Australia East
Oeste de EE. UU. 2			Sudeste asiático

- **Compatibilidad con niveles:** Las opciones de recuperación ante desastres de metadatos están disponibles en todos los niveles de Notification Hubs.

Cost

Notification Hubs no cobra ningún cargo adicional para configurar o usar la recuperación ante desastres geográfica de metadatos. Sin embargo, paga por el ancho de banda entre regiones que se usa para replicar metadatos. Para más información sobre los precios, consulte Precios de ancho de banda y Precios de Notification Hubs.

Configuración de la compatibilidad con varias regiones

- **Habilite la recuperación ante desastres geográfica de metadatos para un nuevo espacio de nombres:** Siga el procedimiento descrito en Creación de un centro de notificaciones de Azure en el portal de Azure. Seleccione la configuración de recuperación ante desastres al crear el espacio de nombres.
- **Habilite o deshabilite la recuperación ante desastres geográfica de metadatos para un espacio de nombres existente:** Siga el procedimiento descrito en Habilitar la recuperación ante desastres para un espacio de nombres Azure Notification Hubs existente.
- **Realice una copia de seguridad de los datos de registro del dispositivo:** Consulte Exportación e importación de registros de Azure Notification Hubs de forma masiva.

Comportamiento cuando todas las regiones están en buen estado

En esta sección se describe qué cabe esperar al configurar un espacio de nombres de Notification Hubs para la recuperación geográfica ante desastres de metadatos, cuando tanto la región primaria como la secundaria están operativas.

- **Operación entre regiones:** La región primaria atiende todas las solicitudes. La región secundaria no atiende solicitudes a menos que se produzca una conmutación por error.
- **Replicación de datos entre regiones:** Los metadatos, como el nombre del espacio de nombres, la configuración del centro de conectividad, las cadenas de conexión y otra información crítica, se replican de forma asincrónica entre regiones. Los datos de registro no se replican. Usted es responsable de exportarlo periódicamente para mantener una copia de seguridad.

Comportamiento durante una falla de región

En esta sección se describe qué cabe esperar al configurar un espacio de nombres de Notification Hubs para la recuperación ante desastres geográficos de metadatos, y cuando se produce una interrupción del servicio en la región principal.

- **Detección y respuesta:** Microsoft es responsable de detectar el error de la región y decidir si se debe desencadenar la conmutación por error en la región secundaria configurada.
- **Notificación:** Microsoft no le notifica automáticamente cuando una región está inactiva. Sin embargo, puede usar Azure Service Health para comprender el estado general del servicio, incluidos los errores de región, y puede configurar alertas de Service Health para notificarle problemas.
- **Solicitudes activas:** Las solicitudes en curso al espacio de nombres de la región primaria pueden producir un error cuando la región se queda sin conexión. Los clientes deben reintentar las operaciones una vez completada la conmutación por error.
- **Pérdida de datos esperada:** Se conservan los metadatos. No se realiza una copia de seguridad automática de los datos de registro, pero puede realizar una copia de seguridad. Para obtener más información, consulte Exportación e importación de registros de Azure Notification Hubs de forma masiva. Si no lo hace, los datos de registro no estarán disponibles hasta que se recupere la región primaria.
- **Tiempo de inactividad previsto:** Microsoft tarda un tiempo en activar la conmutación por error de los metadatos y, posteriormente, en completar dicha conmutación. Aunque el tiempo puede variar, suele tardar varias horas.

  Una vez completada la conmutación por error, es responsabilidad del usuario restaurar cualquier copia de seguridad de los datos de registro.
- **Redistribución:** Tras la conmutación por error, las solicitudes se redirigen a un espacio de nombres de la región secundaria que utiliza los datos replicados desde la región primaria. Una vez completada la conmutación por error, los clientes se conectan automáticamente al espacio de nombres de la región secundaria.

Recuperación de regiones

Si la región primaria se recupera, podría ser posible volver al espacio de nombres primario de la región primaria. El espacio de nombres principal conservaría los datos de registro antes de la interrupción. Este sería un proceso manual y Microsoft se comunicaría con usted para explicar cómo funciona esto.

Una vez que la región primaria se haya recuperado, debes:

- Comprueba el estado de tu espacio de nombres y sus datos.

- Determine si se van a sincronizar los cambios de datos de registro recientes desde la región secundaria a la región primaria.

Prueba de fallos de región

No se puede iniciar una conmutación por error geográfica. Sin embargo, debe probar sus propios procedimientos de recuperación ante desastres. Compruebe que se realiza una copia de seguridad de los registros y que puede restaurarlos en un nuevo espacio de nombres.

Soluciones de varias regiones personalizadas para la resistencia

La recuperación geográfica ante desastres de metadatos administrados por Microsoft solo replica metadatos. La característica puede recuperar esos metadatos en un espacio de nombres secundario, pero es responsable de importar registros de dispositivos en ese espacio de nombres para que la aplicación pueda seguir funcionando. Este enfoque requiere intervención manual durante un desastre e implica tiempo de inactividad.

Si tus objetivos de recuperación requieren menos tiempo de inactividad o intervención manual, puedes implementar una solución multirregional activa-activa personalizada. Implemente un segundo espacio de nombres de Notification Hubs en otra región de Azure con antelación.

> **Nota:**
>
> En esta sección se proporcionan instrucciones básicas para diseñar este tipo de solución. Usted es responsable de diseñar, implementar, probar, implementar, realizar la conmutación por error y administrar la solución.

- **Conmutación por error:** Dado que el segundo espacio de nombres es un recurso operativo, puede implementar lógica para detectar un fallo en una región y cambiar a ese espacio de nombres.

- **Sincronización:** Para mantener un segundo centro de notificaciones sincronizado con el centro de notificaciones principal, use una de las siguientes opciones:

  - **Para las instalaciones:** Use un back-end de aplicación que cree y actualice simultáneamente instalaciones en ambos centros de notificaciones. Las instalaciones permiten especificar su propio identificador de dispositivo único, que admite este escenario de replicación. Para obtener más información, consulte el ejemplo redundantehub .

- **Para registros:** Use un back-end de aplicación que exporte regularmente los registros desde el centro de notificaciones principal como copia de seguridad y los importe en bloque en el centro de notificaciones secundario. Para obtener más información, consulte Exportación e importación de registros de Azure Notification Hubs de forma masiva.

  Como alternativa, si no tiene un back-end, configure la aplicación para crear instalaciones en ambos centros cuando la aplicación se inicie en los dispositivos de destino. Los dispositivos crean nuevos registros en ambos centros de notificaciones. Finalmente, el centro de notificaciones secundario tiene registrados todos los dispositivos activos.

- **Registros e instalaciones expirados:** Es posible que el centro de notificaciones secundario haya expirado los registros y las instalaciones. Cuando se envía una notificación push a un identificador expirado, Notification Hubs elimina automáticamente el registro o registro de instalación asociado en el centro de notificaciones, en función de la respuesta recibida del servidor PNS. Puede limpiar los registros expirados de la solución de copia de seguridad que prefiera agregando lógica personalizada que procesa los comentarios de cada envío y quita los registros y las instalaciones expirados.

- **Aplicaciones sin abrir:** Hay un período de tiempo durante el que los dispositivos con aplicaciones sin abrir no reciben notificaciones.

- **Costo:** Si usa su propio centro secundario para proteger los datos de registro, ese centro incurre en cargos de servicio normales. Del mismo modo, si implementa cualquier otro recurso de Azure en la región secundaria para respaldar la recuperación, pagará por ellos según las tarifas normales del servicio.

Copias de seguridad y restauración

Notification Hubs no proporciona una única característica integrada de copia de seguridad y restauración para todos los datos almacenados en su espacio de nombres. Usted es responsable de combinar los enfoques siguientes:

- Utilice infraestructura como código (IaC), como Bicep, para definir la configuración de su espacio de nombres, hub y directivas. Almacene esas definiciones en el control de código fuente para que pueda volver a implementar los recursos cuando sea necesario.
- Realice una copia de seguridad de los datos de registro del dispositivo exportando los registros de Azure Notification Hubs de manera masiva.

Resistencia al mantenimiento del servicio

Microsoft aplica periódicamente actualizaciones de servicio y realiza otro mantenimiento. La plataforma Azure controla estas actividades automáticamente, lo que garantiza que el mantenimiento sea transparente y sin problemas. No se prevé ningún tiempo de inactividad durante las operaciones de mantenimiento, a menos que se te haya informado a través de el mantenimiento planificado de Azure Service Health.

Acuerdo de nivel de servicio

El acuerdo de nivel de servicio (SLA) para Azure servicios describe la disponibilidad esperada de cada servicio y las condiciones que la solución debe cumplir para lograr esa expectativa de disponibilidad. Para obtener más información, consulte los SLAs de los servicios en línea.

En el caso de Notification Hubs, el acuerdo de nivel de servicio (SLA) de disponibilidad se aplica a los espacios de nombres que utilizan los niveles Básico y Estándar.

Contenido relacionado

- Confiabilidad en Azure

ⓘ**Nota:** El autor ha creado este artículo con ayuda de inteligencia artificial. Más información

---

Last updated on 27/08/2026

Confiabilidad en Azure Automation

Azure Automation es un servicio que ejecuta tareas de administración en su nombre. Puede definir scripts, *denominados runbooks*, que desea ejecutar y Azure Automation proporciona la infraestructura para ejecutar esos scripts. Este artículo se centra en la automatización de procesos, que es la funcionalidad principal del servicio. Los trabajadores híbridos de runbook, que se ejecutan en infraestructura gestionada por el cliente, están fuera del ámbito de este artículo.

Cuando se usa Azure, la confiabilidad es una responsabilidad compartida. Microsoft proporciona una variedad de capacidades para apoyar la resiliencia y la recuperación. Es responsable de comprender cómo funcionan esas funcionalidades dentro de todos los servicios que usa y de seleccionar las funcionalidades que necesita para cumplir los objetivos empresariales y los objetivos de tiempo de actividad.

En este artículo se describe cómo hacer que Azure Automation sea resistente a diversas interrupciones y problemas potenciales, incluidos errores transitorios, interrupciones de zonas de disponibilidad, interrupciones regionales y mantenimiento del servicio. También se describen las opciones de copia de seguridad y restauración, así como información clave sobre el acuerdo de nivel de servicio (SLA) de Azure Automation.

Recomendaciones de implementación de producción para la confiabilidad

Para cargas de trabajo de producción que usan la automatización de procesos, siga estas recomendaciones:

- ✓ Controle los errores transitorios que se producen cuando los runbooks interactúan con Azure servicios y API mediante la adición de la lógica de reintento adecuada a los scripts.
- ✓ Diseñe los runbooks para que sean resistentes a las interrupciones. Use puntos de control para mantener el progreso entre los reinicios del trabajo y, si necesita almacenar el estado, use el almacenamiento externo.

Introducción a la arquitectura de confiabilidad

En esta sección se describen algunos de los aspectos importantes de cómo funciona el servicio que es más relevante desde una perspectiva de confiabilidad. En la sección se presenta la arquitectura lógica, que incluye algunos de los recursos y características que se implementan y usan. También se describe la arquitectura física, que proporciona detalles sobre cómo funciona el servicio en segundo plano.

Arquitectura lógica

Al implementar Azure Automation, se crea una *cuenta de automation*, que es un contenedor lógico para los recursos que ejecutan la automatización.

- *Runbooks*, que representan el trabajo que se va a realizar. *Los runbooks textuales* son scripts escritos en PowerShell o Python. *Los runbooks gráficos* se crean con un editor gráfico.
- Recursos que comparten runbooks, incluidos módulos, conexiones, credenciales, certificados y variables.
- Recursos que inician la ejecución del runbook, incluidas las programaciones y los monitores.

Para obtener más información sobre estos recursos, consulte Ejecución de runbooks en Azure Automation.

En este artículo se tratan la confiabilidad y resistencia de estas funcionalidades, que forman parte de la *automatización de procesos* en Azure Automation.

Arquitectura física

Los runbooks se ejecutan en una infraestructura informática. Existen dos modelos de implementación para la automatización de procesos:

- Trabajos en la nube **(Microsoft administrados):** de forma predeterminada, los runbooks se ejecutan en la infraestructura de nube proporcionada por Microsoft. Microsoft es responsable de la alta disponibilidad y administración de esta infraestructura. Al enviar un trabajo de runbook, Azure Automation asigna un trabajo en la nube desde su grupo de recursos de proceso disponibles, ejecuta el runbook y, a continuación, devuelve el recurso al grupo.

  A veces, es posible que los trabajos en la nube se interrumpan mientras se ejecutan. Diseñe los runbooks en la suposición de que un trabajo podría reiniciarse en una infraestructura diferente y que los datos escritos en el almacenamiento temporal en la instancia anterior ya no son accesibles.

- **Hybrid Runbook Workers:** Opcionalmente, puede configurar su propia infraestructura de proceso (máquinas virtuales en Azure, otras nubes o locales) para ejecutar runbooks. Al usar hybrid runbook workers, es responsable de configurarlos para cumplir los requisitos de confiabilidad. Los trabajadores de runbook híbridos están fuera del alcance de este artículo.

Resistencia a errores transitorios

Los errores transitorios son errores breves e intermitentes en los componentes. Se producen con frecuencia en un entorno distribuido como la nube y son una parte normal de las operaciones. Los errores transitorios se corrigen después de un breve período de tiempo. Es importante que las aplicaciones puedan controlar errores transitorios, normalmente mediante el reintento de solicitudes afectadas.

Todas las aplicaciones hospedadas en la nube deben seguir las instrucciones de control de errores transitorios de Azure cuando se comunican con cualquier API, bases de datos y otros componentes hospedados en la nube. Para obtener más información, consulte Recomendaciones para controlar errores transitorios.

Eres responsable de escribir guías operativas que gestionen errores transitorios en los servicios y las API con los que interactúan. En el caso de los runbooks textuales, implemente la lógica de reintento mediante bucles y control de errores. Para obtener instrucciones y ejemplos, consulte Control de errores transitorios en un script dependiente del tiempo. En los runbooks gráficos, configure el comportamiento de reintento para las actividades de su flujo de trabajo. Para obtener más información sobre la configuración, consulte Actividad de reintento en runbooks gráficos.

El mantenimiento de la infraestructura u otros eventos de la plataforma pueden interrumpir las tareas de runbook. Diseñe los runbooks para controlar estas interrupciones:

- **Implementar puntos de control.** En el caso de los runbooks de flujo de trabajo de PowerShell, use puntos de control para guardar el progreso en los puntos clave del flujo de trabajo. Si un trabajo se interrumpe y se reinicia, puede reanudarse desde el último punto de control en lugar de volver a empezar. Para obtener más información, consulte Uso de puntos de control en un flujo de trabajo.

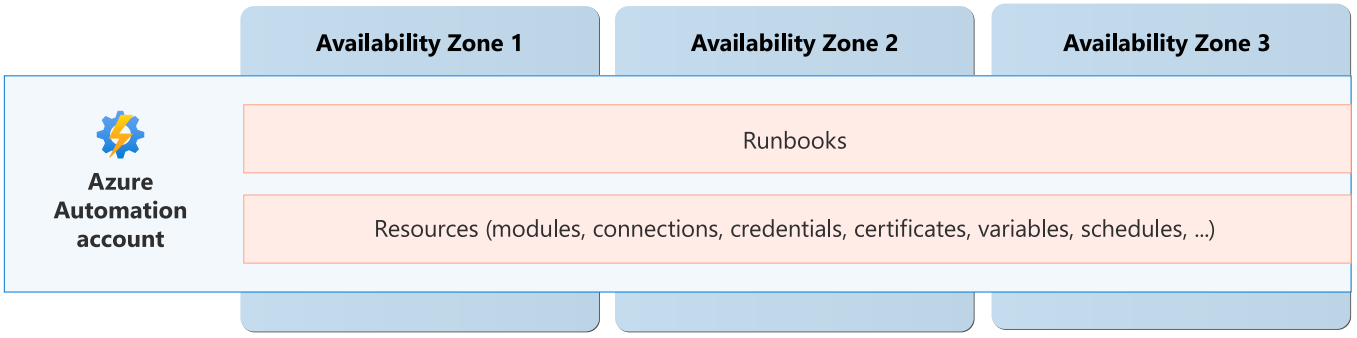
- **Comprenda los límites de las tareas.** Los trabajos en la nube tienen límites de uso equitativo en la duración del tiempo de ejecución. Para obtener más información sobre los límites de ejecución del trabajo y cómo se aplican, consulte Ejecución de runbook.

- **Almacenar el estado persistente externamente.** Los trabajos no mantienen el estado entre ejecuciones. Si se interrumpe un trabajo y se reinicia en otra instancia, es posible que se pierda algo escrito en el almacenamiento temporal por la primera ejecución. Si necesita conservar los datos entre ejecuciones de trabajo, almacénelo en almacenamiento externo, como Azure Blob Storage o una base de datos.

Resistencia a errores de zona de disponibilidad

Las zonas de disponibilidad son grupos físicamente independientes de centros de datos dentro de una región de Azure. Cuando una zona falla, los servicios pueden transferirse a una de las zonas restantes.

En las regiones admitidas, las cuentas de Automation y los trabajos en la nube tienen *redundancia de zona*, lo que significa que el servicio distribuye los recursos entre varias zonas de disponibilidad. Microsoft habilita automáticamente la redundancia de zona y no requiere ninguna configuración.



Requisitos

Compatibilidad regional: Al implementar una cuenta de automatización en una de las siguientes regiones, pasa a tener redundancia de zona automáticamente:

Expandir tabla

Americas	Europe	Oriente Medio	Africa	Asia Pacífico
Brazil South	France Central	Israel Central	Norte de Sudáfrica	Australia East
Canada Central	Centro-oeste de Alemania	Qatar Central		Central India
Central US	Norte de Italia			Norte de

Americas	Europe	Oriente Medio	Africa	Asia Pacífico
				China 3
Este de EE. UU.	Norte de Europa			East Asia
Este de EE. UU. 2	Norway East			Japan East
Centro-sur de EE. UU.	Poland Central			Korea Central
Gobierno de EE.UU. de Virginia	Sweden Central			Southeast Asia
Oeste de EE. UU. 2	UK South			
Oeste de EE. UU. 3	West Europe			

Para obtener la lista actual de regiones admitidas, consulte Compatibilidad con zonas de disponibilidad para Azure Automation.

Costo

No hay ningún cargo adicional por la redundancia de zona. Para la automatización de procesos, la facturación se basa en el tiempo durante el que se ejecutan los trabajos y supervisores. Para obtener más información, consulte precios de Azure Automation.

Configurar soporte de zonas de disponibilidad

Cuando crea una cuenta de automatización en una región admitida, esta es automáticamente redundante entre zonas. No se puede deshabilitar la redundancia de zona. Para obtener más información, consulte Compatibilidad con zonas de disponibilidad para Azure Automation.

Comportamiento cuando todas las zonas están en buen estado

En esta sección se describe qué cabe esperar cuando la cuenta de automatización tiene redundancia de zona y todas las zonas de disponibilidad de la región están operativas.

- **Operación entre zonas:** Las operaciones de administración de cuentas de Automation y los trabajos en la nube se distribuyen automáticamente entre zonas de disponibilidad en la región. Una solicitud o un trabajo pueden ser atendidos por cualquier instancia de cualquier zona de disponibilidad.
- **Replicación de datos entre zonas:** La configuración de la cuenta de Automation, los scripts de runbook y otros recursos que se implementan en la cuenta de Automation se replican

sincrónicamente en varias zonas de disponibilidad.

Comportamiento durante un fallo de zona

En esta sección se describe qué cabe esperar cuando la cuenta de automatización tiene redundancia de zona y se produce una interrupción en una de las zonas de disponibilidad de la región.

- **Detección y respuesta:** La plataforma Azure Automation es responsable de detectar un error en una zona de disponibilidad. No es necesario hacer nada para iniciar una conmutación por error de la zona.

- **Notificación:** Microsoft no le notifica automáticamente cuando una zona está inactiva. Sin embargo, puede usar Azure Service Health para comprender el estado general del servicio, incluidos los errores de zona, y puede configurar alertas de Service Health para notificarle problemas.

- **Solicitudes activas:** Cualquier trabajo que se esté ejecutando en la zona con problemas podría interrumpirse. Azure Automation inicia automáticamente una nueva ejecución de un trabajo utilizando infraestructura de zonas en buen estado. Diseñe los runbooks para que sean resistentes a errores transitorios e interrupciones para que puedan reiniciarse de forma segura.

- **Pérdida de datos esperada:** Las ejecuciones de trabajos no conservan el estado, por lo que no se espera que un error de zona cause pérdida de datos para trabajos en curso. Si un trabajo necesita almacenar los datos que puede usar para recuperarse de la interrupción, como puntos de control, almacene esa información en un servicio de almacenamiento en la nube persistente, como Azure Storage o una base de datos.

  La configuración de la cuenta de Automation y los datos de runbook se replican entre zonas y permanecen accesibles incluso cuando una zona no está disponible.

- **Tiempo de inactividad esperado:** Durante una interrupción de zona, la cuenta de Automation puede experimentar una breve interrupción mientras el servicio detecta el error y redistribuye la carga de trabajo a zonas correctas.

- **Redistribución:** El servicio vuelve a equilibrar automáticamente la capacidad en las zonas correctas restantes. Las nuevas ejecuciones de trabajos, monitores y programaciones siguen ejecutándose en la infraestructura en zonas correctas. La recuperación no depende de que la zona que ha fallado vuelva a estar operativa.

Recuperación de zona

Cuando una zona con error vuelve al servicio, Azure Automation vuelve a reinsertarla automáticamente en la rotación de zona. No se requiere ninguna acción de usted. El servicio supervisa el estado de la zona y redistribuye la carga de trabajo en todas las zonas a medida que se reanudan las operaciones normales.

Prueba de fallos de zona

Azure Automation administra el enrutamiento del tráfico, la conmutación por error y la recuperación de zona para los recursos con redundancia de zona. No es necesario iniciar ninguna acción ni validar los procesos ante fallos de la zona de disponibilidad. Ponga a prueba sus manuales de procedimientos para confirmar que resisten las interrupciones.

Resistencia a errores en toda la región

Azure Automation es un servicio de una sola región. Si la región deja de estar disponible, la cuenta de Automation tampoco está disponible.

Soluciones personalizadas de varias regiones para la resistencia

Puede implementar cuentas de Automation independientes en varias regiones y cambiar entre ellas cuando sea necesario. Usted es responsable de desplegar las cuentas en cada región, configurarlas de forma adecuada, distribuir las solicitudes entre las cuentas y gestionar la conmutación por error si una región no está disponible. Para obtener información detallada sobre los enfoques que puede tener en cuenta, consulte Recuperación ante desastres para Azure Automation.

Copias de seguridad y restauración

Para la mayoría de las soluciones, no debe confiar exclusivamente en copias de seguridad. En su lugar, utilice las otras capacidades descritas en esta guía para apoyar los requisitos de resiliencia. Sin embargo, las copias de seguridad protegen contra algunos riesgos que otros enfoques no. Para más información, consulte ¿Qué son la redundancia, la replicación y la copia de seguridad?.

Azure Automation no proporciona copias de seguridad integradas para la configuración de la cuenta de Automation ni para el contenido del runbook. Mantenga sus propias copias fuera del servicio para que pueda volver a implementarlas si es necesario.

- **Use la infraestructura como código (IaC) para la configuración de la cuenta de Automation.** Defina cuentas de automatización y recursos relacionados en archivos de Bicep, plantillas de ARM o Terraform. Almacene las plantillas en el control de versiones y use su pipeline de implementación para recrear el entorno en la misma región o en otra. Incluya certificados, variables, programaciones y referencias de credenciales en sus artefactos y procesos de despliegue. Almacene secretos en servicios como Azure Key Vault en lugar de insertar valores directamente en el código de runbook.

- **Almacene los scripts del runbook en un sistema de control de código fuente.** Mantenga el origen de PowerShell y Python runbooks en un sistema de control de código fuente como Git. Use el control de versiones, la bifurcación y la revisión de solicitudes de incorporación de cambios para proteger la calidad del script y habilitar la reversión a versiones correctas conocidas.

- **Haga una copia de seguridad del estado desde el almacén de datos correspondiente.** Los trabajos no conservan el estado. Si necesita mantener registros de trabajos detallados o cualquier otro dato que generen los runbooks, almacénelo en otro servicio de almacenamiento o base de datos de Azure y realice una copia de seguridad desde allí.

Resistencia a la eliminación accidental

Si elimina accidentalmente una cuenta de Automation, es posible que pueda restaurarla dentro de un período de tiempo limitado. Para más información, consulte Restauración de una cuenta de Automation eliminada.

Resistencia al mantenimiento del servicio

Microsoft aplica periódicamente actualizaciones de servicio y realiza otro mantenimiento. La plataforma Azure controla estas actividades automáticamente, lo que garantiza que el mantenimiento sea transparente y sin problemas. No se espera ningún tiempo de inactividad durante los eventos de mantenimiento a menos que se le haya informado a través del mantenimiento planeado de Azure Service Health.

Acuerdo de nivel de servicio

El acuerdo de nivel de servicio (SLA) para Azure servicios describe la disponibilidad esperada de cada servicio y las condiciones que la solución debe cumplir para lograr esa expectativa de disponibilidad. Para obtener más información, consulte Acuerdos de Nivel de Servicio para servicios en línea.

Contenido relacionado

- Confiabilidad en Azure
- Recuperación ante desastres para Azure Automation
- Administración de runbooks en Azure Automation
- Ejecución de runbooks en Azure Automation

ⓘ**Nota:** El autor ha creado este artículo con ayuda de inteligencia artificial. Más información

---

Last updated on 07/07/2026

Confiabilidad en Azure Backup

Azure Backup es un servicio de Azure integrado que protege de forma segura las cargas de trabajo locales y en la nube. La copia de seguridad puede escalar su protección en múltiples cargas de trabajo y proporciona integración nativa con las cargas de trabajo de Azure, incluidas las máquinas virtuales (VM) en Azure, SAP HANA en máquinas virtuales en Azure, SQL en máquinas virtuales en Azure, Azure Files, Azure Blob Storage, Azure Data Lake Storage, discos administrados de Azure, volúmenes de Azure Elastic SAN y Azure Kubernetes Service (AKS). No es necesario administrar la automatización ni la infraestructura, escribir scripts ni aprovisionar el almacenamiento.

Al usar Azure, relibilidad es una responsabilidad compartida. Microsoft proporciona una variedad de capacidades para admitir resiliencia y recuperación. Es responsable de comprender cómo funcionan esas funcionalidades dentro de todos los servicios que usa y de seleccionar las funcionalidades que necesita para cumplir los objetivos empresariales y los objetivos de tiempo de actividad.

En este artículo se describe cómo la copia de seguridad puede ser resistente a una variedad de posibles interrupciones y problemas, incluidos errores transitorios, interrupciones de zona de disponibilidad y interrupciones de regiones. También resalta cierta información clave sobre el acuerdo de nivel de servicio (SLA) de copia de seguridad.

> **Nota:**
>
> En este artículo se describe cómo el propio servicio Backup es resistente a varios problemas y cómo puede hacer que sea más resistente. No explica cómo usar backup para proteger las máquinas virtuales, los datos u otros recursos. Para obtener información sobre cómo usar Backup, consulte **Introducción a la copia de seguridad**.

Recomendaciones de implementación de producción para la confiabilidad

Para realizar copias de seguridad de las cargas de trabajo de producción, se recomienda configurar la bóveda de las siguientes maneras:

- ✔ Use el almacenamiento con redundancia de zona (ZRS) como nivel de redundancia mínimo para las copias de seguridad. ZRS replica las copias de seguridad en varias zonas de disponibilidad para poder restaurar las copias de seguridad durante una interrupción de la zona de disponibilidad.
- ✔ Si usa almacenamiento con redundancia geográfica (GRS) para replicar las copias de seguridad en una región de Azure emparejada, habilite la restauración entre regiones (CRR) para orígenes de datos admitidos. CRR le permite restaurar las copias de seguridad en la región emparejada en cualquier momento.

En las secciones siguientes de este artículo se proporcionan más detalles sobre estas configuraciones.

> ⓘ **Nota:**
>
> Estas recomendaciones de redundancia de almacenamiento se aplican a ubicaciones en las que se replican las copias de seguridad, no al servicio Backup o a los recursos de los que realiza una copia de seguridad. La protección de copia de seguridad y la redundancia de almacenamiento se complementan entre sí. Las copias de seguridad protegen contra la pérdida de datos y la redundancia protege frente a errores de infraestructura.

Para obtener una lista de otras recomendaciones para copia de seguridad, incluidas las recomendaciones centradas en la confiabilidad, consulte Copia de seguridad de cargas de trabajo en la nube y cargas de trabajo locales a la nube.

Introducción a la arquitectura de confiabilidad

En esta sección se describen algunos de los aspectos importantes de cómo funciona el servicio que es más relevante desde una perspectiva de confiabilidad. En la sección se presenta la arquitectura lógica, que incluye algunos de los recursos y características que se implementan y usan. También se describe la arquitectura física, que proporciona detalles sobre cómo funciona el servicio en segundo plano.

Arquitectura lógica

La copia de seguridad puede realizar copias de seguridad y restaurar una variedad de *orígenes de datos*. Las copias de seguridad se configuran de forma diferente en función del origen de datos con el que trabaje. Los orígenes de datos siguientes son comunes:

- máquinas virtuales de Azure
- Varias bases de datos
- cuentas de Blob Storage
- Clústeres de AKS
- Servidores locales a través del agente de Microsoft Azure Recovery Services (MARS)

La copia de seguridad almacena los datos de copia de seguridad en *almacenes*. Las bóvedas son entidades de almacenamiento en línea en Azure que contienen datos, como copias de seguridad, puntos de recuperación y políticas de copia de seguridad. Los almacenes de Recovery Services y los almacenes de Backup son dos tipos de almacenes. Puede usar uno o ambos tipos en función de lo que necesite proteger. Para obtener una lista de los orígenes de datos que admite cada tipo de almacén, consulte Preguntas más frecuentes sobre los almacenes admitidos para la copia de seguridad y restauración.

Los trabajos representan la actividad de copia de seguridad o restauración de los datos. Los trabajos de copia de seguridad incluyen operaciones programadas o bajo demanda que copian sus datos desde el origen a la bóveda. Los trabajos de restauración incluyen operaciones que recuperan los datos del almacenamiento de copia de seguridad en una ubicación de destino. Cada trabajo tiene un identificador único y un seguimiento de estado para que pueda supervisar el progreso y solucionar problemas que se producen durante las operaciones de copia de seguridad y restauración. Tú también creas *directivas de copia de seguridad* asociadas a trabajos. Las directivas especifican la configuración como la programación de copia de seguridad y cuánto tiempo desea conservar los datos.

Bóvedas almacenan las directivas de copia de seguridad y la configuración junto con los metadatos de los trabajos, permitiéndole realizar un seguimiento de los trabajos y solucionar problemas.

Arquitectura física

Microsoft administra la infraestructura principal del servicio Backup. Esta infraestructura es responsable de la administración y el funcionamiento del servicio, incluidos los trabajos de desencadenamiento y supervisión.

Los respaldos se almacenan en la bóveda. Las bóvedas se construyen sobre Azure Storage. Las bóvedas replican automáticamente los datos de copias de seguridad, y la durabilidad y resistencia de la copia de seguridad dependen de la redundancia de almacenamiento de la bóveda.

- Almacenamiento con redundancia de ubicación (LRS) replica los datos dentro de su almacén en una o varias zonas de disponibilidad de Azure ubicadas en la región primaria de su elección. No puede elegir la zona de disponibilidad preferida, pero Azure puede mover o expandir cuentas de LRS entre zonas para mejorar el equilibrio de carga. No se garantiza que los datos se repartan entre zonas. Para obtener más información, consulte Introducción a las zonas de disponibilidad.

- ZRS y GRS proporcionan protecciones adicionales. En este artículo se describen estas opciones al detalle.

> ⓘ **Nota:**
>
> Algunas fuentes de datos admiten copias de seguridad de *nivel operativo*, que almacenan datos en otra ubicación en lugar del almacén. Por ejemplo, la **copia de seguridad de discos administrados de Azure** y las **copias de seguridad de AKS** admiten copias de seguridad de nivel operativo, que se almacenan en instantáneas de disco. En este artículo no se describe el almacenamiento de copia de seguridad de nivel operativo, pero puede aplicar las instrucciones de resistencia de este artículo a operaciones y flujos de trabajo de copia de seguridad para estos tipos de copia de seguridad.

Resistencia a errores transitorios

Los errores transitorios son errores breves e intermitentes en los componentes. Se producen con frecuencia en un entorno distribuido como la nube y son una parte normal de las operaciones. Los errores transitorios se corrigen después de un breve período de tiempo. Es importante que las aplicaciones puedan controlar errores transitorios, normalmente mediante el reintento de solicitudes afectadas.

Todas las aplicaciones hospedadas en la nube deben seguir las instrucciones de control de errores transitorios Azure cuando se comunican con cualquier API, bases de datos y otros componentes hospedados en la nube. Para obtener más información, consulte Recomendaciones para controlar errores transitorios.

Cuando se usa Backup, los flujos de trabajo de copia de seguridad y restauración son resistentes a errores intermitentes. El servicio vuelve a intentarlo automáticamente cuando encuentra errores transitorios de red o interrupciones temporales del servicio. Tú no configuras ninguna lógica de reintento. Si experimenta errores repetidos, consulte Solución de problemas de las operaciones de administración de la bóveda de copias de seguridad.

Resistencia a errores de zona de disponibilidad

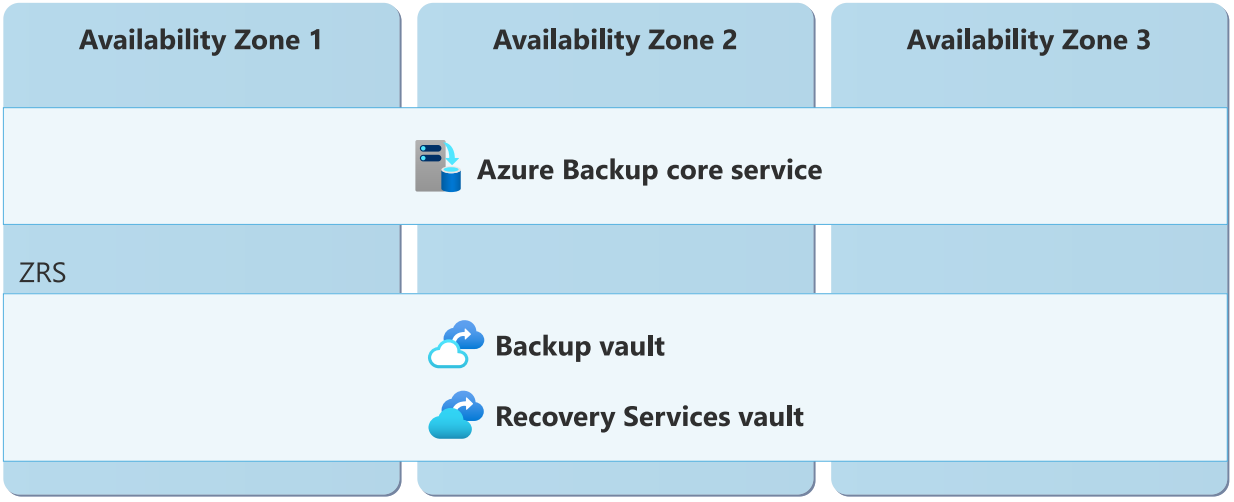
Availability zones son grupos de centros de datos físicamente independientes dentro de una región de Azure. Cuando una zona falla, los servicios pueden transferirse a una de las zonas restantes.

La copia de seguridad administra por separado la configuración de zona de disponibilidad del servicio y de los datos.

- **Servicio:** El servicio de Backup es automáticamente resistente a las zonas en regiones admitidas. Sin embargo, esta resistencia de zona integrada no se aplica a los datos de copia de seguridad.

- **Redundancia de almacenamiento de copia de seguridad:** Seleccione el nivel de redundancia que desea para los datos de copia de seguridad mediante la configuración del almacén de Recovery Services o el almacén de Backup. Si selecciona ZRS, las copias de los datos de copia de seguridad se almacenan automáticamente en varias zonas de disponibilidad en la región de Azure que use.

  Si no usa ZRS, los datos de copia de seguridad se consideran *no zonales* y pueden almacenarse en cualquier zona. Si alguna zona de la región tiene un problema, es posible que los datos de copia de seguridad no zonales no estén disponibles.



Requisitos

- **Compatibilidad con regiones:** El servicio es automáticamente resistente a zonas en todas las regiones que tienen zonas de disponibilidad. Los almacenes de ZRS se admiten en las mismas regiones.

- **Solo bóvedas nuevas:** Configure ZRS en su bóveda antes de la primera copia de seguridad.

Cost

Al habilitar ZRS para las copias de seguridad, se le cobra a una tarifa diferente de LRS debido a la sobrecarga adicional de replicación y almacenamiento. Para obtener más información, consulte Precios de copia de seguridad .

Configurar soporte de zonas de disponibilidad

- **Cree un nuevo almacén que use ZRS:** Configure la redundancia de almacenamiento al crear un almacén. Siga los diferentes pasos según el tipo de bóveda. Para obtener más información, consulte los artículos siguientes:
  - Crear y eliminar almacenes de copias de seguridad
  - Creación y configuración de un almacén de Recovery Services

- **Configuración de ZRS en almacenes existentes:** En el caso de los almacenes de Backup, configure la redundancia de almacenamiento al crear el almacén. Después de crear un almacén de Backup, la configuración está bloqueada y no se puede cambiar.

  En el caso de las bóvedas de Recovery Services, debe configurar la redundancia de almacenamiento antes de proteger cualquier carga de trabajo. Después de proteger una carga de trabajo, la configuración está bloqueada y no se puede cambiar.

  Puede crear una nueva bóveda configurada para usar ZRS y reasignar las cargas de trabajo a la nueva bóveda. Sin embargo, este enfoque requiere tiempo de inactividad. Para obtener más información, consulte Modificación de la configuración predeterminada. También es responsable de eliminar manualmente los puntos de recuperación existentes y otros datos porque las directivas de retención del almacén antiguo ya no se aplican. Para obtener más información, consulte Eliminación de un almacén de Backup o Eliminación de un almacén de Recovery Services.

Comportamiento cuando todas las zonas están en buen estado

En esta sección se describe qué esperar al configurar bóvedas para ZRS cuando todas las zonas están operativas.

- **Operación entre zonas:** Los trabajos de copia de seguridad se ejecutan en la infraestructura replicada entre zonas. Azure administra los trabajos de la infraestructura en cualquier zona.

- **Replicación de datos entre zonas:** ZRS replica los datos de copia de seguridad entre zonas. La replicación se produce de forma sincrónica, lo que significa que varias zonas reconocen cada operación de escritura antes de que se complete.

Comportamiento durante un fallo de zona

En esta sección se describe qué esperar al configurar bóvedas para ZRS cuando se produce una interrupción en una de las zonas.

- **Detección y respuesta:** Para el servicio de copia de seguridad, Microsoft es responsable de detectar fallos en las zonas de disponibilidad y responder. No es necesario hacer nada para iniciar una conmutación por error de la zona.

  > ⓘ **Importante**
  >
  > En el caso de los datos o recursos que no están disponibles debido a una falla de la zona, usted es responsable de detectar la falla y tomar medidas de recuperación, incluida la restauración de copias de seguridad en una zona operativa.

- **Notification**: Microsoft no le notifica automáticamente cuando una zona está inactiva. Sin embargo, puede usar Azure Resource Health para supervisar el estado de un recurso individual y puede configurar Resource Health alertas para notificarle problemas. También puede usar Azure Service Health para comprender el estado general del servicio, incluidos los errores de zona, y puede configurar alertas de Service Health para notificarle problemas.

- **Solicitudes activas:** El comportamiento de las tareas activas depende de qué zona falla.

  - Para cualquier origen de datos en la zona de disponibilidad fallida, el fallo de la zona hace que los orígenes de datos no estén disponibles. Los trabajos activos podrían pausarse o fallar.

  - En el caso de los orígenes de datos en zonas de disponibilidad correctas que ejecutan trabajos activos, una pequeña cantidad de tiempo de inactividad, normalmente unos segundos, puede producirse mientras la plataforma cambia a zonas de disponibilidad correctas para el servicio Backup.

- **Pérdida de datos esperada:** La cantidad esperada de pérdida de datos también se conoce como *objetivo de punto de recuperación (RPO)*. El RPO de los datos de copia de seguridad depende de varios factores, incluida la programación de las copias de seguridad. En general, para una interrupción de zona, no se espera ninguna pérdida de datos de copia de seguridad porque todos los datos se replican de manera sincrónica entre las zonas.

- **Tiempo de inactividad esperado:** La cantidad esperada de tiempo de inactividad también se conoce como *objetivo de tiempo de recuperación (RTO)*. El RTO es diferente para cada uno de los escenarios siguientes:

  - En el caso de los orígenes de datos de la zona de disponibilidad fallida, es posible que estos no estén disponibles hasta que la zona se recupere. Es posible que los trabajos de copia de seguridad no se ejecuten hasta que el origen de datos esté disponible de nuevo. El RTO no está definido.

  - En el caso de los orígenes de datos en zonas de disponibilidad correctas, una pequeña cantidad de tiempo de inactividad, normalmente unos segundos, puede producirse mientras la plataforma cambia a zonas de disponibilidad correctas para el servicio Backup.

- **Redistribución:** Las ejecuciones posteriores de trabajos usan automáticamente la infraestructura en zonas saludables siempre y cuando los orígenes de datos estén disponibles.

  Es responsable de restaurar la copia de seguridad en la infraestructura de una zona saludable y reconfigurar los balanceadores de carga, los clientes y otros sistemas para redirigir el tráfico a una infraestructura saludable en la nueva zona.

Recuperación de zona

Cuando se recupera la zona de disponibilidad, Backup restaura automáticamente las operaciones en la zona de disponibilidad y vuelve a enrutar el tráfico entre las zonas como es normal. Los trabajos siguen ejecutándose y los datos permanecen disponibles.

Prueba de fallos de zona

La plataforma de respaldo gestiona el enrutamiento del tráfico, la replicación de datos, la conmutación por error y la recuperación después de fallos. Esta característica está totalmente administrada, por lo que no es necesario iniciar ni validar los procesos de error de zona de disponibilidad.

Resistencia a errores en toda la región

Backup admite la redundancia geográfica y la conmutación por error a través de GRS y CRR.

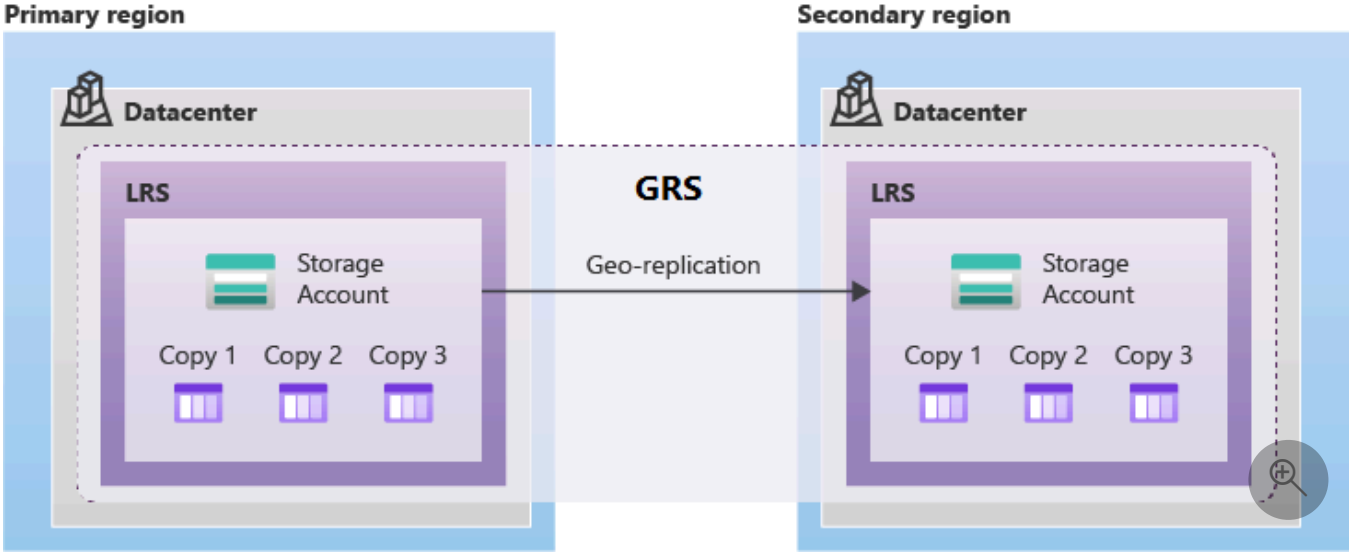
> ⓘ **Importante**
>
> GRS for Backup solo funciona en regiones Azure emparejadas.

Almacenamiento con redundancia geográfica y restauración entre regiones

Para lograr la redundancia regional de los datos de copia de seguridad, use Backup para replicar las copias de seguridad en una región emparejada Azure mediante GRS. GRS protege las copias de seguridad frente a interrupciones regionales.

La región en la que se implementa el almacén se llama *región primaria*. Los orígenes de datos deben encontrarse en la región primaria. No se pueden configurar copias de seguridad en una bóveda de otra región.

La región emparejada también se conoce como *región secundaria*.



Si no configura GRS y se produce una interrupción en la región de la bóveda, es posible que pueda acceder a la bóveda y ver los elementos de copia de seguridad. Sin embargo, sin redundancia regional, los datos de copia de seguridad subyacentes permanecen no disponibles para las operaciones de restauración.

Restauración cruzada de regiones

Al configurar GRS en un almacén, Microsoft pone a disposición las copias de seguridad en la región emparejada después de que se produzca una interrupción en la región primaria. Si el origen de datos admite CRR, puede restaurar desde puntos de recuperación de regiones secundarias incluso cuando no se produzca ninguna interrupción en la región primaria. CRR también le permite ejecutar simulacros para evaluar la resistencia frente a interrupciones regionales. Al activar CRR, Microsoft actualiza el almacenamiento de copia de seguridad de GRS al almacenamiento con redundancia geográfica con acceso de lectura (RA-GRS).

Requisitos

- **Soporte de región:** GRS para Backup solo funciona dentro de regiones emparejadas de Azure.
- **Solo bóvedas nuevas:** Debe configurar GRS en su bóveda antes de realizar la primera copia de seguridad.

Consideraciones

- **CRR:** Después de activar CRR, los elementos de copia de seguridad pueden tardar hasta 48 horas en estar disponibles en la región secundaria.

Cost

Las bóvedas GRS incurren en costos adicionales para la replicación entre regiones y para el almacenamiento en la región secundaria. La transferencia de datos entre regiones de Azure se cobra en función de las tasas estándar de ancho de banda entre regiones. La CRR se cobra a una tarifa distinta porque Microsoft actualiza el almacenamiento del cofre de GRS a RA-GRS. Para obtener más información, consulte Precios de copia de seguridad .

Configuración de la compatibilidad con varias regiones

- **Cree un nuevo almacén que use GRS y CRR:** Al crear un almacén, también debe configurar la redundancia de almacenamiento. Después de seleccionar GRS, opcionalmente puede habilitar CRR en el almacén. Los pasos que siga dependen del tipo de bóveda. Para obtener más información, consulte los artículos siguientes:
  - Crear y eliminar almacenes de copias de seguridad
  - Creación y configuración de un almacén de Recovery Services
- **Configuración de GRS y CRR en almacenes existentes:** En el caso de los almacenes de Backup, debe configurar la redundancia de almacenamiento al crear el almacén.

En el caso de las bóvedas de Recovery Services, debe configurar la redundancia de almacenamiento antes de proteger cualquier carga de trabajo. Una vez protegida una carga de trabajo, la configuración está bloqueada y no se puede cambiar.

Puede habilitar CRR en almacenes GRS existentes. Después de habilitar CRR, no se puede deshabilitar.

Comportamiento cuando todas las regiones están en buen estado

En esta sección se describe qué esperar al configurar bóvedas para el uso de GRS cuando todas las regiones están operativas.

- **Operación entre regiones:** Las copias de seguridad siempre se completan en la región primaria, que es la región donde están desplegados el almacén y el origen de datos.
- **Replicación de datos entre regiones:** Cuando configuras el vault para usar GRS, las copias de seguridad se confirman primero en la región primaria mediante LRS. Después de finalizar correctamente en la región primaria, los datos se replican de forma asincrónica en la región secundaria. La región secundaria usa LRS para almacenar datos. Los datos de copia de seguridad pueden tardar hasta 12 horas en replicarse desde la región primaria a la región secundaria.

Comportamiento durante una falla de región

En esta sección se describe qué esperar al configurar bóvedas para usar GRS en caso de que ocurra una interrupción en la región primaria.

- **Detección y respuesta:** Para los orígenes de datos que admiten CRR y donde CRR está habilitado en la bóveda, puede iniciar su propio CRR en la región emparejada en cualquier momento, incluido durante una caída de la región o un desastre. Es responsable de detectar la interrupción y realizar acciones de recuperación, incluida la restauración de copias de seguridad en una región correcta.

  En todos los demás escenarios, los datos replicados en la región secundaria están disponibles para restaurarse en la región secundaria solo si Azure declara un desastre en la región primaria. Microsoft es responsable de declarar un desastre. La cantidad de tiempo que se tarda en declarar un desastre depende de la gravedad del incidente y del tiempo necesario para evaluar la situación. Microsoft normalmente declara un desastre solo después de un período de tiempo prolongado.

- **Notification**: Microsoft no le notifica automáticamente cuando una región está inactiva. Sin embargo:
  - Puede usar Azure Resource Health para supervisar el estado de un recurso individual y puede configurar Resource Health alertas para notificarle problemas.
  - Puede usar Azure Service Health para comprender el estado general del servicio, incluidos los errores de región, y puede configurar alertas de Service Health para notificarle los problemas.
- **Pérdida de datos esperada:** El RPO de los datos de copia de seguridad depende de varios factores, incluida la programación de copia de seguridad. En términos generales, para una interrupción regional, se puede esperar hasta 36 horas de pérdida de datos porque el RPO en la región primaria es de 24 horas, y puede tardar hasta 12 horas en replicar los datos de copia de seguridad de la región primaria a la región secundaria.
- **Tiempo de inactividad esperado:** El RTO es diferente para cada uno de los escenarios siguientes:
  - Es posible que los orígenes de datos y otros recursos de la región con errores no estén disponibles hasta que se recupere la región, por lo que el RTO no está definido.
  - Es posible que la copia de seguridad no pueda realizar operaciones de copia de seguridad o restauración en la región afectada hasta que la región se recupere, por lo que el RTO no está definido.
  - Si usa CRR, el RTO para iniciar la restauración de las copias de seguridad ya replicadas en la región emparejada es cero. Si no usa CRR, el RTO depende del tiempo que tarda Microsoft en declarar un desastre en la región afectada.
- **Redistribución:** No se pueden ejecutar trabajos de copia de seguridad mientras la región primaria está sin conexión. Puede restaurar datos en la bóveda, pero no puede agregar nuevos.

  Es responsable de restaurar la copia de seguridad en la infraestructura de la región emparejada y de volver a configurar equilibradores de carga, clientes y otros sistemas para redirigir el tráfico a una infraestructura correcta en la región emparejada.

Recuperación de regiones

Cuando se recupera la región primaria, Backup restaura automáticamente las operaciones en la región. Los trabajos se reanudan y los datos permanecen disponibles.

Prueba de fallos de región

Puede usar CRR para realizar una operación de restauración en la región emparejada. Puede usar este enfoque para comprobar la restauración y otros procesos de recuperación.

Resistencia a la pérdida de datos de copia de seguridad

La copia de seguridad proporciona dos características clave de recuperación para evitar la eliminación accidental o malintencionada de los datos de copia de seguridad:

- **La eliminación temporal** permite recuperar los objetos y almacenes eliminados durante un período de retención configurable. De forma predeterminada, este período es de 14 días, pero puede editarlo. Piensa en la eliminación suave como una papelera de reciclaje para tus copias de seguridad y bóvedas. Para obtener más información, consulte Seguro por defecto con la eliminación suave de Backup.

- **Los almacenes inmutables** pueden ayudarle a proteger los datos de copia de seguridad bloqueando las operaciones que podrían provocar la pérdida de puntos de recuperación. Puede bloquear la configuración del almacén inmutable para que sea irreversible. También puede usar el almacenamiento de escribir una vez, leer muchas veces (WORM) para los respaldos, para evitar que usuarios malintencionados deshabiliten la inmutabilidad y eliminen copias de seguridad. Para obtener más información, consulte almacén inmutable para copia de seguridad.

Acuerdo de nivel de servicio

El acuerdo de nivel de servicio (SLA) para Azure servicios describe la disponibilidad esperada de cada servicio y las condiciones que la solución debe cumplir para lograr esa expectativa de disponibilidad. Para obtener más información, vea SLAs for servicios en línea.

El Acuerdo de Nivel de Servicio de copia de seguridad cubre la disponibilidad del servicio para las operaciones de copia de seguridad y restauración. Para estar cubierto por el Acuerdo de Nivel de Servicio, debe reintentar los trabajos de copia de seguridad o restauración con errores al menos una vez cada 30 minutos.

Contenido relacionado

- Confiabilidad en Azure
- Reliability in Azure Virtual Machines
- Reliability in Azure Disk Storage
- Fiabilidad en Azure Site Recovery

ⓘ **Nota:** El autor ha creado este artículo con ayuda de inteligencia artificial. Más información

---

Last updated on 14/04/2026

Confiabilidad en Azure Chaos Studio

En este artículo se describe el soporte de confiabilidad ofrecido por Azure Chaos Studio. Trata cómo configurar zonas de disponibilidad y qué esperar durante una interrupción de toda la zona. Para obtener información general más detallada sobre la confiabilidad en Azure, consulte Azure confiabilidad.

Soporte para zonas de disponibilidad

Availability zones son grupos de centros de datos físicamente independientes dentro de una región de Azure. Cuando una zona falla, los servicios pueden transferirse a una de las zonas restantes.

Azure Chaos Studio admite la redundancia de zona como la configuración predeterminada dentro de una región. Los recursos de Chaos Studio se duplican o distribuyen automáticamente entre diferentes zonas.

Prerrequisitos

Las siguientes regiones admiten zonas de disponibilidad para Chaos Studio:

Expandir tabla

Americas	Europa	Asia Pacífico
Sur de Brasil	Centro de Francia	Australia East
Canada Central	Centro-oeste de Alemania	Japón Oriental
Central US	Norte de Italia	Sudeste Asiático
East US	Norte de Europa	Norte de Emiratos Árabes Unidos
Este de EE. UU. 2	Centro de Suecia	
Oeste de EE. UU. 2	UK South	
Oeste de EE. UU. 3	Oeste de Europa	

Para obtener información detallada sobre el modelo de disponibilidad regional para Azure Chaos Studio vea Disponibilidad regional de Azure Chaos Studio.

Experiencia de reducción de zona

En caso de una interrupción en toda la zona, debe prever una breve degradación del rendimiento y la disponibilidad a medida que el servicio pasa a una zona funcional. Esta interrupción no depende de la restauración de la zona afectada, ya que los servicios administrados por Microsoft mitigan las pérdidas de zona mediante el uso de la capacidad de las zonas alternativas. En caso de interrupción de una zona de disponibilidad, es posible que un experimento de caos encuentre errores o interrupciones, pero los metadatos cruciales del experimento, los datos históricos y los detalles específicos deben permanecer accesibles y el servicio no debe experimentar una interrupción completa.

Recuperación ante desastres entre regiones y continuidad empresarial

Chaos Studio es un servicio de una sola región y no admite la conmutación por error entre regiones habilitada por el servicio.

Pasos siguientes

- Crear y ejecutar el primer experimento.
- Más información sobre la ingeniería de caos.
- Confiabilidad en Azure

Last updated on 14/04/2026

Confiabilidad en Azure Managed Grafana

Azure Managed Grafana proporciona áreas de trabajo de Grafana hospedadas para crear paneles y visualizaciones. Microsoft administra toda la infraestructura subyacente, incluidas las actualizaciones de proceso, redes, almacenamiento y servicio.

Al usar Azure, relibilidad es una responsabilidad compartida. Microsoft proporciona una variedad de capacidades para admitir resiliencia y recuperación. Es responsable de comprender cómo funcionan esas funcionalidades dentro de todos los servicios que usa y de seleccionar las funcionalidades que necesita para cumplir los objetivos empresariales y los objetivos de tiempo de actividad.

En este artículo se describe cómo hacer que Azure Managed Grafana sean resistentes a una variedad de posibles interrupciones y problemas, incluidos errores transitorios, interrupciones de zona de disponibilidad y interrupciones de regiones. También describe cómo puede realizar copias de seguridad y recuperarse de otros tipos de problemas y resalta la información clave sobre el acuerdo de nivel de servicio (SLA) de Azure Managed Grafana.

Recomendaciones de implementación de producción para la confiabilidad

Para aumentar la confiabilidad de las implementaciones de producción mediante Azure Managed Grafana, se recomienda realizar las siguientes acciones:

- ✓ **Habilite la redundancia de zona** al crear un área de trabajo para proporcionar resistencia a los errores de zona de disponibilidad.
- ✓ **Almacena tableros y otros recursos de Grafana como código**, por ejemplo, exportándolos desde la API de Grafana o la CLI y guardándolos en un repositorio de control de versiones como GitHub. Utilice procesos de integración continua y entrega continua (CI/CD) para desplegar paneles en Azure Managed Grafana. Este enfoque admite escenarios de recuperación. También permite la implementación en varias instancias de Grafana, incluidas las instancias de diferentes regiones de Azure si es necesario.

Introducción a la arquitectura de confiabilidad

En esta sección se describen algunos de los aspectos importantes de cómo funciona el servicio que es más relevante desde una perspectiva de confiabilidad. En la sección se presenta la arquitectura lógica, que incluye algunos de los recursos y características que se implementan y usan. También se describe la arquitectura física, que proporciona detalles sobre cómo funciona el servicio en segundo plano.

Arquitectura lógica

El recurso Azure principal que implemente es un *workspace*. Después de implementar el área de trabajo, use el punto de conexión de Grafana del área de trabajo para configurar e interactuar con orígenes de datos, paneles, visualizaciones y otros recursos de Grafana.

Arquitectura física

Al crear un área de trabajo, internamente, la plataforma Azure aprovisiona los siguientes componentes subyacentes:

- **Servidores de Grafana:** Máquinas virtuales dedicadas que ejecutan la aplicación Grafana. De forma predeterminada, se aprovisionan dos servidores para alta disponibilidad y redundancia. Microsoft administra completamente estos servidores. No los ve en su suscripción, no puede acceder a ellos y no es responsable de aplicar actualizaciones, escalarlos ni mantenerlos.

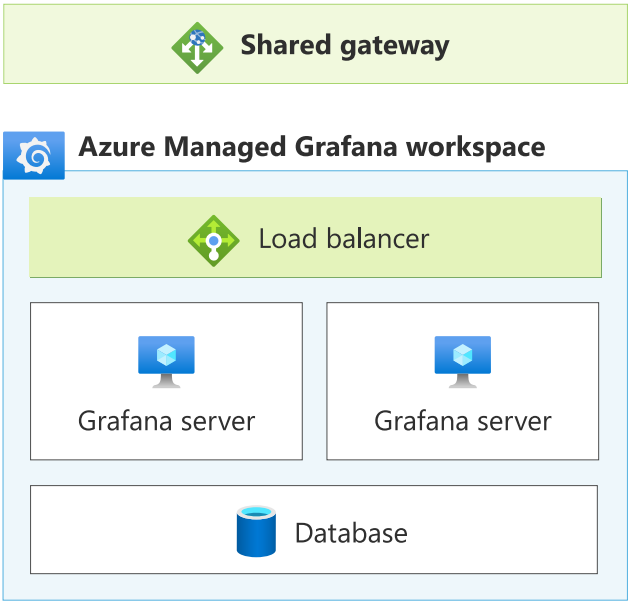
- **Equilibrador de carga:** Equilibrador de carga de red que distribuye las solicitudes entrantes del explorador entre los servidores de Grafana. El equilibrador de carga supervisa el estado del servidor y enruta automáticamente el tráfico fuera de los servidores incorrectos.

- **Backend database:** Una base de datos de Azure Database for PostgreSQL que almacena la configuración del área de trabajo y otros datos persistentes. Todos los servidores de Grafana del área de trabajo comparten esta base de datos. Para obtener más información sobre la resistencia de la base de datos, consulte Reliability in Azure Database for PostgreSQL.

Usted no ve ni administra estos componentes. Microsoft implementa y administra en su nombre.

El equilibrador de carga realiza un seguimiento de qué servidores de Grafana están disponibles. En una configuración de servidor dual, si un servidor se vuelve no operativo, el balanceador de carga envía todas las solicitudes al servidor restante. Ese servidor recoge las

sesiones del explorador que el servidor con errores controló anteriormente, en función de la información de la base de datos compartida. Mientras tanto, Azure Managed Grafana repara o reemplaza el servidor defectuoso.



Resistencia a errores transitorios

Los errores transitorios son errores breves e intermitentes en los componentes. Se producen con frecuencia en un entorno distribuido como la nube y son una parte normal de las operaciones. Los errores transitorios se corrigen después de un breve período de tiempo. Es importante que las aplicaciones puedan controlar errores transitorios, normalmente mediante el reintento de solicitudes afectadas.

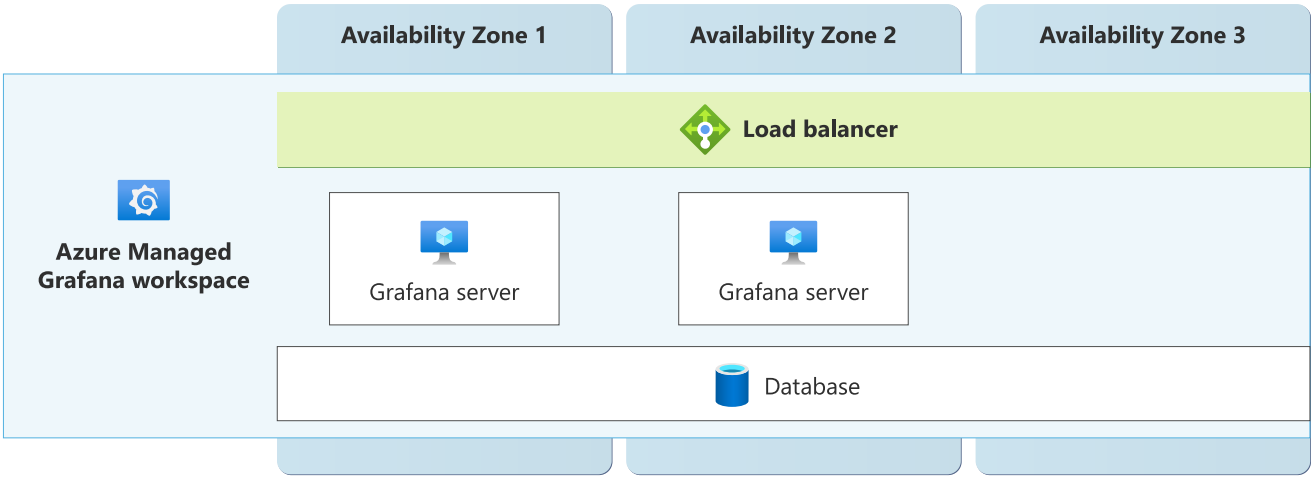
Todas las aplicaciones hospedadas en la nube deben seguir las instrucciones de control de errores transitorios Azure cuando se comunican con cualquier API, bases de datos y otros componentes hospedados en la nube. Para obtener más información, consulte Recomendaciones para controlar errores transitorios.

Puede compilar aplicaciones cliente para interactuar con el área de trabajo de Grafana a través de la API de Grafana. Asegúrese de que esas aplicaciones sigan las pautas de reintento de Azure para las solicitudes fallidas.

Resistencia a errores de zona de disponibilidad

Availability zones son grupos de centros de datos físicamente independientes dentro de una región de Azure. Cuando una zona falla, los servicios pueden transferirse a una de las zonas restantes.

Azure Managed Grafana áreas de trabajo admiten redundancia de zona en regiones de Azure admitidas. Cuando se habilita la redundancia de zona, los servidores Grafana del área de trabajo se distribuyen entre varias zonas de disponibilidad. Microsoft selecciona las zonas que utiliza el entorno de trabajo. Otros recursos, como el equilibrador de carga de red, la base de datos y la puerta de enlace compartida, también están configurados para usar varias zonas de disponibilidad.



Si no habilita la redundancia de zona, el área de trabajo es *no zonal* o *regional*, lo que significa que los servidores y otros componentes pueden colocarse en cualquier zona de disponibilidad dentro de la región o dentro de la misma zona. Si alguna zona de disponibilidad de la región tiene un problema, el área de trabajo podría experimentar tiempo de inactividad.

Requisitos

Compatibilidad con regiones: La compatibilidad con redundancia de zona está disponible en las siguientes regiones.

Expandir tabla

Americas	Europa	Asia Pacífico
East US	Norte de Europa	Australia East
Centro-sur de EE. UU.		Este de Asia
Oeste de EE. UU. 3		

Cost

La redundancia de zona agrega un costo adicional. Para obtener más información, consulte precios Azure Managed Grafana .

Configurar soporte de zonas de disponibilidad

- **Crear un área de trabajo con zonas de disponibilidad habilitadas:** Habilitar la redundancia de zona durante la creación del área de trabajo a través del portal de Azure, las plantillas de CLI de Azure, Bicep o Azure Resource Manager (plantillas de ARM).

  Para obtener más información, consulte Habilitar redundancia de zona en Azure Managed Grafana.

- **Configuración de la redundancia de zona en un área de trabajo existente:** No se puede habilitar ni deshabilitar la redundancia de zona en un área de trabajo existente. En su lugar, debe crear un área de trabajo que use la configuración de redundancia de zona deseada, migrar los paneles y la configuración y, a continuación, eliminar el área de trabajo existente.

Comportamiento cuando todas las zonas están en buen estado

En esta sección se describe qué esperar al configurar un área de trabajo con redundancia de zona y todas las zonas de disponibilidad están operativas.

- **Enrutamiento de tráfico entre zonas:** El equilibrador de carga con redundancia de zona distribuye automáticamente las solicitudes entrantes entre los servidores de Grafana. Ambos servidores pueden procesar el tráfico.

- **Replicación de datos entre zonas:** Los cambios en los datos del área de trabajo se replican sincrónicamente en varias zonas de disponibilidad. Azure Database for PostgreSQL realiza la replicación de datos. Para obtener más información, consulte Reliability en Azure Database for PostgreSQL. Azure Managed Grafana no implementa lógica de replicación personalizada adicional más allá de lo que proporciona la plataforma de base de datos.

Comportamiento durante un fallo de zona

En esta sección se describe qué esperar al configurar un área de trabajo con redundancia zonal, cuando hay una interrupción en una de las zonas.

- **Detección y respuesta:** La plataforma Azure detecta y responde a un error en una zona de disponibilidad. No es necesario iniciar una conmutación por error de zona.

- **Notification**: Microsoft no le notifica automáticamente cuando una zona está inactiva. Sin embargo, puede usar Azure Resource Health para supervisar el estado de un recurso individual y puede configurar Resource Health alertas para notificarle problemas. También puede usar Azure Service Health para comprender el estado general del servicio, incluidos los errores de zona, y puede configurar alertas de Service Health para notificarle problemas.

- **Pérdida de datos esperada:** No se espera ninguna pérdida de datos durante una interrupción de la zona de disponibilidad.

- **Tiempo de inactividad esperado:** El área de trabajo puede experimentar una pequeña cantidad de tiempo de inactividad, normalmente limitado a unos segundos, mientras que el tráfico se redirige a servidores en buen estado. Asegúrese de que las aplicaciones cliente pueden controlar correctamente los errores transitorios para minimizar los efectos del tiempo de inactividad.

- **Reenrutamiento del tráfico:** El tráfico entrante se enruta automáticamente al servidor en la zona correcta. El servicio funciona con capacidad reducida durante la interrupción de la zona. No se aprovisionan servidores de reemplazo en zonas saludables durante la interrupción.

Recuperación de zona

Microsoft administra automáticamente la recuperación de zona, incluida la restauración de la capacidad del servicio cuando la zona afectada vuelve a estar en buen estado.

Prueba de fallos de zona

La plataforma Azure administra el enrutamiento del tráfico, la conmutación por error y la conmutación por recuperación para espacios de trabajo con redundancia zonal. Esta característica está totalmente administrada, por lo que no es necesario iniciar ni validar los procesos de error de zona de disponibilidad.

Resistencia a errores en toda la región

Azure Managed Grafana es un servicio de una sola región. Si la región no está disponible, el área de trabajo tampoco está disponible.

Soluciones personalizadas de varias regiones para la resistencia

Para lograr resistencia a interrupciones regionales, puede implementar varias áreas de trabajo de Grafana en diferentes regiones. En este tipo de solución, es responsable de:

- Replicación de paneles y configuración entre regiones. Por ejemplo, puede aplicar una configuración coherente en varias áreas de trabajo mediante CI/CD y el control de código fuente.
- Implementación del enrutamiento de tráfico y conmutación por error a nivel de aplicación o de cliente.

Copias de seguridad y restauración

Azure Managed Grafana no proporciona funcionalidad integrada de copia de seguridad ni restauración para paneles u otras entidades del plano de datos. Para protegerse contra la eliminación accidental o corrupción:

- Use la API de Grafana o la CLI para exportar paneles y otra configuración de Grafana.
- Almacene paneles exportados en un repositorio de control de código fuente, como GitHub.
- Use la automatización o las canalizaciones de CI/CD para volver a implementar paneles y otra configuración de Grafana.

Para la mayoría de las soluciones, no debe confiar exclusivamente en copias de seguridad. En su lugar, utilice las otras capacidades descritas en esta guía para apoyar los requisitos de resiliencia. Sin embargo, las copias de seguridad protegen contra algunos riesgos que otros enfoques no. Para más información, consulte ¿Qué son la redundancia, la replicación y la copia de seguridad?.

Resistencia al mantenimiento del servicio

Microsoft aplica periódicamente actualizaciones de servicio y realiza otro mantenimiento. La plataforma Azure controla estas actividades automáticamente, lo que garantiza que el mantenimiento sea transparente y sin problemas. Durante las operaciones de mantenimiento, es posible que observe breves interrupciones. Normalmente, estas interrupciones duran unos segundos. Asegúrese de que las aplicaciones cliente están configuradas para controlar errores transitorios para que sean resistentes a interrupciones breves.

Acuerdo de nivel de servicio

El acuerdo de nivel de servicio (SLA) para Azure servicios describe la disponibilidad esperada de cada servicio y las condiciones que la solución debe cumplir para lograr esa expectativa de disponibilidad. Para obtener más información, vea SLAs for servicios en línea.

Contenido relacionado

- Confiabilidad en Azure
- Información general sobre Azure Managed Grafana

ⓘ **Nota:** El autor ha creado este artículo con ayuda de inteligencia artificial. Más información

---

Last updated on 15/04/2026

Confiabilidad en registros de Azure Monitor

Azure Monitor Logs es una plataforma centralizada de software como servicio (SaaS) para recopilar, analizar y tomar medidas en función de los datos generados por el sistema procedentes de recursos y aplicaciones de Azure y no Azure.

Al usar Azure, relibilidad es una responsabilidad compartida. Microsoft proporciona una variedad de capacidades para admitir resiliencia y recuperación. Es responsable de comprender cómo funcionan esas funcionalidades dentro de todos los servicios que usa y de seleccionar las funcionalidades que necesita para cumplir los objetivos empresariales y los objetivos de tiempo de actividad.

Azure Monitor Registros proporciona características de resistencia integradas que protegen los datos y minimizan las interrupciones. En este artículo se describe cómo puede usar esas características para que las áreas de trabajo de Log Analytics sean resistentes a varias posibles interrupciones y problemas, incluidos errores transitorios, interrupciones de zona de disponibilidad y interrupciones de regiones.

Recomendaciones de implementación de producción

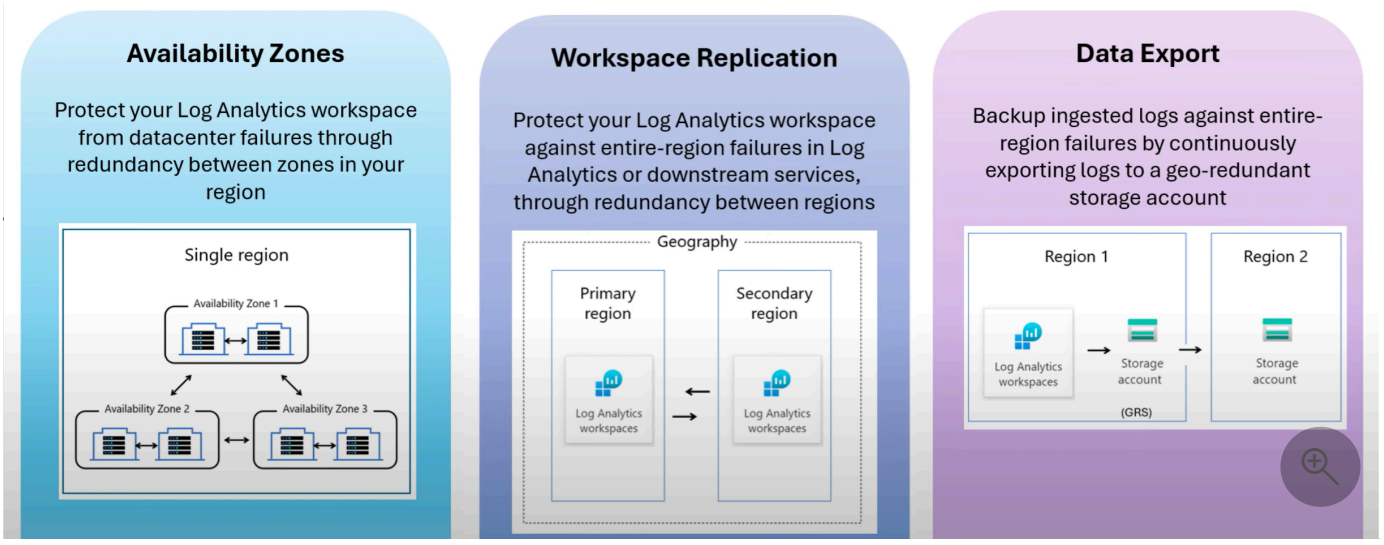
Azure Monitor Logs ofrece varias características que puede usar individualmente o en combinación para ayudar a mejorar la resistencia del espacio de trabajo frente a diversos tipos de problemas.

Expandir tabla

Capacidad	Descripción	Costo
Zonas de disponibilidad	Use redundancia entre zonas de su región para proteger el área de trabajo de Log Analytics frente a errores del centro de datos.	Sin cargo adicional. Incluido en el precio estándar del espacio de trabajo. Los clústeres dedicados deben cumplir los requisitos mínimos del nivel de compromiso.
Replicación del área de trabajo	Use la redundancia entre regiones para proteger el área de trabajo de Log Analytics contra fallos de toda una región en Log Analytics o en servicios dependientes.	Todos los datos ingeridos se replican, lo que agrega cargos de replicación. Para reducir los costos, elija las reglas de recopilación de datos (DCR) que participan en la replicación.

Capacidad	Descripción	Costo
Exportación de datos	Realice una copia de seguridad de los registros ingeridos y expórtelos continuamente al almacenamiento con redundancia geográfica para protegerse frente a errores de toda la región.	Los costos de almacenamiento varían según el volumen y el nivel de redundancia (GRS). No se aplica ningún cargo adicional de Azure Monitor por la exportación en sí.

Para obtener información detallada sobre los precios, consulte Azure Monitor precios.



Algunas características de Log Analytics no son compatibles con todas las características de confiabilidad. Por ejemplo, la replicación del área de trabajo no admite tablas auxiliares. Para más información sobre la compatibilidad de características, consulte Reliability best practices for Log Analytics.

Introducción a la arquitectura de confiabilidad

En esta sección se describen algunos de los aspectos importantes de cómo funciona el servicio que es más relevante desde una perspectiva de confiabilidad. En la sección se presenta la arquitectura lógica, que incluye algunos de los recursos y características que se implementan y usan. También se describe la arquitectura física, que proporciona detalles sobre cómo funciona el servicio en segundo plano.

Arquitectura lógica

Cuando se usan registros de Azure Monitor, se implementa un *área de trabajo de Log Analytics*, que es un almacén de datos que puede recopilar cualquier tipo de datos de registro.

Log Analytics áreas de trabajo están asociadas a un *cluster*, que proporciona los recursos computacionales para el área de trabajo y otras funcionalidades. La mayoría de las áreas de

trabajo se ejecutan en un *clúster compartido*, que varios clientes usan al mismo tiempo. Microsoft implementa y administra clústeres compartidos. Opcionalmente, puede implementar un clúster dedicado, que proporciona recursos dedicados para el área de trabajo. Los clústeres compartidos y dedicados tienen diferentes características de confiabilidad. Las diferencias se describen en este artículo.

Azure Monitor Registros tiene dos rutas de acceso de datos distintas, cada una con sus propias características de confiabilidad:

- **Ingestion** es la ruta de acceso a través de la cual fluyen los datos de registro a un área de trabajo de Log Analytics. El servicio Azure Monitor confirma la recepción de datos solo después de escribirlos en el almacenamiento de forma duradera. La ruta de acceso de ingestión incluye varios componentes:
  - *Los orígenes de datos* generan telemetría. Algunas fuentes, como los registros de plataforma de Azure, envían datos directamente al servicio Azure Monitor a través de *la configuración de diagnóstico*. Otros orígenes requieren que un *agente* recopile y reenvíe datos. Las aplicaciones también pueden usar la API de ingesta de Logs para enviar registros directamente a Azure Monitor.
  - *Las reglas de recopilación de datos (DCR)* definen qué datos se van a recopilar, cómo transformarlos y dónde enviarlos. Las DCR enrutan los datos al servicio Azure Monitor a través de un punto de conexión de recopilación de datos (DCE).
- **La consulta** es la ruta de acceso a través de la que se recuperan y analizan los datos almacenados en el área de trabajo. Las consultas de registro usan el lenguaje de consulta kusto (KQL) y se ejecutan en los datos almacenados del área de trabajo.

La ingesta y la consulta se controlan como operaciones de servicio distintas, por lo que una interrupción de una no afecta necesariamente a la otra. Por ejemplo, durante una degradación del proceso de ingesta, es posible que todavía pueda consultar los datos ingeridos previamente. Del mismo modo, un problema del lado de la consulta no impide que se ingieren y almacenan nuevos datos.

Para obtener más información sobre los componentes principales de los registros de Azure Monitor, consulte Azure Monitor Logs overview.

Arquitectura física

Internamente, los registros de Azure Monitor usan varias *réplicas* de los componentes computacionales y de almacenamiento de los datos de registro. Microsoft administra estas réplicas y no es necesario configurarlas ni administrarlas.

Es responsable de implementar y administrar agentes, así como de la confiabilidad de los recursos de computación en los que se ejecutan.

Resistencia a errores transitorios

Los errores transitorios son errores breves e intermitentes en los componentes. Se producen con frecuencia en un entorno distribuido como la nube y son una parte normal de las operaciones. Los errores transitorios se corrigen después de un breve período de tiempo. Es importante que las aplicaciones puedan controlar errores transitorios, normalmente mediante el reintento de solicitudes afectadas.

Todas las aplicaciones hospedadas en la nube deben seguir las instrucciones de control de errores transitorios Azure cuando se comunican con cualquier API, bases de datos y otros componentes hospedados en la nube. Para obtener más información, consulte Recomendaciones para controlar errores transitorios.

Tenga en cuenta los siguientes tipos de errores transitorios en los registros de Azure Monitor:

- **Fallos transitorios en el servicio Azure Monitor:** El servicio Azure Monitor verifica que cada entrada de registro se ingesta correctamente en el área de trabajo antes de eliminarla del proceso de ingesta. Si la ingesta no está disponible o limita las solicitudes, la configuración de diagnóstico y los agentes de Azure Monitor almacenan temporalmente los datos en un búfer local y vuelven a intentar enviarlos durante varias horas, con retroceso exponencial para evitar sobrecargar el servicio. Por el contrario, una aplicación personalizada que envía solicitudes de ingesta o consultas debe implementar su propia lógica de reintento.

- **Fallos transitorios que afectan a la conectividad de ingestión:** Los agentes y la configuración de diagnóstico de Azure almacenan los datos en búfer localmente y reintentan la entrega cuando el punto de conexión de ingestión no está disponible temporalmente, lo que hace que la ruta de acceso de ingestión sea más resistente a los fallos transitorios.

- **Errores transitorios que afectan al registro de aplicaciones personalizado:** Para validar que la aplicación personalizada controla correctamente los errores transitorios, supervise estas métricas:
  - Latencia de ingesta

    - Recuento de ingestión fallida
    - Errores de exportación
    - Tasa de errores de consulta

    La latencia de ingesta sostenida que dura más de cinco minutos podría indicar un problema más significativo y notransciente. Revise las instrucciones de las otras secciones de este documento para comprender cómo ser resistentes a otros tipos de errores.

- **Errores transitorios durante las consultas y otras operaciones:** Si se producen errores transitorios durante las consultas o al realizar otras operaciones, los clientes son responsables de reintentar.

Resistencia a errores de zona de disponibilidad

Availability zones son grupos de centros de datos físicamente independientes dentro de una región de Azure. Cuando una zona falla, los servicios pueden transferirse a una de las zonas restantes.

Azure Monitor Registros ofrece dos tipos de compatibilidad con zonas de disponibilidad, en función de la región y el tipo de clúster del área de trabajo:

- **La resistencia de los datos** proporciona redundancia de zona para los datos de registro mediante la replicación en varias zonas de disponibilidad.

  La resistencia de los datos está disponible en todas las regiones que admiten zonas de disponibilidad. La mayoría de las regiones que admiten zonas de disponibilidad requieren que el área de trabajo se implemente en un clúster dedicado. Sin embargo, algunas regiones lo admiten mediante la configuración predeterminada del área de trabajo de un clúster compartido. Al pasar a un clúster dedicado en una región que admite zonas de disponibilidad, se protegen los datos ingeridos después del traslado, no los datos históricos.

- **La resistencia del servicio** proporciona redundancia de zona para la ingesta y la continuidad de las consultas durante una interrupción de la zona de disponibilidad.

  Solo algunas regiones ofrecen resiliencia del servicio.

Si un incidente afecta a una zona, Microsoft administra la conmutación por error a otra zona de disponibilidad de la región automáticamente. No es necesario realizar ninguna acción, ya que el cambio de zona se realiza de forma sencilla.

Requisitos

- **Compatibilidad con regiones:** Consulte las regiones admitidas de la zona de disponibilidad para obtener la lista de regiones que admiten la resistencia de datos, la resistencia del servicio y la resistencia de datos en clústeres compartidos y dedicados.
- **Clúster dedicado:** Algunas regiones requieren un clúster dedicado para la resistencia de los datos.

Consideraciones

- **Ingesta de Event Hubs:** En algunas regiones, la ingesta de Event Hubs en el espacio de trabajo no es resiliente a una interrupción de zona. Para obtener una lista de estas regiones, consulte las notas al pie en la lista de regiones admitidas. Evalúe vías de ingesta alternativas para los datos críticos.
- **Traslado a clústeres dedicados:** Los clústeres dedicados solo protegen los nuevos datos. Cuando un área de trabajo se mueve a un clúster dedicado, los datos ingeridos previamente permanecen en el clúster compartido. Todo está disponible en operaciones normales. Sin embargo, durante una interrupción de zona, es posible que solo pueda acceder a los nuevos datos del clúster dedicado.

Costo

La redundancia de zona no afecta a los precios. Las áreas de trabajo en clústeres compartidos se cobran según el precio estándar del área de trabajo y las áreas de trabajo de los clústeres dedicados deben cumplir el nivel de compromiso del clúster. Ninguno de los dos modelos de precios cambia al habilitar la redundancia de zona. Para obtener más información sobre los precios, consulte los precios de Azure Monitor .

Configurar soporte de zonas de disponibilidad

En el caso de las áreas de trabajo que usan clústeres compartidos, al crear un área de trabajo en una región que admita zonas de disponibilidad para registros de Azure Monitor, el servicio habilita automáticamente la redundancia de zona.

Si la región requiere un clúster dedicado para la compatibilidad con zonas, implemente el clúster dedicado y, a continuación, vincule el área de trabajo al clúster dedicado.

Planeamiento y administración de capacidad

Durante una falla de zona, el área de trabajo podría experimentar una mayor carga debido a los reintentos de ingestión y consulta. Si la solución es sensible a los retrasos en la ingesta de registros o las consultas, incluso durante y después de un error de zona, considere la posibilidad de sobreaprovisionar la capacidad de los límites diarios que configure en el área de trabajo. En el caso de los clústeres dedicados, seleccione un nivel de compromiso que tenga espacio para los picos de ingesta temporales. Para más información, consulte Administración de la capacidad mediante el sobreaprovisionamiento.

Comportamiento cuando todas las zonas están en buen estado

En esta sección se describe lo que cabe esperar cuando un área de trabajo de Log Analytics tiene redundancia de zona y todas las zonas están operativas.

- **Operación entre zonas:** Los clústeres que tienen resiliencia de servicio permiten que la ingesta y las consultas usen la infraestructura en cualquier zona. El servicio distribuye automáticamente el trabajo entre zonas, optimizando la localidad y la carga.
- **Replicación de datos entre zonas:** Para los clústeres que tienen resistencia de datos, el servicio Azure Monitor confirma todas las escrituras en varias réplicas en zonas independientes antes de confirmar.

Comportamiento durante un fallo de zona

En esta sección se describe qué cabe esperar cuando un área de trabajo de Log Analytics tiene redundancia de zona y se produce una interrupción en una de las zonas.

- **Detección y respuesta:** La plataforma detecta un error de zona y responde automáticamente. En el caso de los clústeres que tienen resistencia de servicio, la plataforma reasigna la ingesta y el proceso de consulta a zonas supervivientes. En el caso de los clústeres que tienen resiliencia de datos, la plataforma cambia para usar réplicas en zonas saludables. No es necesario iniciar una conmutación por error de zona.
- **Notification:** Microsoft no le notifica automáticamente cuando una zona está inactiva. Sin embargo, puede usar Azure Service Health para comprender el estado general del servicio, incluidos los errores de zona, y puede configurar alertas de Service Health para notificarle problemas.

  También puede supervisar las métricas de salud del área de trabajo y buscar una latencia de ingesta sostenida o una tasa elevada de fallos en las consultas. Puede configurar alertas en estas métricas.

- **Solicitudes activas:** Durante una interrupción de zona, es posible que se produzca un error en las solicitudes activas.

  Es posible que se finalicen las consultas activas. Los clientes pueden reintentar las consultas.

  En los clústeres que no tienen resistencia del servicio, es posible que Azure Monitor deje de procesar los datos no confirmados. Los agentes o las aplicaciones cliente pueden volver a intentarlo cuando la zona está en buen estado.

- **Pérdida de datos esperada:** En el caso de los clústeres que tienen resistencia a los datos, no se espera ninguna pérdida de datos para los datos totalmente confirmados. Las aplicaciones cliente o agente deben reintentar lotes no reconocidos.

- **Tiempo de inactividad esperado:** En el caso de los clústeres que tienen resistencia de servicio, no se espera ningún tiempo de inactividad. Es posible que se produzcan incrementos temporales de latencia o restricciones mientras el servicio reequilibra la infraestructura a través de zonas.

- **Rerouting:** Para los clústeres que tienen resiliencia del servicio, los equilibradores de carga internos dentro del servicio Azure Monitor Logs redirigen el tráfico a nodos de zona saludables. No se requiere ningún cambio de punto de conexión.

Recuperación de zona

Cuando se recupera una zona de disponibilidad, Azure Monitor Logs reintegra automáticamente la zona en la topología de servicio activa. La zona recuperada comienza a procesar las solicitudes de ingesta y consulta en paralelo con las demás zonas. Los datos que se replican en zonas supervivientes durante la interrupción permanecen intactos y la replicación sincrónica normal se reanuda en todas las zonas. No es necesario tomar medidas para la recuperación y la reintegración de zonas.

Prueba de fallos de zona

Azure administra el enrutamiento del tráfico, la conmutación por error y la recuperación de zona en caso de fallos, por lo que no es necesario validar los procesos de fallo de la zona de disponibilidad ni proporcionar más datos.

Resistencia a errores en toda la región

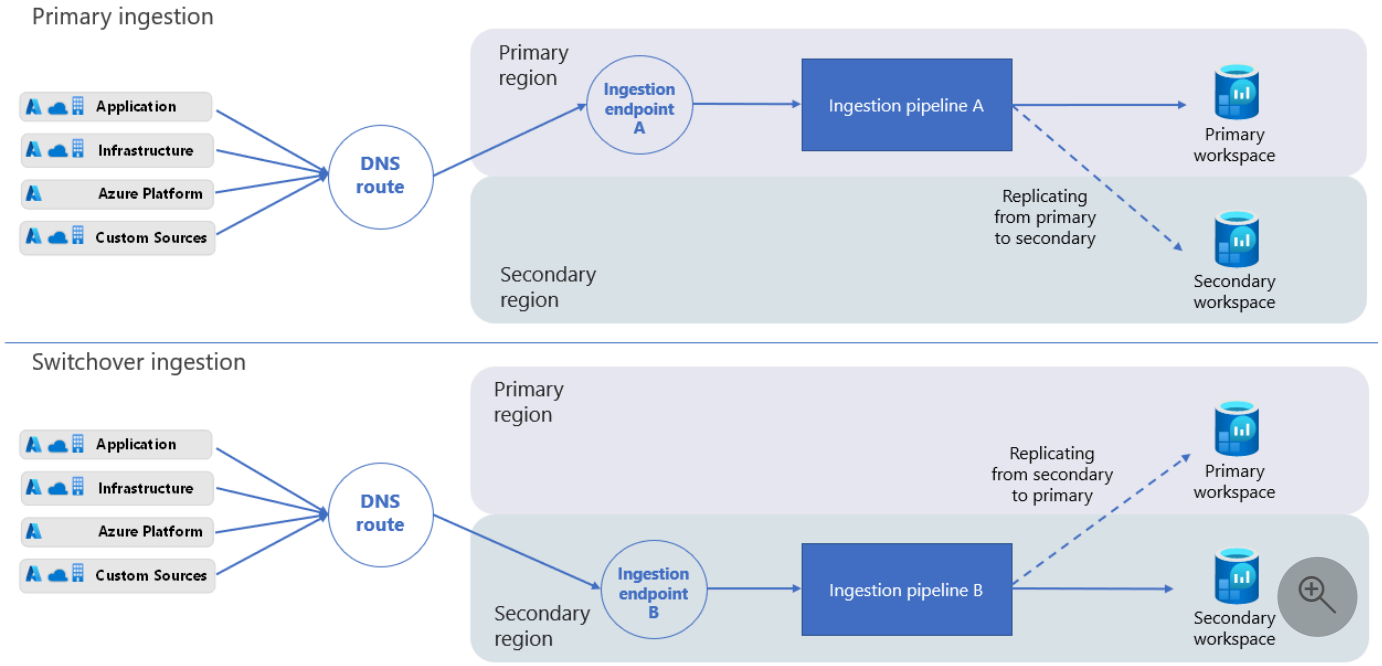
Azure Monitor Registros admite escenarios de varias regiones a través de la replicación del área de trabajo.

Replicación del área de trabajo

La replicación del área de trabajo crea un área de trabajo secundaria en una región secundaria admitida y, a continuación, replica de forma asincrónica los nuevos registros, así como los cambios de esquema y configuración. Los datos que existían antes de configurar la replicación no se rellenan.

No ve el área de trabajo secundaria como un recurso independiente y administrable. Puede acceder a él a través del mismo punto de conexión de recursos del área de trabajo.

Si se produce un error en la región primaria, debe desencadenar una conmutación por medio de una API. La conmutación redirige la ingesta y las consultas al Área de trabajo secundaria. No hay conmutación por error automática. La conmutación es una decisión manual. Debe iniciarse una conmutación de retorno por parte del cliente para restaurar el sistema principal.



En esta sección se resumen aspectos importantes de la replicación del área de trabajo. Revise la documentación completa para comprender exactamente cómo funciona. Para obtener más información, consulte Mejore la resiliencia mediante la replicación del espacio de trabajo de Log Analytics entre regiones.

> ⓘ **Nota:**

La replicación de registros del área de trabajo de Azure Monitor utiliza el término *switchover* porque representa mejor el proceso de conmutación manual al área de trabajo secundaria. También podría ver el término *conmutación por error* utilizado para describir el proceso general.

Requisitos

- **Compatibilidad con regiones:** La replicación del área de trabajo requiere seleccionar una región secundaria dentro del mismo *grupo de regiones*. Algunas combinaciones de regiones no pueden vincularse directamente entre ellas. Para ver los pares primarios y secundarios permitidos y los grupos de regiones, consulte Regiones admitidas. Si la región del área de trabajo no aparece en la lista, no admite actualmente la replicación.

- **Clústeres dedicados:** Si el área de trabajo está vinculada a un clúster dedicado, habilite primero la replicación en el clúster.

Consideraciones

- **Reglas de alertas de búsqueda de registros:** Después de cambiar entre regiones, las reglas de alertas de búsqueda de registros siguen funcionando a menos que el servicio de alertas de la región activa no funcione correctamente o las reglas de alerta no estén disponibles. Esta condición puede ocurrir, por ejemplo, si la región en la que se crearon las reglas de alerta está completamente inactiva. La replicación de reglas de alerta entre regiones no se realiza automáticamente como parte de la replicación del área de trabajo, pero puede replicarlas. Por ejemplo, puede exportar reglas de alertas de la región primaria a la región secundaria.

- **Cambio manual y cambio de vuelta:** Usted es responsable de decidir cuándo cambiar y volver atrás, y de desencadenar las acciones de cambio y cambio de vuelta.

  Para prepararse para un fallo de región, debe:
  - Establezca la monitorización (estado de replicación, latencia de ingesta, métricas de salud del área de trabajo) para la región activa. Audite la región del área de trabajo inactiva antes de una conmutación.
  - Considere si se deben crear runbooks para definir criterios de conmutación, que podrían formar parte del planeamiento de la recuperación ante desastres. Por ejemplo, podría supervisar la latencia de ingesta de toda la región, los incumplimientos del objetivo de nivel de servicio (SLO), las tasas sostenidas de errores en las consultas o los avisos de Azure Service Health. Pruebe los runbooks (guías de ejecución) con regularidad.

- Documente los criterios de conmutación y retorno.
- Compruebe que los usuarios o entidades de servicio que desencadenen la conmutación tienen los permisos de rol necesarios.

- **Tablas auxiliares:** Las tablas auxiliares no se replican. Como procedimiento recomendado, no habilite la replicación en áreas de trabajo que incluyan tablas auxiliares.

- **Microsoft Sentinel:** Las listas de seguimiento y las entradas de inteligencia sobre amenazas de Microsoft Sentinel solo se replican cuando se actualizan, por lo que pueden tardar hasta 12 días en lograr una paridad completa después de la activación inicial.

- **Operaciones de purga:** Las operaciones de purga quitan registros de las áreas de trabajo principales y secundarias. Si alguna de las áreas de trabajo no está disponible, se produce un error en la purga. En este caso, debe planificar los plazos de cumplimiento en consecuencia.

Para más consideraciones, consulte Consideraciones de implementación.

Costo

Al habilitar la replicación del área de trabajo, paga por la replicación de todos los datos que ingiere en el área de trabajo. Las DCR asociadas a la replicación determinan directamente los costos de replicación. Para obtener más información sobre los precios, consulte los precios de Azure Monitor.

Configuración de la compatibilidad con varias regiones

- **Habilitar la replicación del área de trabajo:** Los pasos generales necesarios para habilitar la replicación del área de trabajo son:

  1. Habilite la replicación y especifique la región secundaria.

  2. Valide el estado de aprovisionamiento.

> **Sugerencia**
>
> Incluso cuando el estado de aprovisionamiento es *Correcto*, los esquemas de tabla podrían seguir replicándose. Para comprobar cuándo se replican completamente los esquemas de tabla, **audite el área de trabajo secundaria**.

3. Asocie los DCR con el punto de conexión de recopilación de datos del área de trabajo (DCE) después de habilitar la replicación de esos flujos durante la conmutación.

- **Deshabilitar la replicación del área de trabajo:** Para obtener pasos detallados para deshabilitar la replicación del área de trabajo, consulte Habilitación y deshabilitación de la replicación del área de trabajo.

  Al deshabilitar la replicación, se impide que se repliquen nuevos registros, pero no se quitan las copias replicadas hasta que la característica esté totalmente deshabilitada. Para supervisar el progreso de la operación de deshabilitación, consulte Comprobar el estado de aprovisionamiento del espacio de trabajo.

Planeamiento y administración de capacidad

Durante un fallo en la región u otro evento de conmutación de ida o de vuelta, tu área de trabajo podría experimentar una mayor carga debido a la ingesta y los reintentos de consulta. Si su solución es sensible a los retrasos en la ingesta de registros o en las consultas, incluso durante y después de un fallo de una región, considere sobreaprovisionar la capacidad asociada a cualquier límite diario de su área de trabajo. En el caso de los clústeres dedicados, seleccione un nivel de compromiso con margen para los picos de ingesta temporales. Para más información, consulte Administración de la capacidad mediante el sobreaprovisionamiento.

Comportamiento cuando todas las regiones están en buen estado

En esta sección se describe qué esperar al configurar un área de trabajo de Log Analytics para la replicación del área de trabajo y todas las regiones están operativas.

- **Operación entre regiones:** Todas las consultas y ingestas tienen como destino el área de trabajo de la región primaria mientras está en buen estado. El espacio de trabajo de la región secundaria contiene una réplica pasiva de datos, esquemas y configuración hasta la conmutación.

- **Replicación de datos entre zonas:** Los nuevos registros, esquemas y actualizaciones de configuración se replican de forma asincrónica en lotes en la base de datos secundaria. Dado que es asincrónico, este proceso de replicación no afecta a la latencia de ingesta. Normalmente, los datos se replican en unos dos minutos, pero este período de tiempo no está garantizado.

La replicación del área de trabajo se aplica en el nivel de área de trabajo, por lo que no puede seleccionar tablas individuales para replicar. Sin embargo, puedes controlar la replicación

asociando solo las DCR que contienen flujos de datos críticos con el punto de conexión de recopilación de datos del área de trabajo. Las DCRs que no están asociadas con el DCE del entorno de trabajo no replican sus datos. Para obtener más información, consulte Asocia reglas de recopilación de datos.

Comportamiento durante una falla de región

En esta sección se describe qué esperar al configurar un área de trabajo de Log Analytics para la replicación del área de trabajo y cuando ocurre una interrupción en una de las regiones.

- **Detección y respuesta:** Es responsable de decidir cuándo cambiar el área de trabajo secundaria para convertirse en la nueva área de trabajo principal. Microsoft no toma esta decisión ni inicia el proceso, incluso si hay una interrupción en la región. La alternancia y la reversión son acciones manuales.

  Para obtener información sobre cómo decidir cuándo cambiar, vea ¿Cuándo debo cambiar?. Para obtener información sobre cómo supervisar el estado del área de trabajo durante una posible conmutación, consulte Supervisión del rendimiento del área de trabajo mediante consultas.

  Para obtener más información sobre cómo promover un área de trabajo secundaria a la nueva área de trabajo principal, consulte Cambio a su área de trabajo secundaria.

- **Notifications:** Microsoft no le notifica automáticamente cuando una región está inactiva. Sin embargo, puede usar Azure Service Health para comprender el estado general del servicio, incluidos los errores de región, y puede configurar alertas de Service Health para notificarle problemas.

  Para validar la actualización del área de trabajo secundaria, puede supervisar el estado de replicación y también auditar el área de trabajo secundaria. Para obtener información sobre cómo supervisar el estado del área de trabajo al tomar la decisión de cambiar, consulte Supervisión del rendimiento del área de trabajo mediante consultas.

- **Solicitudes activas:** Es posible que se produzca un error en cualquier ingesta activa en la región con errores y también se produzca un error en las consultas en curso en la región con errores.

  Una vez que el proceso de conmutación cambia las entradas DNS del área de trabajo, el área de trabajo secundaria está disponible para la ingesta y la consulta. Los agentes de Microsoft vuelven a intentar automáticamente los intentos fallidos de ingesta de registros. Las aplicaciones que usan la API de ingesta de registros también deben reintentar.

- **Pérdida de datos esperada:** Los registros u otros cambios que aún no se han replicado no están disponibles en la región secundaria. Normalmente, los datos se replican en unos dos minutos, pero este período de tiempo no está garantizado.

  Después de cambiar a la región secundaria, si la región primaria no puede procesar los datos de registro entrantes, Azure Monitor almacena en búfer los datos de la región secundaria durante un máximo de 11 días. Durante los primeros cuatro días, Azure Monitor vuelve a intentar automáticamente replicar los datos periódicamente.

  La cantidad total de pérdida de datos a veces se denomina objetivo de punto de recuperación (RPO).

- **Tiempo de inactividad esperado:** El proceso de conmutación implica activar la ingesta en el área de trabajo secundaria y actualizar los registros DNS del área de trabajo.

  Aunque el cambio de registro DNS se produce rápidamente, la propagación de DNS puede tardar más tiempo. Si los agentes u otros clientes no respetan el período de vida de DNS (TTL), es posible que sigan intentando la ingesta en el área de trabajo principal. Asegúrese de que los clientes no almacenan en caché las entradas DNS durante más tiempo que su TTL.

  Las consultas se reanudan en el área de trabajo secundaria una vez completada la conmutación.

  La cantidad de tiempo que pasa antes de que el área de trabajo secundaria esté disponible a veces se denomina objetivo de tiempo de recuperación (RTO).

- **Redistribución:** El proceso de conmutación actualiza los registros DNS de los puntos de conexión de ingesta para que apunten a la región secundaria. La plataforma enruta las consultas internamente a la región activa.

  Mientras el espacio de trabajo se encuentra en estado de conmutación, los registros se replican de vuelta a la región primaria mediante replicación asincrónica.

Recuperación de regiones

Switchback es manual. Usted es responsable de decidir cuándo volver. Después de que el servidor principal esté estable y sincronizado, puede iniciar una conmutación de retorno. Compruebe que no queda ninguna cola de registros en el secundario y que la replicación se reanude correctamente desde el principal restaurado.

Para obtener información sobre cómo decidir cuándo volver a cambiar, vea ¿Cuándo debo volver? Para ver los pasos detallados de la conmutación, consulte Cambio al área de trabajo principal.

Prueba de fallos de región

Puede desencadenar una conmutación y una reinversión en cualquier momento, incluidas pruebas o simulacros de recuperación ante desastres.

Si realiza simulacros, se recomienda seguir estos procedimientos recomendados:

- ✓ Use un entorno que no sea de producción o, si prueba en producción, ejecute el proceso en un período de tiempo de bajo riesgo.
- ✓ Si es posible, simule los criterios que desencadenan los criterios de conmutación por error de la región en función de sus propias directivas. Este enfoque le permite probar la detección y automatización, así como el proceso de conmutación y conmutación por recuperación.
- ✓ Mida el RTO (finalización de conmutación) y el RPO (intervalo máximo no replicado). Use conjuntos de datos exportados para validar la paridad de registros de ejemplo entre las áreas de trabajo principal y secundaria.

Soluciones personalizadas de varias regiones para la resistencia

Si la replicación del área de trabajo no está disponible para su región o si debe usar un tipo de tabla que la replicación del área de trabajo no admite, tenga en cuenta estos enfoques alternativos para la resistencia de varias regiones:

- **Escritura dual:** Configura orígenes, como la configuración de diagnóstico, los agentes y la ingestión basada en DCR, para enviar logs a dos áreas de trabajo independientes en regiones diferentes.

  Este enfoque proporciona análisis casi en tiempo real en ambas regiones. Sin embargo, duplica el costo de ingesta e introduce el riesgo de desfase de configuración entre las áreas de trabajo.

- **Exportar y rehidratar:** Si la región está emparejada con otra región de Azure, puede usar data export para exportar continuamente los registros a Azure Blob Storage. Configure la cuenta de almacenamiento para usar uno de los tipos de almacenamiento con redundancia geográfica (GRS). Azure Storage replica automáticamente y de forma asincrónica los registros exportados en la región emparejada.

Durante un desastre, implemente una nueva área de trabajo y ingiera de forma selectiva datos críticos recientes. También puede consultar manualmente el historial a largo plazo directamente desde Blob Storage mediante Azure Data Explorer. Para obtener más información, consulte Consulta de datos exportados.

Copias de seguridad y restauración

Azure Monitor Registros no proporciona una funcionalidad tradicional de copia de seguridad o restauración para un punto específico en el tiempo. En su lugar, la durabilidad y la recuperación dependen de la replicación intraregión (incluida la redundancia de zona), la replicación del área de trabajo opcional y la exportación de datos para mantener una copia externa. Para una recuperación rápida de análisis después de un error, confíe en la replicación de la plataforma en lugar de en las operaciones de restauración.

Puede habilitar data export, que exporta automáticamente copias de los registros a Azure Storage. Si la región de Azure está emparejada, puede considerar la posibilidad de habilitar el almacenamiento con redundancia geográfica (GRS) para replicar los datos de registro en la región emparejada.

Resistencia a eliminaciones accidentales

Para cumplir los estrictos requisitos de cumplimiento y la protección contra alteraciones, considere la posibilidad de usar directivas de inmutabilidad para evitar la eliminación de datos de registro de la cuenta de almacenamiento.

Resistencia al mantenimiento del servicio

Microsoft aplica periódicamente actualizaciones de servicio y realiza otro mantenimiento. La plataforma Azure controla estas actividades automáticamente, lo que garantiza que el mantenimiento sea transparente y sin problemas. No se espera ningún tiempo de inactividad durante los eventos de mantenimiento, a menos que se le haya notificado a través de Azure Service Health de mantenimiento programado.

Para minimizar el efecto del mantenimiento, Azure Monitor realiza el mantenimiento gradual en las réplicas del área de trabajo y las zonas de disponibilidad.

Si tu solución es sensible a la latencia de ingesta y consulta, supervisa la latencia de ingesta y las métricas de salud del área de trabajo tanto durante como después de los eventos de mantenimiento programados.

Algunos cambios de configuración de ejecución prolongada pueden consumir muchos recursos, como habilitar la replicación, vincular clústeres y agregar esquemas grandes. Si necesita realizar estos cambios, evite programarlos cuando también se programe el mantenimiento de la plataforma. Utilice Azure Service Health para el mantenimiento programado y comprenda cuándo se programa el mantenimiento.

Acuerdo de nivel de servicio

El acuerdo de nivel de servicio (SLA) para Azure servicios describe la disponibilidad esperada de cada servicio y las condiciones que la solución debe cumplir para lograr esa expectativa de disponibilidad. Para obtener más información, vea SLAs for servicios en línea.

Registros de Azure Monitor es único entre los servicios en línea porque es la plataforma recomendada para recopilar los datos esenciales y la información necesaria para presentar una reclamación al soporte técnico de Microsoft por un problema relacionado con el SLA.

Dos acuerdos de nivel de servicio cubren los registros de Azure Monitor:

- El **SLA de disponibilidad de consultas de Log Analytics** define las expectativas de disponibilidad para consultar los datos de un área de trabajo de Log Analytics.
- El Acuerdo de Nivel de **Servicio de Azure Monitor** define las expectativas de disponibilidad para las reglas de alertas y los grupos de acciones, incluidas las expectativas para analizar las señales de telemetría y para la entrega de notificaciones.

Contenido relacionado

- Los procedimientos recomendados de fiabilidad para registros de Azure Monitor
- Zonas de disponibilidad en Azure Monitor
- Mejore la resiliencia mediante la replicación del espacio de trabajo de Log Analytics
- Exportación de datos del área de trabajo de Log Analytics
- Reglas de recopilación de datos
- Creación y administración de un clúster dedicado

ⓘ**Nota:** El autor ha creado este artículo con ayuda de inteligencia artificial. Más información

---

Last updated on 24/06/2026

Confiabilidad en Azure Site Recovery

Azure Site Recovery es un servicio de replicación y conmutación por error administrado para máquinas virtuales (VM) que mantiene las cargas de trabajo disponibles durante las interrupciones. Replica continuamente las cargas de trabajo de los sitios primarios a las ubicaciones secundarias y limita la pérdida de datos y el tiempo de inactividad. Durante el mantenimiento planificado o interrupciones inesperadas, gestiona la conmutación por error y el retorno al estado operativo normal. Este servicio admite la recuperación ante desastres (DR) para entornos locales y máquinas virtuales de Azure, lo que ayuda a las organizaciones a mantener la continuidad empresarial.

Al usar Azure, relibilidad es una responsabilidad compartida. Microsoft proporciona una variedad de funcionalidades para admitir resistencia y recuperación. Es responsable de comprender cómo funcionan esas funcionalidades dentro de todos los servicios que usa y de seleccionar las funcionalidades que necesita para cumplir los objetivos empresariales y los objetivos de tiempo de actividad.

En este artículo se describe cómo hacer que Site Recovery sea resistente a varias posibles interrupciones y problemas, incluidos errores transitorios, interrupciones de zona de disponibilidad y interrupciones de regiones. También resalta información clave sobre el acuerdo de nivel de servicio (SLA) de Site Recovery.

> **Nota:**
>
> En este artículo se describe cómo el servicio Site Recovery es resistente o cómo puede hacer que sea resistente a varios problemas. No explica cómo usar Site Recovery para proteger las máquinas virtuales u otros recursos. Para obtener más información, consulte Acerca de Site Recovery.

Recomendaciones de implementación de producción para la confiabilidad

Al usar Site Recovery con cargas de trabajo de producción, se recomienda realizar estas acciones:

- ✔ Implemente la bóveda de Recovery Services en la región de destino para replicación.

- ✓ Para la recuperación ante desastres de Azure a Azure, use la característica de alta renovación de Site Recovery para las máquinas virtuales que tienen una alta tasa de cambios en los datos. El soporte de alta rotación mejora el objetivo de punto de recuperación (RPO) y permite la replicación para muchas cargas de trabajo de bases de datos a gran escala.
- ✓ Para la recuperación ante desastres de Azure a Azure, configure la cuenta de almacenamiento en caché para usar el almacenamiento con redundancia de zona (ZRS).
- ✓ Realice pruebas de conmutación por error periódicamente como parte de los simulacros de recuperación ante desastres. Ejecute simulacros de recuperación ante desastres cada trimestre o semestralmente para comprobar que los procesos de replicación y conmutación por error son correctos.
- ✓ Use reservas de capacidad a petición para asegurar que los recursos de computación están disponibles en la región de destino para la conmutación por falla.
- ✓ Habilite las actualizaciones automáticas para agentes de movilidad.
- ✓ Supervise el estado de la replicación y configure alertas para que se le notifique si se produce un problema.

Introducción a la arquitectura de confiabilidad

Cuando se usa Site Recovery, se define un *origen* y un *destino*, que representan las máquinas virtuales replicadas:

- El *origen* puede ser una máquina virtual de Azure o una máquina virtual o un servidor de otro origen compatible, incluidos los servidores físicos locales, las máquinas virtuales de VMware y las máquinas virtuales de Hyper-V.
- El *target* siempre es una máquina virtual de Azure. Para la replicación de máquinas virtuales de Azure a Azure, el destino puede ser una región o zona de disponibilidad diferente de la máquina virtual de origen.

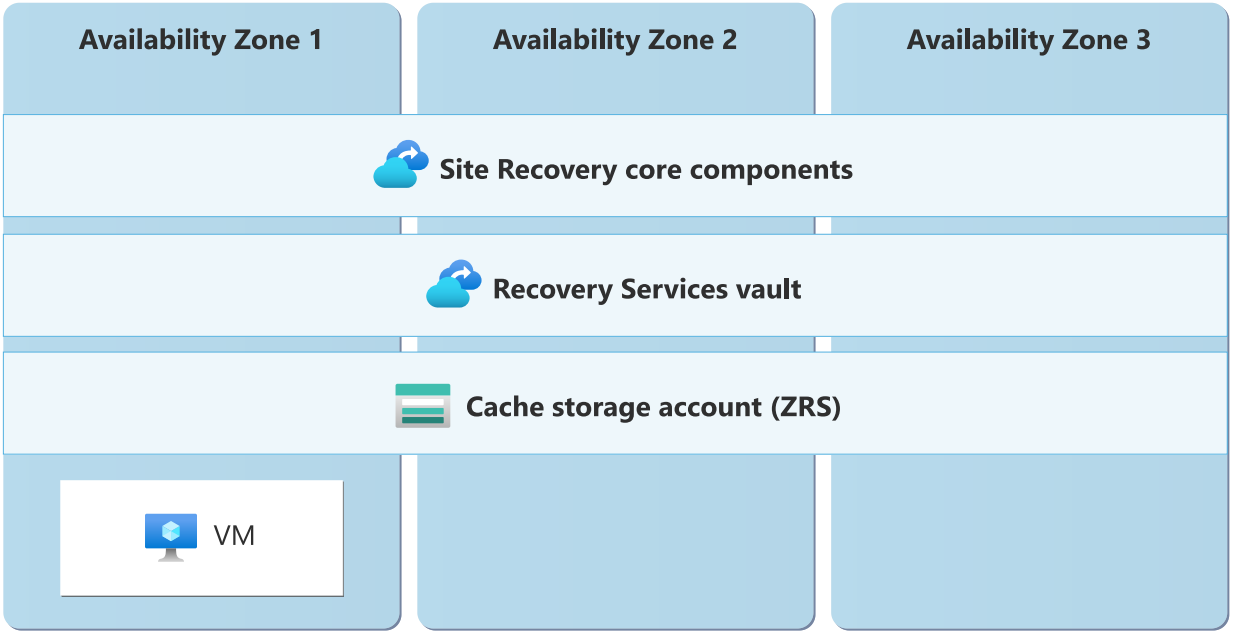
Usted es responsable de implementar y configurar recursos y opciones relacionadas, entre las que se incluyen:

- *Recovery Services vault*, que Site Recovery usa para almacenar las opciones de configuración de replicación. La bóveda no guarda tus datos replicados. La configuración de redundancia del almacén no es importante para Site Recovery, pero es importante si usa el mismo almacén para Azure Backup.

Una bóveda puede incluir configuraciones adicionales, como:

- *Directiva de replicación*, que configura la frecuencia de las instantáneas y la duración de retención.
- *Un plan de recuperación*, que coordina el orden en el que las máquinas conmutan por error y pueden incluir scripts y acciones manuales. Los planes de recuperación son especialmente útiles para las workloads que tienen múltiples niveles, tales como los niveles de aplicación y base de datos, que deben realizar un failover en un orden específico.

- Para la replicación de Azure a Azure, una cuenta de almacenamiento *cache* que almacena una copia de los datos de origen en su región antes de que se replique en el destino. La configuración de redundancia de la cuenta de almacenamiento en caché puede afectar a la confiabilidad durante una interrupción de la zona de disponibilidad.



> **Nota:**
>
> Esta guía se centra en la confiabilidad de los componentes basados en Azure de Site Recovery y en la relación de replicación. Si replica datos o máquinas virtuales desde un entorno local u otro proveedor de nube, considere la confiabilidad de los componentes fuera de Azure.

Para obtener más información sobre los componentes que implemente, consulte los siguientes artículos:

- Arquitectura de DR de Azure a Azure
- Arquitectura de recuperación ante desastres de Hyper-V a Azure
- Arquitectura de recuperación ante desastres de VMware a Azure
- Arquitectura de recuperación ante desastres de servidor físico hacia Azure

El servicio principal Site Recovery se ejecuta en la infraestructura que Administra Microsoft. En este artículo se hace referencia a estos componentes colectivamente como el *servicio principal de Site Recovery*.

Resistencia a errores transitorios

Los errores transitorios son errores breves e intermitentes en los componentes. Se producen con frecuencia en un entorno distribuido como la nube y son una parte normal de las operaciones. Los errores transitorios se corrigen después de un breve período de tiempo. Es importante que las aplicaciones puedan controlar errores transitorios, normalmente mediante el reintento de solicitudes afectadas.

Todas las aplicaciones hospedadas en la nube deben seguir las instrucciones de control de errores transitorios de Azure cuando se comunican con cualquier API, bases de datos y otros componentes hospedados en la nube. Para obtener más información, consulte Recomendaciones para controlar errores transitorios.

Site Recovery controla automáticamente los errores transitorios que se producen durante el proceso de replicación mediante el reintento de sus operaciones. No es necesario configurar el control de errores transitorios para Site Recovery.

Resistencia a errores de zona de disponibilidad

Availability zones son grupos de centros de datos físicamente independientes dentro de una región de Azure. Cuando una zona falla, los servicios pueden transferirse a una de las zonas restantes.

Para comprender cómo se comporta la replicación de Site Recovery durante los errores de zona de disponibilidad, debe tener en cuenta los siguientes componentes de servicio:

- **Servicio Core Site Recovery:** El servicio principal de Site Recovery está diseñado para ser resistente a los errores de zona de disponibilidad en las regiones admitidas. Los componentes internos del servicio apoyan automáticamente la redundancia de zona sin necesidad de configuración por parte del cliente.

- **Bóveda de Recovery Services:** La bóveda almacena los datos de configuración. En las regiones en las que Site Recovery admite la resiliencia de zona, los datos de configuración del almacén también son resilientes a la zona.

- **Cuenta de almacenamiento en caché:** Para la replicación de Azure a Azure, usted es responsable de asegurarse de que la cuenta de almacenamiento de caché sea redundante a nivel de zona mediante su implementación en el nivel ZRS.

  Si usa el nivel de replicación de Azure Storage con almacenamiento redundante local (LRS) para su cuenta de almacenamiento en caché y falla una zona, Site Recovery podría no replicar los datos modificados recientemente a su destino.

> ⓘ **Nota:**
>
> Site Recovery puede ayudarle a conmutar por error entre máquinas virtuales en diferentes zonas de disponibilidad. Para más información, consulte **Habilitar la recuperación ante desastres de máquinas virtuales de Azure entre zonas de disponibilidad**.

Requisitos

Compatibilidad con regiones:

- **Almacenes principales de Site Recovery Service y Recovery Services:** Site Recovery es resistente a la zona en las siguientes regiones.

  Expandir tabla

Americas	Europa	Oriente Medio	Asia Pacífico
Centro de Chile	Este de Austria	Israel Central	Centro de Indonesia
Centro de México	Norte de Italia		Japón Occidental
Oeste de EE. UU. 3	Centro de Polonia		Oeste de Malasia
	Centro de España		Norte de Nueva Zelanda

  Site Recovery está implementando compatibilidad con zonas de disponibilidad en todas las regiones habilitadas para zonas de disponibilidad. En las regiones que no aparecen en la tabla anterior como resistentes a la zona, los errores de zona pueden afectar a las operaciones.

- **Cuenta de almacenamiento en caché:** Puede implementar una cuenta de almacenamiento de ZRS en todas las regiones habilitadas para zonas de disponibilidad.

Cost

Site Recovery se factura en función del número de instancias de máquina virtual protegidas, independientemente de su configuración de zona de disponibilidad. Para obtener más información, consulte Precios de Site Recovery.

Configurar soporte de zonas de disponibilidad

- **Core Site Recovery service:** No se configura la resistencia de zona en el servicio principal Site Recovery. Microsoft proporciona resiliencia de zona en las regiones compatibles.

  Si Microsoft habilita la resistencia de zona en una región más adelante, los recursos de Site Recovery se benefician automáticamente de la resistencia de zona. No es necesario realizar ninguna acción.

- **Almacén de Recovery Services:** Los almacenes de Recovery Services tienen una configuración de redundancia, pero Site Recovery no usa la configuración de redundancia del almacén. No es necesario configurar la bóveda para la redundancia de zona cuando usas Site Recovery.

- **Cache storage account:** Cuando se usa la replicación de Azure a Azure, es responsable de crear la cuenta de almacenamiento de caché y de configurarla con el nivel adecuado de redundancia. Para que sea redundante de zona, configúrelo para el tipo de replicación ZRS. Para obtener más información, consulte Reliability en Azure Blob Storage.

Comportamiento cuando todas las zonas están en buen estado

En esta sección se describe qué esperar cuando se usa Site Recovery en una región con compatibilidad con zonas de disponibilidad para el servicio principal, la cuenta de almacenamiento en caché está configurada para usar ZRS y todas las zonas de disponibilidad están operativas.

- **Operación entre zonas:** El proceso de replicación puede usar la infraestructura en varias zonas de disponibilidad para desencadenar y ejecutar trabajos de replicación. El servicio administra esta infraestructura de forma transparente.

- **Replicación de datos entre zonas:** Site Recovery y Storage controlan la replicación de datos de zona:

- *Configuración de Site Recovery:* Site Recovery replica los datos de configuración entre zonas, incluso si configura el almacén para usar LRS.
- *Cuenta de almacenamiento en caché:* Si configura la cuenta de almacenamiento de caché para que use ZRS, Storage replica de forma sincrónica los datos almacenados en caché entre zonas.

Comportamiento durante un fallo de zona

En esta sección se describe qué esperar cuando se usa Site Recovery en una región con compatibilidad con la zona de disponibilidad para el servicio principal, la cuenta de almacenamiento en caché está configurada para usar ZRS y se produce una interrupción de zona de disponibilidad.

> ⓘ **Nota:**
>
> Si la zona con errores contiene la máquina virtual de origen, usted es responsable de iniciar la conmutación por error al destino. Para obtener más información, consulte los artículos siguientes:
>
> - Conmutar las máquinas virtuales de Azure a una región secundaria
> - Conmutación por error de máquinas virtuales de VMware
> - Conmutar las máquinas virtuales de Hyper-V a Azure

- **Detección y respuesta:** La plataforma Site Recovery detecta automáticamente errores en una zona de disponibilidad e inicia una respuesta. No es necesario iniciar manualmente una conmutación por error de zona para el servicio principal de Site Recovery. Sin embargo, si la interrupción de zona afecta a la máquina virtual de origen, es posible que tenga que iniciar la conmutación por error de la máquina virtual.
- **Notificación:** Microsoft no le notifica automáticamente cuando una zona está inactiva. Sin embargo, puede usar Azure Service Health para comprender el estado general del servicio, incluidos los errores de zona, y puede configurar alertas de Service Health para notificarle problemas.
- **Solicitudes activas:** El efecto en los trabajos de replicación activa depende del tipo de replicación:
  - *Replicación de zona a zona y región a región de máquinas virtuales de Azure:* Si la instancia de origen o de destino está en la zona con errores, la replicación se detiene

hasta que ambas instancias estén disponibles de nuevo.

Si la zona con errores no contiene la máquina virtual de origen o de destino y configuró la cuenta de almacenamiento en caché para usar ZRS, la replicación continúa ejecutándose.

- *On-premises para Azure:* Si la instancia de destino está en la zona con errores, la replicación se detiene hasta que la instancia esté disponible de nuevo.

  Si la zona con errores no contiene la máquina virtual de destino, la replicación continúa ejecutándose.

- **Pérdida de datos esperada:** No se espera ninguna pérdida de datos durante un error de zona.
- **Tiempo de inactividad esperado:** Si la zona con errores contiene la máquina virtual de origen o de destino, la replicación se detiene hasta que ambas instancias estén disponibles de nuevo.
- **Redistribución:** Site Recovery y Storage se adaptan automáticamente a errores de zona:
  - *Servicio Core Site Recovery:* El servicio principal de Site Recovery usa automáticamente la infraestructura en zonas de disponibilidad correctas para realizar la replicación. No es necesario realizar ninguna acción.
  - *Cuenta de almacenamiento en caché:* El almacenamiento enruta automáticamente las solicitudes de datos de caché a zonas correctas.

Recuperación de zona

Cuando se recupera la zona de disponibilidad afectada, Site Recovery reanuda automáticamente los trabajos de replicación que se pausan durante la interrupción de la zona.

Es responsable de iniciar la conmutación por recuperación para servidores o máquinas virtuales conmutadas por error durante la interrupción de la zona. Para obtener más información, consulte los artículos siguientes:

- *Replicación de zona a zona y región a región de máquinas virtuales de Azure:*recuperar una máquina virtual de Azure a la región primaria
- *Replicación local a Azure:*

- *Replicación física a Azure:*arquitectura de DR (recuperación ante desastres) de servidor físico a Azure
- *Replicación de Hyper-V a Azure:*Arquitectura de recuperación ante desastres de Hyper-V a Azure
- *Replicación de VMware a Azure:*Acerca del fallo de recuperación ante desastres y el retorno después de la recuperación en el entorno local

Prueba de fallos de zona

La plataforma Site Recovery administra la resistencia de zona para sus componentes internos. Esta característica está totalmente administrada, por lo que no es necesario iniciar ni validar los procesos de error de zona de disponibilidad.

Es importante realizar simulacros regulares de recuperación ante desastres, que deben probar la conmutación por error de la VM y los procedimientos generales de respuesta. Diseñe los simulacros de recuperación ante desastres para evitar el impacto en el entorno de producción. Para obtener más información, consulte los artículos siguientes:

- *Replicación de zona a zona y de región a región de máquinas virtuales de Azure:*Realizar una prueba de recuperación ante desastres para máquinas virtuales de Azure
- *Replicación local a Azure:*
  - *Replicación física a Azure:*Realizar un simulacro de DR en Azure
  - *Replicación de Hyper-V a Azure:*Ejecución de una prueba de DR en Azure
  - *Replicación de VMware a Azure:*Realizar un simulacro de DR en Azure

Resistencia a errores en toda la región

Para la replicación de Azure a Azure, Site Recovery proporciona resistencia a los fallos de región al habilitar la conmutación por error de las máquinas virtuales (VMs) a una región de destino saludable. Para más información, consulte Replicar máquinas virtuales de Azure a otra región de Azure.

Consideraciones

- **Región del almacén:** Implemente un almacén de Recovery Services en una región de Azure específica que seleccione. La región de la bóveda es importante. La replicación continúa

durante una interrupción en la región de la bóveda. Sin embargo, no puede realizar operaciones de administración de Site Recovery, incluida la conmutación por error y el restablecimiento, hasta que la región se recupere.

La implementación del almacén en la región de destino ayuda a garantizar que las operaciones de conmutación por error y recuperación permanezcan disponibles durante una interrupción de la región de origen. También evita que una interrupción en una tercera región afecte las operaciones de conmutación y recuperación.

> ⓘ **Nota:**
>
> Si su bóveda está en la región que normalmente usa como región de destino, entonces, después de la conmutación por error y restablecimiento de la replicación, esa región se convierte en su nueva región de origen. Si esa región experimenta posteriormente un problema, posiblemente no pueda realizar la reversión de la conmutación hasta que ambas regiones estén funcionales.

- **Reservas de capacidad:** Es responsable de comprobar que la región de destino admite los tipos de máquina virtual que necesita y que tiene capacidad disponible para la carga de trabajo. Se recomienda usar reservas de capacidad a petición para asegurarse de que los recursos de cómputo están disponibles para su carga de trabajo si se produce una conmutación por error.

Configuración de la compatibilidad con varias regiones

- **Almacén de Recovery Services:** Debe seleccionar la región del almacén. Para más información, consulte Consideraciones.

  Las bóvedas de Servicios de Recuperación tienen una configuración de redundancia, pero Site Recovery no utiliza la configuración de redundancia de la bóveda. No es necesario configurar la bóveda para la redundancia geográfica al usar Site Recovery.

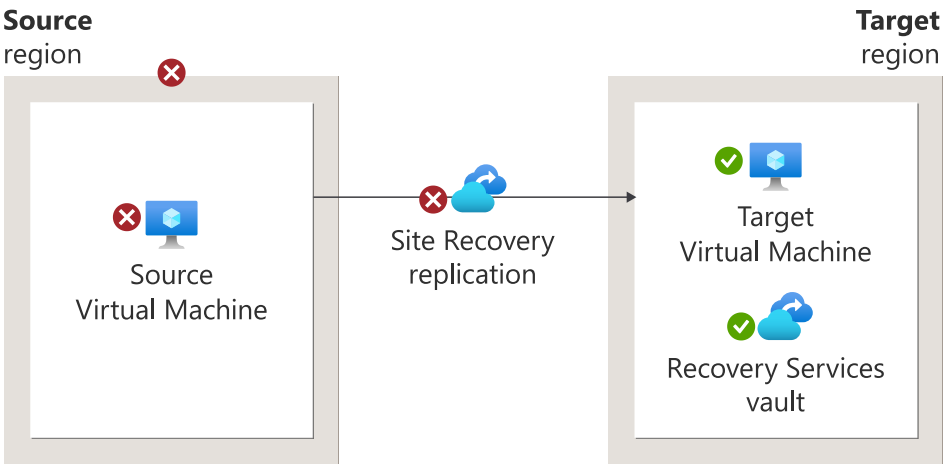
- **Cuenta de almacenamiento en caché:** La cuenta de almacenamiento en caché solo se usa como ubicación temporal para los datos antes de que se replique, por lo que no debe configurarla para usar el almacenamiento con redundancia geográfica (GRS).

Comportamiento durante una falla de región

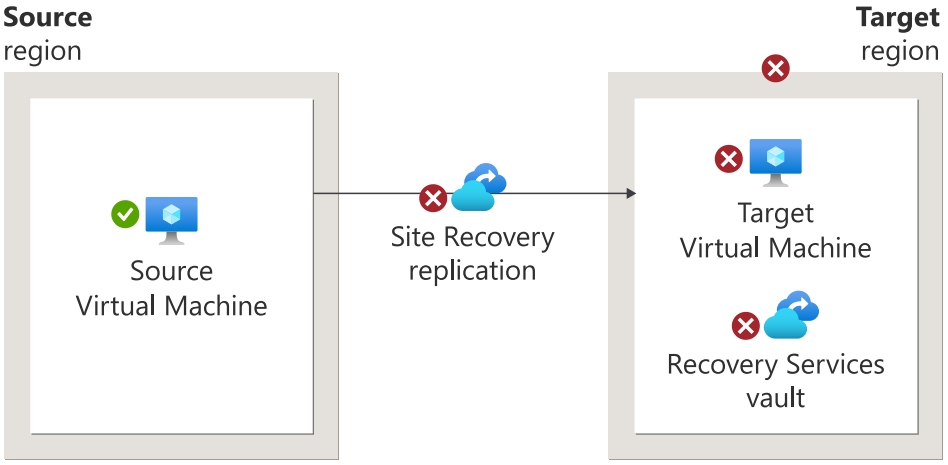
El comportamiento específico del servicio principal de Site Recovery durante un error de región depende de la región en la que experimente el error:

- **Error en la región de origen:** Para la replicación de Azure a Azure, puede activar una conmutación por error automática cuando la región de origen no esté disponible.

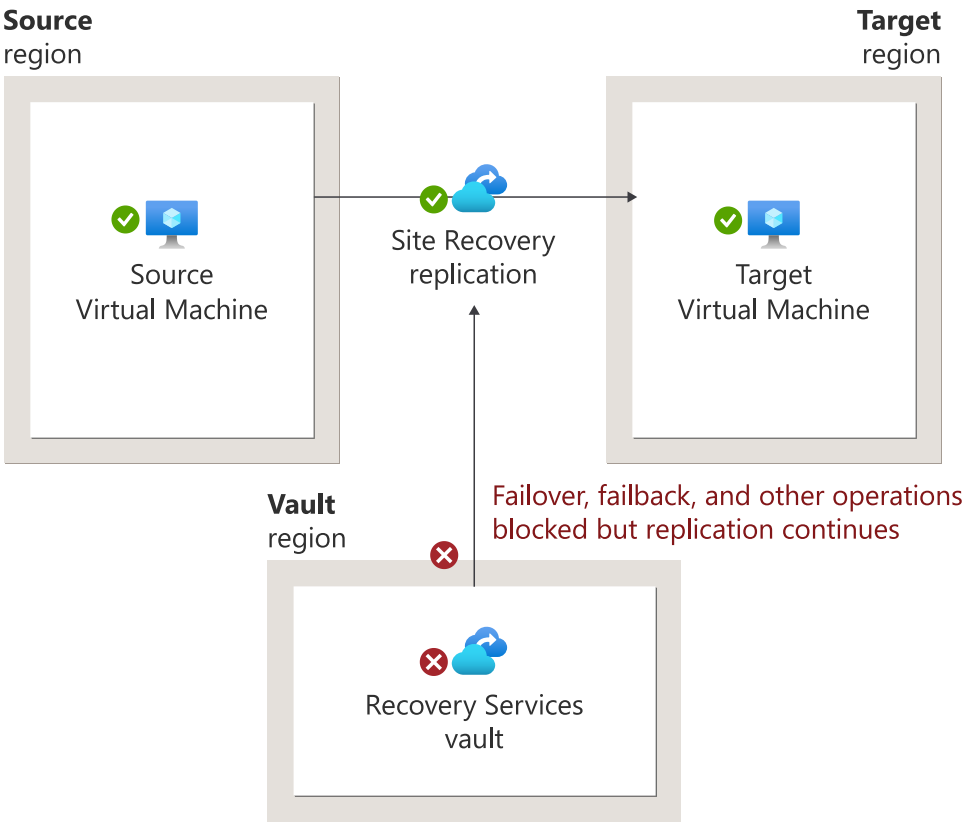
  Dado que la región de origen no está disponible, la replicación se detiene hasta que la máquina virtual de la región de origen está en buen estado.



- **Error en la región de destino:** Dado que la región de destino no está disponible, la replicación se detiene y no se puede recuperar el destino hasta que la región esté operativa.



- **Fallo en la región que contiene la bóveda:** Si despliega la bóveda en una tercera región (no en la región de origen o de destino) y esa región experimenta un fallo, Site Recovery continuará replicando sus datos. Sin embargo, no puede iniciar ninguna operación, incluida la conmutación por error o la reversión, hasta que la bóveda esté en buen estado.



Recuperación de regiones

Es responsable de iniciar la reversión para servidores o máquinas virtuales que conmutó durante la interrupción de la región. Para obtener más información, consulte los artículos siguientes:

- *Replicación de zona a zona y de región a región de máquinas virtuales de Azure:*Recuperar máquina virtual de Azure a la región primaria
- *Replicación local a Azure:*
  - *Replicación de servidor físico a Azure:*arquitectura de recuperación ante desastres
  - *Replicación de Hyper-V a Azure:*Arquitectura de recuperación ante desastres de Hyper-V a Azure
  - *Replicación de VMware a Azure:*conmutación por error y restablecimiento de la recuperación ante desastres en el sitio

Prueba de fallos de región

Es importante realizar simulacros de recuperación ante desastres regulares que prueben la conmutación por error de la VM y los procedimientos generales de respuesta. Diseñe los

simulacros de recuperación ante desastres para evitar el impacto en el entorno de producción. Para obtener más información, consulte los artículos siguientes:

- *Replicación de zona a zona y de región a región de máquinas virtuales de Azure:*Ejecutar un simulacro de recuperación ante desastres para máquinas virtuales de Azure
- *Replicación local a Azure:*
  - *Replicación física a Azure:*Ejecutar un simulacro de DR en Azure
  - *Replicación de Hyper-V a Azure:*Realizar una prueba de DR en Azure
  - *Replicación de VMware a Azure:*Ejecución de un simulacro de recuperación ante desastres en Azure

Resistencia a problemas de configuración y replicación

Una solución de recuperación ante desastres solo es confiable cuando sabe que funciona antes de que se produzca un desastre. Supervise Site Recovery para detectar problemas como errores de configuración o problemas de mantenimiento de replicación de máquinas virtuales. Para obtener más información, consulte Supervisión de Site Recovery.

Se recomienda configurar alertas de Azure Monitor para que se le informe sobre los problemas con el estado de replicación. Para más información, consulte Alertas integradas de Azure Monitor para Site Recovery.

Resistencia al mantenimiento del servicio

Azure administra automáticamente las actualizaciones y el mantenimiento del servicio principal Site Recovery. Las operaciones de mantenimiento no requieren tiempo de inactividad y no interrumpen la replicación de las máquinas virtuales y los servidores.

Sin embargo, es responsable de aplicar actualizaciones a los componentes de Site Recovery en las máquinas virtuales y servidores, incluido el agente de movilidad cuando sea necesario.

> **ⓘ Importante**
>
> Se recomienda encarecidamente habilitar las actualizaciones automáticas para los agentes. Si la versión del agente queda más de cuatro versiones atrás, la replicación se desactiva y la capacidad de recuperación de la carga de trabajo se ve comprometida.

Para obtener más información, consulte actualizaciones de Service en Site Recovery.

Acuerdo de nivel de servicio

El acuerdo de nivel de servicio (SLA) para Azure servicios describe la disponibilidad esperada de cada servicio y las condiciones que la solución debe cumplir para lograr esa expectativa de disponibilidad. Para obtener más información, vea SLAs for online services.

Para Site Recovery, los Acuerdos de Nivel de Servicio independientes cubren:

- **Disponibilidad del servicio**, lo que significa que Site Recovery está disponible para realizar una conmutación por error de las instancias protegidas. Una instancia protegida es una máquina virtual o un servidor físico que se replica en una ubicación secundaria. Para poder optar a este Acuerdo de Nivel de Servicio, debe reintentar los intentos fallidos de conmutación por error al menos cada 30 minutos.

- **Objetivo de tiempo de recuperación (RTO),** que es el tiempo desde que se desencadena una conmutación por fallo (o cuando los scripts lo desencadenan) hasta cuando se ejecuta la máquina virtual de destino. Esta vez excluye acciones manuales o ejecución de scripts.

El Acuerdo de Nivel de Servicio solo proporciona créditos de servicio cuando la región secundaria tiene suficiente capacidad de proceso.

Contenido relacionado

- Acerca de Site Recovery
- Confiabilidad en Azure

ⓘ **Nota:** El autor ha creado este artículo con ayuda de inteligencia artificial. Más información

---

Last updated on 24/04/2026

Confiabilidad en la detección de Azure Storage

Este artículo explica cómo Azure Storage Discovery admite la confiabilidad, incluida la resiliencia intrarregional con zonas de disponibilidad y recuperación ante desastres cruzada de regiones y continuidad empresarial. Para conocer los principios generales de confiabilidad en Azure, consulte Azure confiabilidad.

Azure Storage Discovery es un servicio totalmente administrado que proporciona visibilidad de toda la empresa sobre la Azure Blob Storage y Azure Data lake Storage. Desde un solo panel de vidrio, puede analizar tendencias, optimizar los costos, mejorar la seguridad y mejorar la eficiencia operativa.

Comportamiento del servicio durante interrupciones

Expandir tabla

	Interrupción de la región	Interrupción de zona
Gestión del espacio de trabajo de almacenamiento	Si la región en la que se crea el espacio de trabajo de descubrimiento de almacenamiento experimenta una interrupción, los informes del espacio de trabajo no estarán disponibles durante el periodo.	Si una zona está inactiva, el área de trabajo Detección de almacenamiento permanece disponible. Puede seguir administrando el área de trabajo y acceder a los informes.
Informes de detección de almacenamiento	Si una región deja de funcionar, los datos de las cuentas de almacenamiento de esa región se pueden retrasar en los informes de detección hasta que se restaure la región. En el caso poco frecuente de pérdida completa de datos durante el periodo de interrupción, el área de trabajo de detección no puede recuperar los datos perdidos durante ese periodo.	Si la cuenta de almacenamiento se implementa en una zona con errores, la cuenta deja de estar disponible y la información del almacenamiento afectado no está disponible. Si la cuenta de almacenamiento es redundante por zona, el Discovery workspace continúa proporcionando información sobre la cuenta.

Soporte para zonas de disponibilidad

Availability zones son grupos de centros de datos físicamente independientes dentro de una región de Azure. Cuando una zona falla, los servicios pueden transferirse a una de las zonas

restantes.

- La administración del área de trabajo de Detección de almacenamiento tiene redundancia de zona. Cuando una zona está inactiva, la área de trabajo de Detección de almacenamiento permanece disponible. Puede seguir administrando el área de trabajo y acceder a los informes.
- Los informes de Detección de almacenamiento tienen redundancia de zona. Sin embargo, la interrupción del recurso de almacenamiento puede afectar a la información sobre ese recurso. Si un recurso de almacenamiento tiene redundancia de zona, el flujo de información permanece ininterrumpido.

Recuperación ante desastres entre regiones y continuidad empresarial

La recuperación ante desastres (DR) hace referencia a las prácticas que las organizaciones usan para recuperarse de eventos de alto impacto, como desastres naturales o implementaciones con errores que producen tiempo de inactividad y pérdida de datos. Independientemente de la causa, el mejor remedio para un desastre es un plan de recuperación ante desastres bien definido y probado y un diseño de aplicaciones que apoye activamente la recuperación ante desastres. Antes de empezar a crear el plan de recuperación ante desastres, consulte Recomendaciones para diseñar una estrategia de recuperación ante desastres.

Para la recuperación ante desastres, Microsoft usa el modelo de responsabilidad compartida. En este modelo, Microsoft garantiza que la infraestructura de línea base y los servicios de plataforma estén disponibles. Sin embargo, muchos servicios de Azure no replican automáticamente datos ni se revierten de una región con errores para realizar la replicación cruzada en otra región habilitada. En esos servicios, es responsable de configurar un plan de recuperación ante desastres válido para su carga de trabajo. La mayoría de los servicios que se ejecutan en ofertas de plataforma como servicio (PaaS) de Azure proporcionan características e instrucciones para admitir la DR. Puede usar características específicas del servicio para admitir la recuperación rápida con el fin de contribuir al desarrollo del plan de recuperación ante desastres.

Detección y notificación de interrupciones

La detección de almacenamiento no proporciona notificaciones de interrupción para cada incidente. El servicio De detección de almacenamiento utiliza la misma infraestructura de notificación que se usa en Azure para informar a los clientes sobre interrupciones prolongadas o graves.

> ⓘ **Nota:**
>
> Si necesita acceso ininterrumpido a los informes de detección, puede considerar la posibilidad de crear una segunda área de trabajo en otra región. Esta segunda área de trabajo debe configurarse para analizar los mismos recursos de almacenamiento. Este enfoque le protege de una interrupción regional que afecta al área de trabajo principal. Sin embargo, no protege frente a problemas de disponibilidad en los propios recursos de almacenamiento entre regiones. Obtenga más información sobre los precios de Storage Discovery para evaluar los costos adicionales.

Disponibilidad de datos y cobertura de actualización

Factores clave que influyen en la disponibilidad y actualización de los datos en los informes de detección de almacenamiento:

- Después de crear un área de trabajo de detección de almacenamiento, la mayoría de las conclusiones aparecen en los informes en un plazo de 24 horas.
- La mayoría de las actualizaciones, como agregar o quitar una cuenta de almacenamiento en una suscripción o un grupo de recursos vinculados a un área de trabajo de detección de almacenamiento, se reflejarán en los informes de detección de almacenamiento en un plazo de 24 horas.
- Los cambios relacionados con la adición de un nuevo ámbito o la edición de un ámbito existente en el área de trabajo de detección se reflejan normalmente en los informes en un plazo de 24 horas.
- El efecto de cambiar el plan de precios en un área de trabajo de Detección de almacenamiento existente es inmediato.
- Cuando se elimina un ámbito, se pierde inmediatamente la información recopilada para este ámbito.
- Cuando se elimina un área de trabajo de Discovery, se pierden inmediatamente los análisis recopilados para todos los ámbitos contenidos. Tenga cuidado al eliminar un área de trabajo o un ámbito.

Pasos siguientes

- Planeamiento de la implementación de la Detección de almacenamiento
- Preguntas más frecuentes

ⓘ**Nota:** El autor ha creado este artículo con ayuda de inteligencia artificial. Más información

Last updated on 15/04/2026

Azure Migrate: preguntas comunes

En este artículo se responden a preguntas habituales sobre Azure Migrate. Si tiene preguntas después de leer este artículo, puede publicarlas en el foro de Azure Migrate. También puede revisar estos artículos:

- Preguntas sobre el dispositivo de Azure Migrate
- Preguntas sobre la detección, la evaluación y la visualización de dependencias

¿Qué es Azure Migrate?

Azure Migrate proporciona un centro de conectividad para realizar la detección, valoración y migración a Azure de las aplicaciones y cargas de trabajo locales, así como de las máquinas virtuales en las nubes privadas o públicas. El centro proporciona herramientas de Azure Migrate para la evaluación y la migración, así como ofertas de ISV de terceros. Más información.

¿Qué puedo hacer con Azure Migrate?

Use Azure Migrate para descubrir, valorar y migrar la infraestructura, las aplicaciones y los datos locales a Azure. Azure Migrate admite la valoración y migración de máquinas virtuales de VMware locales, máquinas virtuales de Hyper-V, servidores físicos, otras máquinas virtuales virtualizadas, bases de datos, aplicaciones web y escritorios virtuales.

¿Cuál es la diferencia entre Azure Migrate y Azure Site Recovery?

Azure Migrate proporciona un centro centralizado para la evaluación y la migración a Azure.

- El uso de Azure Migrate proporciona interoperabilidad y extensibilidad futura con herramientas de Azure Migrate, otros servicios de Azure y herramientas de terceros.
- La herramienta Migración y modernización está diseñada específicamente para la migración de servidores a Azure. Está optimizada para la migración. No es necesario obtener información sobre los conceptos y escenarios que no son directamente pertinentes para la migración.
- No hay cargos por uso de herramientas para la migración durante 180 días a partir del momento en que se inicia la replicación de una máquina virtual. Le da tiempo para completar la migración. Solo paga por los recursos de almacenamiento y de red que se

usan en la replicación, y por los cargos de proceso consumidos durante las migraciones de prueba.

- Azure Migrate admite todos los escenarios de migración que admite Site Recovery. Asimismo, en el caso de las VM de VMware, Azure Migrate proporciona una opción de migración sin agente.
- Estamos dando prioridad a las nuevas características de migración solo de la herramienta Migración y modernización. Estas características no están dirigidas a Site Recovery.

Azure Site Recovery solo debe usarse para la recuperación ante desastres.

La herramienta Migración y modernización usa cierta funcionalidad de back-end de Site Recovery para la migración mediante lift-and-shift de algunas máquinas locales.

Tengo un proyecto con la experiencia clásica anterior de Azure Migrate. ¿Cómo empiezo a usar la nueva versión?

Azure Migrate clásico se retira en febrero de 2024. Después de febrero de 2024, ya no se admitirá la versión clásica de Azure Migrate y se eliminarán los metadatos de inventario del proyecto clásico. No se pueden actualizar proyectos ni componentes de la versión anterior a la nueva versión. Debe crear un nuevo proyecto de Azure Migrate y agregarle herramientas de evaluación y migración . Use los tutoriales para aprender a usar las herramientas de valoración y migración disponibles. Si tiene un área de trabajo de Log Analytics asociada a un proyecto clásico y desea adjuntarla a uno de la versión actual, no tiene más que eliminar el proyecto clásico.

¿Cuál es la diferencia entre Azure Migrate: Detección y evaluación y MAP Toolkit?

Server Assessment proporciona una valoración para ayudar con la preparación de la migración y la evaluación de las cargas de trabajo para la migración a Azure. Microsoft Assessment and Planning (MAP) Toolkit ayuda con otras tareas, incluido el planeamiento de la migración para versiones más recientes de los sistemas operativos cliente y servidor de Windows, y el seguimiento de uso de software. Para esos escenarios, siga usando MAP Toolkit.

¿Cuál es la diferencia entre Server Assessment y Site Recovery Deployment Planner?

Server Assessment es una herramienta de planeamiento de la migración. Site Recovery Deployment Planner es una herramienta de planeamiento de recuperación ante desastres.

Elija la herramienta en función de lo que desee hacer:

- **Planear la migración local a Azure**: si planea migrar los servidores locales a Azure, use Server Assessment para planear la migración. Server Assessment evalúa las cargas de trabajo locales y proporciona orientación y herramientas que le ayudarán a migrar. Una vez implementado el plan de migración, puede usar herramientas como Migración y modernización para migrar las máquinas a Azure.
- **Planear la recuperación ante desastres en Azure**: si planea configurar la recuperación ante desastres desde el entorno local a Azure con Site Recovery, use Site Recovery Deployment Planner. Deployment Planner realiza una evaluación profunda y específica de Site Recovery de su entorno local para la recuperación ante desastres. Proporciona recomendaciones relacionadas con la recuperación ante desastres, como la replicación y la conmutación por error.

¿Cómo funciona la herramienta Migración y modernización con Site Recovery?

- Si usa la herramienta de migración y modernización para realizar una migración *sin agente* de máquinas virtuales de VMware locales, la migración es nativa de Azure Migrate y no se utiliza Site Recovery.
- Si usa la herramienta de migración y modernización para realizar una migración *basada en agente* de máquinas virtuales de VMware o si migra Hyper-V máquinas virtuales o servidores físicos, la herramienta de migración y modernización usa el motor de replicación de Azure Site Recovery.

¿Qué zonas geográficas se admiten?

Revisa las zonas geográficas admitidas para nubes públicas y gubernamentales.

¿Qué hace Azure Migrate para garantizar la residencia de datos?

Cuando se crea un proyecto, se selecciona una geografía de su elección. El proyecto y los recursos relacionados se crean en una de las regiones de la geografía, tal y como asigna el

servicio Azure Migrate. Consulte las ubicaciones de almacenamiento de metadatos de cada zona geográfica aquí.

Azure Migrate no mueve ni almacena datos de clientes fuera de la región asignada, lo que garantiza la residencia y resistencia de los datos en la misma geografía.

Azure Migrate proporciona una opción adicional para crear proyectos y garantizar la residencia de datos en las regiones preferidas más allá de las zonas geográficas especificadas. Para aprovechar esta opción, use el vínculo https://aka.ms/migrate/ProjectRegionSelection.

¿Ofrece Azure Migrate copia de seguridad y recuperación ante desastres?

Azure Migrate se clasifica como recuperación ante desastres administrada por el cliente, lo que significa que Azure Migrate no ofrece para recuperar datos de una región alternativa y ofrecerlos a los clientes cuando la región del proyecto no está disponible.

Al usar diferentes funcionalidades, se recomienda exportar el inventario de software, el análisis de dependencias y el informe de evaluación para una copia de seguridad sin conexión.

En caso de error o interrupción regional en la región de Azure en la que se crea el proyecto:

- Es posible que no pueda acceder a los proyectos, evaluaciones y otros informes de Azure Migrate durante la interrupción. Sin embargo, puede usar las copias sin conexión que exportó.
- Cualquier replicación en curso o migración se pausa y es posible que tenga que reiniciarla después de la interrupción.

¿Cómo puedo eliminar un proyecto?

Obtenga información sobre cómo eliminar un proyecto.

¿Se puede mover un recurso de Azure Migrate?

No, Azure Migrate no admite el movimiento de recursos. Para mover recursos creados por Azure Migrate, considere la posibilidad de crear un nuevo proyecto en la región deseada.

¿Azure Migrate sigue admitiendo la migración de VDI?

Sí, Azure Migrate admite migraciones de infraestructura de escritorio virtual (VDI). Para más información, consulte el blog de la comunidad tecnológica.

¿Puedo usar Azure Migrate para mover máquinas virtuales de Azure VMware Solution (AVS) a Azure?

Sí, puede usar Azure Migrate para mover máquinas virtuales (VM) de Azure VMware Solution (AVS) a Azure. Azure Migrate admite la migración de máquinas virtuales de VMware desde entornos locales y AVS a Azure. Pero no admite la migración de máquinas virtuales de Azure de nuevo a AVS.

¿Qué aptitudes y conocimientos ayudan a administrar la migración de forma eficaz con Azure Migrate?

Los clientes que quieran migrar deben estar familiarizados con Azure Portal y comprender los conceptos básicos de migración, como las 6R. También necesitan las siguientes aptitudes técnicas, que varían en función de lo que migran:

- **Migración de** aplicaciones: comprenda la arquitectura de la aplicación, las dependencias y la refactorización de código.
- **Migración** a la nube: conozca los roles, las redes y la seguridad de Azure.
- **Migración de infraestructura**: trabaje con virtualización (VMware, Hyper-V), servidores, almacenamiento y copia de seguridad.
- **Scripting**: use Bash, PowerShell o Python para automatizar partes de la migración y las pruebas.
- **Integración**: Use APIs, middleware y herramientas como MuleSoft o Boomi para mantener la interoperabilidad.

¿Cuáles son los enfoques de migración disponibles en Azure Migrate, como el realojo, cambiar de plataforma o la refactorización?

Azure Migrate ahora ofrece opciones para crear casos empresariales e informes de evaluación para cargas de trabajo que los usuarios desean rehospedar, replatar o refactorizar.

Para la ejecución real de la migración, Azure Migrate admite el rehospedaje (también conocido como migrar mediante lift-and-shift).

Para rehospedaje:

- **Migración del servidor**: ayuda a mover máquinas virtuales (VMware, Hyper-V o servidores físicos).
- **Servicio de migración de bases de datos (DMS):** admite la migración de SQL Server y otras bases de datos.

Este enfoque funciona mejor para:

- Aplicaciones heredadas
- Migraciones rápidas con pruebas mínimas
- Proyectos de migrar mediante lift-and-shift sensibles al tiempo
- Aplicaciones con código complejo que todavía se ejecutan bien en Azure

Pasos siguientes

Lea la introducción a Azure Migrate.

Last updated on 03/09/2025

Confiabilidad en Azure Site Recovery

Azure Site Recovery es un servicio de replicación y conmutación por error administrado para máquinas virtuales (VM) que mantiene las cargas de trabajo disponibles durante las interrupciones. Replica continuamente las cargas de trabajo de los sitios primarios a las ubicaciones secundarias y limita la pérdida de datos y el tiempo de inactividad. Durante el mantenimiento planificado o interrupciones inesperadas, gestiona la conmutación por error y el retorno al estado operativo normal. Este servicio admite la recuperación ante desastres (DR) para entornos locales y máquinas virtuales de Azure, lo que ayuda a las organizaciones a mantener la continuidad empresarial.

Al usar Azure, relibilidad es una responsabilidad compartida. Microsoft proporciona una variedad de funcionalidades para admitir resistencia y recuperación. Es responsable de comprender cómo funcionan esas funcionalidades dentro de todos los servicios que usa y de seleccionar las funcionalidades que necesita para cumplir los objetivos empresariales y los objetivos de tiempo de actividad.

En este artículo se describe cómo hacer que Site Recovery sea resistente a varias posibles interrupciones y problemas, incluidos errores transitorios, interrupciones de zona de disponibilidad y interrupciones de regiones. También resalta información clave sobre el acuerdo de nivel de servicio (SLA) de Site Recovery.

> **Nota:**
>
> En este artículo se describe cómo el servicio Site Recovery es resistente o cómo puede hacer que sea resistente a varios problemas. No explica cómo usar Site Recovery para proteger las máquinas virtuales u otros recursos. Para obtener más información, consulte Acerca de Site Recovery.

Recomendaciones de implementación de producción para la confiabilidad

Al usar Site Recovery con cargas de trabajo de producción, se recomienda realizar estas acciones:

- ✔ Implemente la bóveda de Recovery Services en la región de destino para replicación.

- ✓ Para la recuperación ante desastres de Azure a Azure, use la característica de alta renovación de Site Recovery para las máquinas virtuales que tienen una alta tasa de cambios en los datos. El soporte de alta rotación mejora el objetivo de punto de recuperación (RPO) y permite la replicación para muchas cargas de trabajo de bases de datos a gran escala.
- ✓ Para la recuperación ante desastres de Azure a Azure, configure la cuenta de almacenamiento en caché para usar el almacenamiento con redundancia de zona (ZRS).
- ✓ Realice pruebas de conmutación por error periódicamente como parte de los simulacros de recuperación ante desastres. Ejecute simulacros de recuperación ante desastres cada trimestre o semestralmente para comprobar que los procesos de replicación y conmutación por error son correctos.
- ✓ Use reservas de capacidad a petición para asegurar que los recursos de computación están disponibles en la región de destino para la conmutación por falla.
- ✓ Habilite las actualizaciones automáticas para agentes de movilidad.
- ✓ Supervise el estado de la replicación y configure alertas para que se le notifique si se produce un problema.

Introducción a la arquitectura de confiabilidad

Cuando se usa Site Recovery, se define un *origen* y un *destino*, que representan las máquinas virtuales replicadas:

- El *origen* puede ser una máquina virtual de Azure o una máquina virtual o un servidor de otro origen compatible, incluidos los servidores físicos locales, las máquinas virtuales de VMware y las máquinas virtuales de Hyper-V.
- El *target* siempre es una máquina virtual de Azure. Para la replicación de máquinas virtuales de Azure a Azure, el destino puede ser una región o zona de disponibilidad diferente de la máquina virtual de origen.

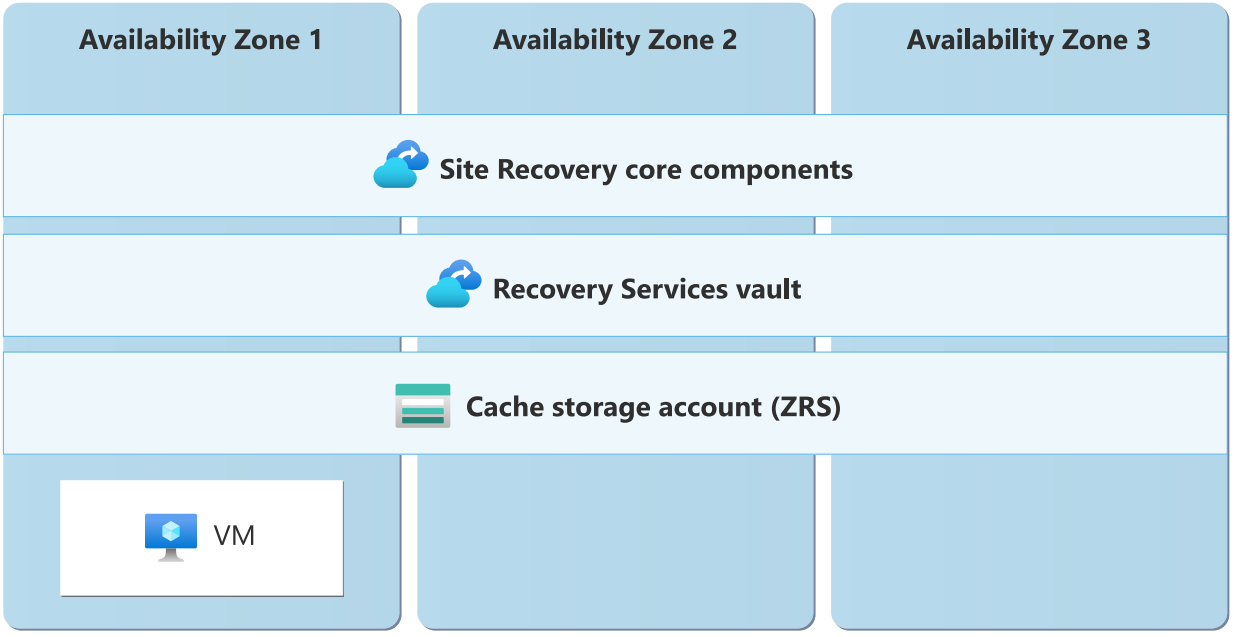
Usted es responsable de implementar y configurar recursos y opciones relacionadas, entre las que se incluyen:

- *Recovery Services vault*, que Site Recovery usa para almacenar las opciones de configuración de replicación. La bóveda no guarda tus datos replicados. La configuración de redundancia del almacén no es importante para Site Recovery, pero es importante si usa el mismo almacén para Azure Backup.

Una bóveda puede incluir configuraciones adicionales, como:

- *Directiva de replicación*, que configura la frecuencia de las instantáneas y la duración de retención.
- *Un plan de recuperación*, que coordina el orden en el que las máquinas conmutan por error y pueden incluir scripts y acciones manuales. Los planes de recuperación son especialmente útiles para las workloads que tienen múltiples niveles, tales como los niveles de aplicación y base de datos, que deben realizar un failover en un orden específico.

- Para la replicación de Azure a Azure, una cuenta de almacenamiento *cache* que almacena una copia de los datos de origen en su región antes de que se replique en el destino. La configuración de redundancia de la cuenta de almacenamiento en caché puede afectar a la confiabilidad durante una interrupción de la zona de disponibilidad.



> **Nota:**
>
> Esta guía se centra en la confiabilidad de los componentes basados en Azure de Site Recovery y en la relación de replicación. Si replica datos o máquinas virtuales desde un entorno local u otro proveedor de nube, considere la confiabilidad de los componentes fuera de Azure.

Para obtener más información sobre los componentes que implemente, consulte los siguientes artículos:

- Arquitectura de DR de Azure a Azure
- Arquitectura de recuperación ante desastres de Hyper-V a Azure
- Arquitectura de recuperación ante desastres de VMware a Azure
- Arquitectura de recuperación ante desastres de servidor físico hacia Azure

El servicio principal Site Recovery se ejecuta en la infraestructura que Administra Microsoft. En este artículo se hace referencia a estos componentes colectivamente como el *servicio principal de Site Recovery*.

Resistencia a errores transitorios

Los errores transitorios son errores breves e intermitentes en los componentes. Se producen con frecuencia en un entorno distribuido como la nube y son una parte normal de las operaciones. Los errores transitorios se corrigen después de un breve período de tiempo. Es importante que las aplicaciones puedan controlar errores transitorios, normalmente mediante el reintento de solicitudes afectadas.

Todas las aplicaciones hospedadas en la nube deben seguir las instrucciones de control de errores transitorios de Azure cuando se comunican con cualquier API, bases de datos y otros componentes hospedados en la nube. Para obtener más información, consulte Recomendaciones para controlar errores transitorios.

Site Recovery controla automáticamente los errores transitorios que se producen durante el proceso de replicación mediante el reintento de sus operaciones. No es necesario configurar el control de errores transitorios para Site Recovery.

Resistencia a errores de zona de disponibilidad

Availability zones son grupos de centros de datos físicamente independientes dentro de una región de Azure. Cuando una zona falla, los servicios pueden transferirse a una de las zonas restantes.

Para comprender cómo se comporta la replicación de Site Recovery durante los errores de zona de disponibilidad, debe tener en cuenta los siguientes componentes de servicio:

- **Servicio Core Site Recovery:** El servicio principal de Site Recovery está diseñado para ser resistente a los errores de zona de disponibilidad en las regiones admitidas. Los componentes internos del servicio apoyan automáticamente la redundancia de zona sin necesidad de configuración por parte del cliente.

- **Bóveda de Recovery Services:** La bóveda almacena los datos de configuración. En las regiones en las que Site Recovery admite la resiliencia de zona, los datos de configuración del almacén también son resilientes a la zona.

- **Cuenta de almacenamiento en caché:** Para la replicación de Azure a Azure, usted es responsable de asegurarse de que la cuenta de almacenamiento de caché sea redundante a nivel de zona mediante su implementación en el nivel ZRS.

  Si usa el nivel de replicación de Azure Storage con almacenamiento redundante local (LRS) para su cuenta de almacenamiento en caché y falla una zona, Site Recovery podría no replicar los datos modificados recientemente a su destino.

> ⓘ **Nota:**
>
> Site Recovery puede ayudarle a conmutar por error entre máquinas virtuales en diferentes zonas de disponibilidad. Para más información, consulte **Habilitar la recuperación ante desastres de máquinas virtuales de Azure entre zonas de disponibilidad**.

Requisitos

Compatibilidad con regiones:

- **Almacenes principales de Site Recovery Service y Recovery Services:** Site Recovery es resistente a la zona en las siguientes regiones.

  Expandir tabla

Americas	Europa	Oriente Medio	Asia Pacífico
Centro de Chile	Este de Austria	Israel Central	Centro de Indonesia
Centro de México	Norte de Italia		Japón Occidental
Oeste de EE. UU. 3	Centro de Polonia		Oeste de Malasia
	Centro de España		Norte de Nueva Zelanda

  Site Recovery está implementando compatibilidad con zonas de disponibilidad en todas las regiones habilitadas para zonas de disponibilidad. En las regiones que no aparecen en la tabla anterior como resistentes a la zona, los errores de zona pueden afectar a las operaciones.

- **Cuenta de almacenamiento en caché:** Puede implementar una cuenta de almacenamiento de ZRS en todas las regiones habilitadas para zonas de disponibilidad.

Cost

Site Recovery se factura en función del número de instancias de máquina virtual protegidas, independientemente de su configuración de zona de disponibilidad. Para obtener más información, consulte Precios de Site Recovery.

Configurar soporte de zonas de disponibilidad

- **Core Site Recovery service:** No se configura la resistencia de zona en el servicio principal Site Recovery. Microsoft proporciona resiliencia de zona en las regiones compatibles.

  Si Microsoft habilita la resistencia de zona en una región más adelante, los recursos de Site Recovery se benefician automáticamente de la resistencia de zona. No es necesario realizar ninguna acción.

- **Almacén de Recovery Services:** Los almacenes de Recovery Services tienen una configuración de redundancia, pero Site Recovery no usa la configuración de redundancia del almacén. No es necesario configurar la bóveda para la redundancia de zona cuando usas Site Recovery.

- **Cache storage account:** Cuando se usa la replicación de Azure a Azure, es responsable de crear la cuenta de almacenamiento de caché y de configurarla con el nivel adecuado de redundancia. Para que sea redundante de zona, configúrelo para el tipo de replicación ZRS. Para obtener más información, consulte Reliability en Azure Blob Storage.

Comportamiento cuando todas las zonas están en buen estado

En esta sección se describe qué esperar cuando se usa Site Recovery en una región con compatibilidad con zonas de disponibilidad para el servicio principal, la cuenta de almacenamiento en caché está configurada para usar ZRS y todas las zonas de disponibilidad están operativas.

- **Operación entre zonas:** El proceso de replicación puede usar la infraestructura en varias zonas de disponibilidad para desencadenar y ejecutar trabajos de replicación. El servicio administra esta infraestructura de forma transparente.

- **Replicación de datos entre zonas:** Site Recovery y Storage controlan la replicación de datos de zona:

- *Configuración de Site Recovery:* Site Recovery replica los datos de configuración entre zonas, incluso si configura el almacén para usar LRS.
- *Cuenta de almacenamiento en caché:* Si configura la cuenta de almacenamiento de caché para que use ZRS, Storage replica de forma sincrónica los datos almacenados en caché entre zonas.

Comportamiento durante un fallo de zona

En esta sección se describe qué esperar cuando se usa Site Recovery en una región con compatibilidad con la zona de disponibilidad para el servicio principal, la cuenta de almacenamiento en caché está configurada para usar ZRS y se produce una interrupción de zona de disponibilidad.

> ⓘ **Nota:**
>
> Si la zona con errores contiene la máquina virtual de origen, usted es responsable de iniciar la conmutación por error al destino. Para obtener más información, consulte los artículos siguientes:
>
> - Conmutar las máquinas virtuales de Azure a una región secundaria
> - Conmutación por error de máquinas virtuales de VMware
> - Conmutar las máquinas virtuales de Hyper-V a Azure

- **Detección y respuesta:** La plataforma Site Recovery detecta automáticamente errores en una zona de disponibilidad e inicia una respuesta. No es necesario iniciar manualmente una conmutación por error de zona para el servicio principal de Site Recovery. Sin embargo, si la interrupción de zona afecta a la máquina virtual de origen, es posible que tenga que iniciar la conmutación por error de la máquina virtual.
- **Notificación:** Microsoft no le notifica automáticamente cuando una zona está inactiva. Sin embargo, puede usar Azure Service Health para comprender el estado general del servicio, incluidos los errores de zona, y puede configurar alertas de Service Health para notificarle problemas.
- **Solicitudes activas:** El efecto en los trabajos de replicación activa depende del tipo de replicación:
  - *Replicación de zona a zona y región a región de máquinas virtuales de Azure:* Si la instancia de origen o de destino está en la zona con errores, la replicación se detiene

hasta que ambas instancias estén disponibles de nuevo.

Si la zona con errores no contiene la máquina virtual de origen o de destino y configuró la cuenta de almacenamiento en caché para usar ZRS, la replicación continúa ejecutándose.

- *On-premises para Azure:* Si la instancia de destino está en la zona con errores, la replicación se detiene hasta que la instancia esté disponible de nuevo.

  Si la zona con errores no contiene la máquina virtual de destino, la replicación continúa ejecutándose.

- **Pérdida de datos esperada:** No se espera ninguna pérdida de datos durante un error de zona.
- **Tiempo de inactividad esperado:** Si la zona con errores contiene la máquina virtual de origen o de destino, la replicación se detiene hasta que ambas instancias estén disponibles de nuevo.
- **Redistribución:** Site Recovery y Storage se adaptan automáticamente a errores de zona:
  - *Servicio Core Site Recovery:* El servicio principal de Site Recovery usa automáticamente la infraestructura en zonas de disponibilidad correctas para realizar la replicación. No es necesario realizar ninguna acción.
  - *Cuenta de almacenamiento en caché:* El almacenamiento enruta automáticamente las solicitudes de datos de caché a zonas correctas.

Recuperación de zona

Cuando se recupera la zona de disponibilidad afectada, Site Recovery reanuda automáticamente los trabajos de replicación que se pausan durante la interrupción de la zona.

Es responsable de iniciar la conmutación por recuperación para servidores o máquinas virtuales conmutadas por error durante la interrupción de la zona. Para obtener más información, consulte los artículos siguientes:

- *Replicación de zona a zona y región a región de máquinas virtuales de Azure:*recuperar una máquina virtual de Azure a la región primaria
- *Replicación local a Azure:*

- *Replicación física a Azure:*arquitectura de DR (recuperación ante desastres) de servidor físico a Azure
- *Replicación de Hyper-V a Azure:*Arquitectura de recuperación ante desastres de Hyper-V a Azure
- *Replicación de VMware a Azure:*Acerca del fallo de recuperación ante desastres y el retorno después de la recuperación en el entorno local

Prueba de fallos de zona

La plataforma Site Recovery administra la resistencia de zona para sus componentes internos. Esta característica está totalmente administrada, por lo que no es necesario iniciar ni validar los procesos de error de zona de disponibilidad.

Es importante realizar simulacros regulares de recuperación ante desastres, que deben probar la conmutación por error de la VM y los procedimientos generales de respuesta. Diseñe los simulacros de recuperación ante desastres para evitar el impacto en el entorno de producción. Para obtener más información, consulte los artículos siguientes:

- *Replicación de zona a zona y de región a región de máquinas virtuales de Azure:*Realizar una prueba de recuperación ante desastres para máquinas virtuales de Azure
- *Replicación local a Azure:*
  - *Replicación física a Azure:*Realizar un simulacro de DR en Azure
  - *Replicación de Hyper-V a Azure:*Ejecución de una prueba de DR en Azure
  - *Replicación de VMware a Azure:*Realizar un simulacro de DR en Azure

Resistencia a errores en toda la región

Para la replicación de Azure a Azure, Site Recovery proporciona resistencia a los fallos de región al habilitar la conmutación por error de las máquinas virtuales (VMs) a una región de destino saludable. Para más información, consulte Replicar máquinas virtuales de Azure a otra región de Azure.

Consideraciones

- **Región del almacén:** Implemente un almacén de Recovery Services en una región de Azure específica que seleccione. La región de la bóveda es importante. La replicación continúa

durante una interrupción en la región de la bóveda. Sin embargo, no puede realizar operaciones de administración de Site Recovery, incluida la conmutación por error y el restablecimiento, hasta que la región se recupere.

La implementación del almacén en la región de destino ayuda a garantizar que las operaciones de conmutación por error y recuperación permanezcan disponibles durante una interrupción de la región de origen. También evita que una interrupción en una tercera región afecte las operaciones de conmutación y recuperación.

> ⓘ **Nota:**
>
> Si su bóveda está en la región que normalmente usa como región de destino, entonces, después de la conmutación por error y restablecimiento de la replicación, esa región se convierte en su nueva región de origen. Si esa región experimenta posteriormente un problema, posiblemente no pueda realizar la reversión de la conmutación hasta que ambas regiones estén funcionales.

- **Reservas de capacidad:** Es responsable de comprobar que la región de destino admite los tipos de máquina virtual que necesita y que tiene capacidad disponible para la carga de trabajo. Se recomienda usar reservas de capacidad a petición para asegurarse de que los recursos de cómputo están disponibles para su carga de trabajo si se produce una conmutación por error.

Configuración de la compatibilidad con varias regiones

- **Almacén de Recovery Services:** Debe seleccionar la región del almacén. Para más información, consulte Consideraciones.

  Las bóvedas de Servicios de Recuperación tienen una configuración de redundancia, pero Site Recovery no utiliza la configuración de redundancia de la bóveda. No es necesario configurar la bóveda para la redundancia geográfica al usar Site Recovery.

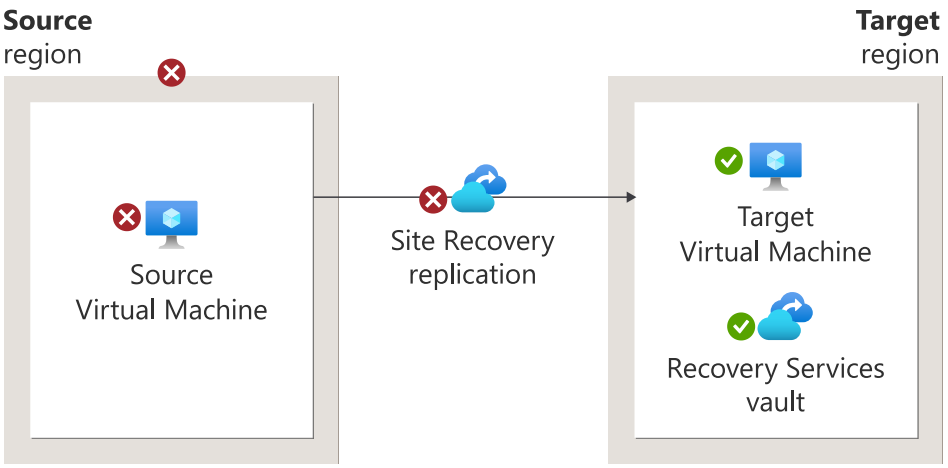
- **Cuenta de almacenamiento en caché:** La cuenta de almacenamiento en caché solo se usa como ubicación temporal para los datos antes de que se replique, por lo que no debe configurarla para usar el almacenamiento con redundancia geográfica (GRS).

Comportamiento durante una falla de región

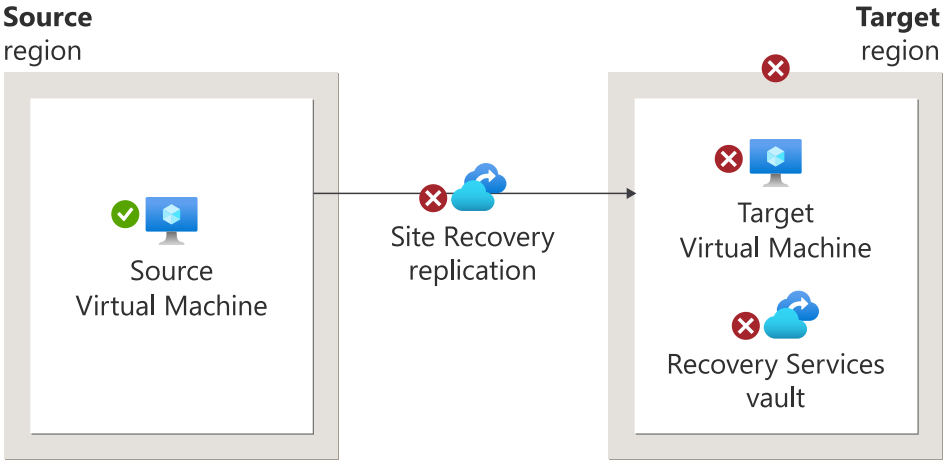
El comportamiento específico del servicio principal de Site Recovery durante un error de región depende de la región en la que experimente el error:

- **Error en la región de origen:** Para la replicación de Azure a Azure, puede activar una conmutación por error automática cuando la región de origen no esté disponible.

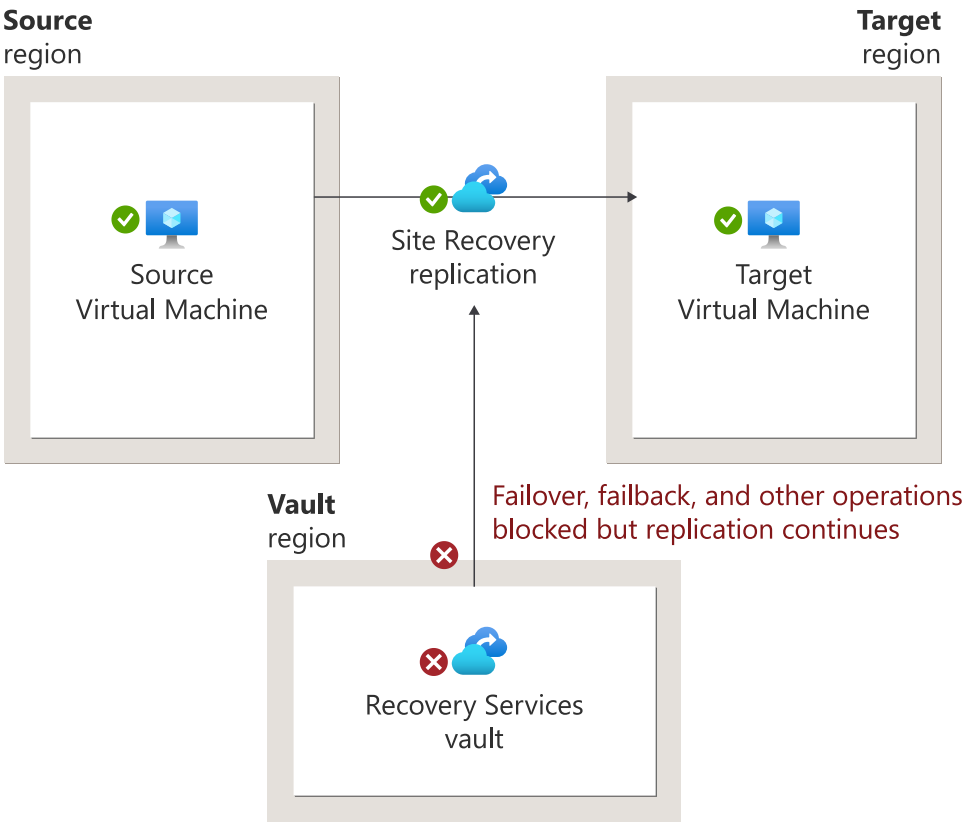
  Dado que la región de origen no está disponible, la replicación se detiene hasta que la máquina virtual de la región de origen está en buen estado.



- **Error en la región de destino:** Dado que la región de destino no está disponible, la replicación se detiene y no se puede recuperar el destino hasta que la región esté operativa.



- **Fallo en la región que contiene la bóveda:** Si despliega la bóveda en una tercera región (no en la región de origen o de destino) y esa región experimenta un fallo, Site Recovery continuará replicando sus datos. Sin embargo, no puede iniciar ninguna operación, incluida la conmutación por error o la reversión, hasta que la bóveda esté en buen estado.



Recuperación de regiones

Es responsable de iniciar la reversión para servidores o máquinas virtuales que conmutó durante la interrupción de la región. Para obtener más información, consulte los artículos siguientes:

- *Replicación de zona a zona y de región a región de máquinas virtuales de Azure:*Recuperar máquina virtual de Azure a la región primaria
- *Replicación local a Azure:*
  - *Replicación de servidor físico a Azure:*arquitectura de recuperación ante desastres
  - *Replicación de Hyper-V a Azure:*Arquitectura de recuperación ante desastres de Hyper-V a Azure
  - *Replicación de VMware a Azure:*conmutación por error y restablecimiento de la recuperación ante desastres en el sitio

Prueba de fallos de región

Es importante realizar simulacros de recuperación ante desastres regulares que prueben la conmutación por error de la VM y los procedimientos generales de respuesta. Diseñe los

simulacros de recuperación ante desastres para evitar el impacto en el entorno de producción. Para obtener más información, consulte los artículos siguientes:

- *Replicación de zona a zona y de región a región de máquinas virtuales de Azure:*Ejecutar un simulacro de recuperación ante desastres para máquinas virtuales de Azure
- *Replicación local a Azure:*
  - *Replicación física a Azure:*Ejecutar un simulacro de DR en Azure
  - *Replicación de Hyper-V a Azure:*Realizar una prueba de DR en Azure
  - *Replicación de VMware a Azure:*Ejecución de un simulacro de recuperación ante desastres en Azure

Resistencia a problemas de configuración y replicación

Una solución de recuperación ante desastres solo es confiable cuando sabe que funciona antes de que se produzca un desastre. Supervise Site Recovery para detectar problemas como errores de configuración o problemas de mantenimiento de replicación de máquinas virtuales. Para obtener más información, consulte Supervisión de Site Recovery.

Se recomienda configurar alertas de Azure Monitor para que se le informe sobre los problemas con el estado de replicación. Para más información, consulte Alertas integradas de Azure Monitor para Site Recovery.

Resistencia al mantenimiento del servicio

Azure administra automáticamente las actualizaciones y el mantenimiento del servicio principal Site Recovery. Las operaciones de mantenimiento no requieren tiempo de inactividad y no interrumpen la replicación de las máquinas virtuales y los servidores.

Sin embargo, es responsable de aplicar actualizaciones a los componentes de Site Recovery en las máquinas virtuales y servidores, incluido el agente de movilidad cuando sea necesario.

> **ⓘ Importante**
>
> Se recomienda encarecidamente habilitar las actualizaciones automáticas para los agentes. Si la versión del agente queda más de cuatro versiones atrás, la replicación se desactiva y la capacidad de recuperación de la carga de trabajo se ve comprometida.

Para obtener más información, consulte actualizaciones de Service en Site Recovery.

Acuerdo de nivel de servicio

El acuerdo de nivel de servicio (SLA) para Azure servicios describe la disponibilidad esperada de cada servicio y las condiciones que la solución debe cumplir para lograr esa expectativa de disponibilidad. Para obtener más información, vea SLAs for online services.

Para Site Recovery, los Acuerdos de Nivel de Servicio independientes cubren:

- **Disponibilidad del servicio**, lo que significa que Site Recovery está disponible para realizar una conmutación por error de las instancias protegidas. Una instancia protegida es una máquina virtual o un servidor físico que se replica en una ubicación secundaria. Para poder optar a este Acuerdo de Nivel de Servicio, debe reintentar los intentos fallidos de conmutación por error al menos cada 30 minutos.

- **Objetivo de tiempo de recuperación (RTO),** que es el tiempo desde que se desencadena una conmutación por fallo (o cuando los scripts lo desencadenan) hasta cuando se ejecuta la máquina virtual de destino. Esta vez excluye acciones manuales o ejecución de scripts.

El Acuerdo de Nivel de Servicio solo proporciona créditos de servicio cuando la región secundaria tiene suficiente capacidad de proceso.

Contenido relacionado

- Acerca de Site Recovery
- Confiabilidad en Azure

ⓘ **Nota:** El autor ha creado este artículo con ayuda de inteligencia artificial. Más información

---

Last updated on 24/04/2026

Confiabilidad en Azure Application Gateway v2

Azure Application Gateway es un equilibrador de carga de tráfico web que puede usar para administrar el tráfico a las aplicaciones web. Proporciona funciones avanzadas como el escalado automático, la redundancia de zona, las direcciones IP virtuales estáticas (VIP) y la integración del firewall de aplicaciones web (WAF) para ofrecer servicios de entrega de aplicaciones de alta disponibilidad y seguros.

Cuando se usa Azure, la confiabilidad es una responsabilidad compartida. Microsoft proporciona una variedad de funcionalidades para admitir resistencia y recuperación. Es responsable de comprender cómo funcionan esas funcionalidades dentro de todos los servicios que usa y de seleccionar las funcionalidades que necesita para cumplir los objetivos empresariales y los objetivos de tiempo de actividad.

En este artículo se describe cómo hacer que Application Gateway v2 sea resistente a una variedad de posibles interrupciones y problemas, incluidos errores transitorios, interrupciones de zona de disponibilidad y interrupciones en regiones. También se resalta cierta información clave sobre el acuerdo de nivel de servicio (SLA) de Application Gateway v2.

> **ⓘ Importante**
>
> La confiabilidad de la solución general depende de la configuración de los servidores back-end a los que Application Gateway enruta el tráfico. En función de la solución, pueden ser máquinas virtuales (VM) de Azure, Microsoft Azure Virtual Machines de Azure, Azure App Services o puntos de conexión externos.
>
> Los servidores back-end no están en el ámbito de este artículo, pero sus configuraciones de disponibilidad afectan directamente a la resistencia de la aplicación. Revise las guías de confiabilidad de todos los servicios de Azure de la solución para comprender cómo admite cada uno los requisitos de confiabilidad. Al asegurarse de que los servidores de back-end también están configurados para alta disponibilidad y redundancia de zona, puede lograr una confiabilidad de un extremo a otro para la aplicación.

Recomendaciones de implementación de producción

Para obtener información sobre cómo implementar Application Gateway para admitir los requisitos de confiabilidad de la solución y cómo afecta la confiabilidad a otros aspectos de la

arquitectura, consulte Procedimientos recomendados de arquitectura para Application Gateway en Azure Well-Architected Framework.

Introducción a la arquitectura de confiabilidad

Application Gateway es un servicio administrado. Es importante comprender algunos elementos clave de la arquitectura de servicio para que pueda tomar decisiones de confiabilidad informadas.

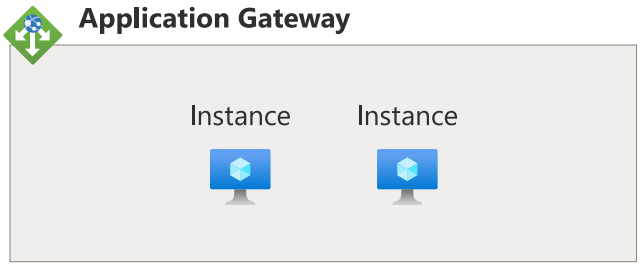
- **Puerta de Enlace:** Al implementar Application Gateway, se implementa una *puerta de enlace*, que es el recurso de Azure con el que trabaja.

- **Instancias:** Una *instancia* es una sola unidad de la puerta de enlace. Una puerta de enlace tiene varias instancias. Cada instancia tiene su propia dirección IP privada, controla el tráfico y enruta las solicitudes.

  > **Importante**
  >
  > Para lograr una alta disponibilidad, cada puerta de enlace siempre se crea con un mínimo de dos instancias. Sin embargo, Azure Portal puede indicar que la puerta de enlace tiene una sola instancia, aunque realmente tenga dos.

  La plataforma administra automáticamente la creación de instancias, la supervisión del estado y la sustitución de instancias incorrectas. También administra la eliminación correcta de instancias durante los eventos de escalado horizontal. Este proceso se conoce como *drenaje de conexiones*.

  En el diagrama siguiente se muestra una puerta de enlace con dos instancias:



Para distribuir instancias entre varias zonas de disponibilidad y, por tanto, aumentar la redundancia y la disponibilidad durante los errores del centro de datos, puede habilitar la redundancia de zona.

- **Escalado:** un aspecto importante de la confiabilidad de una puerta de enlace es cómo se escala para satisfacer la demanda de tráfico. Si una puerta de enlace no se escala

correctamente, el tráfico no puede fluir y la aplicación puede experimentar tiempo de inactividad. Puede configurar Application Gateway para que use uno de los siguientes modos de escalado.

  - *Escalado automático:* ajusta automáticamente el número de instancias dentro de un intervalo que especifique. El escalado automático escala el número de instancias en función de la demanda de tráfico actual.
  - *Escalado manual:* requiere que especifique un número exacto de instancias. Es responsable de seleccionar un recuento de instancias que cumpla la demanda de tráfico.

  Una *unidad de capacidad* representa una cantidad de capacidad que la puerta de enlace puede procesar. Una unidad de capacidad es una medida sintética que incorpora el tráfico, el número de conexiones y los recursos de proceso. Cada instancia puede controlar al menos 10 unidades de capacidad. Para más información, vea Escalado de Application Gateway v2 y WAF v2.

- **Sondeos de estado:** Application Gateway usa sondeos de estado para supervisar continuamente los servidores back-end, como los servidores de aplicaciones individuales. El tráfico se puede redirigir automáticamente a servidores de back-end correctos cuando se detectan servidores incorrectos.

Resistencia a errores transitorios

Los errores transitorios son errores breves e intermitentes en los componentes. Se producen con frecuencia en un entorno distribuido como la nube y son una parte normal de las operaciones. Los errores transitorios se corrigen después de un breve período de tiempo. Es importante que las aplicaciones puedan controlar errores transitorios, normalmente mediante el reintento de solicitudes afectadas.

Todas las aplicaciones hospedadas en la nube deben seguir las instrucciones de control de errores transitorios de Azure cuando se comunican con cualquier API, bases de datos y otros componentes hospedados en la nube. Para obtener más información, consulte Recomendaciones para controlar errores transitorios.

Al usar Application Gateway, tenga en cuenta los procedimientos recomendados siguientes:

- **Implementación de la lógica de reintento.** Los clientes deben implementar mecanismos de reintento adecuados para errores de conexión transitorios.
- **Configure sondeos de estado con tolerancia.** Configure los sondeos de estado a fin de permitir un período de gracia para errores transitorios. Los sondeos de estado se pueden

configurar con un *umbral incorrecto*, que especifica el número de intentos de conexión con error consecutivos que deben desencadenar que el servidor de back-end se marque como incorrecto. El valor predeterminado de tres garantiza que los errores transitorios de los servidores back-end no desencadenen Application Gateway para marcar innecesariamente el servidor como incorrecto.

Resistencia a errores de zona de disponibilidad

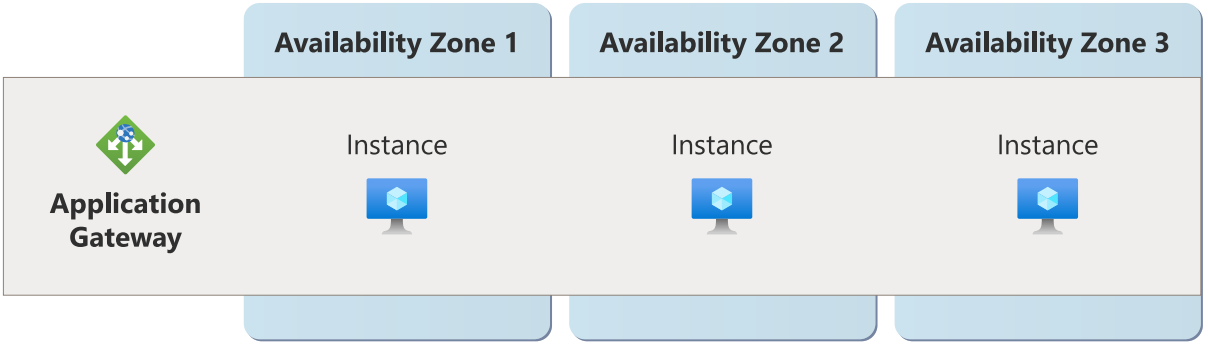
Las zonas de disponibilidad son grupos físicamente independientes de centros de datos dentro de una región de Azure. Cuando una zona falla, los servicios pueden transferirse a una de las zonas restantes.

Application Gateway proporciona dos tipos de compatibilidad con zonas de disponibilidad al implementar una puerta de enlace de Standard_v2 o WAF_v2 en una región admitida.

- *Con redundancia de zona:* Azure distribuye automáticamente las instancias en dos o más zonas de disponibilidad.

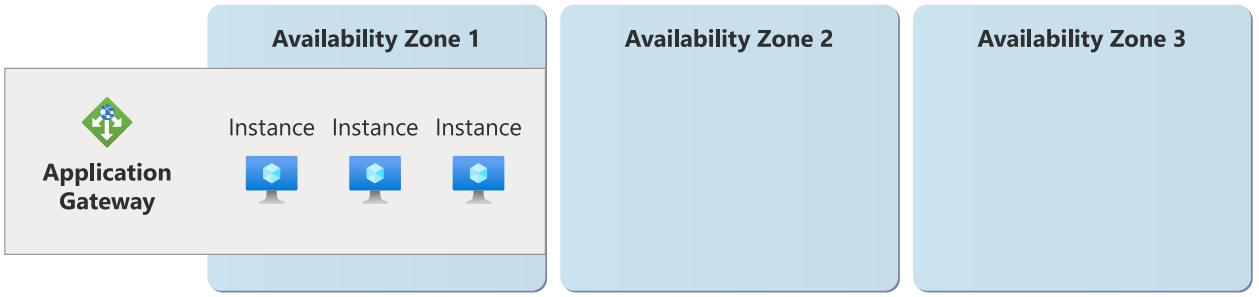
  Al implementar un nuevo gateway, es redundante en zona por defecto.

  En el diagrama siguiente se muestra una puerta de enlace con redundancia de zona con tres instancias distribuidas entre tres zonas de disponibilidad:



- *Zonal:* Azure implementa todas las instancias de Application Gateway en una sola zona que seleccione en la región de Azure elegida.

  En el diagrama siguiente se muestra una puerta de enlace zonal con tres instancias que se implementan en la misma zona de disponibilidad:



> ⓘ **Importante**
>
> La asignación a una sola zona de disponibilidad solo se recomienda cuando la latencia entre zonas es demasiado alta para sus necesidades y cuando se comprueba que la latencia no cumple sus requisitos. Por sí sola, una puerta de enlace zonal no proporciona resiliencia ante una interrupción del servicio en una zona de disponibilidad. Para mejorar la resistencia de una implementación de Application Gateway zonal, debe implementar explícitamente puertas de enlace independientes en varias zonas de disponibilidad y configurar el enrutamiento de tráfico y la conmutación por error.

Requisitos

- **Compatibilidad con regiones:** Application Gateway admite zonas de disponibilidad para los niveles de Standard_v2 y WAF_v2 en todas las regiones de Azure que admiten zonas de disponibilidad.
- **SKU:** Debe usar la Standard_v2 o WAF_v2 SKU para habilitar la compatibilidad con zonas de disponibilidad.
- **Recuento de instancias:** La puerta de enlace debe tener al menos dos instancias.

> ⓘ **Nota:**
>
> Todas las puertas de enlace tienen un mínimo de dos instancias para garantizar alta disponibilidad. Incluso si Azure Portal indica que la puerta de enlace tiene una sola instancia, internamente siempre se crea con un mínimo de dos instancias.

Consideraciones

Las puertas de enlace con redundancia de zona se distribuyen entre dos o más zonas de disponibilidad en la región. Por ejemplo, en una región que tiene tres zonas de disponibilidad, una implementación de Application Gateway v2 con redundancia de zona tiene instancias en al

menos dos de esas zonas. En función de la capacidad regional y las decisiones de plataforma, las instancias se pueden distribuir entre dos zonas o tres zonas.

Cost

La compatibilidad con la zona de disponibilidad para Application Gateway v2 no incurre en cargos adicionales más allá de los precios de la unidad de capacidad estándar. Para obtener más información sobre los precios, consulte Descripción de los precios de Application Gateway y WAF y Application Gateway.

Configurar soporte de zonas de disponibilidad

En esta sección se explica cómo configurar la compatibilidad con zonas de disponibilidad para las puertas de enlace.

- **Cree una nueva puerta de enlace con compatibilidad con la zona de disponibilidad.** El enfoque que se usa para configurar zonas de disponibilidad depende de si desea crear una puerta de enlace con redundancia de zona o zonal.
  - *Con redundancia de zona:* las puertas de enlace se crean como con redundancia de zona de manera predeterminada. Las instancias se distribuyen entre varias zonas de disponibilidad y pueden usar dos o más zonas en la región.

    Para implementar una nueva puerta de enlace, consulte Inicio rápido: tráfico web directo mediante Application Gateway: Azure Portal.

    Si usa la CLI de Azure, Azure PowerShell, Bicep, plantillas de Azure Resource Manager (plantillas de ARM) o Terraform, puede especificar opcionalmente las zonas de disponibilidad en las que implementar la puerta de enlace. Puede implementar una puerta de enlace con redundancia de zona si especifica dos o más zonas. Sin embargo, se recomienda omitir la lista de zonas para que la puerta de enlace pueda usar todas las zonas de disponibilidad, a menos que tenga una razón específica para no usar una zona específica.
  - *Zonal:* Puede implementar una puerta de enlace zonal mediante las siguientes herramientas.
    - *CLI de Azure:* debe seleccionar explícitamente zonas mediante el parámetro `--zones` en el comando `az network application-gateway create`. Para anclar la puerta de enlace a una sola zona, especifique el número de zona lógica.
    - *Azure PowerShell:* use el parámetro `-Zone` en el comando `New-AzApplicationGateway`. Para anclar la puerta de enlace a una sola zona, especifique el

número de zona lógica.

- *Plantillas de Bicep y ARM:* Configure la `zones` propiedad en la definición de recursos. Para anclar la puerta de enlace a una sola zona, especifique el número de zona lógica.

> ⓘ **Nota:**
>
> Cuando selecciona qué zonas de disponibilidad usar, en realidad está seleccionando la *zona de disponibilidad lógica*. Si implementa otros componentes de la carga de trabajo en una suscripción de Azure diferente, podrían usar un número de zona de disponibilidad lógica distinto para acceder a la misma zona de disponibilidad física. Para obtener más información, consulte zonas de disponibilidad física y lógica.

- **Cambie la configuración de zona de disponibilidad de una instancia de Application Gateway v2 existente.** Todas las puertas de enlace tienen redundancia de zona a menos que se configuren para que sean zonales.

  Si necesita pasar de una puerta de enlace con redundancia de zona a una configuración zonal, debe implementar una nueva puerta de enlace.

- **Deshabilite la compatibilidad con la zona de disponibilidad.** No se puede deshabilitar la compatibilidad con la zona de disponibilidad. Todas las puertas de enlace de regiones que admiten zonas de disponibilidad deben ser con redundancia de zona o zonales.

Planeamiento y administración de capacidad

Cuando planifique fallos de zona para una puerta de enlace de aplicaciones redundante por zonas o varias puertas de enlace zonales implementadas en varias zonas, tenga en cuenta que las instancias de las zonas supervivientes pueden experimentar un aumento de la carga a medida que se redistribuye el tráfico. Las conexiones pueden experimentar interrupciones breves que pueden durar unos segundos.

Para administrar la capacidad de forma eficaz, realice las siguientes acciones:

- **Use el escalado automático.** Configure el escalado automático con los recuentos máximos de instancias adecuados para controlar la posible redistribución del tráfico durante las interrupciones de zona.

  Si usa el escalado manual, a fin de prepararse para el error de zona de disponibilidad, considere la posibilidad de *sobreaprovisionar* el número de instancias de la puerta de

enlace. El aprovisionamiento excesivo permite a la solución tolerar cierto grado de pérdida de capacidad, al tiempo que sigue funcionando sin un rendimiento degradado. Para más información, consulte Administración de la capacidad con sobreaprovisionamiento.

- **Responder a los cambios en los patrones de tráfico.** Supervise las métricas de capacidad y ajuste los parámetros de escalado en función de los patrones de tráfico y los requisitos de rendimiento.

Comportamiento cuando todas las zonas están en buen estado

En la siguiente sección se describe qué se puede esperar cuando Application Gateway v2 se configura con compatibilidad con zonas de disponibilidad y todas las zonas de disponibilidad están operativas.

- **Enrutamiento de tráfico entre zonas**: Application Gateway distribuye automáticamente las solicitudes entrantes entre instancias de todas las zonas que usa la puerta de enlace. Esta configuración activa-activa garantiza un rendimiento y una distribución de carga óptimos en condiciones de funcionamiento normales.

- **Administración de instancias:** la plataforma administra automáticamente la colocación de instancias en las zonas que usa la puerta de enlace. Reemplaza las instancias con errores y mantiene el recuento de instancias configurados. El seguimiento de estado garantiza que solo las instancias correctas reciban tráfico.

Comportamiento durante un fallo de zona

En la siguiente sección se describe qué ocurre cuando Application Gateway v2 se configura con compatibilidad con zonas de disponibilidad y una o varias zonas de disponibilidad no están disponibles.

- **Detección y respuesta:** la responsabilidad de la detección y la respuesta depende de la configuración de zona de disponibilidad que usa la puerta de enlace.
  - *Con redundancia de zona:* Microsoft administra la detección de errores de zona e inicia automáticamente la conmutación por error. No se requiere ninguna acción del cliente.
  - *Zonal:* debe detectar la pérdida de una zona de disponibilidad e iniciar una conmutación por error a una puerta de enlace secundaria que cree en otra zona de disponibilidad.

- **Notificación**: Microsoft no le notifica automáticamente cuando una zona está inactiva. Sin embargo, puede usar Azure Resource Health para supervisar el estado de un recurso individual y puede configurar alertas de Resource Health para notificarle problemas. También puede usar Azure Service Health para comprender el estado general del servicio, incluidos los errores de zona, y puede configurar alertas de Service Health para notificarle problemas.

- **Solicitudes activas:** durante una interrupción de zona, se finalizan las solicitudes que procesan las instancias de esa zona. Los clientes deben reintentar las solicitudes siguiendo las instrucciones para controlar errores transitorios.

- **Pérdida de datos esperada:** no se espera que los errores de zona causen pérdida de datos porque Application Gateway es un servicio sin estado.

- **Tiempo de inactividad esperado:** El tiempo de inactividad que puede esperar depende de la configuración de zona de disponibilidad que usa la puerta de enlace.

  - *Con redundancia de zona:* Durante las interrupciones de zona, las conexiones pueden experimentar interrupciones breves que suelen durar unos segundos a medida que se redistribuye el tráfico.

  - *Zonal:* cuando una zona no está disponible, la puerta de enlace no está disponible hasta que se recupere la zona de disponibilidad.

- **Administración de instancias:** el comportamiento de administración de instancias depende de la configuración de zona de disponibilidad que usa la puerta de enlace.

  - *Con redundancia de zona:* la plataforma intenta mantener la capacidad de la puerta de enlace mediante la creación de instancias temporales en otras zonas de disponibilidad.

    Internamente, Application Gateway usa conjuntos de escalado de máquinas virtuales, que realizan el equilibrio de zona más indicado. Debido a este comportamiento, es posible que las operaciones de escalado no se produzcan cuando la capacidad no se pueda dividir uniformemente entre zonas (+/- 1 instancia).

  - *Zonal:* es responsable de crear instancias en zonas correctas si las necesita.

- **Re enrutamiento del tráfico:** El comportamiento se re direcciona del tráfico depende de la configuración de zona de disponibilidad que usa la puerta de enlace.

  - *Con redundancia de zona:* Application Gateway redistribuye inmediatamente el tráfico a las instancias de zonas correctas, incluidas las que se crean temporalmente.

  - *Zonal:* cuando una zona no está disponible, la puerta de enlace no está disponible. Si tiene una puerta de enlace secundaria en otra zona de disponibilidad, es responsable

de redirigir el tráfico a esa puerta de enlace secundaria.

Recuperación de zona

El comportamiento de recuperación de zona depende de la configuración de zona de disponibilidad que usa la puerta de enlace:

- *Con redundancia de zona:* Cuando se recupera la zona de disponibilidad afectada, Application Gateway realiza automáticamente las siguientes acciones:
  - Restaura instancias en la zona recuperada
  - Quita las instancias temporales que se hayan creado en otras zonas durante la interrupción
  - Vuelve a la distribución normal del tráfico en todas las zonas disponibles
- *Zonal:* es responsable de redirigir el tráfico a la puerta de enlace en la zona de disponibilidad original después de que se recupere la zona de disponibilidad.

Prueba de fallos de zona

Las opciones para probar errores de zona dependen de la configuración de zona de disponibilidad que usa la puerta de enlace.

- *Con redundancia de zona:* La plataforma Application Gateway administra completamente el enrutamiento del tráfico, la conmutación por error y la conmutación por recuperación para las puertas de enlace con redundancia de zona. Como Microsoft administra esta característica, no es necesario iniciar ni validar los procesos de error de zona de disponibilidad. La plataforma controla todos los escenarios de error de zona de forma transparente.
- *Zonal:* puede simular algunos aspectos del error de una zona de disponibilidad si detiene explícitamente una puerta de enlace. Al detener la puerta de enlace, puede probar cómo otros sistemas y equilibradores de carga controlan una interrupción en la puerta de enlace. Para más información, consulte Cómo detener e iniciar Application Gateway.

Resistencia a errores en toda la región

Application Gateway v2 es un servicio de una sola región. Si la región deja de estar disponible, la puerta de enlace tampoco está disponible.

Soluciones personalizadas de varias regiones para la resistencia

Para lograr la resistencia de varias regiones mediante Application Gateway v2, debe implementar puertas de enlace independientes en cada región deseada e implementar la administración del tráfico entre regiones. Es responsable de implementar y configurar cada una de las puertas de enlace, así como el enrutamiento del tráfico y la conmutación por error. Considere los siguientes puntos:

- Configure reglas y directivas coherentes de Application Gateway entre regiones. Puede definir la infraestructura como código (IaC) mediante herramientas como Bicep o Terraform para simplificar las implementaciones y configuraciones entre regiones.
- Implemente una solución de equilibrio de carga global que pueda enviar tráfico entre las puertas de enlace regionales. Los servicios globales de equilibrio de carga en Azure son Azure Traffic Manager y Azure Front Door. Cada servicio enruta el tráfico en función de las comprobaciones de estado, la proximidad geográfica o las métricas de rendimiento. Azure Front Door también proporciona una variedad de otras funcionalidades, incluida la protección contra ataques por denegación de servicio distribuido (DDoS), las funcionalidades de WAF y las reglas avanzadas y las características de enrutamiento.
- Más allá de la puerta de enlace, considere la posibilidad de replicar aplicaciones de back-end y datos entre regiones. Consulte las guías de confiabilidad de cada servicio de Azure para comprender los enfoques de implementación de varias regiones.

Para obtener un enfoque de ejemplo, consulte Uso de Application Gateway con Traffic Manager.

Copias de seguridad y restauración

Application Gateway v2 es un servicio sin estado que no requiere operaciones tradicionales de copia de seguridad y restauración. Todos los datos de configuración se almacenan en Resource Manager y se pueden volver a implementar mediante enfoques IaC, como archivos de Bicep o plantillas de ARM.

Para la administración de la configuración y la recuperación ante desastres, debe realizar las siguientes acciones:

- Defina la configuración de la implementación de Application Gateway mediante archivos de Bicep o plantillas de ARM, o bien exporte la configuración de una puerta de enlace existente.

- Almacene certificados de seguridad de la capa de transporte (TLS) en Azure Key Vault para la administración y replicación seguras.
- Directivas, reglas y configuraciones personalizadas de control de versiones y documentos.
- Implemente canalizaciones de implementación automatizadas para el aprovisionamiento coherente de la puerta de enlace.

Para la mayoría de las soluciones, no debe confiar exclusivamente en copias de seguridad. En su lugar, utilice las otras capacidades descritas en esta guía para apoyar los requisitos de resiliencia. Sin embargo, las copias de seguridad protegen contra algunos riesgos que otros enfoques no. Para más información, consulte ¿Qué son la redundancia, la replicación y la copia de seguridad?.

Resistencia al mantenimiento del servicio

Application Gateway v2 realiza actualizaciones de servicio periódicas y otras tareas de mantenimiento. Para mantener la capacidad esperada durante una actualización, la plataforma agrega automáticamente instancias adicionales de la puerta de enlace durante el proceso de actualización. Sin embargo, debe asegurarse de que la subred de la puerta de enlace tenga suficiente espacio de direcciones IP libre para que se creen las instancias temporales. Para obtener más información, consulte ¿Cómo controla Application Gateway el mantenimiento rutinario?.

Acuerdo de nivel de servicio

El contrato de nivel de servicio (SLA) para los servicios de Azure describe la disponibilidad esperada de cada servicio y las condiciones que la solución deberá cumplir para lograr esa expectativa de disponibilidad. Para obtener más información, consulte Acuerdos de Nivel de Servicio para servicios en línea.

Para que una puerta de enlace sea apta para el Acuerdo de Nivel de Servicio de disponibilidad de Application Gateway, debe cumplir los criterios siguientes:

- Debe configurarse correctamente para realizar el equilibrio de carga HTTP.
- Debe tener redundancia de zona o debe configurarse para usar la escalabilidad automática.

Contenido relacionado

- Introducción a la confiabilidad de Azure

ⓘ **Nota:** El autor ha creado este artículo con ayuda de inteligencia artificial. Más información

---

Last updated on 22/01/2026

Confiabilidad en Azure Application Gateway para contenedores

En este artículo se describe la confiabilidad y la compatibilidad con zonas de disponibilidad en Azure Application Gateway for Containers. Aprenda a configurar la redundancia de zona para aumentar la resistencia y la disponibilidad de las cargas de trabajo de contenedor que tienen configuraciones de implementación de alta disponibilidad. Para obtener información general más detallada sobre la confiabilidad de Azure, consulte Confiabilidad de Azure.

Compatibilidad con zonas de disponibilidad y redundancia entre zonas

Las zonas de disponibilidad son grupos físicamente independientes de centros de datos dentro de una región de Azure. Cuando una zona falla, los servicios pueden transferirse a una de las zonas restantes.

Puerta de enlace de aplicaciones para contenedores (AGC) siempre se implementa en una configuración de alta disponibilidad. En el caso de las regiones de Azure que admiten zonas de disponibilidad, AGC se configura automáticamente como redundante de zona. En el caso de las regiones que no admiten zonas de disponibilidad, se usan conjuntos de disponibilidad.

Prerrequisitos

Para implementar con soporte de zonas de disponibilidad, debe elegir una región que admita zonas de disponibilidad. Para ver qué regiones admiten zonas de disponibilidad, consulte la lista de regiones admitidas.

Pasos siguientes

- Confiabilidad en Azure

Last updated on 22/01/2026

Confiabilidad en Azure Bastion

Azure Bastion es una plataforma como servicio (PaaS) totalmente administrada que se aprovisiona para proporcionar conexiones de alta seguridad a máquinas virtuales a través de una dirección IP privada. Proporciona una conectividad RDP/SSH ininterrumpida a las máquinas virtuales directamente sobre TLS desde Azure Portal o a través del cliente nativo SSH o RDP instalado en el equipo local. Cuando se conecta a través de Azure Bastion, las máquinas virtuales no necesitan una dirección IP pública, un agente ni software cliente especial.

Cuando se usa Azure, la confiabilidad es una responsabilidad compartida. Microsoft proporciona una variedad de funcionalidades para admitir resistencia y recuperación. Es responsable de comprender cómo funcionan esas funcionalidades dentro de todos los servicios que usa y de seleccionar las funcionalidades que necesita para cumplir los objetivos empresariales y los objetivos de tiempo de actividad.

En este artículo se describe cómo hacer que Azure Bastion sea resistente a una variedad de posibles interrupciones y problemas, incluidos errores transitorios, interrupciones de zona de disponibilidad y interrupciones de regiones. Resalta cierta información clave sobre el acuerdo de nivel de servicio (SLA) de Azure Bastion.

> **Importante**
>
> La compatibilidad con las zonas de disponibilidad para Azure Bastion actualmente está en versión preliminar. Para conocer los términos legales que se aplican a las características de Azure que se encuentran en la versión beta, en versión preliminar o que todavía no se han publicado para que estén disponibles con carácter general, consulte los **Términos de uso complementarios para las versiones preliminares de Microsoft Azure**.

Recomendaciones de implementación de producción para la confiabilidad

En el caso de las cargas de trabajo de producción, se recomienda que usted:

- ✔ Use el SKU básico o superior.
- ✔ Habilite la redundancia de zona si el host bastión está en una región admitida.

Introducción a la arquitectura de confiabilidad

Al utilizar Azure Bastion, debe implementar un *servidor bastión* en una subred que cumpla con los requisitos de Azure Bastion.

Un host bastión tiene un número definido de *instancias*, que a veces también se denominan *unidades de escalado*. Cada instancia representa una sola máquina virtual dedicada que controla las conexiones de Azure Bastion. La plataforma administra automáticamente la creación de instancias, la supervisión de estado y la sustitución de instancias incorrectas, por lo que no ve ni administra directamente las máquinas virtuales.

La SKU básica admite exactamente dos instancias. Las SKU estándar y Premium admiten el *escalado de host*, donde puede configurar el número de instancias, con un mínimo de dos instancias. Al agregar más instancias, el host bastión puede dar cabida a conexiones de cliente simultáneas adicionales.

Resistencia a errores transitorios

Los errores transitorios son errores breves e intermitentes en los componentes. Se producen con frecuencia en un entorno distribuido como la nube y son una parte normal de las operaciones. Los errores transitorios se corrigen después de un breve período de tiempo. Es importante que las aplicaciones puedan controlar errores transitorios, normalmente mediante el reintento de solicitudes afectadas.

Todas las aplicaciones hospedadas en la nube deben seguir las instrucciones de control de errores transitorios de Azure cuando se comunican con cualquier API, bases de datos y otros componentes hospedados en la nube. Para obtener más información, consulte Recomendaciones para controlar errores transitorios.

Si los errores transitorios afectan a la máquina virtual o al host bastión, los clientes que usan el host de sockets seguros (SSH) y los protocolos del Protocolo de escritorio remoto (RDP) normalmente vuelven a intentarlo automáticamente.

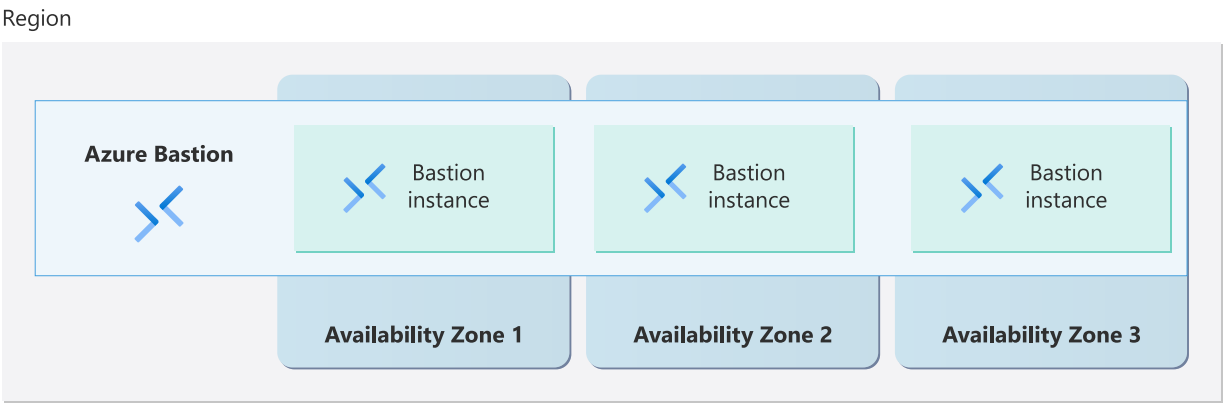
Resistencia a errores de zona de disponibilidad

Las zonas de disponibilidad son grupos físicamente independientes de centros de datos dentro de una región de Azure. Cuando una zona falla, los servicios pueden transferirse a una de las zonas restantes.

Azure Bastion admite zonas de disponibilidad en configuraciones con redundancia de zona y zonales:

- *Con redundancia de zona:* Un host bastión con redundancia de zona logra resistencia y confiabilidad al distribuir sus instancias entre varias zonas de disponibilidad. Usted selecciona las zonas de disponibilidad que desea usar para su host bastión.

En el diagrama siguiente se muestra un host bastión con redundancia de zona, con sus instancias distribuidas entre tres zonas:



Si especifica más zonas de disponibilidad que instancias tiene, Azure Bastion reparte las instancias entre tantas zonas como pueda.

- *Zonal:* Un host bastión zonal y todas sus instancias están en una sola zona de disponibilidad que seleccione.

> **Importante**
>
> La asignación a una sola zona de disponibilidad solo se recomienda cuando la latencia entre zonas es demasiado alta para sus necesidades y después de comprobar que la latencia no cumple sus requisitos. Por sí mismo, un recurso zonal no proporciona resistencia a una interrupción de zona de disponibilidad. Para mejorar la resiliencia de un recurso zonal, debe desplegar explícitamente recursos independientes en múltiples zonas de disponibilidad y configurar el enrutamiento del tráfico y la conmutación por error. Para obtener más información, consulte Recursos zonales y resistencia de zona.

Requisitos

- **Soporte de regiones:** los hosts bastión zonales y con redundancia de zona pueden implementarse en las siguientes regiones:

Expandir tabla

Americas	Europa	Oriente Medio	África	Asia Pacífico
Canada Central	Norte de Europa	Qatar Central	Norte de Sudáfrica	Australia East

Americas	Europa	Oriente Medio	África	Asia Pacífico
Central US	Centro de Suecia	Israel Central		Centro de Corea del Sur
East US	UK South			
Este de EE. UU. 2	West Europe			
Oeste de EE. UU. 2	Norway East			
EUAP de Este de EE. UU. 2	Norte de Italia			
Centro de México	Centro de España			

- **SKU:** Para configurar los hosts bastión para que sean zonales o con redundancia de zona, debe realizar la implementación con los SKU Básico, Estándar o Premium.
- **Dirección IP pública:** Azure Bastion requiere una dirección IP pública con redundancia de zona de SKU estándar.

Cost

No hay ningún costo adicional para usar el soporte de zona de disponibilidad de Azure Bastion. Los cargos se basan en la SKU del host bastión y en el número de instancias que usa. Para más información, consulte Precios de Azure Bastion.

Configurar soporte de zonas de disponibilidad

- **Implemente un nuevo host bastión con compatibilidad con la zona de disponibilidad:** Al implementar un nuevo host bastión en una región que admita zonas de disponibilidad, seleccione las zonas específicas en las que desea implementar.

  Para la redundancia de zona, debe seleccionar varias zonas.

  Cuando selecciona qué zonas de disponibilidad usar, en realidad está seleccionando la *zona de disponibilidad lógica*. Si implementa otros componentes de la carga de trabajo en una suscripción de Azure diferente, podrían usar un número de zona de disponibilidad lógica distinto para acceder a la misma zona de disponibilidad física. Para obtener más información, consulte zonas de disponibilidad física y lógica.

- **Hosts bastión existentes:** No es posible cambiar la configuración de zona de disponibilidad de un host bastión existente. En su lugar, debe crear un servidor bastión

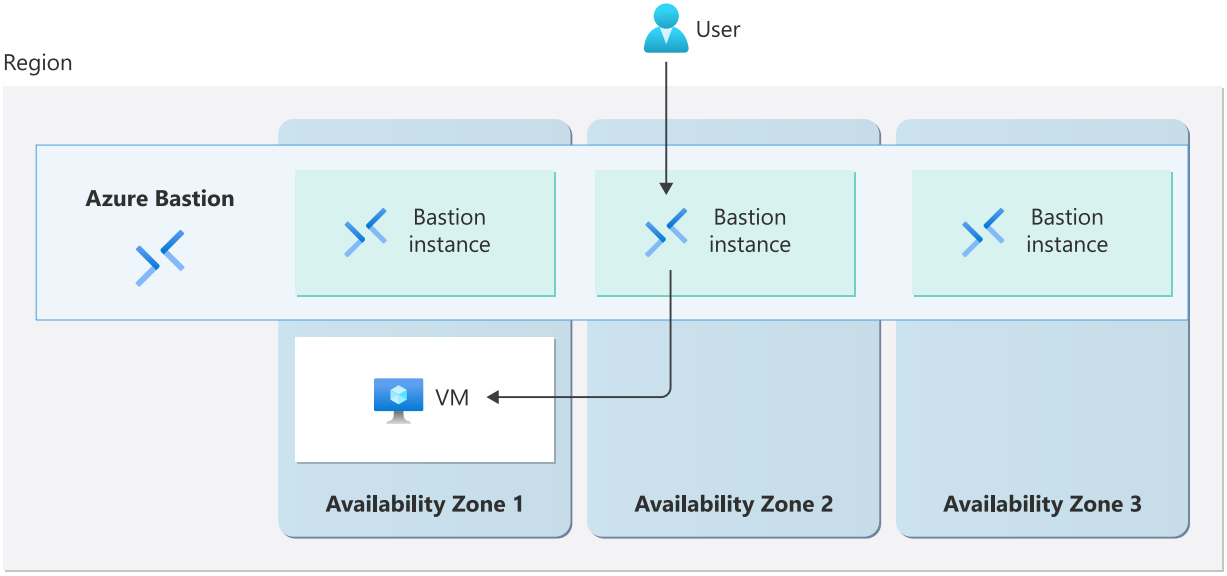
con la nueva configuración y eliminar el anterior.

Comportamiento cuando todas las zonas están en buen estado

En esta sección se describe qué esperar cuando los hosts bastión están configurados para la compatibilidad con zonas de disponibilidad y todas las zonas de disponibilidad están operativas.

- **Enrutamiento de tráfico entre zonas:** Al iniciar una sesión ssh o RDP, se puede enrutar a una instancia de Azure Bastion en cualquiera de las zonas de disponibilidad seleccionadas.

  Si configura la redundancia de zona en el host bastión, es posible que se envíe una sesión a una instancia bastión en una zona de disponibilidad diferente de la máquina virtual a la que se conecta. En el siguiente diagrama, una solicitud del usuario se envía a una instancia de Azure Bastion en la zona 2, mientras que la máquina virtual se encuentra en la zona 1:



> **Sugerencia**
>
> En la mayoría de los escenarios, la cantidad de latencia entre zonas no es significativa. Sin embargo, si tiene requisitos de latencia inusualmente estrictos para sus cargas de trabajo, debe implementar un bastión dedicado de una sola zona en la zona de disponibilidad de la máquina virtual. Tenga en cuenta que esta configuración no proporciona redundancia de zona y no se recomienda para la mayoría de los clientes.

- **Replicación de datos entre zonas:** Dado que Azure Bastion no almacena el estado, no hay datos para replicar entre zonas.

Comportamiento durante un fallo de zona

En esta sección se describe qué esperar cuando los servidores bastión están configurados para admitir zonas de disponibilidad y se produce una interrupción en una zona de disponibilidad.

- **Detección y respuesta:** cuando se usa redundancia de zona, Azure Bastion detecta y responde a errores en una zona de disponibilidad. No necesita hacer nada para iniciar una conmutación por error de una zona de disponibilidad.

  En el caso de las instancias con redundancia de zona, Azure Bastion realiza un intento de mayor esfuerzo para reemplazar las instancias que se pierden debido a un fallo de zona. Sin embargo, no se garantiza que las instancias sean reemplazadas.

- **Notificación**: Microsoft no le notifica automáticamente cuando una zona está inactiva. Sin embargo, puede usar Azure Resource Health para supervisar el estado de un recurso individual y puede configurar alertas de Resource Health para notificarle problemas. También puede usar Azure Service Health para comprender el estado general del servicio, incluidos los errores de zona, y puede configurar alertas de Service Health para notificarle problemas.

- **Solicitudes activas:** cuando una zona de disponibilidad no está disponible, cualquier conexión de RDP o SSH en curso que use una instancia de Azure Bastion en la zona de disponibilidad defectuosa se cancela y es necesario volver a iniciarla.

  Si la máquina virtual a la que se conecta no está en la zona de disponibilidad afectada, continúa ejecutándose. Para más información sobre la experiencia de indisponibilidad de zona de la VM, consulte Confiabilidad en VMs: comportamiento durante un error de zona.

- **Tiempo de inactividad esperado:** El tiempo de inactividad esperado depende de la configuración de zona de disponibilidad que usa el host bastión.

  - *Con redundancia de zona:* Es posible que se produzca un breve período de inactividad mientras el servicio restablece las operaciones. Este tiempo de inactividad suele ser de unos segundos.

  - *Zonal:* La instancia no está disponible hasta que se recupere la zona de disponibilidad.

- **Pérdida de datos esperada:** Dado que Azure Bastion no almacena el estado, no se espera ninguna pérdida de datos durante un error de zona.

- **Reenrutamiento del tráfico:** Al usar redundancia de zona, las nuevas conexiones usan instancias de Azure Bastion en las zonas de disponibilidad correctas. En general, Azure Bastion permanece operativo.

Recuperación de zona

Cuando se recupera la zona de disponibilidad, Azure Bastion restaura automáticamente las instancias de la zona de disponibilidad y vuelve a enrutar el tráfico entre las instancias como normal.

Prueba de fallos de zona

La plataforma Azure Bastion administra el enrutamiento del tráfico, la conmutación por error y la conmutación por recuperación para bastiones con redundancia de zona. Dado que esta característica está totalmente administrada, no es necesario iniciar nada ni validar los procesos de error de zona de disponibilidad.

Resistencia a errores en toda la región

Azure Bastion se implementa en redes virtuales o redes virtuales emparejadas y está asociado a una región de Azure. Azure Bastion es un servicio de una sola región. Si la región no está disponible, el host bastión perderá su disponibilidad.

Azure Bastion admite el acceso a máquinas virtuales en redes virtuales emparejadas globalmente, pero si la región que hospeda el host bastión no está disponible, no podrá usar el host bastión. Para una mayor capacidad de recuperación, si implementa su solución global en varias regiones con redes virtuales independientes en cada región, debería implementar Azure Bastion en cada una de ellas.

Si tiene un sitio de recuperación ante desastres en otra región de Azure, asegúrese de implementar Azure Bastion en la red virtual de esa región.

Acuerdo de nivel de servicio

El contrato de nivel de servicio (SLA) para los servicios de Azure describe la disponibilidad esperada de cada servicio y las condiciones que la solución deberá cumplir para lograr esa expectativa de disponibilidad. Para obtener más información, consulte Acuerdos de Nivel de Servicio para servicios en línea.

Contenido relacionado

- Confiabilidad en Azure

ⓘ **Nota:** El autor ha creado este artículo con ayuda de inteligencia artificial. Más información

---

Last updated on 22/01/2026

Fiabilidad de Azure DNS Private Resolver

Solucionador privado de DNS de Azure es un servicio administrado que permite la resolución de DNS entre entornos locales y zonas DNS privadas de Azure sin necesidad de servidores DNS en máquinas virtuales. Puede configurar el reenvío condicional a servidores DNS locales, multinube y públicos.

Cuando se usa Azure, la confiabilidad es una responsabilidad compartida. Microsoft proporciona una variedad de capacidades para apoyar la resiliencia y la recuperación. Es responsable de comprender cómo funcionan esas funcionalidades dentro de todos los servicios que usa y de seleccionar las funcionalidades que necesita para cumplir los objetivos empresariales y los objetivos de tiempo de actividad.

En este artículo se describe cómo hacer que el resolvedor privado de DNS sea resistente a posibles caídas e interrupciones, incluidos fallos transitorios, fallos en las zonas de disponibilidad y fallos de toda una región. También se describen las consideraciones de copia de seguridad y restauración y el comportamiento de mantenimiento del servicio.

Introducción a la arquitectura de confiabilidad

Al trabajar con la resolución privada de DNS, se implementa una *resolución privada* en una red virtual Azure. A continuación, configure *los puntos de conexión*, que reciben y reenván consultas DNS hacia y desde el entorno local y las *reglas de reenvío*, que definen el comportamiento de reenvío. Para obtener más información, consulte Puntos de conexión y conjuntos de reglas de la Resolución privada de DNS.

Microsoft es responsable de la disponibilidad y el estado del servicio de resolución administrado y sus operaciones de plataforma. Usted es responsable de diseñar la topología de los puntos de conexión y de los conjuntos de reglas, configurar la lógica de reenvío de DNS y garantizar que la trayectoria de conectividad de red cumpla sus requisitos de fiabilidad.

La confiabilidad de un extremo a otro también depende de los servicios dependientes y las decisiones de arquitectura de red. Por ejemplo, las zonas DNS privadas, la configuración de red virtual, las reglas de firewall, la infraestructura local y la configuración de conectividad híbrida afectan directamente a los resultados de la resolución de nombres durante los errores.

Resistencia a errores transitorios

Los errores transitorios son errores breves e intermitentes en los componentes. Se producen con frecuencia en un entorno distribuido como la nube y son una parte normal de las operaciones. Los errores transitorios se corrigen después de un breve período de tiempo. Es importante que las aplicaciones puedan controlar errores transitorios, normalmente mediante el reintento de solicitudes afectadas.

Todas las aplicaciones hospedadas en la nube deben seguir las instrucciones de control de errores transitorios de Azure cuando se comunican con cualquier API, bases de datos y otros componentes hospedados en la nube. Para obtener más información, consulte Recomendaciones para controlar errores transitorios.

Si se produce un error transitorio durante la resolución DNS, el cliente DNS o el solucionador intermedio deben reintentar las consultas DNS. Configurar los valores de tiempo de espera adecuadamente: entre 2 y 5 segundos suele ser un tiempo de espera suficiente para un cliente DNS.

El período de vida de cada registro DNS (TTL) también afecta al modo en que la solución controla los errores. Si el TTL es muy bajo, los clientes deben realizar más solicitudes de resolución dns y hay más oportunidades de que surjan errores transitorios. Si el TTL es muy alto, en caso de un fallo real en un servidor backend que le obligue a redirigir a otra dirección IP, los clientes podrían experimentar retrasos en la conmutación por error hasta que caduque el TTL. Configure las TTL cuidadosamente para equilibrar la disponibilidad, la latencia y la capacidad de respuesta.

Resistencia a errores de zona de disponibilidad

Las zonas de disponibilidad son grupos físicamente independientes de centros de datos dentro de una región de Azure. Cuando una zona falla, los servicios pueden transferirse a una de las zonas restantes.

La resolución privada de DNS es redundante entre zonas automáticamente en las regiones que admiten zonas de disponibilidad. Microsoft aprovisiona el servicio en todas las zonas de esas regiones.

Requisitos

Compatibilidad con regiones: El solucionador privado dns proporciona redundancia de zona automática en regiones que admiten zonas de disponibilidad.

Costo

No hay ningún costo adicional para habilitar la redundancia de zona para la resolución privada de DNS. Para obtener más información sobre los precios, consulte precios de Azure DNS.

Configurar soporte de zonas de disponibilidad

DNS Private Resolver tiene redundancia de zona automáticamente en las regiones admitidas. No hay ninguna configuración necesaria para habilitar la redundancia de zona y no se puede deshabilitar. Para conocer los pasos de implementación, consulte Inicio rápido: Creación de una resolución privada de DNS mediante el portal de Azure.

Comportamiento cuando todas las zonas están en buen estado

En esta sección se describe qué esperar al implementar la resolución privada de DNS en una región con zonas de disponibilidad y todas las zonas están operativas.

- **Operación entre zonas:** El solucionador privado dns ejecuta la infraestructura en varias zonas y enruta las solicitudes automáticamente entre las zonas.
- **Replicación de datos entre zonas:** DNS Private Resolver no replica datos de clientes entre zonas porque el servicio no tiene estado y no mantiene datos de clientes.

Comportamiento durante un fallo de zona

En esta sección se describe qué esperar al implementar la resolución privada de DNS en una región con zonas de disponibilidad y una zona tiene una interrupción.

- **Detección y respuesta:** Microsoft detecta errores de zona de disponibilidad y administra todas las acciones de respuesta. No es necesario realizar ninguna acción para iniciar una conmutación por error de zona.
- **Notificación**: Microsoft no le notifica automáticamente cuando una zona está inactiva. Sin embargo, puede usar Azure Resource Health para supervisar el estado de un recurso individual y puede configurar alertas de Resource Health para notificarle problemas. También puede usar Azure Service Health para comprender el estado general del servicio,

incluidos los errores de zona, y puede configurar alertas de Service Health para notificarle problemas.

- **Solicitudes activas:** Las consultas DNS en curso que usan la infraestructura afectada pueden producir un error o agotar el tiempo de espera y se deben reintentar. Si los clientes controlan correctamente los errores transitorios , normalmente evitan un impacto significativo.
- **Pérdida de datos esperada:** No se espera ninguna pérdida de datos porque el servicio no tiene estado y no almacena los datos del cliente.
- **Tiempo de inactividad esperado:** Es posible que se produzca una pequeña cantidad de tiempo de inactividad mientras el servicio desplaza las solicitudes a la infraestructura de zona correcta. Este tiempo de inactividad suele ser de unos segundos.
- **Redistribución:** Microsoft redirige automáticamente las solicitudes de resolución de DNS a través de las zonas en buen estado.

Recuperación de zona

Cuando se recupera una zona de disponibilidad con errores, el solucionador privado dns restaura automáticamente las operaciones normales sin intervención del cliente.

Prueba de fallos de zona

El resolvedor privado de DNS es un servicio totalmente administrado por Microsoft y con redundancia de zona. Dado que Microsoft gestiona la redundancia de las zonas de disponibilidad, no es necesario probar escenarios de conmutación por error de zonas.

Resistencia a errores en toda la región

DNS Private Resolver es un servicio de región única. Si la región deja de estar disponible, tu resolvedor privado tampoco estará disponible.

Soluciones de varias regiones personalizadas para la resistencia

Para mejorar la resistencia a los errores de toda la región, puede implementar varios solucionadores privados en regiones independientes. A continuación, configure los patrones de reenvío y tráfico para admitir la conmutación por error entre regiones. Para conocer los pasos

detallados, consulte Configuración de la conmutación por error de DNS mediante solucionadores privados.

Copias de seguridad y restauración

El solucionador privado de DNS no almacena los datos de los clientes y no requiere copia de seguridad ni restauración. Para volver a crear configuraciones, considere la posibilidad de mantener plantillas de infraestructura como código (IaC) para los recursos de red. Los solucionadores privados son de solo configuración y no almacenan datos de cliente, por lo que se centran los esfuerzos de copia de seguridad en las plantillas de IaC para volver a implementar rápidamente el solucionador, la configuración del punto de conexión, los conjuntos de reglas de reenvío DNS y otros recursos.

Resistencia al mantenimiento del servicio

Microsoft aplica periódicamente actualizaciones de servicio y realiza otro mantenimiento. La plataforma Azure controla estas actividades automáticamente, lo que garantiza que el mantenimiento sea transparente y sin problemas. No se espera ningún tiempo de inactividad durante los eventos de mantenimiento a menos que se le haya informado a través del mantenimiento planeado de Azure Service Health.

Acuerdo de nivel de servicio

El acuerdo de nivel de servicio (SLA) para Azure servicios describe la disponibilidad esperada de cada servicio y las condiciones que la solución debe cumplir para lograr esa expectativa de disponibilidad. Para obtener más información, consulte Acuerdos de Nivel de Servicio para servicios en línea.

DNS Private Resolver ofrece un SLA de disponibilidad para las respuestas a consultas DNS, siempre que los clientes reintenten repetidamente las solicitudes fallidas.

Contenido relacionado

- ¿Qué es el solucionador privado dns?
- Configura la conmutación por error de DNS mediante resolvedores privados
- Confiabilidad en Azure

Last updated on 15/06/2026

Fiabilidad en las zonas privadas de Azure DNS

zonas privadas de Azure DNS proporcionan una resolución de nombres segura dentro de las redes virtuales de Azure. Puede definir el ámbito de las zonas DNS privadas a una o varias redes virtuales y las organizaciones normalmente las usan para las aplicaciones internas. Los nombres de host que resuelva son nombres DNS locales que no son accesibles públicamente a través de Internet. Las direcciones IP resueltas suelen ser direcciones IP privadas que no son accesibles desde Internet. Azure DNS es un servicio global que no está enlazado a ninguna zona de disponibilidad específica ni a una sola región.

Cuando se usa Azure, la confiabilidad es una responsabilidad compartida. Microsoft proporciona una variedad de capacidades para apoyar la resiliencia y la recuperación. Es responsable de comprender cómo funcionan esas funcionalidades dentro de todos los servicios que usa y de seleccionar las funcionalidades que necesita para cumplir los objetivos empresariales y los objetivos de tiempo de actividad.

En este artículo se describe cómo hacer que las zonas privadas de Azure DNS sean resilientes frente a diversas interrupciones y posibles problemas, incluidos los fallos transitorios y los fallos que afectan a toda una región. También proporciona información clave sobre el acuerdo de nivel de servicio (SLA) del servicio de zonas privadas de Azure DNS.

Recomendaciones de implementación de producción para la confiabilidad

En el caso de las cargas de trabajo de producción, se recomienda seguir estas recomendaciones:

- ✓ **Configure los valores de TTL adecuados:** Establezca valores de período de vida (TTL) que equilibran el rendimiento con el tiempo de recuperación. Los valores de TTL más bajos permiten una conmutación por error más rápida, pero aumentan el volumen de consultas. Considere 300 segundos (5 minutos) como punto de partida para las cargas de trabajo de producción.
- ✓ **Divida las zonas DNS grandes:** Si tiene una zona DNS grande, considere dividir la zona para mejorar la fiabilidad general y la eficiencia operativa.

Introducción a la arquitectura de confiabilidad

En esta sección se describen algunos de los aspectos importantes de cómo funciona el servicio que es más relevante desde una perspectiva de confiabilidad. En la sección se presenta la arquitectura lógica, que incluye algunos de los recursos y características que se implementan y usan. También se describe la arquitectura física, que proporciona detalles sobre cómo funciona el servicio en segundo plano.

Arquitectura lógica

El recurso principal que implemente es una *zona*, que representa un conjunto de registros DNS que asignan nombres de host (nombres de dominio) a direcciones IP. Los nombres de host que resuelve la zona suelen ser nombres DNS locales que no son accesibles públicamente desde Internet.

Las zonas DNS privadas se crean como recursos independientes y se vinculan a redes virtuales específicas mediante la creación de *vínculos de red virtual*. Cuando las solicitudes DNS proceden de clientes dentro de esas redes virtuales, las zonas DNS privadas participan en el proceso de resolución. Puede crear manualmente entradas en una zona DNS o configurar el registro automático de máquinas virtuales en vínculos de red virtual. Las zonas DNS privadas de Azure permiten la resolución de DNS entre redes virtuales de distintas regiones de Azure, incluso sin interconectar de forma explícita las redes virtuales. Sin embargo, todas las redes virtuales están vinculadas a la zona DNS privada.

El proceso de resolución de nombres DNS implica varios componentes, incluidos solucionadores DNS y capas intermedias que procesan solicitudes antes de llegar a los servidores DNS autoritativos. Las zonas privadas usan los mismos protocolos y comportamientos DNS que las zonas públicas, incluidos los valores de TTL y los mecanismos de almacenamiento en caché.

> **ⓘ Importante**
>
> La confiabilidad de la solución general depende de la configuración de los recursos a los que hacen referencia los registros DNS, como máquinas virtuales y equilibradores de carga.
>
> En este artículo no se tratan esos recursos, pero sus configuraciones de disponibilidad afectan directamente a la resistencia de la aplicación. Revise las **guías de confiabilidad de los servicios de Azure de la solución** para obtener información sobre cómo cada servicio admite sus requisitos de confiabilidad.

Arquitectura física

Azure DNS es un servicio no regional. Microsoft implementa su infraestructura en varias zonas de disponibilidad en varias regiones de Azure en todo el mundo. Este diseño permite que Azure DNS permanezcan resistentes durante una interrupción de zona o región de disponibilidad porque la infraestructura de otra zona o región sigue respondiendo a las solicitudes de resolución.

Los protocolos globales de Internet, como Anycast, DNS y Border Gateway Protocol (BGP), enrutan automáticamente las solicitudes entrantes de resolución de DNS a la infraestructura de Azure DNS en buen estado más próxima.

Resistencia a errores transitorios

Los errores transitorios son errores breves e intermitentes en los componentes. Se producen con frecuencia en un entorno distribuido como la nube y son una parte normal de las operaciones. Los errores transitorios se corrigen después de un breve período de tiempo. Es importante que las aplicaciones puedan controlar errores transitorios, normalmente mediante el reintento de solicitudes afectadas.

Todas las aplicaciones hospedadas en la nube deben seguir las instrucciones de control de errores transitorios de Azure cuando se comunican con cualquier API, bases de datos y otros componentes hospedados en la nube. Para más información, vea Recomendaciones para gestionar errores temporales.

Azure DNS controla los errores transitorios a través de su infraestructura DNS global.

Si se produce un error transitorio durante la resolución DNS, el cliente o el solucionador intermedio deben reintentar la solicitud. Configure los valores de tiempo de espera adecuadamente. Normalmente, un tiempo de espera de 2 a 5 segundos es suficiente para un cliente DNS.

El período de vida de cada registro DNS (TTL) también afecta al modo en que la solución controla los errores. Si el TTL es muy bajo, los clientes realizan más solicitudes a Azure DNS, lo que crea más oportunidades para errores transitorios. Si el TTL es muy alto, en caso de un fallo real en un servidor backend que le obligue a redirigir a otra dirección IP, los clientes podrían experimentar retrasos en la conmutación por error hasta que caduque el TTL. Configure las TTL cuidadosamente para equilibrar la disponibilidad, la latencia y la capacidad de respuesta.

Resistencia a errores de zona de disponibilidad

Las zonas de disponibilidad son grupos físicamente independientes de centros de datos dentro de una región de Azure. Cuando una zona falla, los servicios pueden transferirse a una de las zonas restantes.

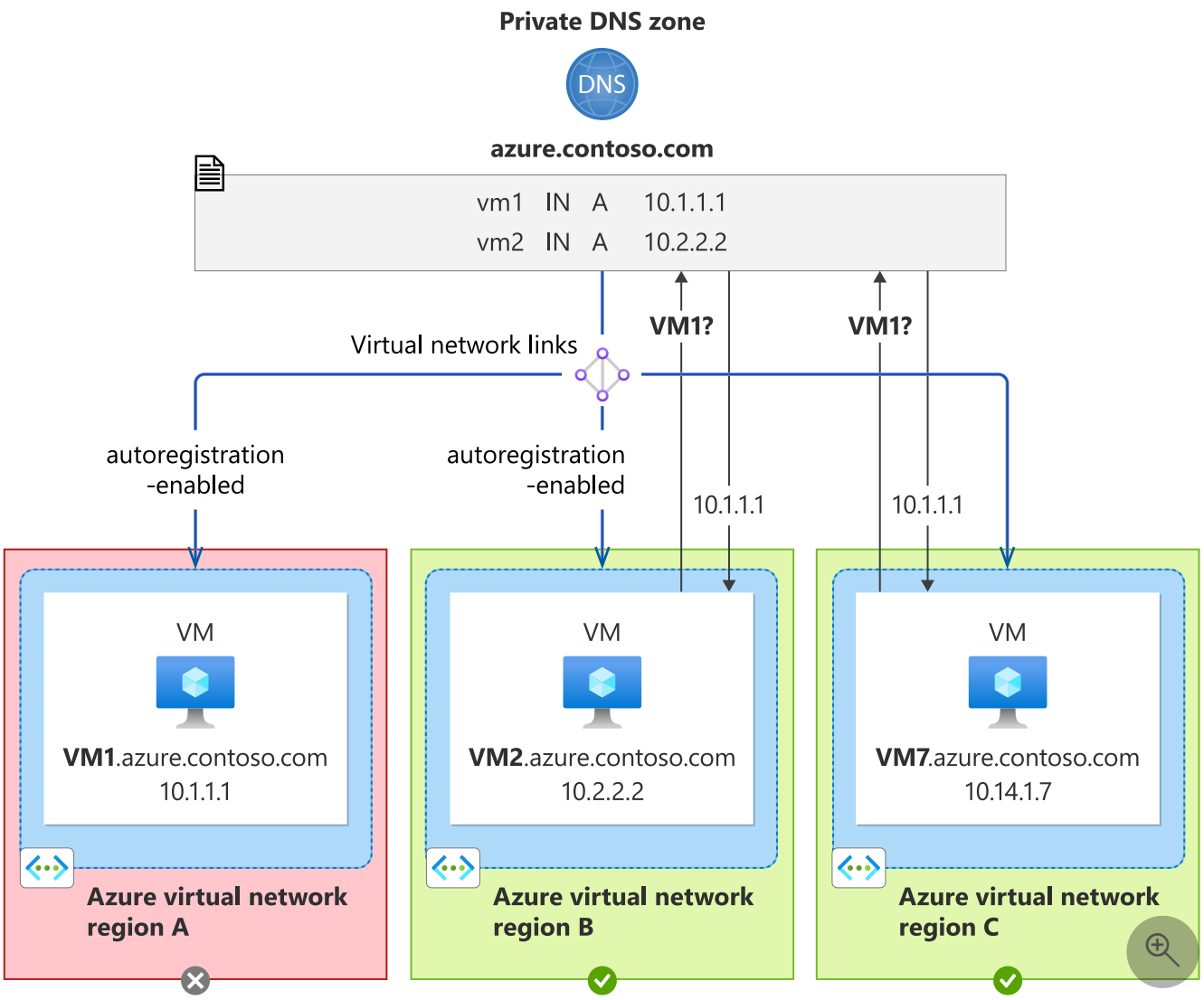
Azure DNS funciona como un servicio no regional. Microsoft distribuye su infraestructura en varias zonas de disponibilidad en varias regiones de Azure y replica los cambios en las zonas DNS privadas en esa infraestructura. No selecciona zonas de disponibilidad ni configura redundancia de zona. Durante una interrupción de zona de disponibilidad, la infraestructura de otra zona o región sigue respondiendo a las solicitudes de resolución.

Si un recurso que se implementa en una sola zona de disponibilidad, como una máquina virtual, deja de estar disponible durante un error de zona, Azure DNS continúa devolviendo la dirección IP configurada del recurso porque no supervisa el estado del punto de conexión. Si realiza una conmutación por error a un recurso en una zona en buen estado, usted es responsable de actualizar el registro DNS para que los clientes usen el recurso en buen estado. Como alternativa, sitúe los recursos detrás de un balanceador de carga con redundancia de zona que redirija el tráfico a las máquinas virtuales de las zonas en buen estado.

Resistencia a errores en toda la región

Las zonas privadas de Azure DNS resisten las interrupciones regionales porque los datos de la zona están disponibles globalmente. Si una región tiene una interrupción, sus redes virtuales y recursos, como las máquinas virtuales, podrían no estar disponibles, pero la resolución de nombres sigue funcionando.

En el ejemplo siguiente se muestra cómo permanecen los datos de zona privada disponibles en varias regiones. La zona `azure.contoso.com` privada está vinculada a redes virtuales en tres regiones: región A, región B y región C. El registro automático está habilitado en las regiones A y B. En el diagrama se muestra la región A que experimenta una interrupción:



Supongamos que se produce una interrupción temporal en la región A. Las máquinas virtuales de las regiones B y C todavía pueden consultar nombres DNS en la zona privada, incluidos los nombres que se registran automáticamente desde la región A. Pueden seguir resolviendo la dirección IP de VM1 en la región A, aunque VM1 no esté disponible. La interrupción del servicio en la región A no afecta a la resolución de nombres en las otras regiones.

En el ejemplo anterior no se muestra un escenario de recuperación ante desastres en el que la solución conmuta por error a un reemplazo de VM1 en otra región. Sin embargo, dado que las zonas privadas son globales, puede volver a crear VM1 en la red virtual de otra región para asumir la carga de trabajo.

Si crea redes virtuales y recursos de red en varias regiones, debe planificar e implementar su estrategia multirregional para las aplicaciones que requieren conmutación por error entre regiones.

Resistencia a amenazas de seguridad y configuración incorrecta

Los ataques de seguridad y los errores de configuración son dos de los riesgos de confiabilidad más significativos para las zonas DNS. Varias clases de ataques se dirigen específicamente a la resolución de DNS, y un error de configuración accidental puede interrumpir sus cargas de trabajo con la misma gravedad.

Para obtener instrucciones de seguridad completas específicas de zonas DNS privadas, consulte Protección de registros y zonas DNS privadas.

Resistencia a interrupciones del servicio

Azure DNS es un servicio altamente resistente, con un Acuerdo de Nivel de Servicio de disponibilidad de 100% cuando la aplicación cumple ciertas condiciones. Las interrupciones del servicio son extremadamente inusuales, pero los problemas de red u otros problemas de infraestructura pueden interrumpir la conectividad con el servicio Azure DNS.

Supervisión de interrupciones de servicio

Microsoft no le notifica automáticamente cuando una región está inactiva. Sin embargo, puede usar Azure Service Health para comprender el estado general del servicio, incluidos los errores de región, y puede configurar alertas de Service Health para notificarle problemas.

Prueba de interrupciones del servicio

Azure Chaos Studio proporciona un conjunto de errores para simular problemas con la resolución de DNS. Por ejemplo, el agente de Chaos Studio proporciona el tipo de error de DNS Failure y Azure Kubernetes Service (AKS) Chaos Mesh proporciona la funcionalidad Chaos de DNS. Puede usar estos tipos de error para probar cómo responden las aplicaciones y la infraestructura cuando se produce un error en las solicitudes de resolución DNS, lo que puede producirse durante un error de red parcial.

Resiliencia ante interrupciones del portal y herramientas de gestión

Si administra la zona DNS en el portal de Azure, prepárese para escenarios en los que no pueda acceder a ella, especialmente si necesita volver a configurar la zona DNS durante una

interrupción de la plataforma.

Puede usar varias herramientas para implementar y administrar Azure DNS zonas privadas. Aprenda a usar CLI de Azure o Azure PowerShell para administrar la zona privada. Como alternativa, use la infraestructura como código (IaC), como Bicep o Terraform, para implementar y configurar la zona privada. Estas herramientas permanecen operativas incluso si el portal de Azure está degradado.

Copias de seguridad y restauración

Azure DNS es un servicio sin estado. No proporciona copias de seguridad administradas ni restauración a un momento dado para zonas DNS privadas.

Para conservar la configuración completa de recursos Azure, defina las zonas DNS privadas mediante IaC, como Bicep o Terraform, y almacene las definiciones en el control de código fuente. Pruebe periódicamente las definiciones para poder usarlas para volver a implementar la configuración.

Resistencia al mantenimiento del servicio

Microsoft aplica periódicamente actualizaciones de servicio y realiza otro mantenimiento. La plataforma Azure controla estas actividades automáticamente, lo que garantiza que el mantenimiento sea transparente y sin problemas. No se espera ningún tiempo de inactividad durante los eventos de mantenimiento a menos que se le haya informado a través del mantenimiento planeado de Azure Service Health.

Acuerdo de nivel de servicio

El acuerdo de nivel de servicio (SLA) para Azure servicios describe la disponibilidad esperada de cada servicio y las condiciones que la solución debe cumplir para lograr esa expectativa de disponibilidad. Para obtener más información, consulte los SLAs de los servicios en línea.

Azure DNS proporciona un Acuerdo de Nivel de Servicio de disponibilidad de 100% para respuestas de consulta DNS válidas cuando cumple ciertas condiciones. Estas condiciones incluyen reintentar las solicitudes fallidas durante al menos 60 segundos consecutivos. Revise el documento del Acuerdo de Nivel de Servicio para conocer las condiciones detalladas.

Contenido relacionado

- Documentación sobre zonas privadas de Azure DNS

- Preguntas más frecuentes sobre Azure DNS
- Protección de registros y zonas DNS privadas: Azure DNS
- Fiabilidad de las zonas públicas de Azure DNS

ⓘ**Nota:** El autor ha creado este artículo con ayuda de inteligencia artificial. Más información

---

Last updated on 31/08/2026

Fiabilidad en las zonas públicas de Azure DNS

Azure DNS proporciona resolución de nombres mediante la infraestructura de Microsoft Azure. Este artículo se centra en las zonas DNS públicas, que normalmente se crean para dominios que posee y se usan para publicar registros para aplicaciones y servicios que están disponibles en Internet. Los nombres de host que resuelva son nombres DNS accesibles públicamente y las direcciones IP resueltas suelen ser direcciones IP públicas accesibles desde Internet.

Azure DNS es un servicio no regional que no está enlazado a una zona de disponibilidad específica o Azure región.

Cuando se usa Azure, la confiabilidad es una responsabilidad compartida. Microsoft proporciona una variedad de capacidades para apoyar la resiliencia y la recuperación. Es responsable de comprender cómo funcionan esas funcionalidades dentro de todos los servicios que usa y de seleccionar las funcionalidades que necesita para cumplir los objetivos empresariales y los objetivos de tiempo de actividad.

En este artículo se describe cómo las zonas públicas de Azure DNS responden a fallos transitorios, fallos de zona de disponibilidad, fallos que afectan a toda una región, interrupciones del servicio, amenazas de seguridad y errores de configuración, interrupciones del portal y de las herramientas de administración, y tareas de mantenimiento del servicio. También se describe cómo proteger y restaurar la configuración de la zona y se explican los requisitos clave del acuerdo de nivel de servicio (SLA).

Recomendaciones de implementación de producción para la confiabilidad

Para las implementaciones de producción de Azure DNS zonas públicas, siga estas recomendaciones para mejorar la confiabilidad:

- ✓ **Delegue en todos los servidores de nombres:** Azure DNS asigna cuatro servidores de nombres a cada zona DNS pública. Configure la delegación de dominio para usar los cuatro servidores de nombres. Esta configuración proporciona aislamiento de errores y es necesario para calificar para el Acuerdo de Nivel de Servicio de Azure DNS.

- ✓ **Configure los valores de TTL adecuados:** Establezca los valores de período de vida (TTL) que equilibran el volumen de consultas con la rapidez con la que los clientes reciben cambios de registro. Los valores de TTL inferiores permiten a los clientes recibir cambios antes, pero aumentar el volumen de consultas. Los valores de TTL más altos reducen el volumen de consultas, pero pueden retrasar la conmutación por error después de cambiar un registro.
- ✓ **Use registros de alias para recursos de Azure admitidos:**los registros de alias reflejan automáticamente los cambios en un recurso de Azure subyacente durante la resolución DNS y ayudan a evitar registros DNS obsoletos.

Introducción a la arquitectura de confiabilidad

En esta sección se describen algunos de los aspectos importantes de cómo funciona el servicio que es más relevante desde una perspectiva de confiabilidad. En la sección se presenta la arquitectura lógica, que incluye algunos de los recursos y características que se implementan y usan. También se describe la arquitectura física, que proporciona detalles sobre cómo funciona el servicio en segundo plano.

Arquitectura lógica

El recurso principal que implemente es una *zona*, que contiene los conjuntos de registros DNS para un dominio. Un conjunto de registros asocia un nombre DNS a un valor, como una dirección IP o un punto de conexión. Los nombres que resuelve una zona DNS pública son accesibles a través de Internet.

Para que Azure DNS sea autoritativo para su dominio, delegue el dominio en los servidores de nombres que Azure asigna cuando crea la zona. Una vez que se haya implementado la delegación, se crean conjuntos de registros para los tipos de registro DNS que Azure DNS admite. También puede crear registros de alias que hagan referencia a recursos Azure, como direcciones IP públicas, perfiles de Traffic Manager y puntos de conexión de Azure Front Door, de modo que el registro DNS permanezca sincronizado con el recurso de destino.

Durante la resolución DNS, los solucionadores DNS recursivos siguen la jerarquía del DNS para llegar a los servidores de nombres autoritativos de Azure DNS de su zona.

> ⓘ **Importante**
>
> Azure DNS resuelve los nombres, pero no supervisa el estado del punto de conexión ni enruta el tráfico de la aplicación. La confiabilidad de la solución general depende de la

configuración de los recursos a los que hacen referencia los registros DNS, como máquinas virtuales y equilibradores de carga.

En este artículo no se tratan esos recursos, pero sus configuraciones de disponibilidad afectan directamente a la resistencia de la aplicación. Revise las **guías de confiabilidad de los servicios de Azure de la solución** para obtener información sobre cómo cada servicio admite sus requisitos de confiabilidad.

Arquitectura física

Azure DNS funciona como un servicio no regional e implementa su infraestructura en varias zonas de disponibilidad en varias regiones de Azure en todo el mundo. Este diseño permite que Azure DNS permanezcan resistentes durante una interrupción de zona o región de disponibilidad porque la infraestructura de otra zona o región sigue respondiendo a las solicitudes de resolución.

Los protocolos globales de Internet, como Anycast, DNS y BGP, dirigen automáticamente las solicitudes entrantes de resolución de DNS a la infraestructura de Azure DNS en buen estado más cercana.

El plano de servicio de Azure DNS opera en una configuración activa-activa en dos entornos de servicio independientes: uno que se ejecuta en Linux y otro que se ejecuta en Windows. Estas pilas no comparten código ni hardware subyacente. Como son independientes, un error, una vulnerabilidad o un fallo que afecta a una pila no afecta a la otra. Esta independencia reduce el riesgo de una interrupción completa del servicio causada por un único punto de error y ayuda a protegerse contra determinadas clases de vulnerabilidades de día cero.

Resistencia a errores transitorios

Los errores transitorios son errores breves e intermitentes en los componentes. Se producen con frecuencia en un entorno distribuido como la nube y son una parte normal de las operaciones. Los errores transitorios se corrigen después de un breve período de tiempo. Es importante que las aplicaciones puedan controlar errores transitorios, normalmente mediante el reintento de solicitudes afectadas.

Todas las aplicaciones hospedadas en la nube deben seguir las instrucciones de control de errores transitorios de Azure cuando se comunican con cualquier API, bases de datos y otros componentes hospedados en la nube. Para más información, vea Recomendaciones para gestionar errores temporales.

Azure DNS controla los errores transitorios a través de su infraestructura DNS global.

Si se produce un error transitorio durante la resolución de DNS, el cliente o el resolutor intermedio debe reintentar según su comportamiento de reintento de DNS configurado. Entre 2 y 5 segundos suele ser un tiempo de espera suficiente para un cliente DNS.

El período de vida de cada registro DNS (TTL) también afecta al modo en que la solución controla los errores. Si el TTL es muy bajo, los clientes deben realizar más solicitudes para Azure DNS y hay más oportunidades potenciales para que surjan errores transitorios. Si el TTL es muy alto, en caso de un fallo real en un servidor backend que le obligue a redirigir a otra dirección IP, los clientes podrían experimentar retrasos en la conmutación por error hasta que caduque el TTL. Configure las TTL cuidadosamente para equilibrar la disponibilidad, la latencia y la capacidad de respuesta.

Resistencia a errores de zona de disponibilidad

Las zonas de disponibilidad son grupos físicamente independientes de centros de datos dentro de una región de Azure. Cuando una zona falla, los servicios pueden transferirse a una de las zonas restantes.

Azure DNS funciona como un servicio no regional. Microsoft distribuye su infraestructura en varias zonas de disponibilidad en varias regiones de Azure y replica los cambios en las zonas DNS públicas en esa infraestructura. No selecciona zonas de disponibilidad ni configura redundancia de zona. Durante una interrupción de zona de disponibilidad, la infraestructura de otra zona o región sigue respondiendo a las solicitudes de resolución.

Si un recurso que se implementa en una sola zona de disponibilidad, como una máquina virtual, deja de estar disponible durante un error de zona, Azure DNS continúa devolviendo la dirección IP configurada del recurso porque no supervisa el estado del punto de conexión. Si realiza una conmutación por error a un recurso en una zona en buen estado, usted es responsable de actualizar el registro DNS para que los clientes usen el recurso en buen estado. Como alternativa, sitúe los recursos detrás de un balanceador de carga con redundancia de zona que redirija el tráfico a las máquinas virtuales de las zonas en buen estado.

Resistencia a errores en toda la región

Las zonas DNS son resistentes a las interrupciones de regiones porque los datos de zona están disponibles globalmente e implementados en varias regiones de Azure. Si una región tiene una interrupción, los recursos que implementó en esa región, como las redes virtuales y las máquinas

virtuales, podrían no estar disponibles, pero Azure DNS continúa resolviendo los registros de la zona.

Si tiene una solución que necesita cambiar entre varias regiones, como con fines de recuperación ante desastres, considere la posibilidad de usar Azure Traffic Manager o Azure Front Door. Estos servicios proporcionan capacidades de conmutación por error automática, que puede utilizar si una región presenta problemas.

Resistencia a amenazas de seguridad y configuración incorrecta

Los ataques de seguridad y los errores de configuración son dos de los riesgos de confiabilidad más significativos para las zonas DNS. Varias clases de ataques se dirigen específicamente a la resolución de DNS, y un error de configuración accidental puede interrumpir sus cargas de trabajo con la misma gravedad.

Para obtener instrucciones de seguridad completas específicas de las zonas DNS públicas, consulte Protección de la implementación de Azure DNS y Protección de zonas y registros DNS.

Resistencia a interrupciones del servicio

Azure DNS es un servicio altamente resistente, con un Acuerdo de Nivel de Servicio de disponibilidad de 100% cuando la aplicación cumple ciertas condiciones. Las interrupciones del servicio son extremadamente inusuales, pero los problemas de red o problemas con otra infraestructura pueden interrumpir la conectividad con el servicio Azure DNS.

La resiliencia de Azure DNS se debe en parte a su arquitectura de plano de servicio activo-activo distribuida globalmente.

Uso de varios servidores de nombres

Azure DNS asigna cuatro servidores de nombres a cada zona DNS pública. Al delegar el dominio, configure los cuatro servidores de nombres. Si un solucionador no puede llegar a un servidor de nombres, puede consultar a otro.

Supervisión de interrupciones de servicio

Use Azure Service Health para supervisar el estado de Azure DNS. Configure las alertas de Service Health para notificarle los incidentes de servicio.

Prueba de interrupciones del servicio

Azure Chaos Studio proporciona fallos que simulan errores de resolución de DNS desde dentro de algunos tipos de cargas de trabajo de prueba. Estos errores no desencadenan una interrupción en Azure DNS. El agente de Chaos Studio proporciona el error de Error de DNS, y AKS Chaos Mesh proporciona la funcionalidad Caos de DNS. Use estos errores para probar cómo responden las aplicaciones y la infraestructura cuando se produce un error en la resolución DNS, como durante un error de red parcial.

Resiliencia ante interrupciones del portal y herramientas de gestión

Si administra la zona DNS pública en el portal de Azure, prepare una ruta de acceso de administración alternativa para escenarios en los que no pueda acceder al portal, especialmente si es posible que tenga que volver a configurar la zona durante una interrupción.

Si el portal de Azure no está disponible, use el CLI de Azure, el Azure PowerShell o la infraestructura como código (IaC), como Bicep o Terraform para administrar la zona DNS pública. Estas herramientas permanecen operativas incluso si el portal de Azure está degradado.

Copias de seguridad y restauración

Azure DNS es un servicio sin estado. No proporciona copias de seguridad administradas ni restauración a un punto concreto en el tiempo para las zonas DNS públicas.

Para conservar la configuración completa de recursos Azure, defina las zonas DNS públicas mediante IaC, como Bicep o Terraform, y almacene las definiciones en el control de código fuente. Pruebe las definiciones periódicamente para poder usarlas para volver a implementar la configuración.

Como opción de recuperación de nivel de registro adicional, exporte un archivo de zona compatible con BIND. La importación de archivos de zona tiene limitaciones y no conserva todas las opciones de recursos específicas de Azure, por lo que no use un archivo de zona exportado como único artefacto de recuperación. Revise las limitaciones de importación documentadas y compruebe los registros después de restaurar una zona.

Resistencia al mantenimiento del servicio

Microsoft aplica periódicamente actualizaciones de servicio y realiza otro mantenimiento. La plataforma Azure controla estas actividades automáticamente, lo que garantiza que el mantenimiento sea transparente y sin problemas. No se espera ningún tiempo de inactividad durante los eventos de mantenimiento a menos que se le haya informado a través del mantenimiento planeado de Azure Service Health.

Acuerdo de nivel de servicio

El acuerdo de nivel de servicio (SLA) para Azure servicios describe la disponibilidad esperada de cada servicio y las condiciones que la solución debe cumplir para lograr esa expectativa de disponibilidad. Para obtener más información, consulte los SLAs de los servicios en línea.

Azure DNS proporciona un Acuerdo de Nivel de Servicio de disponibilidad de 100% para respuestas de consulta DNS válidas, siempre y cuando se cumplan determinadas condiciones. Estas condiciones incluyen volver a intentar solicitudes con errores repetidamente durante al menos 60 segundos consecutivos y usar todos los servidores de nombres que Azure DNS asignan a la zona. Revise el documento del Acuerdo de Nivel de Servicio para conocer las condiciones detalladas.

Contenido relacionado

- Introducción a las zonas públicas de Azure DNS
- Preguntas más frecuentes sobre Azure DNS
- Importación y exportación de un archivo de zona de Azure DNS
- Proteja la implementación de Azure DNS
- Protección de zonas y registros de Azure DNS
- Información general de Azure DNS privado

ⓘ**Nota:** El autor ha creado este artículo con ayuda de inteligencia artificial. Más información

---

Last updated on 26/08/2026

Confiabilidad en la protección contra DDoS de Azure

Azure DDoS Protection es una funcionalidad fundamental de red Azure que ayuda a proteger las aplicaciones frente a ataques distribuidos por denegación de servicio (DDoS). Los ataques DDoS intentan sobrecargar las aplicaciones con el tráfico para denegar el servicio a los usuarios legítimos. DDoS Protection ayuda a proteger las aplicaciones mediante la supervisión de patrones de tráfico de red y la mitigación automática del tráfico anómalo que podría afectar a la disponibilidad.

Al usar Azure, relibilidad es una responsabilidad compartida. Microsoft proporciona una variedad de capacidades para admitir resiliencia y recuperación. Es responsable de comprender cómo funcionan esas funcionalidades dentro de todos los servicios que usa y de seleccionar las funcionalidades que necesita para cumplir los objetivos empresariales y los objetivos de tiempo de actividad.

En este artículo se describe cómo DDoS Protection es resistente a varias posibles interrupciones y problemas, incluidos errores transitorios, interrupciones de zona de disponibilidad y interrupciones de regiones. También resalta cierta información clave sobre el acuerdo de nivel de servicio (SLA) de DDoS Protection.

> ⓘ **Nota:**
>
> En este artículo se describe cómo el propio servicio DDoS Protection es resistente a varios problemas. No explica cómo usar DDoS Protection para proteger las máquinas virtuales (VM) u otros recursos. Para obtener información sobre cómo usar DDoS Protection para proteger las cargas de trabajo, consulte **Procedimientos recomendados fundamentales de DDoS Protection**.

Introducción a la arquitectura de confiabilidad

Puede implementar DDoS Protection en uno de estos niveles:

- **DDoS Network Protection:** Implementar un plan de protección *DDoS*, que define un conjunto de redes virtuales que tienen habilitada la protección contra redes DDoS, incluso si

están en distintas suscripciones y regiones de Azure. Cada dirección IP pública asociada a los recursos de esas redes virtuales está protegida por el plan.

- **DDoS IP Protection:** Implemente una dirección IP pública. DDoS IP Protection solo está habilitado en esa dirección IP.

Para obtener más información sobre la diferencia entre DDoS Network Protection y DDoS IP Protection, consulte Acerca de la comparación de niveles de DDoS Protection.

Resistencia a errores transitorios

Los errores transitorios son errores breves e intermitentes en los componentes. Se producen con frecuencia en un entorno distribuido como la nube y son una parte normal de las operaciones. Los errores transitorios se corrigen después de un breve período de tiempo. Es importante que las aplicaciones puedan controlar errores transitorios, normalmente mediante el reintento de solicitudes afectadas.

Todas las aplicaciones hospedadas en la nube deben seguir las instrucciones de control de errores transitorios Azure cuando se comunican con cualquier API, bases de datos y otros componentes hospedados en la nube. Para obtener más información, consulte Recomendaciones para controlar errores transitorios.

La habilitación de la protección contra DDoS no cambia la forma en que las aplicaciones controlan los errores transitorios.

Resistencia a errores de zona de disponibilidad

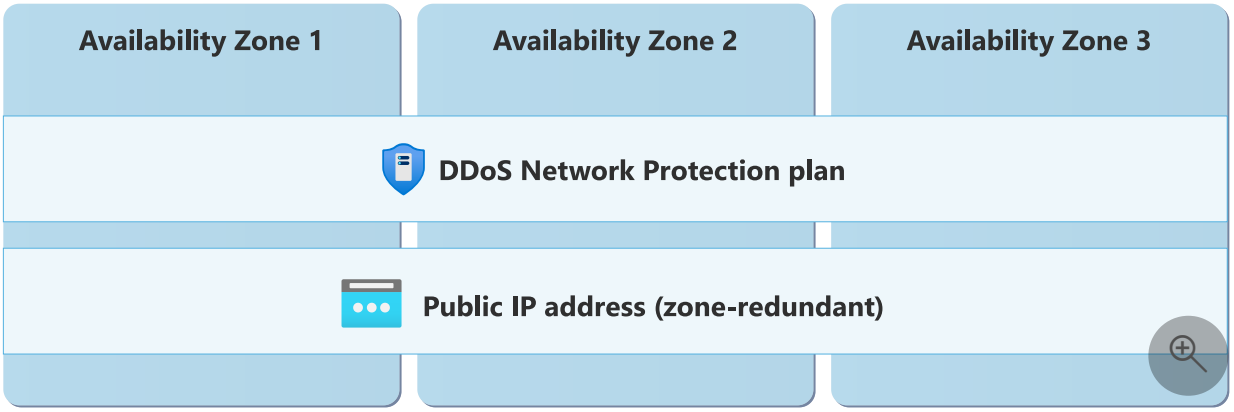
Availability zones son grupos de centros de datos físicamente independientes dentro de una región de Azure. Cuando una zona falla, los servicios pueden transferirse a una de las zonas restantes.

DDoS Protection tiene redundancia de zona de forma predeterminada en las regiones que admiten zonas de disponibilidad. El servicio abarca todas las zonas de disponibilidad automáticamente y no requiere ninguna configuración del cliente para habilitar la redundancia de zona. Microsoft administra la distribución de la infraestructura de DDoS Protection entre zonas.

DDoS Protection está diseñado para proteger las direcciones IP públicas frente a ataques DDoS. Para lograr una resistencia completa a los errores de zona de disponibilidad, también debe

asegurarse de que las direcciones IP públicas tienen redundancia de zona. También debe revisar la resistencia de zona de toda la carga de trabajo, incluidos otros servicios de Azure que use.

En el diagrama siguiente se muestra un plan de protección de red DDoS con redundancia de zona y varias direcciones IP públicas con redundancia de zona protegidas:



Requisitos

Compatibilidad regional: DDoS Protection tiene redundancia de zona en cualquier región que admita zonas de disponibilidad.

Costo

No hay ningún costo adicional para habilitar la redundancia de zona para DDoS Protection. Para más información, consulte Precios de Azure DDoS Protection.

Configurar soporte de zonas de disponibilidad

DDoS Protection tiene automáticamente redundancia de zona en las regiones admitidas. No se requiere ninguna configuración para habilitar la redundancia de zona y no se puede deshabilitar.

Comportamiento cuando todas las zonas están en buen estado

En esta sección se describe qué esperar al implementar un plan de DDoS Protection en una región con zonas de disponibilidad, la dirección IP pública es con redundancia de zona y todas las zonas de disponibilidad están operativas.

- **Operación entre zonas:** La inspección del tráfico puede tener lugar en todas las zonas y el tráfico se enruta entre zonas de forma transparente como parte de las operaciones de red de Azure.

- **Replicación de datos de operación entre zonas:** DDoS Protection no replica los datos del cliente entre zonas porque el servicio no tiene estado y no mantiene los datos del cliente.

Comportamiento durante un fallo de zona

En esta sección se describe qué esperar al implementar un plan de DDoS Protection en una región con zonas de disponibilidad, la dirección IP pública es con redundancia de zona y hay una interrupción en una de las zonas de disponibilidad.

- **Detección y respuesta:** Microsoft detecta errores de zona de disponibilidad y administra todas las acciones de respuesta. No es necesario realizar ninguna acción para iniciar una conmutación por error de zona.

- **Notification:** Microsoft no le notifica automáticamente cuando una zona está inactiva. Sin embargo, puede usar Azure Service Health para comprender el estado general del servicio, incluidos los errores de zona, y puede configurar alertas de Service Health para notificarle problemas.

- **Solicitudes activas:** El tráfico activo se sigue procesando automáticamente. No es necesario realizar ninguna acción.

- **Pérdida de datos esperada:** No se espera ninguna pérdida de datos porque el servicio no tiene estado y no almacena los datos del cliente.

- **Tiempo de inactividad esperado:** Durante un error de zona, la disponibilidad de la *ruta de acceso de datos* no se ve afectada. La ruta de acceso de datos representa el camino del tráfico desde el borde de Azure, a través de la plataforma Azure, hasta tu aplicación. La aplicación permanece protegida por el plan de protección contra DDoS durante una interrupción de zona.

  Sin embargo, si realiza operaciones de administración en el plan de protección contra DDoS durante un fallo de zona, es posible que estas operaciones se retrasen hasta que la plataforma realice internamente una conmutación por error.

- **Redistribution:** Microsoft vuelve a enrutar automáticamente la protección del tráfico a través de las zonas saludables.

Recuperación de zona

Cuando se recupera una zona de disponibilidad con errores, DDoS Protection restaura automáticamente las operaciones normales sin su intervención.

Prueba de fallos de zona

DDoS Protection es un servicio con redundancia de zona totalmente administrado por Microsoft. Dado que Microsoft gestiona la redundancia de las zonas de disponibilidad, no es necesario probar escenarios de conmutación por error de zonas.

Resistencia a errores en toda la región

El comportamiento de DDoS Protection durante los errores de toda la región difiere en función del tipo de protección contra DDoS que use:

- **DDoS Network Protection plans** se implementan en una región de Azure que seleccione. Sin embargo, el plan también protege las direcciones IP públicas en otras regiones.

  Si la región que hospeda un plan DDoS Network Protection deja de estar disponible, las direcciones IP públicas protegidas de otras regiones siguen estando protegidas. Sin embargo, es posible que las operaciones de administración del plan no estén disponibles hasta que se recupere la región.

- **DDoS IP Protection** está configurado en una sola dirección IP pública.

  En el caso de las direcciones IP públicas regionales, si se produce un error en toda la región, es probable que la dirección IP y sus servidores no estén disponibles.

  En el caso de las direcciones IP públicas globales, la dirección IP permanece protegida por DDoS IP Protection incluso cuando se produce un error en una región.

Acuerdo de nivel de servicio

El acuerdo de nivel de servicio (SLA) para Azure servicios describe la disponibilidad esperada de cada servicio y las condiciones que la solución debe cumplir para lograr esa expectativa de disponibilidad. Para obtener más información, vea SLAs for servicios en línea.

DDoS Protection proporciona un Acuerdo de Nivel de Servicio que cubre la disponibilidad del servicio de mitigación de DDoS para protegerse contra un ataque.

Contenido relacionado

- Información general de Azure DDoS Protection
- Fiabilidad en Azure

ⓘ**Nota:** El autor ha creado este artículo con ayuda de inteligencia artificial. Más información

---

Last updated on 11/06/2026

Confiabilidad en puertas de enlace de red virtual de Azure

Una puerta de enlace de red virtual de Azure proporciona conectividad segura entre una red virtual de Azure y otras redes, ya sea una red local u otra red virtual en Azure.

Azure proporciona dos tipos de puertas de enlace de red virtual:

- Las puertas de enlace de Azure ExpressRoute usan conexiones privadas que no atraviesan la red pública de Internet.
- Las puertas de enlace de red privada virtual (VPN) de Azure usan túneles cifrados a través de Internet.

Como componente de Azure, una puerta de enlace de red virtual proporciona diversas funcionalidades para admitir los requisitos de confiabilidad.

Cuando se usa Azure, la confiabilidad es una responsabilidad compartida. Microsoft proporciona una variedad de funcionalidades para admitir resistencia y recuperación. Es responsable de comprender cómo funcionan esas funcionalidades dentro de todos los servicios que usa y de seleccionar las funcionalidades que necesita para cumplir los objetivos empresariales y los objetivos de tiempo de actividad.

En este artículo se describe cómo hacer que una puerta de enlace de red virtual sea resistente a varias posibles interrupciones y problemas, incluidos errores transitorios, interrupciones de zona de disponibilidad, interrupciones de regiones y mantenimiento planeado del servicio. También resalta información clave sobre el acuerdo de nivel de servicio (SLA) de la puerta de enlace de red virtual.

Para ver información sobre Azure VPN Gateway, seleccione el tipo de puerta de enlace de red virtual adecuado al principio de esta página.

> **ⓘ Importante**
>
> En este artículo se describe la confiabilidad de las puertas de enlace de red virtual de ExpressRoute, que son los componentes basados en Azure del sistema ExpressRoute.
>
> Pero cuando se usa ExpressRoute, debe diseñar toda la *arquitectura de red* (no solo la puerta de enlace) para satisfacer los requisitos de resistencia. Normalmente, se utilizan varios sitios, también conocidos como *ubicaciones de emparejamiento*, y se habilita la alta disponibilidad y la conmutación por error rápida para los componentes locales. Para obtener más información, vea **Diseño y arquitectura de ExpressRoute para la resistencia**.

Azure Well-Architected Framework proporciona recomendaciones sobre confiabilidad, rendimiento, seguridad, costo y operaciones. Para comprender cómo estas áreas influyen entre sí y contribuyen a una solución confiable de ExpressRoute, consulte Procedimientos recomendados de arquitectura para ExpressRoute en Well-Architected Framework.

Introducción a la arquitectura de confiabilidad

Con ExpressRoute, debe implementar componentes en el entorno local, los puntos de conexión y en Azure. Estos componentes incluyen los siguientes elementos:

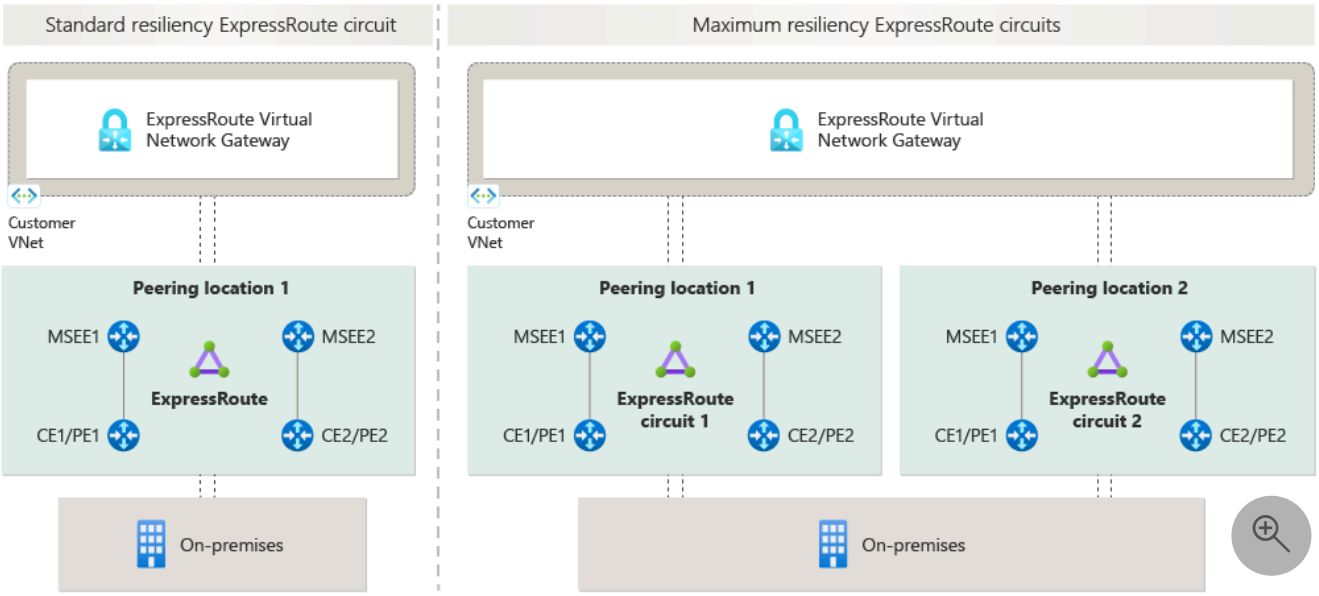
- **Circuitos y conexiones:** Un *circuito* ExpressRoute consta de dos *conexiones* a través de una sola ubicación de emparejamiento a Microsoft Enterprise Edge. Al usar dos conexiones, puede lograr la conectividad activa-activa. Pero esta configuración no protege frente a errores de nivel de sitio.

- **Equipo local del cliente (CPE):** Este equipo incluye enrutadores perimetrales y dispositivos cliente. Asegúrese de que el CPE está diseñado para ser resistente y puede recuperarse rápidamente cuando otras partes de la infraestructura de ExpressRoute experimentan problemas.

- **Sitios:** Los circuitos se establecen a través de un *sitio*, que es una ubicación de emparejamiento físico. Los sitios están diseñados para ser de alta disponibilidad y tienen redundancia integrada en todas las capas. Pero los sitios representan una sola ubicación física, por lo que pueden producirse problemas. Para mitigar el riesgo de interrupciones del sitio, ExpressRoute proporciona opciones de resistencia del sitio que tienen distintos niveles de protección.

- **Puerta de enlace de red virtual de Azure:** En Azure, creará una *puerta de enlace de red virtual* que actúa como punto de terminación para uno o varios circuitos ExpressRoute dentro de la red virtual de Azure.

En el diagrama siguiente se muestran dos configuraciones de ExpressRoute diferentes, cada una con una sola puerta de enlace de red virtual, configurada para distintos niveles de resistencia entre sitios.



Puerta de enlace de red virtual

Una puerta de enlace de ExpressRoute contiene dos o más *máquinas virtuales (VM) de puerta* de enlace, que son las máquinas virtuales subyacentes que usa la puerta de enlace para procesar el tráfico de ExpressRoute.

No ves ni gestionas directamente las máquinas virtuales de la puerta de enlace. La plataforma gestiona automáticamente la creación de máquinas virtuales de puerta de enlace, el monitoreo del estado y la sustitución de máquinas virtuales de puerta de enlace no saludables. Para protegerse frente a fallos de servidor y de bastidor de servidores, Azure distribuye automáticamente las máquinas virtuales de puerta de enlace en varios dominios de fallo dentro de una región. Si ocurre una falla en un rack de servidores, la plataforma Azure migra automáticamente las máquinas virtuales de puerta de enlace de ese clúster a otro clúster.

Usted configura la SKU de la puerta de enlace. Cada SKU admite un nivel de rendimiento diferente y un número diferente de circuitos. Cuando se usa la SKU de ErGwScale (versión preliminar), ExpressRoute escala automáticamente la puerta de enlace agregando más máquinas virtuales de puerta de enlace. Para más información, consulte Puertas de enlace de red virtual de ExpressRoute.

Una puerta de enlace funciona en modo *activo-activo* de forma predeterminada, lo que admite alta disponibilidad para su circuito. Opcionalmente, puede cambiar al modo *activo-pasivo* , pero esta configuración aumenta el riesgo de un error que afecte a la conectividad. Para más información, consulte Conexiones activas-activas.

Normalmente, el tráfico se enruta a través de la puerta de enlace de red virtual. Pero si usa FastPath, el tráfico del entorno local omite la puerta de enlace. Este enfoque mejora el rendimiento y reduce la latencia. La puerta de enlace sigue siendo esencial porque configura el enrutamiento para su tráfico.

Puede protegerse contra fallos en zonas de disponibilidad distribuyendo las máquinas virtuales de la puerta de enlace entre varias zonas. Esta distribución proporciona conmutación automática dentro de la región y mantiene la conectividad durante el mantenimiento o las interrupciones en la zona. Para obtener más información, consulte Resilience to availability zone failures (Resistencia a errores de zona de disponibilidad).

Resistencia a errores transitorios

Los errores transitorios son errores breves e intermitentes en los componentes. Se producen con frecuencia en un entorno distribuido como la nube y son una parte normal de las operaciones. Los errores transitorios se corrigen después de un breve período de tiempo. Es importante que las aplicaciones puedan controlar errores transitorios, normalmente mediante el reintento de solicitudes afectadas.

Todas las aplicaciones hospedadas en la nube deben seguir las instrucciones de control de errores transitorios de Azure cuando se comunican con cualquier API, bases de datos y otros componentes hospedados en la nube. Para obtener más información, consulte Recomendaciones para controlar errores transitorios.

En el caso de las aplicaciones que se conectan a través de una puerta de enlace de red virtual, implemente la lógica de reintento con retroceso exponencial para controlar posibles problemas de conexión transitorios. La naturaleza con estado de las puertas de enlace de red virtual garantiza que las conexiones legítimas se mantengan durante breves interrupciones de red.

En un entorno de red distribuido, los errores transitorios pueden producirse en varias capas, incluidas las siguientes ubicaciones:

- Su entorno local
- Un sitio perimetral
- Azur

ExpressRoute reduce el efecto de los errores transitorios mediante rutas de conexión redundantes, detección rápida de errores y conmutación por error automatizada. Pero debe configurar las aplicaciones y los componentes locales correctamente para ser resistentes a varios problemas. Para obtener estrategias completas de control de errores, consulte Diseño de alta disponibilidad mediante ExpressRoute.

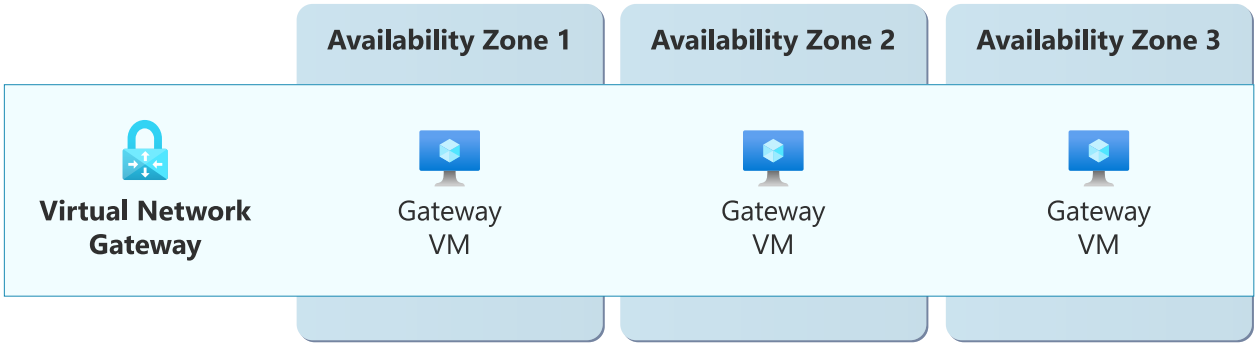
Resistencia a errores de zona de disponibilidad

Las zonas de disponibilidad son grupos físicamente independientes de centros de datos dentro de una región de Azure. Cuando una zona falla, los servicios pueden transferirse a una de las zonas restantes.

Las puertas de enlace de red virtual tienen automáticamente *redundancia de zona* cuando cumplen los requisitos. La redundancia de zona elimina cualquier zona única como punto de error y proporciona el nivel más alto de resistencia de zona. Las puertas de enlace con redundancia de zona proporcionan conmutación automática por error dentro de la región y mantienen la conectividad durante el mantenimiento o las interrupciones de zona.

Las máquinas virtuales de puerta de enlace de ExpressRoute con redundancia de zona se distribuyen automáticamente entre al menos tres zonas de disponibilidad.

En el diagrama siguiente se muestra una puerta de enlace de red virtual con redundancia de zona con tres máquinas virtuales de puerta de enlace distribuidas en distintas zonas de disponibilidad.



> ⓘ **Nota:**
>
> Los circuitos o las conexiones no incluyen la configuración de zona de disponibilidad. Estos recursos residen en las instalaciones perimetrales de red, que no están diseñadas para usar zonas de disponibilidad.

Al usar una SKU compatible, las puertas de enlace recién creadas tienen automáticamente redundancia de zona. Se recomienda la redundancia de zona para todas las cargas de trabajo de producción.

Requisitos

- **Compatibilidad con regiones:** Las puertas de enlace de red virtual con redundancia de zona están disponibles en todas las regiones que admiten zonas de disponibilidad.
- **SKU:** Para que una puerta de enlace de red virtual tenga redundancia de zona, debe usar una SKU que admita redundancia de zona. En la tabla siguiente se muestran las SKU que

admiten redundancia de zona.

Expandir tabla

Nombre de la SKU	Admite zonas de disponibilidad
Estándar	No
HighPerformance	No
UltraPerformance	No
ErGw1Az	Sí
ErGw2Az	Sí
ErGw3Az	Sí
ErGwScale (versión preliminar)	Sí

Cost

Las puertas de enlace con redundancia de zona para ExpressRoute requieren SKU específicas, que pueden tener mayores tasas por hora en comparación con las SKU de puerta de enlace estándar, gracias a sus funcionalidades mejoradas y características de rendimiento. Para obtener más información, consulte Precios de ExpressRoute.

Configurar soporte de zonas de disponibilidad

En esta sección se explica cómo configurar la redundancia de zona para las puertas de enlace de red virtual.

- **Cree una nueva puerta de enlace de red virtual que admita zonas de disponibilidad.** Las nuevas puertas de enlace de red virtual son automáticamente redundantes de zona si cumplen los requisitos enumerados anteriormente. Para más información, consulte Creación de una puerta de enlace de red virtual con redundancia de zona en zonas de disponibilidad.

- **Cambie la configuración de zona de disponibilidad de una puerta de enlace de red virtual existente.** Es posible que las puertas de enlace de red virtual que ya haya creado no sean redundantes de zona. Puede migrar una puerta de enlace no zonal a una puerta de enlace con redundancia de zona con un tiempo de inactividad mínimo. Para más información, consulte Migración de puertas de enlace de ExpressRoute a SKU habilitadas para zona de disponibilidad.

Comportamiento cuando todas las zonas están en buen estado

En la sección siguiente se describe qué esperar cuando la puerta de enlace de red virtual está configurada para la redundancia de zona y todas las zonas de disponibilidad están operativas.

- **Enrutamiento de tráfico entre zonas:** El tráfico del entorno local se distribuye entre las máquinas virtuales de puerta de enlace en todas las zonas que usa la puerta de enlace. Esta configuración activa-activa garantiza un rendimiento y una distribución de carga óptimos en condiciones de funcionamiento normales.

  Si usa FastPath para optimizar el rendimiento, el tráfico del entorno local omite la puerta de enlace, lo que mejora el rendimiento y reduce la latencia. Para obtener más información, consulte FastPath de ExpressRoute.

- **Replicación de datos entre zonas:** No se produce ninguna replicación de datos entre zonas porque la puerta de enlace de red virtual no almacena datos persistentes del cliente.

- **Administración de máquinas virtuales de puerta de enlace:** La plataforma selecciona automáticamente las zonas de las máquinas virtuales de puerta de enlace y administra la ubicación en todas las zonas. La supervisión del estado garantiza que solo las máquinas virtuales de puerta de enlace correctas reciban tráfico.

Comportamiento durante un fallo de zona

En la sección siguiente se describe qué esperar cuando la puerta de enlace de red virtual está configurada para zona de redundancia y hay una interrupción en la zona de disponibilidad.

- **Detección y respuesta:** La plataforma Azure detecta y responde a un error en una zona de disponibilidad. No es necesario iniciar una conmutación por error de zona.

- **Notificación**: Microsoft no le notifica automáticamente cuando una zona está inactiva. Sin embargo, puede usar Azure Resource Health para supervisar el estado de un recurso individual y puede configurar alertas de Resource Health para notificarle problemas. También puede usar Azure Service Health para comprender el estado general del servicio, incluidos los errores de zona, y puede configurar alertas de Service Health para notificarle problemas.

- **Solicitudes activas:** Se finalizan las solicitudes activas conectadas a través de máquinas virtuales de puerta de enlace en la zona que falla. Las aplicaciones cliente deben reintentar las solicitudes siguiendo las instrucciones para controlar errores transitorios.

- **Pérdida de datos esperada:** Los errores de zona no deben provocar pérdida de datos porque las puertas de enlace de red virtual no almacenan datos persistentes de los clientes.
- **Tiempo de inactividad esperado:** Durante las interrupciones de zona, las conexiones pueden experimentar interrupciones breves que suelen durar hasta un minuto cuando se redistribuye el tráfico. Las aplicaciones cliente deben reintentar las solicitudes siguiendo las instrucciones para controlar errores transitorios.
- **Reenrutamiento del tráfico:** La plataforma distribuye automáticamente el tráfico a las máquinas virtuales de puerta de enlace en zonas correctas.

  Las conexiones habilitadas para FastPath mantienen el enrutamiento optimizado a lo largo del proceso de conmutación por error, lo que garantiza un efecto mínimo en el rendimiento de la aplicación.

Recuperación de zona

Cuando se recupera la zona de disponibilidad afectada, Azure restaura automáticamente las máquinas virtuales de puerta de enlace en la zona recuperada y vuelve a la distribución de tráfico normal en todas las zonas que usa la puerta de enlace.

Prueba de fallos de zona

La plataforma Azure gestiona el enrutamiento del tráfico, la conmutación por error y la conmutación por recuperación para las puertas de enlace de red virtual con redundancia de zona. Esta característica está totalmente administrada, por lo que no es necesario iniciar ni validar los procesos de error de zona de disponibilidad.

Resistencia a errores en toda la región

Una puerta de enlace de red virtual es un recurso de una sola región. Si la región deja de estar disponible, la puerta de enlace tampoco está disponible.

> ⓘ **Nota:**
>
> Puede usar la SKU premium de ExpressRoute cuando los recursos de Azure se distribuyen entre varias regiones. Pero la SKU Premium no afecta cómo se configura su puerta de enlace y todavía se implementa en una sola región. Para más información, consulte Introducción a ExpressRoute.

Soluciones personalizadas de varias regiones para la resistencia

Puede crear rutas de acceso de conectividad independientes a su entorno de Azure mediante uno o varios de los enfoques siguientes:

- Cree varios circuitos ExpressRoute que se conecten a puertas de enlace en diferentes regiones de Azure.
- Use una VPN de sitio a sitio como respaldo para el tráfico de interconexión privada.
- Use la conectividad a Internet como copia de seguridad del tráfico de emparejamiento de Microsoft.

Para obtener más información, consulte Diseño para la recuperación ante desastres mediante el emparejamiento privado de ExpressRoute.

Resistencia al mantenimiento del servicio

Azure realiza un mantenimiento normal en las puertas de enlace de red virtual para garantizar un rendimiento y seguridad óptimos. Durante estas ventanas de mantenimiento, se pueden producir algunas interrupciones del servicio, pero Azure diseña estas actividades para minimizar el efecto en la conectividad.

Durante las operaciones de mantenimiento planificado en las puertas de enlace de red virtual, el proceso se ejecuta secuencialmente en las máquinas virtuales de la puerta de enlace, no simultáneamente. Este proceso garantiza que una máquina virtual de puerta de enlace siempre permanezca activa durante el mantenimiento, lo que minimiza el impacto en las conexiones activas.

Para reducir la probabilidad de interrupciones inesperadas, puede configurar las ventanas de mantenimiento del gateway para que coincidan con los requisitos operativos.

Para más información, consulte Configuración del mantenimiento controlado por el cliente para puertas de enlace de ExpressRoute.

Acuerdo de nivel de servicio

El contrato de nivel de servicio (SLA) para los servicios de Azure describe la disponibilidad esperada de cada servicio y las condiciones que la solución deberá cumplir para lograr esa expectativa de disponibilidad. Para obtener más información, consulte Acuerdos de Nivel de Servicio para servicios en línea.

ExpressRoute proporciona un Acuerdo de Nivel de Servicio de alta disponibilidad que garantiza un tiempo de actividad elevado para las conexiones. Se aplican diferentes acuerdos de nivel de servicio de disponibilidad si despliega en varias ubicaciones de emparejamiento (sitios), si utiliza ExpressRoute Metro o si tiene una configuración de sitio único.

Contenido relacionado

- Introducción a ExpressRoute
- Diseñar y planificar ExpressRoute para resiliencia
- Diseño para alta disponibilidad mediante ExpressRoute
- Diseño para la recuperación ante desastres mediante el emparejamiento privado de ExpressRoute

ⓘ **Nota:** El autor ha creado este artículo con ayuda de inteligencia artificial. Más información

---

Last updated on 22/01/2026

Confiabilidad en Azure Firewall

Azure Firewall es un servicio de seguridad de red administrado basado en la nube que protege los recursos de Azure Virtual Network. Es un servicio de firewall con estado completo que incluye alta disponibilidad integrada y escalabilidad de nube sin restricciones.

Cuando se usa Azure, la confiabilidad es una responsabilidad compartida. Microsoft proporciona una variedad de funcionalidades para admitir resistencia y recuperación. Es responsable de comprender cómo funcionan esas funcionalidades dentro de todos los servicios que usa y de seleccionar las funcionalidades que necesita para cumplir los objetivos empresariales y los objetivos de tiempo de actividad.

En este artículo se describe cómo hacer que Azure Firewall sea resistente a una variedad de posibles interrupciones y problemas, como errores transitorios, interrupciones de zona de disponibilidad y interrupciones de regiones. También describe la resistencia durante el mantenimiento del servicio y resalta cierta información clave sobre el acuerdo de nivel de servicio (SLA) de Firewall.

Recomendaciones de implementación de producción

Para obtener información sobre cómo implementar Azure Firewall para admitir los requisitos de confiabilidad de la solución y cómo afecta la confiabilidad a otros aspectos de la arquitectura, consulte Procedimientos recomendados de arquitectura para Azure Firewall en Azure Well-Architected Framework.

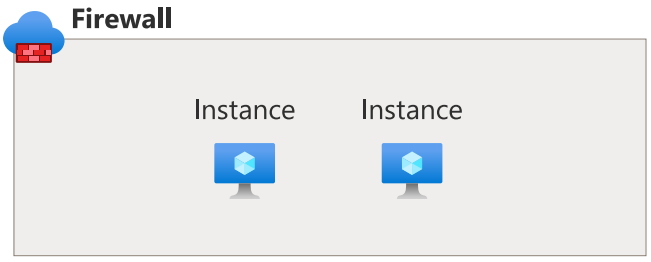
Introducción a la arquitectura de confiabilidad

Una *instancia* hace referencia a una unidad de nivel de máquina virtual (VM) del firewall. Cada instancia representa la infraestructura que controla el tráfico y realiza comprobaciones de firewall.

Para lograr una alta disponibilidad de un firewall, Azure Firewall proporciona automáticamente al menos dos instancias, sin necesidad de intervención o configuración. El firewall se escala horizontalmente automáticamente cuando el rendimiento promedio, el consumo de CPU y el uso de la conexión alcanzan umbrales predefinidos. Para más información, consulte Rendimiento de Azure Firewall. La plataforma administra automáticamente la creación de instancias, la supervisión del estado y el reemplazo de instancias incorrectas.

Para protegerse frente a errores de servidor y bastidor de servidores, Azure Firewall distribuye automáticamente las instancias entre varios dominios de error dentro de una región.

En el diagrama siguiente se muestra un firewall con dos instancias:



Para aumentar la redundancia y la disponibilidad durante los errores del centro de datos, Azure Firewall habilita automáticamente la redundancia de zona en regiones que admiten varias zonas de disponibilidad, distribuyendo instancias entre al menos dos zonas de disponibilidad.

Resistencia a errores transitorios

Los errores transitorios son errores breves e intermitentes en los componentes. Se producen con frecuencia en un entorno distribuido como la nube y son una parte normal de las operaciones. Los errores transitorios se corrigen después de un breve período de tiempo. Es importante que las aplicaciones puedan controlar errores transitorios, normalmente mediante el reintento de solicitudes afectadas.

Todas las aplicaciones hospedadas en la nube deben seguir las instrucciones de control de errores transitorios de Azure cuando se comunican con cualquier API, bases de datos y otros componentes hospedados en la nube. Para obtener más información, consulte Recomendaciones para controlar errores transitorios.

En el caso de las aplicaciones que se conectan a través de Azure Firewall, implemente la lógica de reintento con retroceso exponencial para controlar posibles problemas de conexión transitorios. La naturaleza con estado de Azure Firewall garantiza que las conexiones legítimas permanezcan activas durante breves interrupciones de red.

Durante las operaciones de escalado, que tardan cinco a siete minutos, el firewall conserva las conexiones existentes mientras agrega nuevas instancias de firewall para controlar la mayor carga.

Resistencia a errores de zona de disponibilidad

Las zonas de disponibilidad son grupos físicamente independientes de centros de datos dentro de una región de Azure. Cuando una zona falla, los servicios pueden transferirse a una de las zonas restantes.

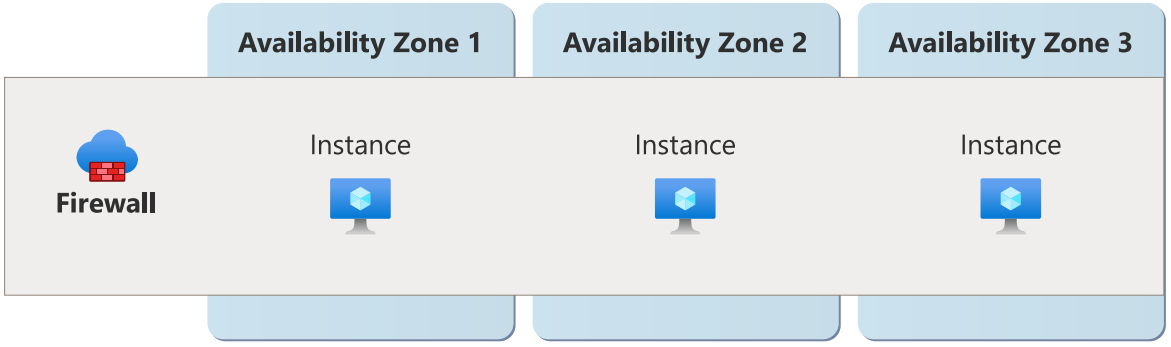
Azure Firewall se implementa automáticamente con redundancia de zona en regiones que admiten múltiples zonas de disponibilidad. Un firewall tiene redundancia de zona al implementarlo en al menos dos zonas de disponibilidad.

Azure Firewall admite modelos de implementación con redundancia de zona y zonales:

- **Con redundancia de zona:** En regiones que admiten zonas de disponibilidad, Azure distribuye automáticamente las instancias de firewall entre varias zonas de disponibilidad (al menos dos). Azure administra automáticamente el equilibrio de carga y la conmutación por error entre zonas. Este modelo de implementación es el valor predeterminado para todos los firewalls nuevos.

  Los firewalls con redundancia de zona logran el acuerdo de nivel de servicio (SLA) de tiempo de actividad más alto. Úselas para cargas de trabajo de producción que requieran la máxima disponibilidad.

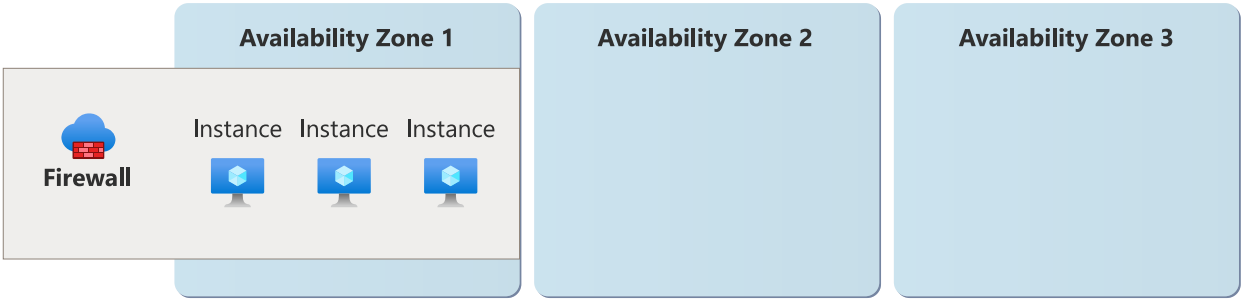
  En el diagrama siguiente se muestra un firewall con redundancia de zona con tres instancias distribuidas entre tres zonas de disponibilidad:



> ⓘ **Nota:**
>
> Todas las implementaciones de firewall en regiones con varias zonas de disponibilidad son automáticamente con redundancia de zona. Esta regla se aplica a las implementaciones a través de Azure Portal y implementaciones basadas en API (CLI de Azure, PowerShell, Bicep, plantillas de ARM, Terraform).

- **Zonal:** En escenarios específicos en los que existen restricciones de capacidad o los requisitos de latencia son críticos, puede implementar Azure Firewall en una zona de disponibilidad específica mediante herramientas basadas en API (CLI de Azure, PowerShell, Bicep, plantillas de ARM, Terraform). Implemente todas las instancias de un firewall zonal dentro de esa zona.

  En el diagrama siguiente se muestra un firewall zonal con tres instancias que se implementan en la misma zona de disponibilidad:



> **Importante**
>
> Solo puede crear implementaciones zonales a través de herramientas basadas en API. No se pueden configurar a través de Azure Portal. Las implementaciones de firewall zonales existentes se migrarán a implementaciones con redundancia de zona en el futuro. Use implementaciones con redundancia de zona siempre que sea posible para alcanzar el mayor nivel de disponibilidad según el SLA (Acuerdo de Nivel de Servicio). Un firewall zonal solo no proporciona resistencia a una interrupción de zona de disponibilidad.

Migración de implementaciones existentes

Anteriormente, las implementaciones de Azure Firewall que no están configuradas para que sean con redundancia de zona o zonales son *no zonales* o *regionales*. A lo largo del año natural 2026, Azure está migrando todas las implementaciones de firewall no zonales existentes a implementaciones con redundancia de zona en regiones que admiten varias zonas de disponibilidad.

Soporte para regiones

Azure Firewall admite zonas de disponibilidad en todas las regiones que admiten zonas de disponibilidad, donde el servicio Azure Firewall está disponible.

Requisitos

- Todos los niveles de Azure Firewall admiten zonas de disponibilidad.
- Los firewalls con redundancia de zona requieren direcciones IP públicas estándar configuradas para tener redundancia de zona.
- Los firewalls zonales (implementados mediante herramientas basadas en API) requieren direcciones IP públicas estándar y se pueden configurar para que sean con redundancia de zona o zonales en la misma zona que el firewall.

Cost

No hay ningún costo adicional para las implementaciones de firewall con redundancia de zona.

Configurar soporte de zonas de disponibilidad

En esta sección se explica la configuración de zona de disponibilidad para los firewalls.

- **Cree un nuevo firewall:** Todas las nuevas implementaciones de Azure Firewall en regiones con varias zonas de disponibilidad son automáticamente con redundancia de zona de forma predeterminada. Esta regla se aplica a las implementaciones basadas en el portal y basadas en API.

  - *Con redundancia de zona (valor predeterminado):* Al implementar un nuevo firewall en una región con varias zonas de disponibilidad, Azure distribuye automáticamente las instancias entre al menos dos zonas de disponibilidad. No se requiere ninguna configuración adicional. Para más información, consulte Implementar Azure Firewall mediante Azure Portal.
    - **Azure Portal:** Implementa automáticamente firewalls con redundancia de zona. No puede seleccionar una zona de disponibilidad específica a través del portal.
    - **Herramientas basadas en API (CLI de Azure, PowerShell, Bicep, plantillas de ARM, Terraform):** Implemente firewalls con redundancia de zona de forma predeterminada. Opcionalmente, puede especificar zonas para la implementación.

    Para más información sobre cómo implementar un firewall con redundancia de zona, consulte Implementación de Azure Firewall con zonas de disponibilidad.

  - *Zonal (solo herramientas basadas en API):* Para implementar un firewall en una zona de disponibilidad específica (por ejemplo, debido a restricciones de capacidad en una región), use herramientas basadas en API, como la CLI de Azure, PowerShell, Bicep, plantillas de ARM o Terraform. Especifique una sola zona en la configuración de implementación. Esta opción no está disponible a través de Azure Portal.

    > ⓘ **Nota:**
    >
    > Cuando selecciona qué zonas de disponibilidad usar, en realidad está seleccionando la *zona de disponibilidad lógica*. Si implementa otros componentes de la carga de trabajo en una suscripción de Azure diferente, podrían usar un número de zona de disponibilidad lógica distinto para acceder a la misma zona de disponibilidad física. Para obtener más información, consulte **zonas de disponibilidad física y lógica**.

- **Firewalls existentes:** Todas las implementaciones de firewall no zonales (regionales) existentes se migran automáticamente a implementaciones con redundancia de zona en regiones que admiten varias zonas de disponibilidad. Las implementaciones de firewall zonales existentes (ancladas a una zona específica) se migran a implementaciones con redundancia de zona en una fecha futura.

- **Restricciones de capacidad:** Si una región no tiene capacidad para una implementación con redundancia de zona (que requiere al menos dos zonas de disponibilidad), se produce un error en la implementación. En este escenario, puede implementar un firewall zonal en una zona de disponibilidad específica mediante herramientas basadas en API.

Comportamiento cuando todas las zonas están en buen estado

En esta sección se describe qué esperar cuando Azure Firewall está configurado con compatibilidad con la zona de disponibilidad y todas las zonas de disponibilidad están operativas.

- **Enrutamiento de tráfico entre zonas:** El comportamiento de enrutamiento del tráfico depende de la configuración de zona de disponibilidad que usa el firewall.

  - *Con redundancia de zona:* Azure Firewall distribuye automáticamente las solicitudes entrantes entre instancias de todas las zonas que usa el firewall. Esta configuración activa-activa garantiza un rendimiento y una distribución de carga óptimos en condiciones de funcionamiento normales.

  - *Zonal:* Si implementa varias instancias zonales en distintas zonas, debe configurar el enrutamiento de tráfico mediante soluciones de equilibrio de carga externas, como Azure Load Balancer o Azure Traffic Manager.

- **Administración de instancias:** La plataforma administra automáticamente la ubicación de la instancia en las zonas que usa el firewall. Reemplaza las instancias con errores y mantiene el recuento de instancias configurados. El seguimiento de estado garantiza que solo las instancias correctas reciban tráfico.

- **Replicación de datos entre zonas:** Azure Firewall no necesita sincronizar el estado de conexión entre zonas de disponibilidad. La instancia que procesa la solicitud mantiene el estado de cada conexión.

Comportamiento durante un fallo de zona

En esta sección se describe qué esperar cuando Azure Firewall está configurado con compatibilidad con la zona de disponibilidad y una o varias zonas de disponibilidad no están disponibles.

- **Detección y respuesta:** La responsabilidad de la detección y la respuesta depende de la configuración de la zona de disponibilidad que usa el firewall.
  - *Con redundancia de zona:* En el caso de las instancias configuradas para usar la redundancia de zona, la plataforma Azure Firewall detecta y responde a un error en una zona de disponibilidad. No es necesario iniciar una conmutación por error de zona.
  - *Zonal:* Para que los firewalls configurados sean zonales, debe detectar la pérdida de una zona de disponibilidad e iniciar una conmutación por error a un firewall secundario que cree en otra zona de disponibilidad.
- **Notificación:** Microsoft no le notifica automáticamente cuando una zona está inactiva. Sin embargo, puede usar Azure Service Health para comprender el estado general del servicio, incluidos los errores de zona, y puede configurar alertas de Service Health para notificarle problemas.
- **Conexiones activas:** Cuando una zona de disponibilidad no está disponible, las solicitudes en curso que se conectan a una instancia de firewall en la zona de disponibilidad errónea pueden finalizar y requerir reintentos.
- **Pérdida de datos esperada:** No se espera ninguna pérdida de datos durante la conmutación por error de zona porque Azure Firewall no almacena datos persistentes del cliente.
- **Tiempo de inactividad esperado:** El tiempo de inactividad depende de la configuración de zona de disponibilidad que use el firewall.
  - *Con redundancia de zona:* Espere un tiempo de inactividad mínimo (normalmente unos segundos) durante una interrupción de la zona de disponibilidad. Las aplicaciones cliente deben seguir los procedimientos para el control de errores transitorios, incluida la implementación de directivas de reintento con retroceso exponencial.
  - *Zonal:* Cuando una zona no está disponible, el firewall sigue sin estar disponible hasta que se recupere la zona de disponibilidad.
- **Redireccionamiento del tráfico:** El comportamiento de redireccionar el tráfico depende de la configuración de zona de disponibilidad que usa el firewall.
  - *Con redundancia de zona:* El tráfico se vuelve a enrutar automáticamente a zonas de disponibilidad correctas. Si es necesario, la plataforma crea nuevas instancias de

firewall en zonas correctas.

- *Zonal:* Cuando una zona no está disponible, el firewall zonal tampoco está disponible. Si tiene un firewall secundario en otra zona de disponibilidad, es responsable de redirigir el tráfico a ese firewall.

Failback

El comportamiento de la conmutación por recuperación depende de la configuración de zona de disponibilidad que usa el firewall.

- *Con redundancia de zona:* Una vez recuperada la zona de disponibilidad, Azure Firewall redistribuye automáticamente las instancias de todas las zonas que usa el firewall y restaura el equilibrio de carga normal entre zonas.
- *Zonal:* Una vez recuperada la zona de disponibilidad, es responsable de redirigir el tráfico al firewall en la zona de disponibilidad original.

Prueba de fallos de zona

Las opciones para las pruebas de errores de zona dependen de la configuración de la zona de disponibilidad del firewall.

- *Con redundancia de zona:* La plataforma Azure Firewall administra el enrutamiento del tráfico, la conmutación por error y la conmutación por recuperación para los recursos de firewall con redundancia de zona. Esta característica está totalmente administrada, por lo que no es necesario iniciar ni validar los procesos de error de zona de disponibilidad.
- *Zonal:* Puede simular aspectos del error de una zona de disponibilidad deteniendo un firewall. Use este enfoque para probar cómo otros sistemas y equilibradores de carga controlan una interrupción en el firewall. Para más información, consulte Detener e iniciar Azure Firewall.

Resistencia a errores en toda la región

Azure Firewall es un servicio de una sola región. Si la región no está disponible, el recurso firewall tampoco está disponible.

Soluciones personalizadas de varias regiones para la resistencia

Para implementar una arquitectura de varias regiones, use firewalls independientes. Este enfoque requiere implementar un firewall independiente en cada región, enrutar el tráfico al firewall regional adecuado e implementar la lógica de conmutación por error personalizada. Considere los siguientes puntos:

- **Use Azure Firewall Manager** para la administración centralizada de directivas en varios firewalls. Use el método de Directiva de firewall para la administración centralizada de reglas en varias instancias de firewall.
- **Implemente el enrutamiento de tráfico** mediante Traffic Manager o Azure Front Door.

Para obtener una arquitectura de ejemplo que muestra arquitecturas de seguridad de red de varias regiones, consulte Equilibrio de carga de varias regiones mediante Traffic Manager, Azure Firewall y Application Gateway.

Resistencia al mantenimiento del servicio

Azure Firewall realiza periódicamente actualizaciones de servicio y otras formas de mantenimiento.

Puede configurar ventanas de mantenimiento diarias para alinear las programaciones de actualización con sus necesidades operativas. Para más información, consulte Configuración del mantenimiento controlado por el cliente para Azure Firewall.

Acuerdo de nivel de servicio

El contrato de nivel de servicio (SLA) para los servicios de Azure describe la disponibilidad esperada de cada servicio y las condiciones que la solución deberá cumplir para lograr esa expectativa de disponibilidad. Para obtener más información, consulte Acuerdos de Nivel de Servicio para servicios en línea.

Azure Firewall proporciona un Acuerdo de Nivel de Servicio de mayor disponibilidad para firewalls con redundancia de zona implementados en dos o más zonas de disponibilidad.

Contenido relacionado

- Información general de Azure Firewall
- Características de Azure Firewall
- Procedimientos recomendados de Azure Firewall para el rendimiento
- Firewall Manager

ⓘ **Nota:** El autor ha creado este artículo con ayuda de inteligencia artificial. Más información

Last updated on 10/02/2026

Confiabilidad en Azure Load Balancer

Azure Load Balancer es un servicio de equilibrio de carga de nivel 4 para el tráfico del Protocolo de control de transmisión (TCP) y el protocolo de datagramas de usuario (UDP) que distribuye las solicitudes entrantes entre las instancias correctas de los servicios. Load Balancer proporciona alta disponibilidad y rendimiento de red de latencia ultra baja.

Cuando se usa Azure, la confiabilidad es una responsabilidad compartida. Microsoft proporciona una variedad de funcionalidades para admitir resistencia y recuperación. Es responsable de comprender cómo funcionan esas funcionalidades dentro de todos los servicios que usa y de seleccionar las funcionalidades que necesita para cumplir los objetivos empresariales y los objetivos de tiempo de actividad.

En este artículo se describe cómo hacer que Load Balancer sea resistente a una variedad de posibles interrupciones y problemas, incluidos errores transitorios, interrupciones de zona de disponibilidad y interrupciones de regiones. También resalta información clave sobre el acuerdo de nivel de servicio (SLA) de Load Balancer.

> **ⓘ Importante**
>
> La confiabilidad de la solución general depende de la configuración de las instancias de back-end (servidores), como máquinas virtuales (VM) de Azure o conjuntos de escalado de máquinas virtuales a los que el equilibrador de carga enruta el tráfico.
>
> En este artículo no se tratan las instancias de back-end, pero sus configuraciones de disponibilidad afectan directamente a la resistencia de la aplicación. Revise las **guías de confiabilidad de los servicios de Azure de la solución** para obtener información sobre cómo cada servicio admite sus requisitos de confiabilidad. Al configurar las instancias de back-end para alta disponibilidad y redundancia de zona, puede lograr una confiabilidad completa para la aplicación.

Recomendaciones de implementación de producción

Azure Well-Architected Framework proporciona recomendaciones sobre confiabilidad, seguridad, costo, operaciones y rendimiento. Para obtener información sobre cómo estas áreas influyen entre sí y contribuyen a una solución confiable de Load Balancer, consulte Procedimientos recomendados de arquitectura para Load Balancer.

Introducción a la arquitectura de confiabilidad

Un equilibrador de carga puede ser público o interno. Se puede acceder a un equilibrador de carga público desde Internet a través de un recurso de dirección IP pública. Un equilibrador de carga interno solo es accesible desde dentro de la red virtual y otras redes que se conectan a la red virtual.

Cada equilibrador de carga consta de varios componentes:

- *Configuraciones ip de front-end* que reciben tráfico. Un equilibrador de carga público recibe tráfico en una dirección IP pública. Un equilibrador de carga interno recibe tráfico en una dirección IP dentro de la red virtual.
- *Grupos de back-end* que contienen una colección de *instancias de back-end* que pueden recibir tráfico, como máquinas virtuales individuales que ejecutan la aplicación.
- *Reglas del equilibrador de carga* que definen cómo el equilibrador de carga distribuye el tráfico desde un front-end a un grupo de instancias en el back-end.
- *Sondas de estado* que supervisan la disponibilidad de las instancias de servidor del back-end.

Para obtener más información, consulte Componentes de Load Balancer.

En el caso de las soluciones implementadas globalmente, puede implementar un *equilibrador de carga global*, que es un tipo único de equilibrador de carga público para enrutar el tráfico entre diferentes implementaciones regionales de la solución. Un equilibrador de carga global proporciona una única dirección IP de anycast y enruta el tráfico al equilibrador de carga regional más cercano basándose en la proximidad del cliente y el estado de salud regional. Para obtener más información, consulte Resilience to region-wide failures (Resistencia a errores en toda la región).

Resistencia a errores transitorios

Los errores transitorios son errores breves e intermitentes en los componentes. Se producen con frecuencia en un entorno distribuido como la nube y son una parte normal de las operaciones. Los errores transitorios se corrigen después de un breve período de tiempo. Es importante que las aplicaciones puedan controlar errores transitorios, normalmente mediante el reintento de solicitudes afectadas.

Todas las aplicaciones hospedadas en la nube deben seguir las instrucciones de control de errores transitorios de Azure cuando se comunican con cualquier API, bases de datos y otros componentes hospedados en la nube. Para obtener más información, consulte Recomendaciones para controlar errores transitorios.

Al usar Load Balancer, tenga en cuenta los procedimientos recomendados siguientes para minimizar el riesgo de errores transitorios que afectan a la aplicación:

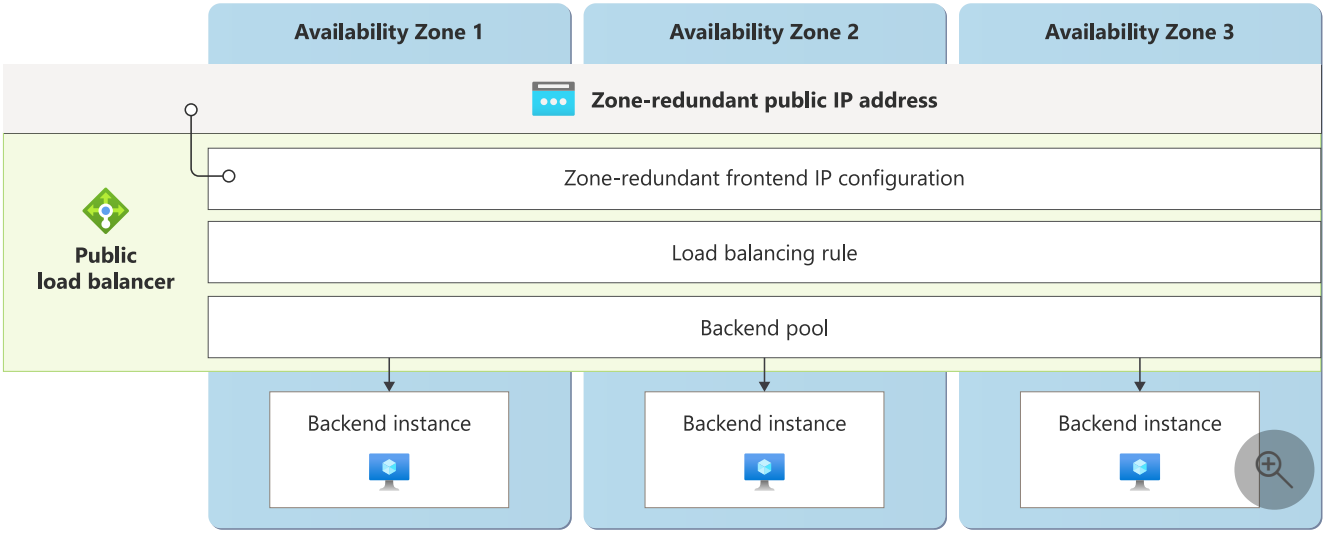
- **Implementación de la lógica de reintento.** Los clientes deben implementar mecanismos de reintento adecuados para errores de conexión transitorios, que incluyen estrategias de retroceso exponenciales.
- **Configure sondeos de estado con tolerancia.** Configure los sondeos de estado para equilibrar la detección rápida de errores con la necesidad de evitar falsos positivos durante problemas transitorios.
- **Supervisar la asignación de puertos SNAT.** Para las conexiones salientes, supervise la asignación de puertos de traducción de direcciones de red de origen (SNAT) y configure reglas de salida para evitar errores de conexión transitorios que se producen debido al agotamiento de puertos.

Resistencia a errores de zona de disponibilidad

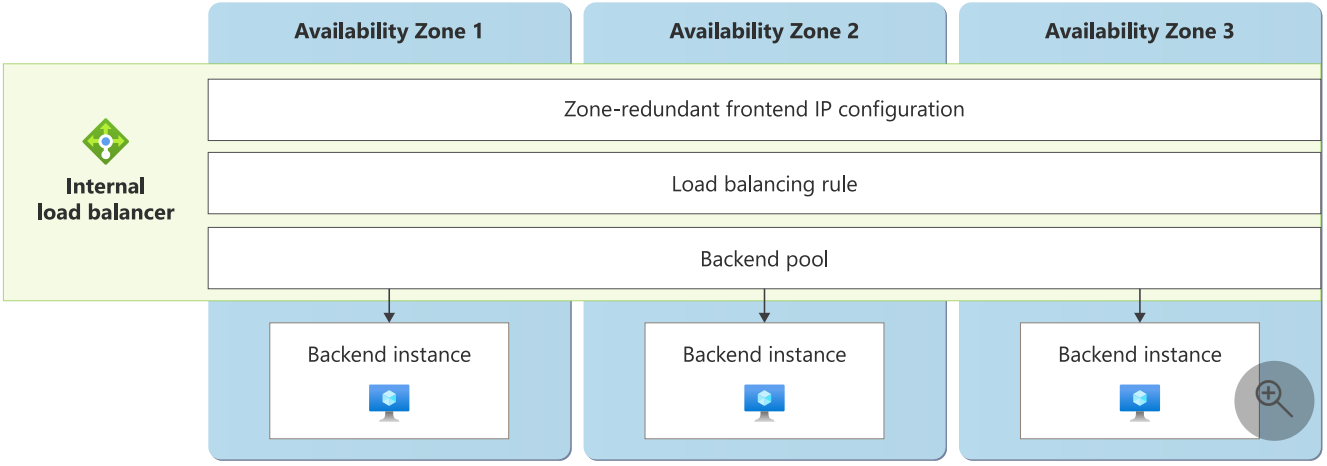
Las zonas de disponibilidad son grupos físicamente independientes de centros de datos dentro de una región de Azure. Cuando una zona falla, los servicios pueden transferirse a una de las zonas restantes.

Puede implementar el Load Balancer como *redundante por zona* configurando cada configuración de IP de front-end que usted cree. Una configuración de IP de front-end con redundancia de zona usa una infraestructura independiente en varias zonas para atender el tráfico simultáneamente. Esta configuración garantiza que los errores de zona no afecten a la capacidad del equilibrador de carga para recibir y distribuir el tráfico.

En el diagrama siguiente se muestra un equilibrador de carga público con redundancia de zona, que se configura mediante la creación de una dirección IP pública con redundancia de zona.



En el diagrama siguiente se muestra un equilibrador de carga interno que usa una configuración similar con redundancia de zona.



> **Nota:**
>
> Puede implementar equilibradores de carga zonales, pero se recomienda usar equilibradores de carga con redundancia de zona para todas las cargas de trabajo, incluidas las cargas de trabajo que implemente en una sola zona.

En regiones sin zonas de disponibilidad, Azure crea todos los equilibradores de carga en una configuración *regional* o no *zonal* mediante una configuración de front-end sin zona configurada. Si la región experimenta una interrupción, los equilibradores de carga nozonales pueden experimentar tiempo de inactividad. Las direcciones IP públicas creadas en una región que se actualice más adelante para tener zonas de disponibilidad serán redundantes a nivel de zona una vez que la región esté en estado de disponibilidad general con varias zonas de disponibilidad.

Instancias de back-end y zonas de disponibilidad

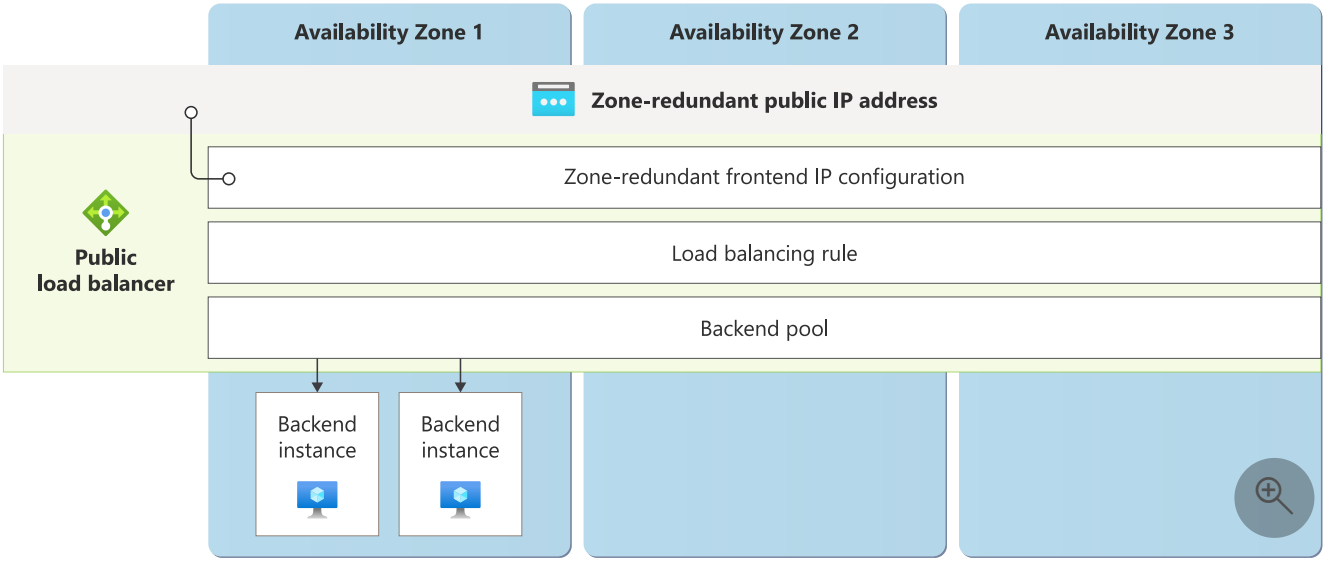
La configuración de zona de disponibilidad de las instancias de back-end funciona independientemente de la configuración de IP de front-end del equilibrador de carga.

Para distribuir las instancias de back-end entre zonas, configure el servicio pertinente según sus características de confiabilidad y sus requisitos de arquitectura.

> **Nota:**
>
> Para lograr resiliencia, distribuya las instancias de back-end entre varias zonas de disponibilidad. Si todas las instancias de back-end se ejecutan en una sola zona, una interrupción en esa zona hace que la aplicación no esté disponible, incluso si usa un equilibrador de carga con redundancia de zona.

Por ejemplo, cuando se usan máquinas virtuales, un enfoque de diseño común para cargas de trabajo de producción consiste en colocar varias máquinas virtuales zonales en las zonas 1, 2 y 3 para lograr resistencia de zona. Para realizar el equilibrio de carga, puede crear un equilibrador de carga redundante por zonas y configurar esas máquinas virtuales como instancias de back-end para el equilibrador. Los sondeos de estado del balanceador de carga eliminan automáticamente las máquinas virtuales no saludables de la rotación independientemente de su ubicación de zona.

Si implementa las máquinas virtuales en la misma zona de disponibilidad, puede implementar una configuración ip de front-end con redundancia de zona en el equilibrador de carga, como se muestra en el diagrama siguiente.



Requisitos

Compatibilidad con regiones: Puede implementar equilibradores de carga con redundancia de zona en cualquier región que admita zonas de disponibilidad.

Cost

La configuración de la zona de disponibilidad no cambia la forma en que Azure factura un equilibrador de carga. Azure basa la facturación en el número de reglas que configure y los datos que procesa, independientemente de la configuración de zona. Para más información, consulte los precios de Azure Load Balancer.

Configurar soporte de zonas de disponibilidad

Al utilizar un Load Balancer, se activa la compatibilidad de zona de disponibilidad en la configuración del IP de front-end.

- **Cree un nuevo equilibrador de carga con soporte para zonas de disponibilidad.**
  - En el caso de *los equilibradores de carga públicos*, la configuración de IP de front-end adopta automáticamente la configuración de zona de disponibilidad del recurso de dirección IP pública que se asocia a él. Para que la zona de configuración de IP de front-end sea redundante, cree o elija una dirección IP pública con redundancia de zona. Azure configura las direcciones IP públicas como redundantes de zona de forma predeterminada. Para más información, consulte Creación de un equilibrador de carga público para equilibrar la carga de máquinas virtuales mediante Azure Portal.
  - En el caso de *los equilibradores de carga internos*, al configurar la dirección IP de front-end del equilibrador de carga, se establece el tipo de compatibilidad de zona de disponibilidad en la configuración ip de front-end. Para más información, consulte Creación de un equilibrador de carga interno para equilibrar la carga de las máquinas virtuales mediante Azure Portal.
- **Cambie la configuración de zona de disponibilidad de un equilibrador de carga existente.** Para cambiar la configuración de zona de disponibilidad de un equilibrador de carga existente, reemplace la configuración de IP de front-end. Puede usar este enfoque para pasar de una configuración zonal a una configuración ip de front-end con redundancia de zona:

  1. Cree una nueva configuración de IP de front-end que tenga la configuración de zona de disponibilidad deseada.

     En el *caso de los equilibradores de carga públicos*, cree una nueva dirección IP pública que use la configuración de zona de disponibilidad deseada. A continuación, vuelva a configurar el equilibrador de carga para agregar una configuración ip de front-end que haga referencia a esa dirección IP pública.

En el *caso de los equilibradores de carga internos*, vuelva a configurar el equilibrador de carga para agregar una nueva configuración ip de front-end con la configuración de zona de disponibilidad deseada. Este paso asigna una nueva dirección IP privada desde la subred.

2. Vuelva a configurar las reglas de equilibrio de carga para usar la nueva configuración de IP de front-end.

   > **Importante**
   >
   > Esta operación requiere que vuelva a configurar los clientes para enviar tráfico a la nueva dirección IP de front-end. En función de los clientes, el proceso puede requerir tiempo de inactividad.

3. Quite la configuración anterior de IP de front-end.

Comportamiento cuando todas las zonas están en buen estado

En esta sección se describe qué esperar cuando un equilibrador de carga usa una configuración ip de front-end con redundancia de zona y todas las zonas de disponibilidad funcionan normalmente.

- **Enrutamiento de tráfico entre zonas:** El equilibrio de carga puede producirse en cualquier zona de disponibilidad. El equilibrador de carga envía tráfico a instancias de back-end saludables que especificáis en el grupo de back-end, sin tener en cuenta la región de disponibilidad que contiene la instancia de back-end.

- **Replicación de datos entre zonas:** Load Balancer funciona como un servicio de paso a través de red que no almacena ni replica los datos de la aplicación. Incluso si configura la persistencia de sesión en el equilibrador de carga, el equilibrador de carga no almacena ningún estado. La persistencia de la sesión ajusta el proceso de hash para dirigir las solicitudes a la misma instancia de back-end, aunque el equilibrador de carga no garantiza dicha persistencia. Cuando cambia el grupo de back-end, el equilibrador de carga vuelve a calcular la distribución de las solicitudes de cliente y no almacena ni sincroniza el estado.

  El servicio mantiene su estado de configuración a través de la replicación sincrónica entre zonas, lo que garantiza la coherencia inmediata de las reglas de equilibrio de carga, las configuraciones de sondeo de estado y la pertenencia al grupo de back-end en todas las zonas.

Comportamiento durante un fallo de zona

En esta sección se describe qué esperar cuando un balanceador de carga usa una configuración IP frontal con redundancia de zona y ocurre una falla en la zona de disponibilidad.

- **Detección y respuesta:** La plataforma Azure es responsable de detectar un error en una zona de disponibilidad y responder. No es necesario hacer nada para iniciar una conmutación por error de la zona.

- **Notificación**: Microsoft no le notifica automáticamente cuando una zona está inactiva. Sin embargo, puede usar Azure Resource Health para supervisar el estado de un recurso individual y puede configurar alertas de Resource Health para notificarle problemas. También puede usar Azure Service Health para comprender el estado general del servicio, incluidos los errores de zona, y puede configurar alertas de Service Health para notificarle problemas.

- **Solicitudes activas:** Cuando se produce un error en una zona, los flujos TCP y UDP existentes dentro de la zona se restablecen automáticamente y los clientes deben reintentar esos flujos. Los clientes deben implementar un control de errores transitorio suficiente, incluidos los reintentos automatizados.

- **Pérdida de datos esperada:** Load Balancer es un servicio de red sin estado, por lo que no almacena los datos de la aplicación y no se produce ninguna pérdida de datos en la capa del equilibrador de carga.

- **Tiempo de inactividad esperado:** No se espera ningún tiempo de inactividad para los equilibradores de carga con redundancia zonal ya que los equilibradores de carga siguen funcionando desde zonas saludables.

  Si el error afecta a los servicios de proceso de la zona, es posible que las máquinas virtuales u otros recursos de la zona afectada no estén disponibles. Los sondeos de estado del equilibrador de carga detectan estos errores y enrutan el tráfico a instancias alternativas en otra zona, según el algoritmo de equilibrio de carga y el estado de salud de las instancias de back-end.

- **Reenrutamiento del tráfico:** El equilibrador de carga sigue funcionando desde las zonas correctas. El Balanceador de Carga mantiene la misma dirección IP de front-end durante fallos de zona. Este comportamiento significa que no es necesario aplicar actualizaciones del sistema de nombres de dominio (DNS) ni volver a configurar clientes. La plataforma Azure establece nuevas conexiones de cliente a través de zonas correctas restantes.

Recuperación de zona

Cuando se recupera una zona de disponibilidad, Load Balancer reanuda automáticamente las operaciones normales. El front-end con redundancia de zona comienza automáticamente a atender el tráfico desde la zona recuperada junto con otras zonas operativas. Los sondeos de estado de la zona recuperada reanudan la evaluación de las instancias de back-end.

Si el error de zona también afecta a los servicios de proceso de esa zona, Load Balancer agrega automáticamente instancias de back-end al grupo de instancias de back-end saludable después de que las instancias se recuperen y pasen las comprobaciones de estado. La distribución del tráfico se rebalancea en todas las zonas disponibles en función del algoritmo de equilibrio de carga y del estado de salud de las instancias de fondo.

Prueba de fallos de zona

La plataforma Azure administra el enrutamiento del tráfico, la respuesta de zona inactiva y la recuperación, por lo que no es necesario iniciar ni validar los procesos de error de zona de disponibilidad.

Puede usar Azure Chaos Studio para simular el error de una máquina virtual en una sola zona. Chaos Studio proporciona errores integrados para las máquinas virtuales, incluido un error que apaga una máquina virtual. Puede usar estas funcionalidades para simular los fallos de zona y probar los procesos de failover.

Resistencia a errores en toda la región

Los equilibradores de carga públicos e internos se implementan en una sola región de Azure. Si la región deja de estar disponible, los equilibradores de carga de esa región también no estarán disponibles. Load Balancer proporciona compatibilidad nativa con varias regiones a través de un equilibrador de carga global, que admite el equilibrio de carga entre regiones de Azure. Además, puede desplegar otros servicios de balanceo de carga para enrutar y realizar conmutación por error entre las regiones de Azure.

Equilibradores de carga globales

Los equilibradores de carga globales proporcionan una única dirección IP anycast estática que automáticamente enruta el tráfico a la implementación regional óptima, basada en la proximidad del cliente y la salud regional. Los equilibradores de carga globales mejoran la confiabilidad y el rendimiento de la aplicación.

Con los equilibradores de carga globales, se implementan varios equilibradores de carga públicos en distintas regiones y un equilibrador de carga global actúa como front-end global. Si los servidores back-end de una región tienen un problema, el tráfico cambia

automáticamente a regiones saludables sin cambios de DNS porque la dirección IP anycast permanece constante y enruta el tráfico a otra región.

Para más información, consulte Equilibrador de carga global.

Soluciones personalizadas de varias regiones para la resistencia

Azure proporciona una gama de servicios de equilibrio de carga que se ajustan a diferentes requisitos. Elija un equilibrador de carga que cumpla los requisitos de resistencia y se adapte al tipo de aplicación. Para obtener más información, consulte Opciones de equilibrio de carga.

Acuerdo de nivel de servicio

El contrato de nivel de servicio (SLA) para los servicios de Azure describe la disponibilidad esperada de cada servicio y las condiciones que la solución deberá cumplir para lograr esa expectativa de disponibilidad. Para obtener más información, consulte Acuerdos de Nivel de Servicio para servicios en línea.

El Acuerdo de Nivel de Servicio de Load Balancer se aplica cuando tiene al menos dos máquinas virtuales correctas configuradas como instancias de back-end. El Acuerdo de Nivel de Servicio excluye el tiempo de inactividad debido al agotamiento de puertos SNAT o a los grupos de seguridad de red (NSG) que bloquean el tráfico.

Contenido relacionado

- Documentación de Load Balancer
- Introducción a la confiabilidad de Azure

ⓘ **Nota:** El autor ha creado este artículo con ayuda de inteligencia artificial. Más información

---

Last updated on 20/03/2026

Confiabilidad en Azure NAT Gateway

Azure NAT Gateway es un servicio de traducción de direcciones de red (NAT) totalmente administrado que proporciona conectividad saliente a Internet para los recursos conectados a la red virtual privada. El servicio proporciona la traducción de direcciones de red de origen (SNAT) para las conexiones salientes y la traducción de direcciones de red de destino (DNAT) solo para los paquetes de respuesta a las conexiones originadas desde el exterior. Dado que Azure NAT Gateway controla el tráfico de los recursos críticos de red virtual, está diseñado para proporcionar alta resistencia.

Cuando se usa Azure, la confiabilidad es una responsabilidad compartida. Microsoft proporciona una variedad de funcionalidades para admitir resistencia y recuperación. Es responsable de comprender cómo funcionan esas funcionalidades dentro de todos los servicios que usa y de seleccionar las funcionalidades que necesita para cumplir los objetivos empresariales y los objetivos de tiempo de actividad.

En este artículo se describe cómo hacer que azure NAT Gateway sea resistente a una variedad de posibles interrupciones y problemas, incluidos los errores transitorios y las interrupciones de zona de disponibilidad. También resalta información clave sobre el acuerdo de nivel de servicio (SLA) de Azure NAT Gateway.

> **Importante**
>
> Cuando considere la confiabilidad de una puerta de enlace NAT, considere también la confiabilidad de las máquinas virtuales (VM), discos, otra infraestructura de red y aplicaciones que se ejecutan en máquinas virtuales. Mejorar la resistencia de la puerta de enlace NAT podría tener un impacto limitado si los otros componentes no son igualmente resistentes. En función de los requisitos de resistencia, puede realizar cambios de configuración en varios componentes.

Recomendaciones de implementación de producción para la confiabilidad

En el caso de las cargas de trabajo de producción, se recomiendan los procedimientos siguientes:

- ✓ **Utiliza la SKU StandardV2** para obtener redundancia automática de zona en regiones compatibles.

> ⓘ **Nota:**
>
> Antes de la implementación, revise las limitaciones clave de StandardV2 Azure NAT Gateway para asegurarse de que admite la configuración.

- ✓ **Asigne suficientes direcciones IP públicas** para los requisitos máximos de conexión. Las direcciones IP insuficientes pueden provocar problemas de disponibilidad y agotamiento de puertos SNAT.
- ✓ **Utilice direcciones IP públicas de SKU StandardV2** con StandardV2 Azure NAT Gateway. StandardV2 Azure NAT Gateway no admite direcciones IP públicas de SKU estándar.

Introducción a la arquitectura de confiabilidad

En esta sección se describen algunos de los aspectos importantes de cómo funciona el servicio que es más relevante desde una perspectiva de confiabilidad. En la sección se presenta la arquitectura lógica, que incluye algunos de los recursos y características que se implementan y usan. También se describe la arquitectura física, que proporciona detalles sobre cómo funciona el servicio en segundo plano.

Arquitectura lógica

Una *puerta de enlace NAT* es un recurso que se implementa. Para usar la puerta de enlace NAT como ruta predeterminada para el tráfico saliente de Internet, asóciela a una o varias subredes de la red virtual. No es necesario configurar ninguna ruta personalizada ni otro enrutamiento.

Arquitectura física

Internamente, una puerta de enlace NAT consta de una o varias *instancias*, que representan la infraestructura subyacente necesaria para operar el servicio.

Azure NAT Gateway implementa una arquitectura distribuida mediante redes definidas por software (SDN) para proporcionar alta confiabilidad y escalabilidad. El servicio funciona en varios dominios de error para que pueda resistir varios errores de componentes de infraestructura sin impacto en el servicio. Azure administra las operaciones de servicio subyacentes, incluida la distribución entre dominios de error y redundancia de infraestructura.

Para más información sobre la arquitectura y la redundancia de Azure NAT Gateway, consulte Recursos de azure NAT Gateway.

Resistencia a errores transitorios

Los errores transitorios son errores breves e intermitentes en los componentes. Se producen con frecuencia en un entorno distribuido como la nube y son una parte normal de las operaciones. Los errores transitorios se corrigen después de un breve período de tiempo. Es importante que las aplicaciones puedan controlar errores transitorios, normalmente mediante el reintento de solicitudes afectadas.

Todas las aplicaciones hospedadas en la nube deben seguir las instrucciones de control de errores transitorios de Azure cuando se comunican con cualquier API, bases de datos y otros componentes hospedados en la nube. Para obtener más información, consulte Recomendaciones para controlar errores transitorios.

El agotamiento de puertos SNAT se produce cuando las aplicaciones realizan varias conexiones independientes a la misma dirección IP y puerto, lo que agota los puertos SNAT disponibles para la dirección IP saliente. El agotamiento de puertos SNAT puede aparecer como un error transitorio en la aplicación. Para reducir la probabilidad de errores transitorios relacionados con NAT, realice las siguientes tareas:

- **Minimice la probabilidad de agotamiento de puertos SNAT.** Configure las aplicaciones para controlar SNAT correctamente mediante la implementación de la agrupación de conexiones y la administración adecuada del ciclo de vida de la conexión.
- **Implemente suficientes direcciones IP públicas.** Una sola puerta de enlace NAT admite varias direcciones IP públicas y cada dirección IP pública proporciona un conjunto independiente de puertos SNAT.
- **Supervise la métrica de disponibilidad de la ruta de acceso de datos de la puerta de enlace NAT.** Use Azure Monitor para detectar posibles problemas de conectividad al principio. Configure alertas para errores de conexión y agotamiento de puertos SNAT para identificar y solucionar de forma proactiva las condiciones de error transitorias antes de que afecten a la conectividad saliente de las aplicaciones. Para más información, consulte Métricas y alertas de Azure NAT Gateway.
- **Evite establecer valores de tiempo de espera de inactividad elevados.** Al establecer valores de tiempo de espera de inactividad significativamente mayores que los 4 minutos

predeterminados para las conexiones de puerta de enlace NAT, el agotamiento de puertos SNAT puede producirse durante grandes volúmenes de conexión.

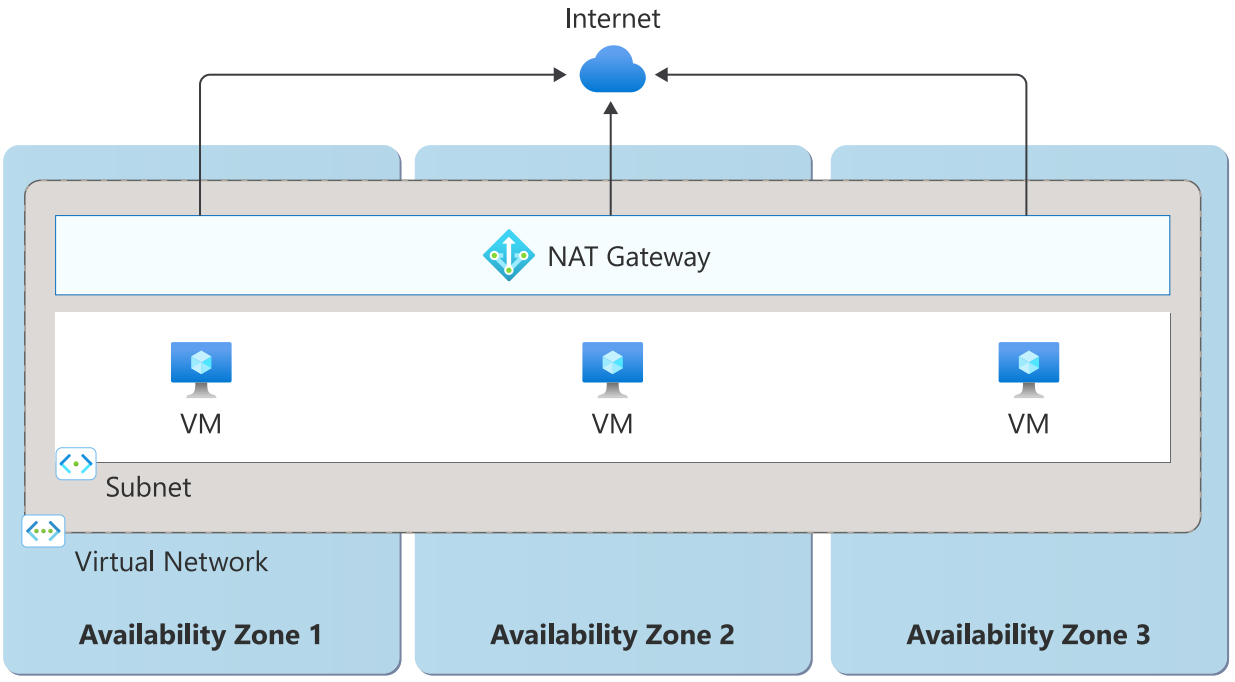
Para más información sobre la administración de conexiones y la solución de problemas en Azure NAT Gateway, consulte Solución de problemas de conectividad de Azure NAT Gateway.

Resistencia a errores de zona de disponibilidad

Las zonas de disponibilidad son grupos físicamente independientes de centros de datos dentro de una región de Azure. Cuando una zona falla, los servicios pueden transferirse a una de las zonas restantes.

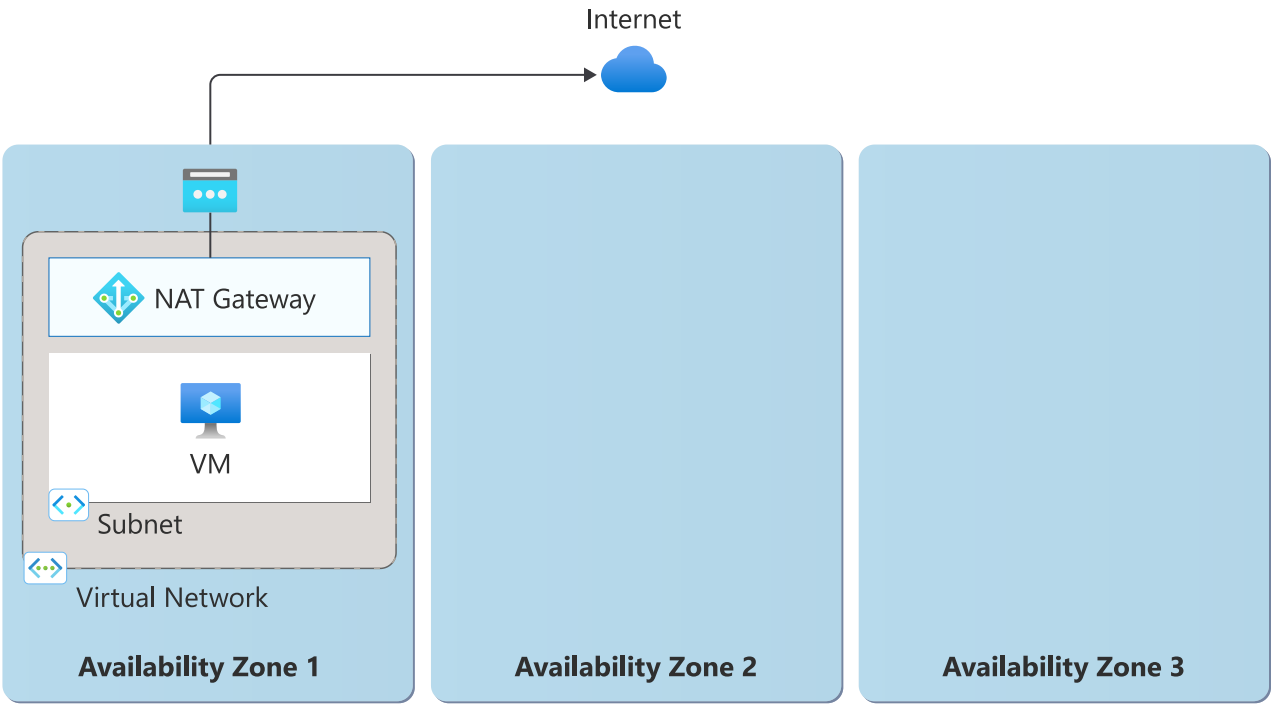
Azure NAT Gateway admite zonas de disponibilidad en configuraciones con redundancia de zona y zonales:

- **Con redundancia de zona:** La SKU de StandardV2 para Azure NAT Gateway proporciona redundancia de zona automática. La redundancia de zona propaga las instancias de puerta de enlace NAT en todas las zonas de disponibilidad de una región. Una configuración con redundancia de zona mejora la resistencia y la confiabilidad de las cargas de trabajo de producción.



- **Zonal:** Si usa la SKU Estándar (v1), puede crear opcionalmente una configuración zonal. Implemente una puerta de enlace NAT zonal en una zona de disponibilidad que seleccione. Al implementar una puerta de enlace NAT en una zona específica, proporciona conectividad saliente a Internet explícitamente desde esa zona. No se permiten direcciones IP públicas

zonales de una zona de disponibilidad diferente. Todo el tráfico de subredes conectadas se enruta a través de la puerta de enlace NAT, incluso si los recursos de subred residen en una zona de disponibilidad diferente.

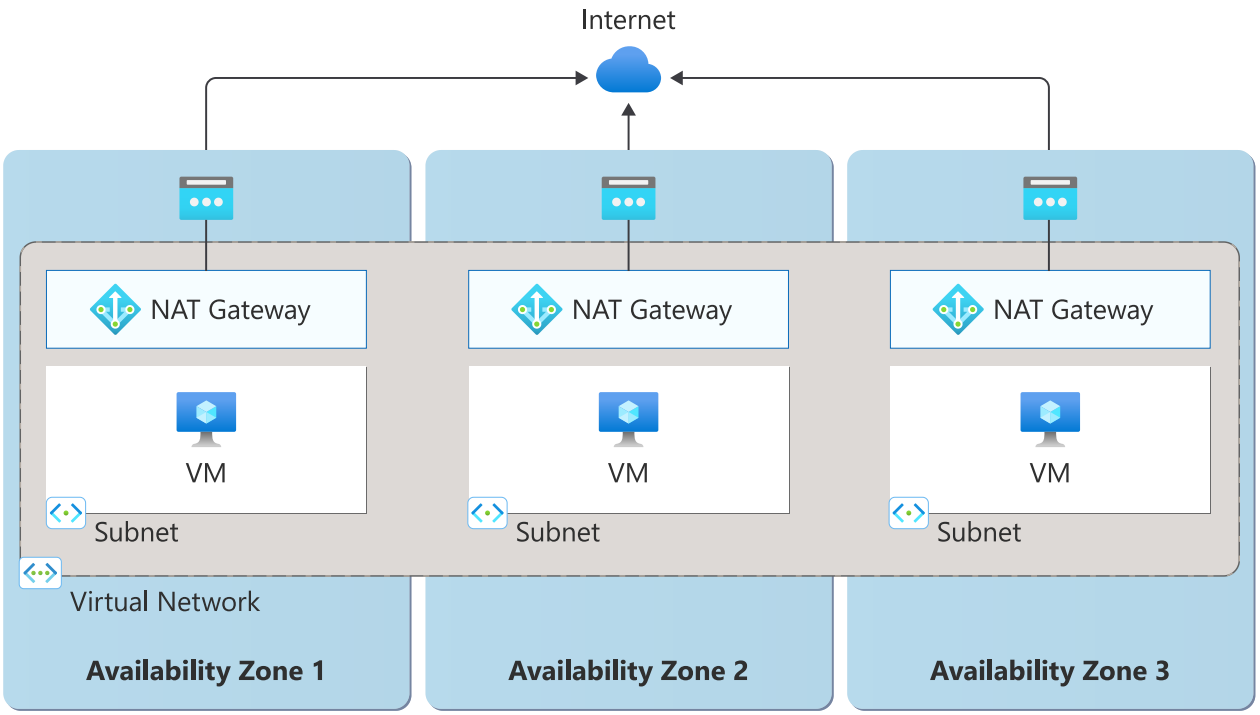


Si una puerta de enlace NAT dentro de una zona de disponibilidad experimenta una interrupción, todas las máquinas virtuales de las subredes conectadas no se pueden conectar a Internet, incluso si esas máquinas virtuales residen en zonas de disponibilidad correctas.

> **Importante**
>
> La asignación a una sola zona de disponibilidad solo se recomienda cuando la latencia entre zonas es demasiado alta para sus necesidades y después de comprobar que la latencia no cumple sus requisitos. Por sí mismo, un recurso zonal no proporciona resistencia a una interrupción de zona de disponibilidad. Para mejorar la resiliencia de un recurso zonal, debe desplegar explícitamente recursos independientes en múltiples zonas de disponibilidad y configurar el enrutamiento del tráfico y la conmutación por error. Para obtener más información, consulte Recursos zonales y resistencia de zona.

Si implementa máquinas virtuales en varias zonas de disponibilidad y necesita usar puertas de enlace NAT zonales, puede crear *pilas zonales* en cada zona de disponibilidad. Para crear pilas zonales, implemente los siguientes recursos:

- *Varias subredes:* Cree una subred independiente para cada zona de disponibilidad en lugar de usar una subred que abarque zonas.
- *Puertas de enlace NAT zonales:* Implemente una puerta de enlace NAT en la misma zona de disponibilidad que su subred conectada.
- *Asignación manual de máquinas virtuales:* Coloque cada máquina virtual en la zona de disponibilidad correcta y la subred correspondiente de la zona de disponibilidad.



Si implementa una puerta de enlace NAT estándar (v1) y no especifica una zona de disponibilidad, la puerta de enlace NAT es *no zonal*, lo que significa que Azure selecciona la zona de disponibilidad. Si alguna zona de disponibilidad de la región tiene una interrupción, podría afectar a la puerta de enlace NAT. No se recomienda una configuración no zonal porque no proporciona protección contra interrupciones de zona de disponibilidad.

Requisitos

- **Compatibilidad con regiones:** Puede implementar puertas de enlace NAT zonales y con redundancia de zona en cualquier región que admita zonas de disponibilidad.
- **SKU:** Para implementar una puerta de enlace NAT con redundancia de zona, use la SKU StandardV2. Para implementar una puerta de enlace NAT zonal, use la SKU estándar. Se recomienda la SKU StandardV2.

- **Direcciones IP públicas:** Los requisitos para las direcciones IP públicas asociadas a una puerta de enlace NAT dependen de la configuración de implementación y SKU.

Expandir tabla

SKU de Puerta de enlace NAT de Azure	Tipo de soporte de zona de disponibilidad	Requisitos de dirección IP pública
StandardV2	Con redundancia de zona	Dirección IP pública standardV2
Estándar	Zonal	Dirección IP pública estándar con redundancia de zona o en la misma zona que la puerta de enlace NAT
Estándar	No zonal	Dirección IP pública estándar con redundancia de zona o en cualquier zona

Cost

La compatibilidad con la zona de disponibilidad para Azure NAT Gateway no tiene ningún costo adicional. Para más información, consulte Precios de Azure NAT Gateway.

Configurar soporte de zonas de disponibilidad

- **Nuevos recursos:** Los pasos de implementación dependen de la configuración de la zona de disponibilidad:
  - *Con redundancia de zona:* Para implementar una nueva puerta de enlace NAT con redundancia de zona mediante la SKU standardV2, consulte Creación de una puerta de enlace NAT de Azure StandardV2.
  - *Zonal:* Para implementar una nueva puerta de enlace NAT zonal mediante la SKU estándar, consulte Creación de una puerta de enlace NAT. Al crear la puerta de enlace NAT, seleccione su zona de disponibilidad en lugar de seleccionar **Sin zona**.
- **Active la compatibilidad con las zonas de disponibilidad:** No se puede cambiar la configuración de zona de disponibilidad de Azure NAT Gateway una vez implementado. Para modificar la configuración de la zona de disponibilidad, debe implementar una nueva puerta de enlace NAT con la configuración de zona deseada.

  Para actualizar de una puerta de enlace NAT Estándar a StandardV2, debe crear una nueva dirección IP pública que use la SKU standardV2.

Comportamiento cuando todas las zonas están en buen estado

En esta sección se describe qué esperar cuando las puertas de enlace NAT están configuradas para la compatibilidad con la zona de disponibilidad y todas las zonas de disponibilidad están operativas.

- **Enrutamiento de tráfico entre zonas:** La forma en que el tráfico de la máquina virtual se enruta a través de la puerta de enlace NAT depende de la configuración de la zona de disponibilidad que use la puerta de enlace NAT.
  - *Con redundancia de zona:* El tráfico puede enrutarse a través de una instancia de puerta de enlace NAT en cualquier zona de disponibilidad.
  - *Zonal:* Cada instancia de puerta de enlace NAT funciona independientemente dentro de su zona de disponibilidad asignada. El tráfico saliente de los recursos de subred se enruta a través de la zona de la puerta de enlace NAT, incluso si la máquina virtual reside en una zona diferente.
- **Replicación de datos entre zonas:** Azure NAT Gateway no replica datos entre zonas porque es un servicio sin estado para la conectividad saliente. Cada instancia de puerta de enlace NAT funciona independientemente dentro de su zona de disponibilidad y no requiere sincronización con instancias de otras zonas.

Comportamiento durante un fallo de zona

En esta sección se describe qué ocurre cuando una puerta de enlace NAT está configurada para ser compatible con zonas de disponibilidad y no hay suficientes zonas de disponibilidad.

- **Detección y respuesta:** La responsabilidad de la detección y respuesta depende de la configuración de zona de disponibilidad que usa la puerta de enlace NAT.
  - *Con redundancia de zona:* Azure NAT Gateway detecta e responde a fallos en una zona de disponibilidad. No necesita hacer nada para iniciar una conmutación por error de una zona de disponibilidad.
  - *Zonal:* Es responsable de implementar la conmutación por error a nivel de aplicación en métodos de conectividad alternativos o puertas de enlace NAT en otras zonas.
- **Notificación:** Microsoft no le notifica automáticamente cuando una zona está inactiva. Sin embargo, puede usar Azure Resource Health para supervisar el estado de un recurso individual y puede configurar alertas de Resource Health para notificarle problemas.

También puede usar Azure Service Health para comprender el estado general del servicio, incluidos los errores de zona, y puede configurar alertas de Service Health para notificarle problemas.

También puede usar la métrica de disponibilidad de la ruta de acceso de datos de la puerta de enlace NAT para supervisar el estado de la puerta de enlace NAT. Configure alertas en la métrica de disponibilidad de la ruta de datos para detectar problemas de conectividad.

- **Solicitudes activas:** El comportamiento de la solicitud activa depende de la configuración de zona de disponibilidad que usa la puerta de enlace NAT.
  - *Zona redundante:* Las instancias en la zona defectuosa interrumpen las conexiones salientes activas. Los clientes deben reintentar sus solicitudes de conexión, y los intentos posteriores se enrutan a través de una instancia de NAT en otra zona de disponibilidad.
  - *Zonal:* Una puerta de enlace NAT zonal fallida interrumpe las conexiones salientes activas. Debe decidir si y cómo volver a establecer la conectividad a través de rutas de conectividad alternativas. Las aplicaciones deben implementar la lógica de reintento para controlar los errores de conexión.

    Si vuelve a enrutar el tráfico, la dirección IP pública saliente cambia, por lo que es posible que sea necesario volver a establecer las sesiones del Protocolo de control de transmisión (TCP).
- **Pérdida de datos esperada:** No se produce ninguna pérdida de datos porque Azure NAT Gateway es un servicio sin estado para la conectividad saliente. El estado de conexión se vuelve a crear cuando se vuelven a establecer las conexiones.
- **Tiempo de inactividad esperado:** El tiempo de inactividad esperado depende de la configuración de zona de disponibilidad que use la puerta de enlace NAT.
  - *Con redundancia en la zona:* Es posible que las conexiones existentes desde la zona fallida se desconecten. Los clientes pueden reintentar las conexiones inmediatamente y las solicitudes se enrutan a una instancia en otra zona.
  - *Zonal:* La conectividad saliente sigue sin estar disponible hasta que la zona se recupere o hasta que vuelva a enrutar el tráfico a través de métodos de conectividad alternativos o puertas de enlace NAT en otras zonas.
- **Reenrutamiento del tráfico:** El comportamiento de reenrutamiento del tráfico depende de la configuración de zona de disponibilidad que usa la puerta de enlace NAT.

- *Con redundancia de zona:* Las nuevas solicitudes de conexión se enrutan a través de una instancia de puerta de enlace NAT en una zona de disponibilidad saludable.

  Durante un error de zona, las máquinas virtuales de esa zona también suelen dejar de funcionar. Pero si un error de zona parcial afecta solo a la puerta de enlace NAT mientras las máquinas virtuales siguen ejecutándose, las conexiones salientes de esas máquinas virtuales se enrutan a través de una instancia de puerta de enlace NAT en otra zona.

- *Zonal:* Debe implementar la conmutación por error de nivel de aplicación, como usar métodos de conectividad alternativos o redirigir el tráfico a puertas de enlace NAT en otras zonas.

Recuperación de zona

Azure NAT Gateway es un servicio sin estado, por lo que las operaciones de recuperación no requieren intervención manual.

Cuando se recupera una zona de disponibilidad, las instancias de puerta de enlace NAT de esa zona estarán disponibles automáticamente para las nuevas conexiones salientes. Las conexiones establecidas a través de instancias de puerta de enlace NAT en otras zonas durante la interrupción siguen usando sus rutas de conectividad existentes hasta que se cierren esas conexiones.

Prueba de fallos de zona

Las opciones para probar errores de zona dependen de la configuración de zona de disponibilidad que usa la instancia.

- *Con redundancia de zona:* la plataforma de Azure NAT Gateway administra el enrutamiento del tráfico, la recuperación frente a errores y la restauración de puertas de enlaces NAT Gateways con redundancia de zona. Estas características administradas no requieren que inicie ninguna acción manual ni valide los procesos de error de zona de disponibilidad.

- *Zonal:* Es responsable de preparar y probar planes de recuperación ante desastres para gestionar una posible falla de zona.

Resistencia a errores en toda la región

Azure NAT Gateway es un servicio de una sola región que funciona dentro de los límites de una región de Azure específica. El servicio no incluye funciones nativas multirregionales ni

recuperación automática frente a errores entre regiones. Si una región deja de estar disponible, las puertas de enlace NAT de esa región tampoco están disponibles.

Si diseña un enfoque de red que abarca varias regiones, implemente puertas de enlace NAT independientes en cada región.

Acuerdo de nivel de servicio

El contrato de nivel de servicio (SLA) para los servicios de Azure describe la disponibilidad esperada de cada servicio y las condiciones que la solución deberá cumplir para lograr esa expectativa de disponibilidad. Para obtener más información, consulte Acuerdos de Nivel de Servicio para servicios en línea.

El Acuerdo de Nivel de Servicio de NAT de red virtual de Azure cubre Azure NAT Gateway. El acuerdo de nivel de servicio (SLA) de disponibilidad solo se aplica cuando tiene dos o más máquinas virtuales operativas. También excluye el agotamiento de puertos SNAT de los cálculos de tiempo de inactividad.

Contenido relacionado

- Introducción a Azure NAT Gateway
- Solución de problemas de conectividad de Azure NAT Gateway
- Azure NAT Gateway y Resource Health

ⓘ**Nota:** El autor ha creado este artículo con ayuda de inteligencia artificial. Más información

---

Last updated on 16/06/2026

Preguntas más frecuentes (P+F) sobre Network Watcher

En este artículo se proporcionan respuestas a las preguntas más frecuentes sobre Azure Network Watcher.

General

¿Qué es Network Watcher?

Network Watcher proporciona un conjunto de herramientas para supervisar. diagnosticar, ver métricas y habilitar o deshabilitar registros para recursos IaaS (Infraestructura como servicio), lo que incluye máquinas virtuales, redes virtuales, puerta de enlace de aplicación, equilibradores de carga y otros recursos en una red virtual de Azure. No es una solución para supervisar la infraestructura de PaaS (plataforma como servicio) u obtener análisis web/móviles.

¿Qué herramientas proporciona Network Watcher?

Network Watcher proporciona tres conjuntos principales de funcionalidades:

- Supervisión
  - Vista de la topología muestra los recursos de la red virtual y las relaciones entre ellos.
  - Connection Monitor le permite supervisar la conectividad y la latencia entre puntos de conexión dentro y fuera de Azure.
- Herramientas de diagnóstico de red
  - Comprobación del flujo de IP permite detectar incidencias en el filtrado del tráfico a nivel de máquina virtual.
  - Diagnóstico de NSG le permite detectar incidencias en el filtrado de tráfico en el nivel de máquina virtual, conjunto de escalado de máquinas virtuales o puerta de enlace de aplicación.
  - Próximo salto le ayuda a comprobar las rutas de tráfico y a detectar problemas de enrutamiento.
  - Solucionar problemas de conexión habilita una comprobación de conectividad y latencia que se realiza una única vez entre una máquina virtual y un host de Bastion, puerta de enlace de aplicación u otra máquina virtual.

    - Captura de paquetes le permite capturar el tráfico de su máquina virtual.
    - Solución de problemas de VPN ejecuta varias comprobaciones de diagnóstico en las puertas de enlace de VPN y las conexiones para facilitar la depuración de problemas.
- Tráfico
    - Los registros de flujo de red virtual y los registros de flujo de grupo de seguridad de red permiten registrar el tráfico de red que pasa por las redes virtuales y los grupos de seguridad de red (NSG) respectivamente.
    - Análisis de tráfico procesa los datos del registro de flujo del registro de flujo de grupo de seguridad de red para que pueda visualizar, consultar, analizar y comprender el tráfico.

Para información más detallada, consulte la Información general de Network Watcher.

¿Cómo son los precios de Network Watcher?

Consulte Precios de Network Watcher para más detalles sobre el precio de los distintos componentes de Network Watcher.

¿En qué regiones está admitido y disponible Network Watcher ahora?

Consulte Regiones de Network Watcher para saber las regiones que admiten Network Watcher.

¿Qué permisos se necesitan para usar Network Watcher?

Consulte Permisos de Azure RBAC requeridos para usar Network Watcher para una lista detallada de los permisos de cada capacidad de Network Watcher.

¿Cómo puedo habilitar Network Watcher?

El servicio Network Watcher se habilita automáticamente en cada suscripción. Debe activar manualmente Network Watcher si deshabilitó la activación automática de Network Watcher. Para obtener más información, consulte Habilitar o deshabilitar Azure Network Watcher.

¿Qué es el modelo de implementación de Network Watcher?

El recurso primario de Network Watcher se implementa con una instancia única en todas las regiones. Formato de nomenclatura predeterminado: NetworkWatcher_RegionName. Ejemplo: NetworkWatcher_centralus es el recurso de Network Watcher para la región "Central US" (Centro de EE. UU.). Puede personalizar el nombre de la instancia de Network Watcher mediante PowerShell o la API REST.

¿Por qué Azure permite una única instancia de Network Watcher por región?

Network Watcher solo necesita ser habilitado una vez por región y suscripción para que funcionen sus características. Network Watcher se habilita en una región al crear una instancia de Network Watcher en esa misma región.

¿Cómo puedo administrar el recurso Network Watcher?

El recurso Network Watcher representa el servicio back-end para Network Watcher, que está totalmente administrado por Azure. Sin embargo, puede crear o eliminar el recurso Network Watcher para habilitarlo o deshabilitarlo en una región determinada. Para obtener más información, consulte Habilitar o deshabilitar Azure Network Watcher.

¿Puedo trasladar mi instancia de Network Watcher de una región a otra?

No, no se admite el traslado del recurso Network Watcher o cualquiera de sus recursos secundarios entre regiones. Para más información, vea Compatibilidad con la operación de traslado para recursos de red.

¿Puedo trasladar mi instancia de Network Watcher de un grupo de recursos a otro?

Sí, se admite el traslado de recurso de Network Watcher entre grupos de recursos. Para más información, vea Compatibilidad con la operación de traslado para recursos de red.

¿Qué es NetworkWatcherRG?

NetworkWatcherRG es un grupo de recursos que se crea automáticamente para los recursos de Network Watcher. Por ejemplo, los recursos instancias regionales de Network Watcher y registro de flujo de grupo de seguridad de red se crean en el grupo de recursos **NetworkWatcherRG**. Puede personalizar el nombre del grupo de recursos de Network Watcher mediante PowerShell, la CLI de Azure o la API REST.

¿Almacena Network Watcher los datos de los clientes?

Azure Network Watcher no almacena los datos de los clientes, excepto para Connection Monitor. El monitor de conexión almacena los datos de los clientes, que Network Watcher almacena automáticamente en una sola región para satisfacer los requisitos de residencia de datos en la región.

¿Cuáles son los límites de recursos en Network Watcher?

Network Watcher tiene los límites siguientes:

Expandir tabla

Recurso	Límite
Instancias de Network Watcher por región por suscripción	1 (una instancia de una región para habilitar el acceso al servicio en la región)
Monitores de conexión por región por suscripción	100
Máximo de grupos de prueba por monitor de conexión	20
Máximo de orígenes y destinos por monitor de conexión	100
Máximo de configuraciones de prueba por monitor de conexión	20
Sesiones de captura de paquetes por región	10.000 (solo número de sesiones, sin capturas

Recurso	Límite
por suscripción	guardadas)
Operaciones de solución de problemas de VPN por suscripción	1 (número de operaciones a la vez)

Disponibilidad y redundancia del servicio

¿Es resistente a las zonas Network Watcher?

Sí, el servicio Network Watcher es resistente a las zonas de manera predeterminada.

¿Cómo se configura el servicio Network Watcher para que sea resistente a las zonas?

No se necesita configurar nada para habilitar la resistencia de zonas. La resistencia de zonas para los recursos de Network Watcher está disponible de manera predeterminada y la administra el propio servicio.

Agente de Network Watcher

¿Por qué es necesario instalar el agente Network Watcher?

El agente Network Watcher es necesario para cualquier función de Network Watcher que genere o intercepte tráfico desde una máquina virtual.

¿Qué características requiere el agente Network Watcher?

Las características Connection Monitor, Captura de paquetes y Solución de problemas de conexión (prueba de conectividad) requieren que la extensión Network Watcher esté presente.

¿Cuál es la última versión del agente de Network Watcher?

La versión más reciente de la extensión Network Watcher es `1.4.3914.3`.. Para más información, vea Actualización de la extensión Azure Network Watcher a la versión más reciente.

¿Qué puertos usa el agente de Network Watcher?

- **Linux**: el agente de Network Watcher usa puertos disponibles a partir de `port 50000` y posteriores hasta que alcanza `port 65535`.
- **Windows**: el agente de Network Watcher usa los puertos con los que responde el sistema operativo cuando se consulta para los puertos disponibles.

¿Con qué direcciones IP se comunica el agente de Network Watcher?

El agente de Network Watcher requiere conectividad TCP saliente para `169.254.169.254` a través de `port 80` y `168.63.129.16` a través de `port 8037`. El agente usa estas direcciones IP para comunicarse con la plataforma Azure.

Monitor de conexión

¿Admite Connection Monitor máquinas virtuales clásicas?

No, Connection Monitor no admite máquinas virtuales clásicas. Para más información, consulte Migración de recursos de IaaS de la implementación clásica a la de Resource Manager.

¿Qué pasa si mi topología no está decorada o falta información en mis saltos?

La topología solo se puede decorar desde fuera de Azure a Azure si el recurso de Azure de destino y el recurso de Connection Monitor están en la misma región.

¿Qué ocurre si en Connection Monitor la creación genera el siguiente error: "No se permite la creación de puntos de conexión diferentes para la misma máquina virtual"?

No se puede usar la misma máquina virtual de Azure con configuraciones diferentes en el mismo monitor de conexión. Por ejemplo, no se admite el uso de la misma VM con filtro y sin filtro en el mismo monitor de conexión.

¿Qué sucede si el motivo del error de la prueba es "No hay nada para mostrar"?

Los problemas que se muestran en el panel de Connection Monitor aparecen durante la detección de la topología o la exploración de saltos. Puede haber casos en los que el umbral establecido para el porcentaje de pérdida o RTT se incumpla, pero no se encuentre ningún problema en los saltos.

Al migrar una instancia de Connection Monitor existente (clásico) a la más reciente, ¿qué ocurre si las pruebas de punto de conexión externo se migran solo con el protocolo TCP?

No hay ninguna selección de protocolo en Connection Monitor (clásico). Las pruebas de Connection Monitor (clásico) solo usan el protocolo TCP. Por eso, durante la migración, creamos una configuración TCP en pruebas en la nueva instancia de Connection Monitor.

¿Existen limitaciones en el uso de Azure Monitor y agentes de Arc con el monitor de conexión?

Actualmente hay un límite regional cuando un punto de conexión usa agentes de Azure Monitor y Arc con el área de trabajo de Log Analytics asociada. Como resultado de esta limitación, el área de trabajo de Log Analytics asociada debe estar en la misma región que el punto de conexión de Arc. Los datos ingeridos en áreas de trabajo individuales se pueden unir para una sola vista, consulte Consulta de datos en áreas de trabajo, aplicaciones y recursos de Log Analytics en Azure Monitor.

Registros de flujos

¿Qué hace el registro de flujo?

Los registros de flujo le permiten registrar información en 5 tuplas sobre el tráfico de IP de Azure que pasa a través del grupo de seguridad de red o la red virtual de Azure. Los registros de flujo sin procesar se escriben en una cuenta de Azure Storage. Desde allí, puede procesarlos, analizarlos, consultarlos o exportarlos según necesite.

¿Los registros de flujo afectan a la latencia o el rendimiento de la red?

Los datos de los registros de flujo se recopilan fuera de la ruta del tráfico de red y, por tanto, no afectan al rendimiento o la latencia de la red. Puede crear o eliminar registros de flujo sin riesgo de que afecte al rendimiento de la red.

¿Cuál es la diferencia entre los registros de flujo de NSG y los diagnósticos de NSG?

Los registros de flujo del grupo de seguridad de red registran el tráfico que fluye a través de un grupo de seguridad de red. Por otro lado, los diagnósticos de NSG devuelven todos los grupos de seguridad de red que el tráfico atraviesa y las reglas de cada grupo de seguridad de red que se aplican a este tráfico. Use diagnósticos de NSG para comprobar que las reglas del grupo de seguridad de red se aplican según lo previsto.

¿Puedo registrar el tráfico que usa los protocolos ESP y AH mediante los registros de flujo de grupo de seguridad de red?

No, los registros de flujo de grupo de seguridad de red no admiten los protocolos ESP y AH.

¿Se puede registrar el tráfico ICMP mediante los registros de flujo?

No, los registros de flujo de grupo de seguridad de red y los registros de flujo de red virtual no admiten el protocolo ICMP.

¿Puedo eliminar un grupo de seguridad de red que tenga habilitado el registro de flujo?

Sí. El recurso de registro de flujo asociado también se eliminará. Los datos del registro de flujo se conservan en la cuenta de almacenamiento durante el período de retención configurado en el registro de flujo.

¿Puedo mover un grupo de seguridad de red que tiene un registro de flujo habilitado a otro grupo de recursos o suscripción?

Sí, pero debe eliminar el recurso de registro de flujo asociado. Después de migrar el grupo de seguridad de red, puede volver a crear los registros de flujo para habilitar el registro de flujo en él.

¿Puedo usar una cuenta de almacenamiento en una suscripción diferente del grupo de seguridad de red o la red virtual para la que está habilitado el registro de flujo?

Sí, puede usar una cuenta de almacenamiento de una suscripción diferente siempre que esta suscripción esté en la misma región del grupo de seguridad de red y asociada al mismo inquilino de Microsoft Entra del grupo de seguridad de red o la suscripción de la red virtual.

¿Cómo puedo asegurarme de que Traffic Analytics tiene visibilidad de los recursos dentro de una suscripción?

Debe habilitar al menos un registro de flujo de red virtual en la suscripción. Análisis de tráfico solo procesa datos de registros de flujo habilitados, por lo que la configuración de un registro de flujo garantiza que Traffic Analytics pueda acceder a los recursos de esa suscripción y analizar la actividad de red.

¿Cómo se usan los registros de flujo con una cuenta de almacenamiento detrás de un firewall?

Para usar una cuenta de almacenamiento detrás de un firewall, debe permitir que los *servicios de Azure de confianza* accedan a la cuenta de almacenamiento:

1. Vaya a la cuenta de almacenamiento introduciendo el nombre de la cuenta de almacenamiento en el cuadro de búsqueda en la parte superior del portal.
2. En **Seguridad y redes**, seleccione **Redes**, y después **Firewalls y redes de virtuales**.
3. En **Acceso de red pública** seleccione **Habilitado desde redes virtuales y direcciones IP seleccionadas**. Entonces, en **Excepciones**, marque la casilla junto a **Permitir que los servicios de Azure de la lista de servicios de confianza accedan a esta cuenta de almacenamiento.**
4. Habilite los registros de flujo mediante la creación de un registro de flujo para el recurso de destino mediante la cuenta de almacenamiento. Para obtener más información, consulte Crear un registro de flujo.

Puede comprobar los registros de almacenamiento tras unos minutos. Debería ver una marca de tiempo actualizada o un nuevo archivo JSON creado.

¿Cuántas reglas de directiva de retención pueden tener una cuenta de almacenamiento?

Actualmente, una cuenta de almacenamiento admite 100 reglas y cada una puede acomodar 10 prefijos de blob. Si alcanza el límite, puede establecer la directiva de retención en ***0*** al habilitar los nuevos registros de flujo de red virtual y, a continuación, agregar manualmente una regla de retención para la suscripción.

Para crear una regla de suscripción de directiva de retención, siga estos pasos:

1. Vaya a la cuenta de almacenamiento introduciendo el nombre de la cuenta de almacenamiento en el cuadro de búsqueda en la parte superior del portal.
2. En **Administración de datos**, seleccione **Administración del ciclo de vida**.
3. Seleccione + **Agregar regla**.
4. En la sección **Detalles**, seleccione los siguientes ajustes: **Ámbito de regla:***Limitar blobs con filtros*, **Tipo de blob:***Blobs de bloque* y **Subtipo de blob:***Blobs base*.
5. En la sección **Blobs base** , configure las opciones de retención.
6. En la sección **Conjunto de filtros** , dé formato al **prefijo Blob** como se indica a continuación: `"insights-logs-flowlogflowevent/flowLogResourceID=/<yourSubscriptionId>_NETWORKWATCHERRG"`
7. Selecciona **Agregar**.

¿Qué ocurre si tengo una regla de directiva de retención de registros de flujo y una regla de nivel de

suscripción al mismo tiempo?

La regla con un período de retención más corto tiene prioridad.

¿Por qué veo algunos errores 403 en los registros de actividad de la cuenta de almacenamiento?

Network Watcher tiene un mecanismo de reserva integrado que usa al conectarse a una cuenta de almacenamiento detrás de un firewall (firewall habilitado). Intenta conectarse a la cuenta de almacenamiento mediante una clave y, si se produce un error, cambia a un token. En este caso, se registra un error 403 en el registro de actividad de la cuenta de almacenamiento.

¿Puede Network Watcher enviar datos de registros de flujo a una cuenta de almacenamiento habilitada con el punto de conexión privado?

Sí, Network Watcher admite el envío de datos de registros de flujo a una cuenta de almacenamiento habilitada con un punto de conexión privado.

¿Cómo se usan los registros de flujo con una cuenta de almacenamiento detrás de un punto de conexión de servicio?

Los registros de flujo son compatibles con los puntos de conexión de servicio sin necesidad de ninguna configuración adicional. Para más información, consulte Habilitación de un punto de conexión de servicio.

¿Cuál es la diferencia entre la versión 1 y la versión 2 de los registros de flujo?

La versión 2 de los registros de flujo introduce el concepto de *estado del flujo* y almacena información sobre los bytes y los paquetes transmitidos. Para más información, consulte Formato de registros de flujo de grupo de seguridad de red.

¿Puedo crear un registro de flujo para un grupo de seguridad de red que tenga un bloqueo de

solo lectura?

No, un bloqueo de **solo lectura** en un grupo de seguridad de red impide la creación del registro de flujo de grupo de seguridad de red correspondiente.

¿Puedo crear un registro de flujo para un grupo de seguridad de red que tiene un bloqueo que no puede eliminar?

Sí, un bloqueo que **impide la eliminación** en el grupo de seguridad de red no impide la creación o modificación del registro de flujo de grupo de seguridad de red correspondiente.

¿Se pueden automatizar los registros de flujo de grupo de seguridad de red?

Sí, los registros de flujo de grupo de seguridad de red se pueden automatizar mediante plantillas de Azure Resource Manager (plantillas de ARM). Para más información, vea Configuración de registros de flujo de NSG mediante una plantilla de Azure Resource Manager (ARM).

Análisis de tráfico

¿Puedo habilitar los registros de flujo para los recursos que se encuentran en regiones diferentes de mi región de área de trabajo?

Sí, las redes virtuales y los grupos de seguridad de red pueden estar en diferentes regiones que la región del área de trabajo de Log Analytics.

¿Se pueden configurar varios grupos de seguridad de red dentro de una sola área de trabajo?

Sí.

¿Por qué el análisis de tráfico no muestra los datos de mis grupos de seguridad de red habilitados para

análisis de tráfico?

En la lista desplegable de selección de recursos del panel de análisis de tráfico, se debe seleccionar el grupo de recursos del recurso de **red virtual** , no el grupo de recursos de la máquina virtual o el grupo de seguridad de red.

¿Por qué algunos recursos aparecen como "Desconocidos" en el análisis de tráfico?

El análisis de tráfico realiza un examen de detección de recursos cada 6 horas para identificar nuevas máquinas virtuales, NIC, redes virtuales y subredes. Cuando se crea una nueva máquina virtual o NIC después del ciclo de detección más reciente y los datos de flujo se recopilan antes de la siguiente detección, el análisis de tráfico aún no puede asociar el tráfico a un recurso conocido. Como resultado, esos recursos se etiquetan temporalmente como *desconocidos* en la vista de análisis.

¿Puedo restringir el acceso público al recurso de punto de conexión de recopilación de datos (DCE) creado por el análisis de tráfico?

Sí, puede deshabilitar el acceso público al recurso del punto de conexión de recopilación de datos (DCE) para bloquear el tráfico entrante público. La ingesta sigue funcionando sin asociar el recurso DCE con un ámbito de Private Link de Azure Monitor.

¿Puedo usar un área de trabajo existente?

Sí. Si selecciona un área de trabajo existente, asegúrese de que se migró al nuevo lenguaje de consulta. Si no desea actualizar el área de trabajo, debe crear una nueva. Para obtener más información sobre el lenguaje de consultas Kusto (KQL), consulte Consultas de registro en Azure Monitor.

¿Mi cuenta de Almacenamiento de Azure puede estar en una suscripción y mi área de trabajo de Log Analytics está en otra suscripción?

Sí, la cuenta de Azure Storage puede estar en una suscripción y el área de trabajo de Log Analytics puede estar en otra suscripción.

¿Puedo almacenar registros sin procesar en una suscripción diferente a la que se usa para grupos de seguridad de red o redes virtuales?

Sí. Puede configurar los registros de flujo que se enviarán a una cuenta de almacenamiento ubicada en una suscripción diferente, siempre que tenga los privilegios adecuados y que la cuenta de almacenamiento se encuentre en la misma región que el grupo de seguridad de red (registros de flujo del grupo de seguridad de red) o la red virtual (registros de flujo de red virtual). La cuenta de almacenamiento de destino debe compartir el mismo inquilino de Microsoft Entra del grupo de seguridad de red o la red virtual.

¿Mis recursos de registro de flujo y las cuentas de almacenamiento pueden estar en distintos inquilinos?

No. Todos los recursos deben estar en el mismo inquilino, incluidos los grupos de seguridad de red (registros de flujo del grupo de seguridad de red), las redes virtuales (registros de flujo de red virtual), los registros de flujo, las cuentas de almacenamiento y las áreas de trabajo de Log Analytics (si el análisis de tráfico está habilitado).

¿Puedo configurar una directiva de retención diferente para la cuenta de almacenamiento que el área de trabajo de Log Analytics?

Sí.

Al ver mis libros de trabajo de Análisis de Tráfico, mis datos no se cargan debido a un "error de memoria baja". ¿Cómo puedo corregir esto para ver mis datos en el libro de trabajo?

Los libros de Análisis de tráfico tienen tecnología de consultas de Log Analytics. Si la consulta supera los límites de Log Analytics, el libro podría presentar errores de memoria baja. Para mejorar el rendimiento y reducir los errores de memoria baja, los usuarios pueden usar clústeres de Log Analytics dedicados.

¿Se perderán los datos almacenados en el área de trabajo de Log Analytics si elimino la cuenta de almacenamiento usada para el registro de flujo?

No. Si elimina la cuenta de almacenamiento que se usa para los registros de flujo, los datos almacenados en el área de trabajo de Log Analytics no se verán afectados. Todavía puede ver datos históricos en el área de trabajo de Log Analytics (algunas métricas se verán afectadas), pero el análisis de tráfico ya no procesará ningún registro de flujo nuevo hasta que actualice los registros de flujo para usar una cuenta de almacenamiento diferente.

¿Qué ocurre si no puedo configurar un grupo de seguridad de red para el análisis de tráfico debido a un error "No encontrado"?

Seleccione una región compatible. Si selecciona una región no admitida, recibirá un error "No encontrado". Para más información, consulte Regiones compatibles con Análisis de tráfico.

¿Qué ocurre si obtengo el estado: "No se pudo cargar" en la página de registros de flujo?

El proveedor debe estar registrado para que el `Microsoft.Insights` registro de flujo funcione correctamente. Si no está seguro de si el `Microsoft.Insights` proveedor está registrado para su suscripción, consulte las instrucciones de Administración de registros de flujo de NSG para saber cómo registrarlo.

He configurado la solución. ¿Por qué no veo nada en el panel?

El panel puede tardar hasta 30 minutos en mostrar informes por primera vez. La solución primero debe agregar suficientes datos para que derive información significativa y, a continuación, genere informes.

¿Qué ocurre si obtengo este mensaje: "No hemos encontrado ningún dato en esta área de trabajo durante el intervalo de tiempo seleccionado. Pruebe a

cambiar el intervalo de tiempo o seleccione una área de trabajo diferente".?

Pruebe las siguientes opciones:

- Cambie el intervalo de tiempo en la barra superior.
- Seleccione otro área de trabajo de Log Analytics en la barra superior.
- Pruebe a acceder al análisis de tráfico después de 30 minutos, si se ha habilitado recientemente.

¿Qué sucede si aparece este mensaje: "Analizando registros de flujo de grupos de seguridad de red por primera vez. Este proceso puede tardar entre 20 y 30 minutos en completarse. ¿Vuelve después de algún tiempo."?

Es posible que vea este mensaje porque:

- El análisis de tráfico se ha habilitado recientemente y es posible que aún no haya agregado suficientes datos para que derive información significativa.
- Está utilizando la versión gratuita del área de trabajo de Log Analytics, que ha superado los límites de cuota. Es posible que tenga que usar un área de trabajo con una capacidad mayor.

Pruebe las soluciones sugeridas para la pregunta anterior. Si los problemas persisten, generen preocupaciones en Q&A de Microsoft.

Qué ocurre si obtengo este mensaje: "Parece que tenemos datos de recursos (topología) y no hay información de flujos. Para obtener más información, haga clic aquí para ver los datos de recursos y consulte preguntas más frecuentes".

Está viendo la información de los recursos en el panel; sin embargo, no hay estadísticas relacionadas con el flujo. Es posible que los datos no estén presentes debido a que no hay flujos de comunicación entre los recursos. Espere 60 minutos y vuelva a comprobar el estado. Si el

problema persiste y está seguro de que existen flujos de comunicación entre los recursos, plantee sus dudas en Microsoft Q&A.

¿Cómo se determina el precio del análisis de tráfico?

El análisis de tráfico se puede medir. La medición se basa en el procesamiento de los datos de registro de flujo sin procesar por el servicio. Para más información, consulte Precios de Network Watcher.
Los registros mejorados ingeridos en el área de trabajo de Log Analytics se pueden conservar sin cargo por un máximo de 31 días (o 90 días si Microsoft Sentinel está habilitado en el área de trabajo). Para obtener más información, consulte Precios de Azure Monitor.

¿Por qué veo transacciones de almacenamiento adicionales cuando se habilitan los registros de flujo de NSG y Análisis de tráfico?

Cuando los registros de flujo de NSG están habilitados, los registros de flujo se escriben en la cuenta de Azure Storage configurada. Para generar información de Análisis de tráfico, Azure Network Watcher realiza operaciones de back-end en estos datos almacenados para procesar, agregar y enriquecer los registros de registro de flujo. Estas interacciones de almacenamiento de back-end son necesarias para la funcionalidad de análisis y se producen incluso si no consulta activamente los datos. Puede controlar el impacto general en el costo ajustando el ámbito del registro de flujo, la retención, y determinando si Traffic Analytics está habilitado. Para más información, consulte Precios de los registros de flujo de NSG.

¿Con qué frecuencia procesa el análisis de tráfico los datos?

El intervalo de procesamiento predeterminado del análisis de tráfico es de 60 minutos; sin embargo, puede seleccionar el procesamiento acelerado en intervalos de 10 minutos. Para más información, consulte Agregación de datos en análisis de tráfico.

¿Cuáles son los demás recursos creados con mi área de trabajo?

Análisis de tráfico crea y administra recursos de regla de recopilación de datos y punto de conexión de recopilación de datos en el mismo grupo de recursos que el área de trabajo, con el

prefijo `NWTA`. Si realiza alguna operación en estos recursos, es posible que el análisis de tráfico no funcione según lo previsto. Para más información, consulte Agregación de datos en análisis de tráfico. Para más información, consulte Reglas de recopilación de datos en Azure Monitor y puntos de conexión de recopilación de datos en Azure Monitor.

¿Puedo aplicar bloqueos a los recursos DCE y DCR creados por el análisis de tráfico?

No se recomienda aplicar bloqueos a los recursos de DCR y DCE creados por el análisis de tráfico, ya que el servicio administra estos recursos. Los recursos bloqueados no se limpian tras la eliminación de los registros de flujo relacionados. Si realiza alguna operación en estos recursos, es posible que el análisis de tráfico no funcione según lo previsto. Para más información, consulte Agregación de datos en análisis de tráfico.

¿Cómo decide el análisis de tráfico que una dirección IP es malintencionada?

El análisis de tráfico se basa en sistemas de inteligencia sobre amenazas internos de Microsoft para considerar una dirección IP como malintencionada. Estos sistemas usan diversos orígenes de telemetría, como productos y servicios de Microsoft, la Unidad de crímenes digitales de Microsoft (DCU), el Centro de respuesta de seguridad de Microsoft (MSRC) y fuentes externas y crean inteligencia sobre él. Algunos de estos datos son Internos de Microsoft. Si una dirección IP conocida se marca como malintencionada, cree una incidencia de soporte técnico para conocer los detalles.

¿Cómo puedo establecer alertas sobre los datos de análisis de tráfico?

El análisis de tráfico no tiene compatibilidad integrada con las alertas. Sin embargo, dado que los datos de análisis de tráfico se almacenan en Log Analytics, puede escribir consultas personalizadas y establecer alertas en ellos. Siga estos pasos:

- Puede usar el vínculo de Log Analytics en análisis de tráfico.
- Use el esquema de análisis de tráfico para escribir las consultas.
- Seleccione **Nueva regla de alertas** para crear la alerta.
- Consulte Creación de una nueva regla de alertas para crear la alerta.

¿Se admite el análisis de tráfico en un área de trabajo de Log Analytics con el perímetro de seguridad de red habilitado?

No, actualmente no es compatible. Si se implementa un área de trabajo de Log Analytics detrás de un perímetro de seguridad de red, el análisis de tráfico no podrá ingerir datos de él.

¿Por qué mis flujos de red aparecen como "Desconocido" en lugar de estar correctamente etiquetados

Traffic Analytics (TA) solo puede clasificar y etiquetar correctamente flujos para recursos en suscripciones donde al menos un registro de flujo está habilitado y se envían datos al mismo espacio de trabajo de Log Analytics (LA) que utiliza Traffic Analytics. Si un flujo se origina o termina en un recurso de una suscripción que no tiene registro de flujo configurado contra el espacio de trabajo de Log Analytics habilitado por Traffic Analytics, Traffic Analytics no tiene visibilidad sobre los recursos de esa suscripción y marcará los flujos asociados como "Desconocidos".

Herramientas de diagnóstico de red

¿Existen limitaciones con las herramientas de diagnóstico de red para Application Gateway?

Sí. La solución de problemas de conexión y los diagnósticos de NSG no admiten implementaciones privadas de Application Gateway. Para obtener más información, consulte Implementación de Application Gateway privado.

Last updated on 19/08/2026

Confiabilidad en el servicio Azure Private Link

Azure Private Link servicio le ayuda a exponer de forma privada sus propias aplicaciones, como las que se ejecutan en máquinas virtuales (VM), dentro de una red virtual Azure. El servicio Private Link ayuda a clientes de Azure u otros clientes en las redes a conectarse de forma segura sin direcciones IP públicas, lo que garantiza que el tráfico permanezca dentro de la red de Azure.

Al usar Azure, relibilidad es una responsabilidad compartida. Microsoft proporciona una variedad de capacidades para admitir resiliencia y recuperación. Es responsable de comprender cómo funcionan esas funcionalidades dentro de todos los servicios que usa y de seleccionar las funcionalidades que necesita para cumplir los objetivos empresariales y los objetivos de tiempo de actividad.

Este artículo se centra en el servicio Private Link y los puntos de conexión privados asociados como mecanismo de conectividad. Describe el comportamiento del nivel de plataforma y del plano de control durante errores transitorios, interrupciones de zona de disponibilidad y interrupciones en toda la región.

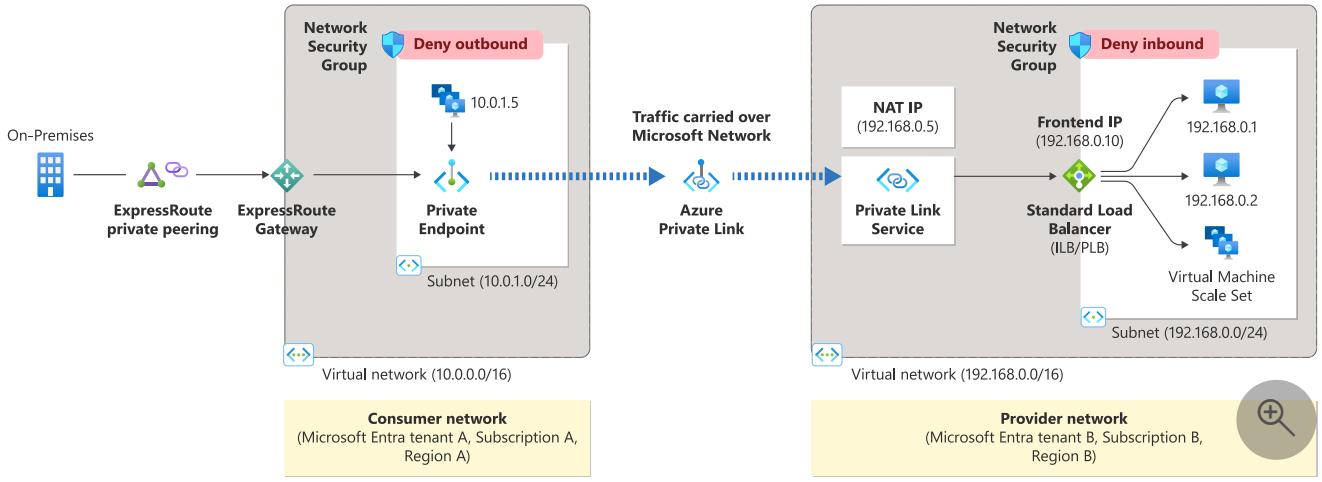
> **Nota:**
>
> Este artículo se centra en el servicio Private Link, que facilita la conectividad privada con las aplicaciones que ejecutas en tus propias máquinas virtuales. Si usa puntos de conexión privados con otros servicios de Azure, por ejemplo, Azure Storage o Azure SQL Database, debe revisar en su lugar las guías de confiabilidad de esos servicios para obtener información de confiabilidad sobre sus puntos de conexión privados.

> **Importante**
>
> La confiabilidad de la solución general depende de la configuración de los servidores back-end a los que se conecta Private Link servicio. Estos servidores backend pueden ser máquinas virtuales de Azure, conjuntos de escalado de máquinas virtuales de Azure o puntos de conexión externos. La confiabilidad de la solución también depende de la configuración de equilibradores de carga y de otros componentes de red.

Los servidores back-end no están en el ámbito de este artículo, pero sus configuraciones de disponibilidad afectan directamente a la resistencia de la aplicación. Para comprender cómo cada servicio admite sus requisitos de confiabilidad, revise las guías de confiabilidad de todos los servicios Azure de la solución. Puede lograr la confiabilidad de un extremo a otro para la aplicación asegurándose de que los servidores back-end también están configurados para alta disponibilidad y redundancia de zona.

Introducción a la arquitectura de confiabilidad

El servicio Private Link ayuda a sus clientes a conectarse de forma privada a sus cargas de trabajo en Azure. Como proveedor de servicios, usted implementa un recurso de servicio Private Link. *Los consumidores de servicios* crean *puntos de conexión privados* en sus propias redes virtuales de Azure. Estos puntos de conexión se conectan de forma segura y privada a las aplicaciones a través de Azure Private Link. Esta configuración no expone direcciones IP públicas, incluso cuando un consumidor usa el punto de conexión privado desde un entorno local a través de Azure ExpressRoute u otro método de conectividad privado.



Normalmente, un servicio de Private Link se asocia a un equilibrador de carga de Azure que se coloca delante de los recursos de backend, como máquinas virtuales o conjuntos de escalado de máquinas virtuales. También puede usar Private Link servicio Direct Connect (versión preliminar), lo que facilita la conectividad a cualquier dirección IP enrutable privada dentro de la red virtual. Si usa Private Link servicio Direct Connect, revise la documentación cuidadosamente para comprender los requisitos, la disponibilidad de regiones y las limitaciones.

> **ⓘ Importante**
>
> El servicio Private Link Direct Connect está actualmente en fase de vista previa.

Consulte los **Términos Suplementarios de Uso para las Vistas Previas de Microsoft Azure** para conocer los términos legales que se aplican a las funcionalidades de Azure que están en beta, en vista previa, o que de otro modo no están generalmente disponibles.

Resistencia a errores transitorios

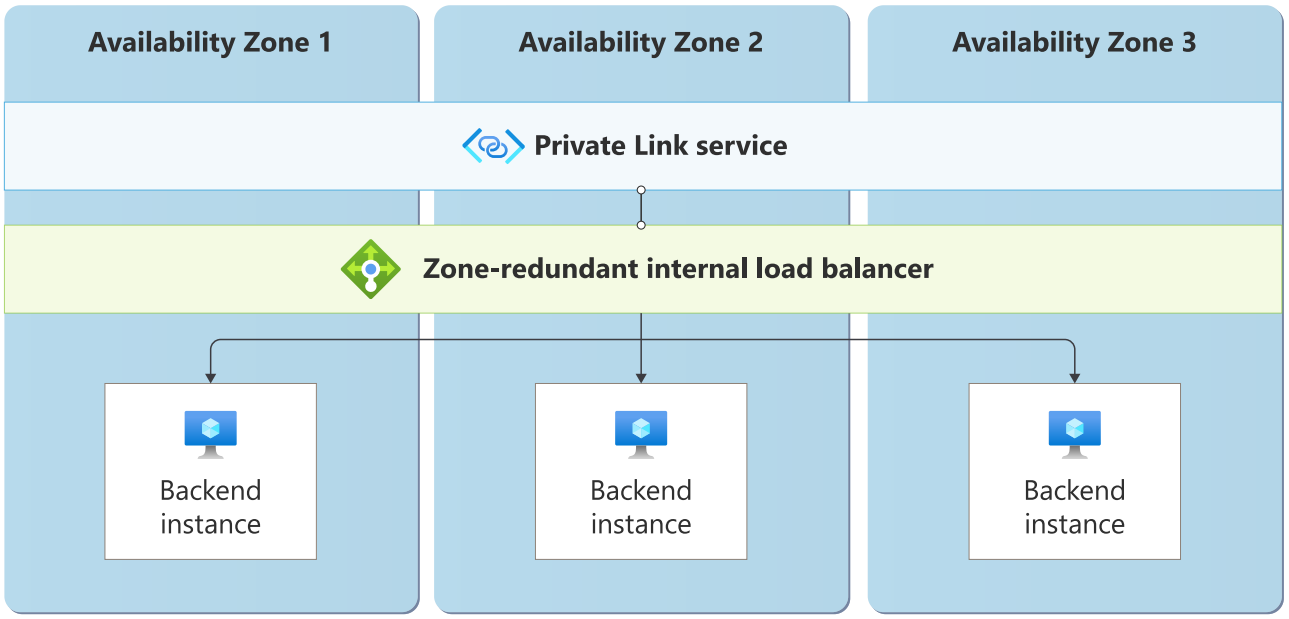
Los errores transitorios son errores breves e intermitentes en los componentes. Se producen con frecuencia en un entorno distribuido como la nube y son una parte normal de las operaciones. Los errores transitorios se corrigen después de un breve período de tiempo. Es importante que las aplicaciones puedan controlar errores transitorios, normalmente mediante el reintento de solicitudes afectadas.

Todas las aplicaciones hospedadas en la nube deben seguir las instrucciones de control de errores transitorios Azure cuando se comunican con cualquier API, bases de datos y otros componentes hospedados en la nube. Para obtener más información, consulte Recomendaciones para controlar errores transitorios.

Al implementar un servicio de Private Link con Standard Load Balancer, revise las recomendaciones de control de errores transient para Azure Load Balancer y asegúrese de que el equilibrador de carga está configurado para controlar errores transitorios.

Resistencia a errores de zona de disponibilidad

El servicio de Private Link es automáticamente resistente a los fallos de zona de disponibilidad cuando se implementa en una región que admite zonas de disponibilidad. Los proveedores de servicios no necesitan configurar nada para habilitar este comportamiento.



Los puntos de conexión privados se distribuyen automáticamente entre zonas de disponibilidad en la región. Los consumidores de servicios no necesitan crear puntos de conexión privados independientes en distintas zonas.

Requisitos

- **Compatibilidad de la región:** Puede implementar servicios de Private Link con redundancia de zona en cualquier región que admita zonas de disponibilidad.
- **Dependencia del equilibrador de carga:** Si usa el servicio Private Link con un equilibrador de carga en el backend, configure también el equilibrador de carga con redundancia de zona para garantizar la resistencia de la zona de un extremo a otro. Para obtener más información, consulte Reliability in Load Balancer.

Cost

No hay ningún costo adicional asociado a la compatibilidad con las zonas de disponibilidad para el servicio Private Link.

Configurar soporte de zonas de disponibilidad

La compatibilidad con las zonas de disponibilidad se habilita automáticamente al implementar el servicio Private Link en una región que admita zonas de disponibilidad.

Comportamiento cuando todas las zonas están en buen estado

En esta sección se describe qué esperar cuando los servicios Private Link y los puntos de conexión privados admiten zonas de disponibilidad y todas las zonas están operativas.

- **Operación entre zonas:** El tráfico a través de un punto de conexión privado y un servicio de "Private Link" puede enrutarse a través de cualquier zona de disponibilidad.
- **Replicación de datos entre zonas:** Private Link no realiza la replicación de datos entre zonas porque es un servicio sin estado para la conectividad.

Comportamiento durante un fallo de zona

En esta sección se describe qué esperar cuando los servicios de Private Link y los puntos de conexión privados admiten zonas de disponibilidad y hay una interrupción en una de las zonas.

- **Detección y respuesta:** Microsoft es responsable de detectar errores de zona de disponibilidad y administrar la respuesta del servicio.
- **Notification:** Microsoft no le notifica automáticamente cuando una zona está inactiva. Sin embargo, puede usar Azure Service Health para comprender el estado general del servicio, incluidos los errores de zona, y puede configurar alertas de Service Health para notificarle problemas.
- **Solicitudes activas:** Es posible que las solicitudes activas se finalicen durante un error de zona de disponibilidad. Los consumidores del servicio deben reintentar las solicitudes con errores después de interrupciones transitorias, de forma similar a otros errores transitorios.
- Pérdida de datos esperada: No se produce ninguna pérdida de datos porque Private Link es un servicio sin estado para la conectividad.
- **Tiempo de inactividad esperado:** Es posible que se produzca un error en las conexiones existentes que se conectan a través de la zona con errores. Si los componentes backend, como el equilibrador de carga y los servidores de aplicaciones, siguen estando disponibles, los consumidores del servicio pueden reintentar inmediatamente la conexión y las solicitudes se enrutan a través de la infraestructura de otra zona.
- **Redistribución:** Cuando se produce un error en una sola zona de disponibilidad, el servicio enruta el tráfico nuevo a través de zonas correctas, lo que continúa funcionando.

  Es poco probable que las máquinas virtuales de la zona de disponibilidad afectada permanezcan operativas durante una interrupción de zona. Sin embargo, si un error de

zona parcial hace que Private Link no esté disponible en la zona afectada mientras las máquinas virtuales de esa zona siguen funcionando, las conexiones salientes a las máquinas virtuales de la zona afectada se redirigen a través de la infraestructura de Private Link en otra zona.

El tiempo de inactividad de la aplicación también puede producirse si los componentes dependientes, como equilibradores de carga o máquinas virtuales de back-end, no son resilientes a la zona.

Recuperación de zona

Cuando se recupera la zona de disponibilidad afectada, Microsoft administra automáticamente el proceso de conmutación inversa. No se requiere ninguna acción del cliente.

Prueba de fallos de zona

La plataforma Private Link administra el enrutamiento del tráfico, la conmutación por error y la conmutación por recuperación para los servicios de Private Link y puntos de conexión privados en zonas de disponibilidad. Dado que esta característica está totalmente administrada, no es necesario validar los procesos de error de zona de disponibilidad.

Resistencia a errores en toda la región

El servicio Private Link es un servicio regional único. El servicio no incluye funciones nativas multirregionales ni recuperación automática frente a errores entre regiones. Si una región de Azure deja de estar disponible, Private Link servicios de esa región tampoco están disponibles.

Soluciones personalizadas de varias regiones para la resistencia

Si diseña un enfoque de red que abarca varias regiones, debe implementar servicios de Private Link independientes en cada región. Es responsable de la implementación y administración del servicio Private Link. Los consumidores de servicios son responsables de la configuración del punto de conexión privado en los servicios Private Link de cada región. Los consumidores del servicio también son responsables de enrutar el tráfico al servicio Private Link adecuado.

Copias de seguridad y restauración

Private Link servicio no almacena los datos del cliente y no requiere copia de seguridad ni restauración. Para volver a crear configuraciones, considere la posibilidad de mantener plantillas de infraestructura como código (IaC) para los recursos de red. Los servicios Private Link son de solo configuración y no almacenan datos de clientes, por lo que los esfuerzos de respaldo deben centrarse en las plantillas de IaC para reimplementación rápida.

Acuerdo de nivel de servicio

El acuerdo de nivel de servicio (SLA) para Azure servicios describe la disponibilidad esperada de cada servicio y las condiciones que la solución debe cumplir para lograr esa expectativa de disponibilidad. Para obtener más información, vea SLAs for servicios en línea.

Contenido relacionado

- Confiabilidad en Azure
- Relibilidad en Load Balancer

ⓘ **Nota:** El autor ha creado este artículo con ayuda de inteligencia artificial. Más información

---

Last updated on 14/04/2026

Direcciones IP públicas

Las direcciones IP públicas permiten a los recursos de Internet la comunicación entrante a los recursos de Azure. Las direcciones IP públicas permiten a los recursos de Azure comunicarse con Internet y a los servicios de Azure orientados al público. Dedique la dirección al recurso hasta que anule la asignación. Un recurso sin una dirección IP pública asignada todavía puede comunicarse de forma saliente. Azure asigna automáticamente una dirección IP dinámica disponible para la comunicación saliente. Esta dirección no está dedicada al recurso y puede cambiar con el tiempo. Para más información acerca de las conexiones salientes en Azure, consulte Comprender las conexiones salientes en Azure.

En el Administrador de recursos de Azure, una dirección IP pública es un recurso que cuenta con propiedades específicas.

Los siguientes recursos se pueden asociar a una dirección IP pública:

- Interfaces de red de máquinas virtuales
- Virtual Machine Scale Sets
- Equilibradores de carga de Azure (público)
- Puertas de enlace de red virtual (VPN/ER)
- Puertas de enlace NAT
- Puertas de enlace de aplicaciones
- Instancias de Azure Firewall
- Hosts bastión
- Servidores de rutas
- API Management

En Virtual Machine Scale Sets, use prefijos de IP pública.

De un vistazo

En la tabla siguiente se muestra la propiedad que una dirección IP pública puede asociarse a un recurso y a los métodos de asignación. La compatibilidad con IPv6 pública no está disponible para todos los tipos de recursos en este momento.

Expandir tabla

Recurso de nivel superior	Asociación de dirección IP	IPv4 dinámico	IPv4 estático	IPv6 dinámico	IPv6 estático
Máquina virtual	Interfaz de red	Sí	Sí	Sí	Sí

Recurso de nivel superior	Asociación de dirección IP	IPv4 dinámico	IPv4 estático	IPv6 dinámico	IPv6 estático
Equilibrador de carga público	Configuración de front-end	Sí	Sí	Sí	Sí
Puerta de enlace de red virtual (VPN)	Configuración de IP de puerta de enlace	Sí (solo sin AZ)	Sí	No	No
Puerta de enlace de red virtual (ER)	Configuración de IP de puerta de enlace	Sí	Sí	Sí (versión preliminar)	No
NAT Gateway	Configuración de IP de puerta de enlace	No	Sí	No	No
Application Gateway	Configuración de front-end	Sí (solo V1)	Sí (solo V2)	No	No
Azure Firewall	Configuración de front-end	No	Sí	No	No
Host bastión	Configuración de IP pública	No	Sí	No	No
Servidor de Rutas	Configuración de front-end	No	Sí	No	No
API Management	Configuración de front-end	No	Sí	No	No

Versión de la dirección IP

Las direcciones IP públicas se pueden crear con una dirección IPv4 o IPv6. Es posible que tenga la opción de crear una implementación de doble pila con una dirección IPv4 e IPv6.

SKU

> **Importante**
>
> El 30 de septiembre de 2025, se retiraron las direcciones IP públicas de SKU básicas. Para obtener más información, consulte el anuncio oficial. Si actualmente usa direcciones IP públicas de SKU básicas, asegúrese de actualizar a la SKU estándar lo antes posible. Para obtener instrucciones sobre la actualización, visite Actualización de una dirección IP pública básica al SKU Estándar: Guía.

Las direcciones IP públicas se crean con una SKU de **Estándar** o **Básico**. La SKU determina su funcionalidad, lo que incluye el método de asignación, la compatibilidad con características y los recursos a los que se pueden asociar.

Los detalles completos se muestran en la tabla siguiente:

Expandir tabla

Dirección IP pública	**Estándar**	**Básico**
Método de asignación	Estática	Para IPv4: dinámica o estática; para IPv6: dinámica.
Tiempo de espera de inactividad	Tiene un tiempo espera de inactividad del flujo originado de entrada ajustable de entre 4 y 30 minutos, y un valor predeterminado de 4 minutos, y el valor predeterminado del tiempo de espera del flujo originado es de 4 minutos.	Tiene un tiempo espera de inactividad del flujo originado de entrada ajustable de entre 4 y 30 minutos, y un valor predeterminado de 4 minutos, y el valor predeterminado del tiempo de espera del flujo originado es de 4 minutos.
Seguridad	Proteja de forma predeterminada el modelo y se cierre al tráfico entrante cuando se usa como front-end. Se requiere permitir el tráfico con grupo de seguridad de red (NSG) (por ejemplo, en la NIC de una máquina virtual con una dirección IP pública de SKU Estándar conectada).	Abra de forma predeterminada. Se recomiendan grupos de seguridad de red, pero opcionales para restringir el tráfico entrante o saliente.
Zonas de disponibilidad	Compatible. Las direcciones IP Estándar pueden ser no zonales, zonales o con redundancia de zona. **Las direcciones IP con redundancia de zona solo se pueden crear en regiones de en las que haya 3 zonas de disponibilidad activas.** Las direcciones IP creadas antes que las zonas de disponibilidad no son redundantes de zona.	No se admite.
Preferencia de enrutamiento	Se admite para habilitar un control más pormenorizado de cómo se enruta el tráfico entre Azure e Internet.	No se admite.
Nivel global	Compatible mediante equilibradores de carga entre regiones	No se admite.

Las máquinas virtuales conectadas a un grupo de back-end no necesitan una dirección IP pública para asociarse a un equilibrador de carga público. Pero, si lo hacen, se requieren SKU

coincidentes para el equilibrador de carga y los recursos de IP pública. No puede tener una combinación de recursos de SKU básica y recursos de SKU estándar. No se pueden asociar máquinas virtuales independientes, máquinas virtuales en un recurso de conjunto de disponibilidad o conjuntos de escalado de máquinas virtuales a ambas SKU simultáneamente. Para los nuevos diseños, deberá considerar usar los recursos de SKU Estándar. Para obtener más información sobre un equilibrador de carga estándar, consulte Standard Load Balancer.

Asignación de dirección IP

Las direcciones IP públicas tienen dos tipos de asignaciones:

- **Dinámicas**: La dirección IP **no** se asigna al recurso en el momento de crearse cuando se selecciona el método dinámico. La dirección IP se asigna cuando se asocia la dirección IP pública a un recurso. La dirección IP se libera cuando se detiene o elimina el recurso. Las direcciones IP públicas dinámicas se usan normalmente para cuando no haya ninguna dependencia en la dirección IP. Por ejemplo, se libera un recurso de dirección IP pública desde una máquina virtual tras la detención y, a continuación, se inicia. Cualquier dirección IP asociada se libera si el método de asignación es **dinámico**. Si no desea que cambie la dirección IP, establezca el método de asignación en **estático** para asegurarse de que la dirección IP siga siendo la misma.

- **Estático**: Al recurso se le asigna una dirección IP en el momento en que se crea. La dirección IP se libera cuando se elimina el recurso y no se puede recuperar. Incluso cuando se establece el método de asignación en **estático**, no se puede especificar la dirección IP real asignada al recurso de dirección IP pública. Azure asigna la dirección IP desde un grupo de direcciones IP disponibles en la ubicación de Azure cuando se crea el recurso.

Las direcciones IP públicas estáticas se usan normalmente en los escenarios siguientes:

- Cuando debe actualizar las reglas de firewall para comunicarse con los recursos de Azure.
- Resolución de nombres DNS, donde un cambio en la dirección IP requeriría actualizar registros A.
- Los recursos de Azure se comunican con otras aplicaciones o servicios que usan un modelo de seguridad basado en direcciones IP.
- Los certificados TLS/SSL se usan vinculados a una dirección IP.

Expandir tabla

Recurso	Estática	Dinámica
IPv4 pública estándar	✅	x

Recurso	Estática	Dinámica
IPv6 pública estándar	✅	x
IPv4 pública básica	✅	✅
IPv6 pública básica	x	✅

Zona de disponibilidad

> **Importante**
>
> Todas las IPs estándar no zonales ahora son redundantes de zona en todas las regiones que admiten zonas de disponibilidad. Significa que las direcciones IP que se muestran como "zonas 1 2 3" o que no tienen zonas tienen redundancia de zona equivalente.

Las direcciones IP públicas de SKU estándar se pueden crear como zonales o con redundancia de zona en regiones que admiten zonas de disponibilidad. Las IP públicas SKU Básico no tienen zonas y se crean como no zonales. Una vez creada, una dirección IP pública no puede cambiar su zona de disponibilidad.

En regiones sin zonas de disponibilidad, todas las direcciones IP públicas se crean como no zonales. Las direcciones IP públicas creadas en una región que se actualice más adelante para tener zonas de disponibilidad serán redundantes a nivel de zona una vez que la región esté en estado de disponibilidad general con varias zonas de disponibilidad.

Expandir tabla

Valor	Comportamiento
Con redundancia de zona	Se crea una dirección IP con redundancia de zona en todas las zonas de una región y puede sobrevivir a cualquier error de zona única.
Zonal	Una dirección IP zonal está vinculada a una zona de disponibilidad específica y comparte el destino del estado de la zona.
No zonal	Solo es válido en regiones que no admiten zonas de disponibilidad.

Etiqueta de nombre de dominio

Seleccione esta opción para especificar una etiqueta de DNS para un recurso de dirección IP pública. Esta funcionalidad funciona tanto para direcciones IPv4 (registros A de 32 bits) como para direcciones IPv6 (registros AAAA de 128 bits). Esta selección crea una asignación para

domainnamelabel.ubicación.cloudapp.azure.com a la dirección IP pública en el DNS administrado por Azure.

Por ejemplo, la creación de una dirección IP pública con la siguiente configuración:

- **contoso** como **domainnamelabel**
- **Ubicación** de Azure **Oeste de EE. UU.**

El nombre de dominio completo (FQDN) **contoso.westus.cloudapp.azure.com** se resuelve en la dirección IP pública del recurso. Cada etiqueta de nombre de dominio creada debe ser única dentro de su ubicación de Azure.

Si se desea un dominio personalizado para los servicios que usan una dirección IP pública, puede usar Azure DNS o un proveedor DNS externo para el registro DNS.

Ámbito de etiqueta de nombre de dominio (versión preliminar)

Las direcciones IP públicas también tienen un parámetro opcional para **Ámbito de etiqueta de nombre de dominio**, que define qué etiqueta de dominio usa un objeto con el mismo nombre. Esta característica puede ayudar a evitar que los actores malintencionados puedan reutilizar los nombres DNS pendientes. Cuando se elige esta opción, el nombre DNS de la dirección IP pública tiene otra cadena entre el **domainnamelabel** y los campos **ubicación**, por ejemplo, **contoso.fjdng2acavhkevd8.westus.cloudapp.Azure.com**. (Esta cadena es un hash generado a partir de la entrada específica de la suscripción, el grupo de recursos, la etiqueta de nombre de dominio y otras propiedades).

El ámbito de la etiqueta de nombre de dominio solo se puede especificar al crear una dirección IP pública.

> **Importante**
>
> El ámbito de etiqueta de nombre de dominio está actualmente en versión preliminar pública. Se proporciona sin un contrato de nivel de servicio y no se recomienda para cargas de trabajo de producción. Para obtener más información, consulte Términos de uso complementarios para versiones preliminares de Microsoft Azure.

El valor del **Ámbito de etiqueta de nombre de dominio** debe coincidir con una de las opciones siguientes:

Expandir tabla

Valor	Comportamiento
TenantReuse	El objeto con el mismo nombre en el mismo inquilino recibe la misma etiqueta de dominio
SubscriptionReuse	Un objeto con el mismo nombre en la misma suscripción recibe la misma etiqueta de dominio
ResourceGroupReuse	El objeto con el mismo nombre en el mismo grupo de recursos recibe la misma etiqueta de dominio.
NoReuse	El objeto con el mismo nombre recibe una nueva etiqueta de dominio para cada nueva instancia

Por ejemplo, si la opción **SubscriptionReuse** está seleccionada, y un cliente que tiene la etiqueta de nombre de dominio de ejemplo **contoso.fjdng2acavhkevd8.westus.cloudapp.Azure.com** elimina y vuelve a implementar una dirección IP pública utilizando la misma plantilla que antes, la etiqueta de nombre de dominio sigue siendo la misma. Si el cliente implementa una dirección IP pública con esta misma plantilla en otra suscripción, la etiqueta de nombre de dominio cambiaría (por ejemplo, **contoso.c9ghbqhhbxevhzg9.westus.cloudapp.Azure.com**).

Otras características de dirección IP pública

Hay otros atributos que se pueden usar para una dirección IP pública (solo SKU Estándar).

- La opción de **Nivel** global crea una dirección IP de difusión global que se puede usar con equilibradores de carga entre regiones.
- La opción **Preferencia de enrutamiento** por Internet minimiza el tiempo que el tráfico pasa en la red de Microsoft, lo que reduce el costo de transferencia de los datos de salida.

Límites

Los límites de una dirección IP se encuentran en el conjunto completo de límites de red de Azure. Los límites son por región y por suscripción.

Póngase en contacto con el servicio de soporte técnico para aumentar los límites predeterminados según sus necesidades empresariales.

Precios

Las direcciones IPv4 públicas tienen un cargo nominal; Las direcciones IPv6 públicas no tienen ningún cargo.

Para más información sobre los precios de las direcciones IP en Azure, revise la página Precios de las direcciones IP.

Limitaciones para IPv6

- Las puertas de enlace de red VPN no se pueden usar en una red virtual que tiene IPv6 habilitado, ya sea directamente o emparejadas con "UseRemoteGateway".
- Azure no admite la comunicación de IPv6 para contenedores.
- No se admite el uso de máquinas virtuales ni conjuntos de escalado de máquinas virtuales solo con IPv6. Cada NIC debe incluir al menos una configuración ip IPv4 (pila doble).
- Al agregar IPv6 a las implementaciones IPv6 existentes, los rangos de IPv4 no se pueden añadir a una red virtual con vínculos de navegación de recursos existentes.
- El reenvío de DNS para IPv6 es compatible con DNS público de Azure. No se admite DNS inverso.
- No se admiten preferencias de enrutamiento de Internet.

Para más información sobre IPv6 en Azure, consulte aquí.

Pasos siguientes

- Obtener información sobre las Direcciones IP privadas en Azure
- Implementar una máquina virtual con una dirección IP pública estática mediante Azure Portal

---

Last updated on 20/10/2025

Preguntas más frecuentes sobre Azure Route Server

En este artículo se proporcionan respuestas a preguntas frecuentes sobre Azure Route Server, que incluyen información general del servicio, funcionalidades de enrutamiento y limitaciones actuales.

General

¿Qué es Azure Route Server?

Azure Route Server es un servicio totalmente administrado que permite administrar fácilmente el enrutamiento entre la aplicación virtual de red (NVA) y la red virtual.

¿Azure Route Server es solo una máquina virtual?

No. Azure Route Server es un servicio diseñado con alta disponibilidad. Route Server tiene redundancia de nivel de zona si la implementa en una región de Azure que admite Availability Zones.

¿Es necesario emparejar cada NVA con ambas instancias de Azure Route Server?

Sí, debe emparejar cada instancia de NVA con ambas instancias de Route Server para asegurarse de que Route Server recibe correctamente las rutas y para configurar la alta disponibilidad. También debe anunciar las mismas rutas a ambas instancias. También se recomienda emparejar al menos dos instancias de NVA con ambas instancias de Route Server.

> **Nota:**
>
> Durante los eventos de mantenimiento de Route Server, el emparejamiento BGP podría bajar entre la aplicación virtual de red y una de las instancias de Route Server. Si configura la aplicación virtual de red para emparejarse con ambas instancias de Route Server, la conectividad permanece disponible durante los eventos de mantenimiento.

¿Puedo recibir notificaciones avanzadas de mantenimiento?

Actualmente, no puede recibir notificaciones avanzadas para el mantenimiento de Azure Route Server.

¿Azure Route Server almacena los datos de los clientes?

No. Azure Route Server solo intercambia rutas de BGP con la aplicación virtual de red (NVA) y, a continuación, las propaga a la red virtual.

¿Admite Azure Route Server el emparejamiento de red virtual?

Sí. Si empareja una red virtual que hospeda Azure Route Server con otra red virtual y habilita la opción **Usar la puerta de enlace o el Route Server de la red virtual remota** en la segunda red virtual, Azure Route Server aprende los espacios de direcciones de la red virtual emparejada y los envía a todas las appliances virtuales de red (NVA) emparejadas. Route Server también programa las rutas de las NVA en la tabla de enrutamiento de las máquinas virtuales de la red virtual emparejada.

¿Por qué Azure Route Server requiere una dirección IP pública con puertos abiertos?

Estos puntos de conexión públicos son necesarios para que las redes definidas por software (SDN) subyacentes de Azure y la plataforma de administración se comuniquen con Azure Route Server. Dado que Route Server se considera parte de la red privada del cliente, la plataforma subyacente de Azure no puede acceder directamente y administrar Route Server a través de sus puntos de conexión privados debido a los requisitos de cumplimiento. La conectividad a los puntos de conexión públicos de Route Server se autentica a través de certificados y Azure realiza auditorías de seguridad rutinarias de estos puntos de conexión públicos. Como resultado, no constituyen una exposición de seguridad de la red virtual.

> ⓘ **Nota:**
>
> Azure firma estos certificados con una entidad de certificación interna, por lo que esta cadena de certificados parece no estar firmada por una entidad de confianza conocida. Esto no representa una vulnerabilidad ssl.

¿Azure Route Server admite IPv6?

No. Si implementa una red virtual con un espacio de direcciones IPv6 y posteriormente implementa una instancia de Azure Route Server en la misma red virtual, esto interrumpe la conectividad para el tráfico IPv6.

Puede emparejar una red virtual con un espacio de direcciones IPv6 a la red virtual de Route Server, y la conectividad IPv4 con esta red virtual de pila doble asociada seguirá funcionando. No se admite la conectividad IPv6 con esta red virtual emparejada.

¿Qué son las unidades de infraestructura de enrutamiento?

De forma predeterminada, azure Route Server se implementa con una capacidad de dos unidades de infraestructura de enrutamiento. Esta implementación predeterminada admite 4000 máquinas virtuales conectadas implementadas en la red virtual de Route Server y en todas las redes virtuales emparejadas.

Puede especificar más unidades de infraestructura de enrutamiento para aumentar la capacidad de Route Server en incrementos de 1,000 máquinas virtuales (VMs). Cada unidad de infraestructura de enrutamiento adicional cuesta 0,10 USD por hora en Los Estados Unidos. Los precios de otras regiones pueden variar. Los precios completos, incluidas las diferencias regionales, estarán disponibles a mediados de diciembre en la página oficial de precios de Azure.

Enrutamiento

¿Azure Route Server enruta el tráfico de datos entre la NVA y las máquinas virtuales?

No. Azure Route Server solo intercambia rutas de BGP con la aplicación virtual de red (NVA). El tráfico de datos va directamente desde la NVA a la máquina virtual (VM) de destino y viceversa.

¿Qué protocolos de enrutamiento admite Azure Route Server?

Azure Route Server solo admite el Protocolo de puerta de enlace de borde (BGP). La aplicación virtual de red (NVA) debe admitir el BGP externo de salto múltiple porque tiene que implementar la instancia de Route Server en una subred dedicada de la red virtual. Al configurar BGP en la NVA, el ASN que elija debe ser diferente del ASN del servidor de rutas.

¿Conserva Azure Route Server la ruta de acceso BGP AS de la ruta que recibe?

Sí, Azure Route Server propaga la ruta con la ruta de acceso BGP AS tal y como está.

Si el preprendido de ruta de acceso AS está configurado en una aplicación virtual de red hacia Route Server, ¿el circuito ExpressRoute pasa la información de preprendido de la ruta de acceso de AS al entorno local?

Cuando ExpressRoute anuncia rutas a un entorno local, quita la información privada de ASN de BGP. El entorno local recibe el prefijo con AS 12076.

¿Conserva Azure Route Server las comunidades BGP de la ruta que recibe?

Sí, Azure Route Server propaga la ruta con las comunidades BGP intactas.

¿Cuál es la configuración del temporizador BGP de Azure Route Server?

El temporizador keepalive de Azure Route Server es de 60 segundos y el temporizador de retención es de 180 segundos.

¿Puede Azure Route Server filtrar las rutas de las aplicaciones virtuales de red?

Azure Route Server admite **NO_ADVERTISE** comunidad de BGP. Si una aplicación virtual de red (NVA) anuncia rutas con esta cadena de comunidad a Route Server, este no las anuncia a otros elementos del mismo nivel, incluida la puerta de enlace de ExpressRoute. Esta característica puede ayudar a reducir el número de rutas que se envían desde Azure Route Server a ExpressRoute.

¿Puedo usar mapas de rutas para controlar el enrutamiento en Azure Route Server?

Sí. Los mapas de rutas permiten filtrar rutas, agrupar prefijos y modificar atributos BGP como AS-PATH y Community en emparejamientos BGP, conexiones de puerta de enlace de ExpressRoute y conexiones de puerta de enlace de VPN. Para más información, consulta Acerca de mapas de rutas para Azure Route Server y Configurar mapas de rutas.

Cuando se crea un emparejamiento de red virtual entre mis redes virtuales de hub y spoke, ¿esto provoca un restablecimiento suave de BGP entre Azure Route Server y las NVAs emparejadas?

Sí. Si se establece un emparejamiento de red virtual entre su red virtual hub y su red virtual spoke, Azure Route Server realiza un restablecimiento suave de BGP enviando solicitudes de actualización de rutas a todas sus NVA emparejadas. Si las NVA no admiten la actualización de rutas BGP, Azure Route Server realiza un restablecimiento duro de BGP con las NVA emparejadas, lo que podría provocar una interrupción de la conectividad para el tráfico que atraviesa las NVA.

¿Cómo se calcula el límite de rutas 4000 en una sesión de emparejamiento BGP entre una NVA y Azure Route Server?

Actualmente, Route Server puede aceptar un máximo de 4000 rutas desde un único peer BGP. Cuando el Servidor de Rutas procesa las actualizaciones de rutas BGP, este límite se calcula como el número de rutas actuales aprendidas de un vecino BGP más el número de rutas incluidas en la actualización de rutas BGP. Por ejemplo, si una NVA anuncia inicialmente rutas 2001 a Route Server y, posteriormente, vuelve a anunciar estas rutas 2001 en una actualización de rutas BGP, Route Server calcula esto como rutas 4,002 y derriba la sesión BGP.

¿Qué números de sistema autónomo (ASN) puedo usar?

Puede utilizar sus propios ASN públicos o ASN privados en la aplicación virtual de red (NVA). No puede usar los ASN reservados por Azure ni por la Autoridad de Números Asignados de Internet (IANA).

ASN reservados por Azure:

- ASN públicos: 8074, 8075, 12076
- ASN privados: 65515, 65517, 65518, 65519, 65520

ASNs reservados por IANA:

- 23456, 64496-64511, 65535-65551

¿Se pueden usar ASN de 32 bits (4 bytes)?

No, Azure Route Server solo admite ASN de 16 bits (2 bytes).

Si Azure Route Server recibe la misma ruta de más de una aplicación virtual de red, ¿cómo las administra?

Si la ruta tiene la misma longitud de la ruta AS, Azure Route Server programa varias copias de la ruta, cada una con un siguiente salto diferente, a las máquinas virtuales (MV) de la red virtual. Cuando una VM envía tráfico al destino de esta ruta, los hosts de la VM usan un enrutamiento multidireccional de igual costo (ECMP). Sin embargo, si una aplicación virtual de red envía la ruta con una longitud de ruta de acceso de AS más corta que otras aplicaciones virtuales de red, Azure Route Server solo programa la ruta que tenga el próximo salto establecido en esta aplicación virtual de red para las máquinas virtuales de la red virtual.

¿La creación de un servidor de rutas afecta al funcionamiento de las puertas de enlace de red virtual existentes (VPN o ExpressRoute)?

Sí. Cuando cree o elimine un servidor de rutas en una red virtual que contenga una puerta de enlace de red virtual (ExpressRoute o VPN), espere que el tiempo de inactividad dure 10 minutos. La implementación real de Route Server puede tardar entre 30 y 60 minutos en completarse, por lo que se recomienda programar una ventana de mantenimiento de 60 minutos para la implementación. Si tiene un circuito ExpressRoute conectado a la red virtual en la que está creando o eliminando el servidor de rutas, el tiempo de inactividad no afecta a las conexiones del circuito ExpressRoute a otras redes virtuales.

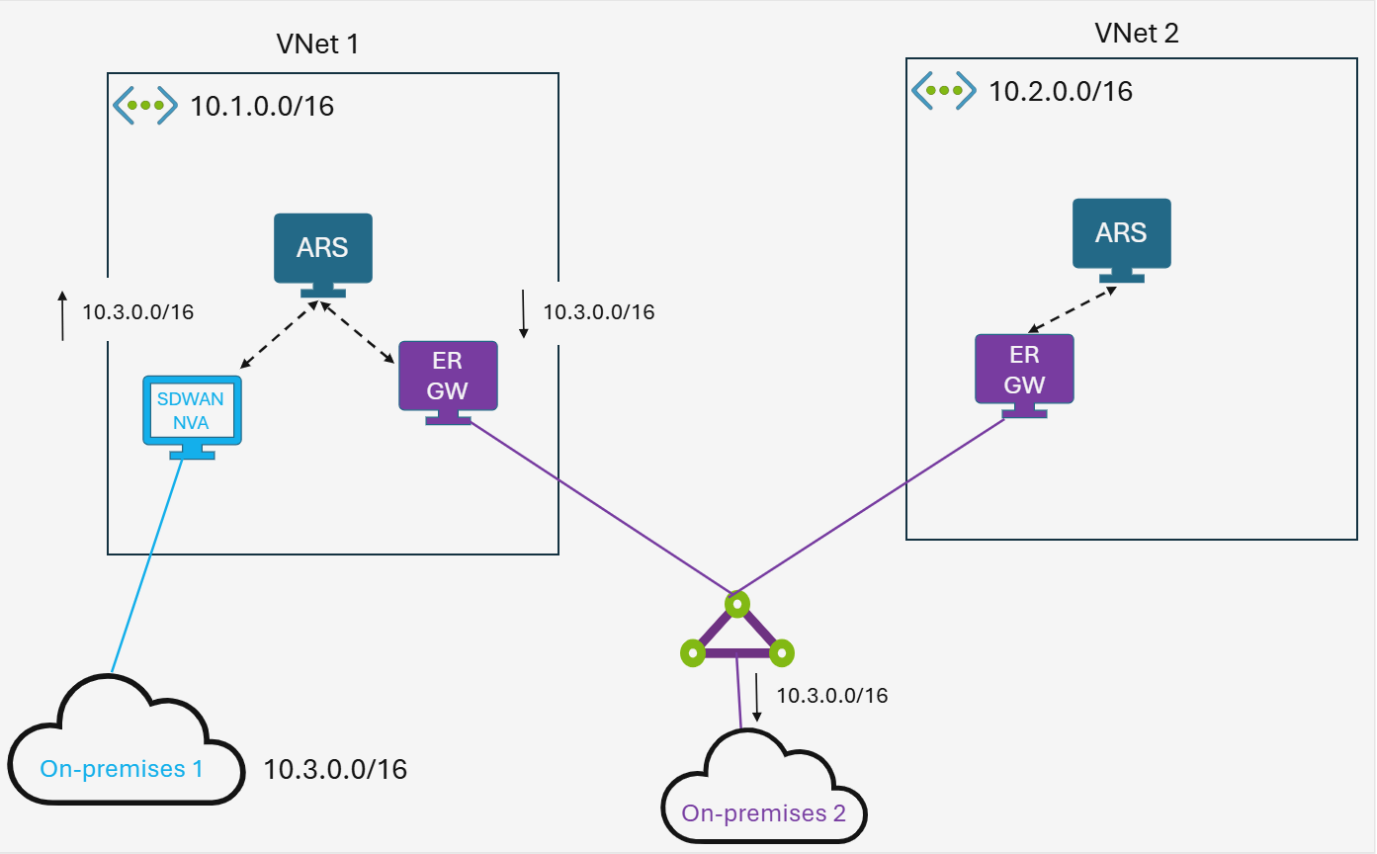
¿Azure Route Server intercambia rutas de forma predeterminada entre las NVA y las puertas de enlace de red virtual (VPN o ExpressRoute)?

No. De forma predeterminada, Azure Route Server no propaga las rutas que recibe de una NVA y una puerta de enlace de red virtual entre sí. Route Server intercambia estas rutas después de habilitar la **rama a rama** en ella.

¿Las rutas anunciadas por NVA se propagan de un servidor de rutas a otro a través de ExpressRoute MSEE?

No. Si la **conexión de sucursal a sucursal** está habilitada, las rutas anunciadas por NVA se anunciarán al entorno local conectado a ExpressRoute. Sin embargo, el segundo servidor de rutas (Route Server) eliminará las rutas anunciadas por NVA. Para ilustrarlo, en el diagrama siguiente se muestra la publicidad local de SDWAN 10.3.0.0/16 a la NVA de SDWAN. Esta ruta se aprende en

el primer servidor de rutas, por lo que todas las cargas de trabajo de la red virtual 1 (o emparejadas con la red virtual 1) también aprenden esta ruta. Además, la infraestructura local conectada a ExpressRoute aprende la ruta 10.3.0.0/16 si se habilita la opción de **conexión entre sucursales**. Sin embargo, el segundo servidor de rutas (en la red virtual 2) no aprende la ruta 10.3.0.0/16, por lo que ninguna carga de trabajo de la red virtual 2 no aprenderá esta ruta. Como resultado, la red local 10.3.0.0/16 será inalcanzable desde cualquier carga de trabajo en la red virtual 2 y desde las que estén emparejadas con la red virtual 2.



Cuando se aprende la misma ruta a través de ExpressRoute, VPN o SD-WAN, ¿qué red se prefiere?

De forma predeterminada, las rutas aprendidas sobre ExpressRoute tienen prioridad sobre las aprendidas a través de VPN o SD-WAN. Puede configurar la preferencia de enrutamiento para influir en la selección de rutas de Route Server. Para obtener más información, consulte Preferencia de enrutamiento.

¿Cuáles son los requisitos para que una puerta de enlace de VPN de Azure funcione con Azure Route Server?

Azure VPN Gateway debe estar configurado en modo activo-activo y tener el valor de ASN establecido en 65515.

¿Es necesario habilitar BGP en la puerta de enlace de VPN?

No. No es un requisito tener BGP habilitado en la puerta de enlace de VPN para comunicarse con el servidor de rutas.

¿Puedo emparejar dos servidores de ruta de Azure en dos redes virtuales emparejadas y habilitar los NVA conectados a los servidores de ruta para que se comuniquen entre sí?

Topología: NVA1 -> RouteServer1 -> (a través del emparejamiento de red virtual) -> RouteServer2 -> NVA2

No, Azure Route Server no desvía el tráfico de datos. Para permitir la conectividad de tránsito a través de la NVA, configure una conexión directa (por ejemplo, un túnel IPsec) entre las NVA y use los servidores de rutas para la propagación dinámica de rutas.

¿Se puede usar Azure Route Server para dirigir el tráfico entre subredes de la misma red virtual para que el tráfico entre las subredes fluya por la NVA?

No. Azure Route Server usa el BGP para anunciar rutas. Las rutas de sistema para el tráfico relacionado con la red virtual, los emparejamientos de la red virtual o los puntos de conexión de servicio de red virtual son las rutas preferidas, aunque las rutas BGP sean más específicas. Tendrá que seguir usando rutas definidas por el usuario (UDR) para invalidar las rutas del sistema y no puede utilizar el BGP para conmutarlas por error rápidamente. Tendrá que seguir usando una solución de terceros para actualizar las UDR mediante la API en caso de conmutación por error, o bien usar una instancia de Azure Load Balancer con el modo de puertos de alta disponibilidad para dirigir el tráfico.

Todavía puede usar Route Server para dirigir el tráfico entre subredes de distintas redes virtuales para que fluyan mediante la NVA. Un diseño que podría funcionar es tener una subred para cada red virtual "spoke" y que todas las redes virtuales "spoke" estén conectadas a una red virtual "hub". Este diseño es muy limitado y debe tener en cuenta las consideraciones de escalado y los límites máximos de Azure en las redes virtuales en comparación con las subredes.

¿Puede Azure Route Server proporcionar tránsito entre ExpressRoute y una conexión de puerta de enlace de VPN de punto a sitio (P2S) al habilitar la configuración de rama a rama?

No, Azure Route Server proporciona tránsito solo entre ExpressRoute y conexiones de puerta de enlace de VPN de sitio a sitio (S2S) al habilitar la configuración *de rama a rama*.

¿Puedo crear un servidor de rutas de Azure en una red virtual de satélite que se conecte a un hub de Virtual WAN?

No. La red virtual de spoke no puede tener un Route Server si está conectado al centro de conectividad de Virtual WAN.

¿Qué ocurre si una NVA anuncia una ruta que contiene el ASN 65515 en la AS_PATH?

Azure Route Server utiliza ASN 65515 por defecto. Si un NVA anuncia una ruta a Azure Route Server que ya contiene 65515 en el atributo BGP AS_PATH, Azure Route Server identifica su propio ASN en la ruta AS y rechaza la ruta como parte del comportamiento estándar de prevención de bucles BGP. Como resultado, la ruta no se instala ni propaga por Azure Route Server.

Para evitar este problema, asegúrese de que las rutas anunciadas por los NVA no contengan el ASN 65515 reservado por Azure en el AS_PATH.

Ejemplo: Si un NVA anuncia el prefijo 10.10.0.0/16 con un AS_PATH 65001 65515, Azure Route Server elimina la ruta porque el AS_PATH ya contiene su propio ASN (65515), lo que crearía un posible bucle de enrutamiento. Este comportamiento es coherente con los mecanismos estándar de prevención de bucles BGP.

Limitaciones

¿Cuántos servidores de Azure Route Server puedo crear en una red virtual?

Solo puede crear un servidor de rutas en una red virtual. Debe implementar Route Server en una subred dedicada denominada *RouteServerSubnet*.

¿Puedo asociar una UDR a RouteServerSubnet?

No, Azure Route Server no admite la configuración de una ruta definida por el usuario (UDR) en la subred *RouteServerSubnet* . Azure Route Server no enruta ningún tráfico de datos entre aplicaciones virtuales de red (NVA) y máquinas virtuales (VM).

¿Puedo asociar un grupo de seguridad de red (NSG) a RouteServerSubnet?

No, Azure Route Server no admite la asociación de grupos de seguridad de red a la subred *RouteServerSubnet* .

¿Cuáles son los límites de Azure Route Server?

Azure Route Server tiene los límites siguientes (por implementación).

Expandir tabla

Recurso	Límite
Número de pares BGP	16
Número de rutas que puede anunciar cada par BGP a Azure Route Server [1]	4.000
Número de máquinas virtuales de la red virtual (incluidas las redes virtuales emparejadas) que Azure Route Server puede admitir.	50,000
Número de redes virtuales que admite Azure Route Server	500
Número total de prefijos locales y de Azure Virtual Network que Azure Route Server puede admitir	10 000

[1] Si NVA anuncia más rutas que el límite, la sesión BGP se anula.

> **Nota:**
>
> El número total de rutas anunciadas desde el espacio de direcciones de red virtual y el Servidor de Rutas hacia el circuito ExpressRoute, cuando está habilitada la opción Rama a Rama, no debe exceder de 1 000. Para más información, consulte los límites de anuncios de rutas de ExpressRoute.

Para obtener información sobre cómo solucionar problemas de enrutamiento en una máquina virtual, consulte Diagnóstico de un problema de enrutamiento de máquinas virtuales de Azure.

¿Por qué veo un error sobre el ámbito y la autorización no válidos para realizar operaciones en los recursos de Route Server?

Si ve un error en el formato siguiente, asegúrese de que tiene los permisos siguientes configurados: Roles y permisos de Route Server.

Formato del mensaje de error: "El cliente con el identificador {} de objeto no tiene autorización para realizar acciones {} en el ámbito {} o el ámbito no es válido. Para más información sobre los permisos necesarios, visite {}. Si el acceso se ha concedido recientemente, actualice las credenciales."

Pasos siguientes

Consulte Configuración de una instancia de Azure Route Server.

Last updated on 30/07/2026

Confiabilidad en Azure Traffic Manager

Azure Traffic Manager es un equilibrador de carga de tráfico basado en DNS que distribuye el tráfico de forma óptima entre back-end distribuidos globalmente. Traffic Manager proporciona alta disponibilidad y capacidad de respuesta rápida para las aplicaciones orientadas al público mediante DNS para dirigir las solicitudes de cliente a los puntos de conexión de servicio adecuados en función de los métodos de enrutamiento del tráfico y la supervisión del estado del punto de conexión.

Cuando se usa Azure, la confiabilidad es una responsabilidad compartida. Microsoft proporciona una variedad de capacidades para apoyar la resiliencia y la recuperación. Es responsable de comprender cómo funcionan esas funcionalidades dentro de todos los servicios que usa y de seleccionar las funcionalidades que necesita para cumplir los objetivos empresariales y los objetivos de tiempo de actividad.

En este artículo se describen las funcionalidades de confiabilidad de Azure Traffic Manager en respuesta a una serie de posibles interrupciones, incluidos los errores transitorios y los errores de toda la región. También resalta las consideraciones clave para mantener la resistencia y la preparación para la recuperación, y proporciona información general sobre el acuerdo de nivel de servicio (SLA) de Azure Traffic Manager.

> ⓘ **Nota:**
>
> En este artículo se describe cómo el servicio Traffic Manager es resistente o cómo puede hacer que sea resistente a varios problemas. No explica cómo usar Traffic Manager para realizar la conmutación por error entre aplicaciones o regiones. Para obtener un ejemplo de arquitectura de conmutación por error, consulte **Aplicación web de varios niveles creada para alta disponibilidad y recuperación ante desastres**.

Recomendaciones de implementación de producción

Azure Well-Architected Framework proporciona recomendaciones para la confiabilidad, el rendimiento, la seguridad, el costo y las operaciones. Para obtener información sobre cómo estas áreas influyen entre sí y contribuyen a una solución confiable de Traffic Manager, consulte Architecture best practices for Azure Traffic Manager in the Well-Architected Framework.

Introducción a la arquitectura de confiabilidad

En esta sección se describen algunos de los aspectos importantes de cómo funciona el servicio que es más relevante desde una perspectiva de confiabilidad. En la sección se presenta la arquitectura lógica, que incluye algunos de los recursos y características que se implementan y usan. También se describe la arquitectura física, que proporciona detalles sobre cómo funciona el servicio en segundo plano.

Arquitectura lógica

Cuando se usa Traffic Manager, se implementa un *perfil*, que especifica los puntos de conexión de back-end de la aplicación y configura cómo Traffic Manager debe enrutar las solicitudes a esos puntos de conexión. Para obtener más información, consulte Puntos de conexión de Traffic Manager y métodos de enrutamiento de Traffic Manager.

Un perfil de Traffic Manager se presenta como un registro CNAME de DNS. Cuando recibe una solicitud de resolución de un cliente o resolución DNS, Traffic Manager resuelve dinámicamente la dirección IP en función de las reglas que especifique en el perfil. La responsabilidad de Traffic Manager es proporcionar a los clientes la dirección IP de un punto de conexión para llegar al servicio. Después de la resolución de nombres, el tráfico de la aplicación no fluye a través de Traffic Manager. Para obtener más información, consulte Funcionamiento de Traffic Manager.

Traffic Manager supervisa el estado de los puntos de conexión y enruta las solicitudes entrantes a puntos de conexión correctos, a la vez que evita puntos de conexión incorrectos. Para más información, consulte Supervisión de puntos de conexión de Traffic Manager.

> **ⓘ Importante**
>
> La confiabilidad de la solución general depende de la configuración de los puntos de conexión a los que el administrador de tráfico enruta el tráfico.
>
> En este artículo no se tratan los puntos de conexión, pero sus configuraciones de disponibilidad afectan directamente a la resistencia de la aplicación. Revise las **guías de confiabilidad de los servicios de Azure de la solución** para obtener información sobre cómo cada servicio admite sus requisitos de confiabilidad.

Arquitectura física

Traffic Manager funciona como un servicio no regional e implementa su infraestructura en varias zonas de disponibilidad en varias regiones de Azure en todo el mundo. Este diseño permite que Traffic Manager siga siendo resistente durante una interrupción de la región o zona de disponibilidad, ya que la infraestructura de otra zona o región sigue respondiendo a las solicitudes de resolución.

Los protocolos globales de Internet como Anycast, DNS y BGP enrutan automáticamente solicitudes de resolución DNS entrantes a la infraestructura de Traffic Manager más cercana y adecuada.

Resistencia a errores transitorios

Los errores transitorios son errores breves e intermitentes en los componentes. Se producen con frecuencia en un entorno distribuido como la nube y son una parte normal de las operaciones. Los errores transitorios se corrigen después de un breve período de tiempo. Es importante que las aplicaciones puedan controlar errores transitorios, normalmente mediante el reintento de solicitudes afectadas.

Todas las aplicaciones hospedadas en la nube deben seguir las instrucciones de control de errores transitorios de Azure cuando se comunican con cualquier API, bases de datos y otros componentes hospedados en la nube. Para obtener más información, consulte Recomendaciones para controlar errores transitorios.

Traffic Manager funciona en el nivel DNS y usa sondeos de estado para supervisar la disponibilidad del punto de conexión. El servicio controla los errores transitorios a través de sus capacidades globales de supervisión de puntos de conexión e infraestructura DNS.

Al usar Traffic Manager, tenga en cuenta los siguientes tipos de errores transitorios por separado:

- **Errores transitorios durante la resolución dns:** Si se produce un error transitorio durante la resolución DNS, el cliente o el solucionador intermedio deben reintentar.
- **Errores transitorios que afectan a los puntos de conexión de back-end:**El monitoreo de puntos de conexión por parte de Traffic Manager verifica periódicamente el estado de estos puntos de conexión. Un error transitorio dentro de un punto de conexión o en la ruta de acceso de red a un punto de conexión puede detectarse como un punto de conexión incorrecto. Configure la supervisión de puntos de conexión para detectar problemas consecutivos en un periodo determinado.

El tiempo de vida del registro DNS (TTL) determina cómo controla la solución los errores. Si el TTL es muy bajo, los clientes deben realizar más solicitudes a Traffic Manager y hay más oportunidades de que surjan errores transitorios. Si el TTL es elevado, en caso de una falla real en un punto de conexión, los clientes podrían experimentar retrasos en el proceso de conmutación por error hasta que expire el TTL. Configure las TTL cuidadosamente para equilibrar la disponibilidad, la latencia y la capacidad de respuesta. Cuando se usa Azure DNS, se puede configurar automáticamente el TTL del registro para que coincida con el valor de TTL del perfil, que es de 60 segundos de forma predeterminada. Para obtener más información, consulte Consideraciones de rendimiento para Traffic Manager.

Resistencia a errores de zona de disponibilidad

Las zonas de disponibilidad son grupos físicamente independientes de centros de datos dentro de una región de Azure. Cuando una zona falla, los servicios pueden transferirse a una de las zonas restantes.

Traffic Manager funciona como un servicio no regional e implementa su infraestructura en varias zonas de disponibilidad en varias regiones de Azure en todo el mundo. Replica los cambios en el perfil de forma sincrónica en estas zonas y regiones. Este diseño permite que Traffic Manager permanezca resistente durante una interrupción de zona de disponibilidad, ya que la infraestructura de otra zona o región sigue respondiendo a las solicitudes de resolución.

Resistencia a errores en toda la región

Traffic Manager funciona como un servicio no regional e implementa su infraestructura en varias zonas de disponibilidad en varias regiones de Azure en todo el mundo. Este diseño permite que Traffic Manager permanezca resistente durante una interrupción de la región, ya que la infraestructura de otra zona o región sigue respondiendo a las solicitudes de resolución.

Resiliencia ante interrupciones del portal y herramientas de gestión

Si administra el perfil de Traffic Manager en el portal de Azure, prepárese para escenarios en los que no pueda acceder a él, especialmente si necesita volver a configurar el perfil durante una interrupción de la plataforma.

Al igual que otros servicios de Azure, Traffic Manager admite la implementación y la administración a través de una variedad de herramientas. Se recomienda familiarizarse con cómo usar CLI de Azure o Azure PowerShell para administrar el perfil. Como alternativa, implemente y

configure su perfil mediante el uso de tecnologías de infraestructura como código, como Bicep o Terraform. Estas herramientas permanecen operativas incluso si el portal de Azure está degradado.

Copias de seguridad y restauración

Traffic Manager es un servicio DNS sin estado. No conserva los datos y no tiene capacidad de copia de seguridad ni restauración.

Para proteger la configuración de recursos, defina los perfiles de Traffic Manager y otros recursos mediante la infraestructura como código (como Bicep o plantillas de ARM) y almacene esas definiciones en el control de código fuente. Si necesita volver a crear un recurso, vuelva a implementarlo desde la configuración almacenada.

Resistencia al mantenimiento del servicio

Microsoft aplica periódicamente actualizaciones de servicio y realiza otro mantenimiento. La plataforma Azure controla estas actividades automáticamente, lo que garantiza que el mantenimiento sea transparente y sin problemas. No se espera ningún tiempo de inactividad durante los eventos de mantenimiento a menos que se le haya informado a través del mantenimiento planeado de Azure Service Health.

Acuerdo de nivel de servicio

El acuerdo de nivel de servicio (SLA) para Azure servicios describe la disponibilidad esperada de cada servicio y las condiciones que la solución debe cumplir para lograr esa expectativa de disponibilidad. Para obtener más información, consulte Acuerdos de Nivel de Servicio para servicios en línea.

Azure Traffic Manager proporciona un Acuerdo de Nivel de Servicio de disponibilidad de 100% para las respuestas de consulta DNS, siempre y cuando los clientes vuelvan a intentar solicitudes con errores repetidamente.

Contenido relacionado

- Información general de Azure Traffic Manager
- Documentación sobre confiabilidad de Azure

Last updated on 12/05/2026

Confiabilidad en Azure Virtual Network

Azure Virtual Network es una representación lógica de la red en la nube. Puede usar una red virtual para definir su propio espacio de direcciones IP privadas y segmentar la red en subredes. Las redes virtuales sirven como límite de confianza para hospedar los recursos de proceso, como Azure Virtual Machines y equilibradores de carga.

Al usar Azure, relibilidad es una responsabilidad compartida. Microsoft proporciona una variedad de funcionalidades para admitir resistencia y recuperación. Es responsable de comprender cómo funcionan esas funcionalidades dentro de todos los servicios que usa y de seleccionar las funcionalidades que necesita para cumplir los objetivos empresariales y los objetivos de tiempo de actividad.

En este artículo se describe cómo hacer que Virtual Network sean resistentes a una variedad de posibles interrupciones y problemas, incluidos errores transitorios, interrupciones de zona de disponibilidad y interrupciones de regiones. También se describe cómo puede usar copias de seguridad para recuperarse de otros tipos de problemas y se resalta cierta información clave sobre el acuerdo de nivel de servicio (SLA) de Virtual Network.

Recomendaciones de implementación de producción

Para obtener más información sobre cómo implementar redes virtuales para admitir los requisitos de confiabilidad de la solución y cómo afecta la confiabilidad a otros aspectos de la arquitectura, consulte Architecture best practices for Azure Virtual Network in the Azure Well-Architected Framework.

Introducción a la arquitectura de confiabilidad

Una red virtual es uno de los componentes principales de red en Azure. Al crear una red virtual, se crea un conjunto de recursos que definen colectivamente la configuración de red. Estos recursos incluyen los siguientes componentes de red:

- Grupos de seguridad de red y grupos de seguridad de aplicaciones, que restringen la comunicación entre partes de su red
- Rutas definidas por el usuario, que controlan cómo fluye el tráfico
- Equilibradores de carga, que distribuyen el tráfico dentro de la red

- Direcciones IP públicas, que proporcionan conectividad con y desde Internet
- Tarjetas de interfaz de red, que proporcionan conectividad de red a máquinas virtuales de Azure (VM)
- Puntos de conexión privados, que proporcionan conectividad privada a servicios Azure y a recursos fuera de la red virtual

Una red virtual permite la comunicación ip privada directa entre los recursos que hospeda. Para habilitar escenarios de nube híbrida y ampliar de forma segura el centro de datos a Azure, puede vincular una red virtual a una red local a través de Azure VPN Gateway o Azure ExpressRoute.

También puede implementar *dispositivos*, como puertas de enlace de ExpressRoute, puertas de enlace de VPN y firewalls. Los dispositivos proporcionan servicios para admitir los requisitos de red, como conectarse a entornos locales o proporcionar controles sofisticados sobre el flujo de tráfico.

Por último, implemente sus propios componentes, como las máquinas virtuales que ejecutan aplicaciones o bases de datos, y otros servicios de Azure que proporcionan integración de red virtual.

Para obtener más información sobre las redes en Azure, consulte Diseño de arquitectura de red.

Resistencia a errores transitorios

Los errores transitorios son errores breves e intermitentes en los componentes. Se producen con frecuencia en un entorno distribuido como la nube y son una parte normal de las operaciones. Los errores transitorios se corrigen después de un breve período de tiempo. Es importante que las aplicaciones puedan controlar errores transitorios, normalmente mediante el reintento de solicitudes afectadas.

Todas las aplicaciones hospedadas en la nube deben seguir las instrucciones de control de errores transitorios Azure cuando se comunican con cualquier API, bases de datos y otros componentes hospedados en la nube. Para obtener más información, consulte Recomendaciones para controlar errores transitorios.

Los errores transitorios no suelen afectar a las redes virtuales. Sin embargo, los errores transitorios pueden afectar a los recursos implementados dentro de una red virtual. Revise la guía de confiabilidad de cada recurso que use para comprender sus comportamientos transitorios de control de errores.

Resistencia a errores de zona de disponibilidad

Availability zones son grupos de centros de datos físicamente independientes dentro de una región de Azure. Cuando una zona falla, los servicios pueden transferirse a una de las zonas restantes.

Una red virtual y sus subredes abarcan todas las zonas de disponibilidad dentro de la región donde se implementan, por lo que son redundantes de zona automáticamente. No es necesario configurar nada para habilitar la redundancia de zona.

No es necesario dividir las redes virtuales o subredes por zonas de disponibilidad para dar cabida a los recursos zonales. Por ejemplo, si configura una máquina virtual zonal, no tiene que tener en cuenta la red virtual al seleccionar la zona de disponibilidad de la máquina virtual. Lo mismo sucede con otros recursos zonales.

Requisitos

Compatibilidad con regiones: Puede implementar redes virtuales con redundancia de zona en cualquier región que admita zonas de disponibilidad.

Cost

No hay costo adicional por la redundancia de zona en redes virtuales de Azure.

Configurar soporte de zonas de disponibilidad

La redundancia de zona se configura automáticamente cuando se implementa una red virtual en una región que admite zonas de disponibilidad.

Comportamiento cuando todas las zonas están en buen estado

En esta sección se describe qué esperar al implementar una red virtual en una región que admita zonas de disponibilidad y todas las zonas están operativas.

- **Operación entre zonas:** Una red virtual y sus subredes abarcan todas las zonas de disponibilidad de la región. Los recursos implementados en cualquier zona usan la misma red virtual y espacio de direcciones. El tráfico IP privado fluye directamente entre los recursos, incluso cuando los recursos están en diferentes zonas de disponibilidad.

- **Replicación de datos entre zonas:** Virtual Network es un servicio de red sin estado. La configuración se replica sincrónicamente entre zonas, lo que garantiza una configuración

coherente en cada zona.

Comportamiento durante un fallo de zona

En esta sección se describe qué esperar al implementar una red virtual en una región que admite zonas de disponibilidad y hay una interrupción en una de las zonas.

> **Importante**
>
> Tenga en cuenta cada recurso dentro de la red virtual individualmente, ya que cada recurso puede tener un conjunto diferente de comportamientos durante la pérdida de una zona de disponibilidad. Revise la **guía de confiabilidad de cada recurso** que use para comprender su compatibilidad y comportamiento de la zona de disponibilidad cuando una zona no está disponible.

- **Detección y respuesta:** La plataforma Azure detecta fallos de zona de disponibilidad y responde automáticamente. No es necesario hacer nada para iniciar una conmutación por error de zona para la propia red virtual.

- **Notificación:** Microsoft no le notifica automáticamente cuando una zona está inactiva. Sin embargo, puede usar Azure Service Health para comprender el estado general del servicio, incluidos los errores de zona, y puede configurar alertas de Service Health para notificarle problemas.

- **Active requests:** Virtual Network es una construcción lógica y no finaliza las conexiones. Los flujos activos entre recursos ubicados en zonas en buen estado siguen funcionando. Los flujos hacia o desde los recursos de la zona afectada se interrumpen porque los recursos subyacentes no están disponibles, no debido a la red virtual. Los clientes deben implementar el control de errores transitorios, incluidos los reintentos automatizados.

- **Pérdida de datos esperada:** La red virtual no almacena datos, por lo que no se produce ninguna pérdida de datos en la capa de red virtual.

- **Tiempo de inactividad esperado:** No se espera ningún tiempo de inactividad para la propia red virtual. La red virtual sigue funcionando desde las zonas correctas y su configuración sigue estando disponible.

  Sin embargo, los recursos implementados dentro de la red virtual podrían no estar disponibles en función de su propia configuración de zona de disponibilidad.

- **Redistribution:** Cuando una zona deja de estar disponible, Virtual Network vuelve a enrutar automáticamente las solicitudes de virtual network a las zonas restantes. Este proceso no requiere ninguna acción de usted.

  El comportamiento de redistribución de los recursos implementados en la red virtual depende de cada tipo de recurso. Para obtener más información, consulte la guía de confiabilidad de cada recurso.

Recuperación de zona

Cuando la zona se recupera, Microsoft inicia un proceso de reactivación para asegurarse de que las redes virtuales sigan funcionando en la zona recuperada. El proceso de conmutación por recuperación es automático y no requiere ninguna acción de usted.

Sin embargo, debe verificar los comportamientos de recuperación ante fallos de los recursos que implemente en la red virtual. Para obtener más información, consulte la guía de confiabilidad de cada recurso.

Prueba de fallos de zona

La plataforma Virtual Network gestiona el enrutamiento del tráfico, la conmutación por error y el restablecimiento de redes virtuales entre zonas de disponibilidad. Dado que esta característica está totalmente administrada, no es necesario validar los procesos de error de zona de disponibilidad.

Resistencia a errores en toda la región

Virtual Network es un servicio de una sola región. Si la región deja de estar disponible, la red virtual tampoco está disponible.

Soluciones personalizadas de varias regiones para la resistencia

Puede crear redes virtuales en varias regiones. También puede optar por conectar esas redes *emparejandolas*.

Al crear redes virtuales y otros recursos en varias regiones, puede ser resistente a interrupciones regionales. Sin embargo, debe tener en cuenta los siguientes factores:

- **Enrutamiento del tráfico:** Si hospeda servicios accesibles desde Internet en la red virtual, debe decidir cómo enrutar el tráfico entrante entre las regiones y los componentes. Con

servicios como Azure Traffic Manager y Azure Front Door, puede enrutar el tráfico de Internet en función de las reglas que especifique.

- **Failover:** Si una región de Azure no está disponible, normalmente debe realizar una conmutación por error procesando el tráfico en regiones funcionales. Traffic Manager y Azure Front Door proporcionan funcionalidades de conmutación por error para las aplicaciones de Internet.
- **Administración:** Cada red virtual es un recurso independiente y debe configurarse y administrarse independientemente de otras redes virtuales.
- **Espacio de direcciones IP:** Determine cómo asignar direcciones IP al crear varias redes virtuales. Puede crear varias redes virtuales mediante el mismo espacio de direcciones IP privadas en regiones diferentes. Sin embargo, no puede emparejar ni conectar dos redes virtuales con el mismo espacio de direcciones a la red local porque causa problemas de enrutamiento. Si tiene previsto crear un diseño de varias redes, el planeamiento de direcciones IP es una consideración importante.

Las redes virtuales no requieren una gran cantidad de recursos para ejecutarse. Puede invocar Azure API para crear una red virtual con el mismo espacio de direcciones en otra región. Sin embargo, para volver a crear un entorno similar al que existe en la región afectada, debe volver a implementar las máquinas virtuales y otros recursos. Si tiene conectividad local, como en una implementación híbrida, debe implementar una nueva instancia de VPN Gateway y conectarse a la red local.

Para obtener más información sobre una arquitectura de red de varias regiones para aplicaciones web, consulte Equilibrio de carga multirregional con Traffic Manager, Azure Firewall y Azure Application Gateway.

Copias de seguridad y restauración

Las redes virtuales no almacenan ningún dato que requiera copia de seguridad. Sin embargo, puede usar Bicep, Azure Resource Manager plantillas o Terraform para tomar una instantánea de la configuración de una red virtual si necesita volver a crearla. Para obtener más información, consulte Create an Azure virtual network.

Resistencia al mantenimiento del servicio

Microsoft aplica periódicamente actualizaciones de servicio y realiza otro mantenimiento. La plataforma Azure controla estas actividades automáticamente, lo que garantiza que el

mantenimiento sea transparente y sin problemas. No se espera ningún tiempo de inactividad durante los eventos de mantenimiento a menos que se le haya informado a través del mantenimiento planeado de Azure Service Health.

Acuerdo de nivel de servicio

El acuerdo de nivel de servicio (SLA) para Azure servicios describe la disponibilidad esperada de cada servicio y las condiciones que la solución debe cumplir para lograr esa expectativa de disponibilidad. Para obtener más información, vea SLAs for servicios en línea.

Debido a la naturaleza del servicio proporcionado, no hay un acuerdo de nivel de servicio definido para Virtual Network.

Contenido relacionado

- Zonas de disponibilidad

ⓘ**Nota:** El autor ha creado este artículo con ayuda de inteligencia artificial. Más información

---

Last updated on 11/06/2026

Confiabilidad en Azure Virtual Network Manager

Azure Virtual Network Manager es un servicio de administración centralizado que ayuda a agrupar, configurar, implementar y administrar redes virtuales en las suscripciones de Azure y los inquilinos de Microsoft Entra. Puede usarlo para aplicar configuraciones de conectividad, seguridad, administración de direcciones IP y enrutamiento a escala.

Cuando se usa Azure, la confiabilidad es una responsabilidad compartida. Microsoft proporciona una variedad de capacidades para apoyar la resiliencia y la recuperación. Es responsable de comprender cómo funcionan esas funcionalidades dentro de todos los servicios que usa y de seleccionar las funcionalidades que necesita para cumplir los objetivos empresariales y los objetivos de tiempo de actividad.

En este artículo se describe cómo hacer que Azure Virtual Network Manager sea resistente a errores transitorios, fallos en las zonas de disponibilidad, fallos en toda una región y tareas de mantenimiento del servicio. También resalta el acuerdo de nivel de servicio (SLA).

Introducción a la arquitectura de confiabilidad

Administrador de redes virtuales es un servicio de administración de red centralizado. Cree una *instancia* de Administrador de redes virtuales, defina su ámbito y coloque redes virtuales en *grupos de red*, incluso si esas redes virtuales se distribuyen entre varias regiones de Azure.

Después de crear la instancia de Administrador de redes virtuales, implemente configuraciones de conectividad, seguridad y enrutamiento en las redes dentro de los grupos de red. Una *implementación* representa un conjunto de cambios que se aplican a las redes virtuales. El servicio administra cómo se aplican las implementaciones en los recursos de red.

Como servicio de administración, Administrador de redes virtuales almacena y administra los datos de configuración, pero no procesa el tráfico de la carga de trabajo. Las cargas de trabajo siguen usando las redes virtuales subyacentes, los emparejamientos, las puertas de enlace y los recursos de enrutamiento que configure.

Para obtener más información sobre la arquitectura y las funcionalidades del servicio, consulte ¿Qué es Azure Virtual Network Manager?.

Resistencia a errores transitorios

Los errores transitorios son errores breves e intermitentes en los componentes. Se producen con frecuencia en un entorno distribuido como la nube y son una parte normal de las operaciones. Los errores transitorios se corrigen después de un breve período de tiempo. Es importante que las aplicaciones puedan controlar errores transitorios, normalmente mediante el reintento de solicitudes afectadas.

Todas las aplicaciones hospedadas en la nube deben seguir las instrucciones de control de errores transitorios de Azure cuando se comunican con cualquier API, bases de datos y otros componentes hospedados en la nube. Para obtener más información, consulte Recomendaciones para controlar errores transitorios.

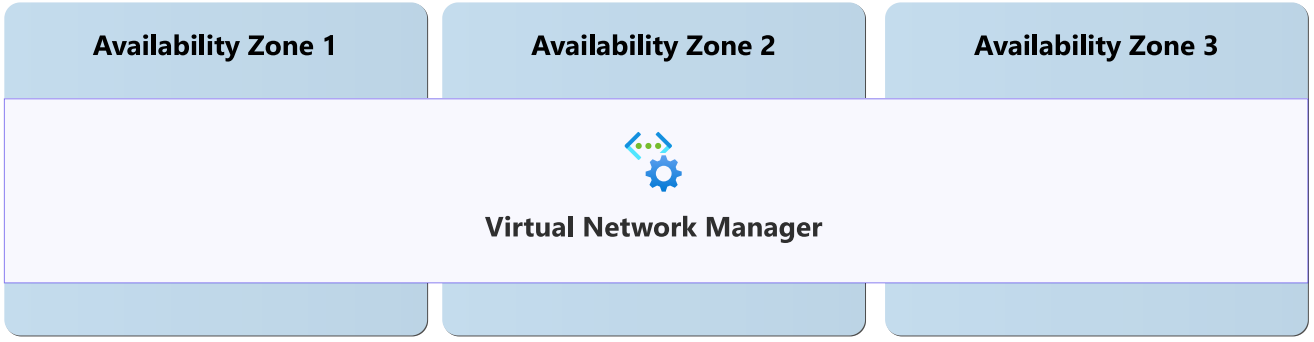
Interactúe con Administrador de redes virtuales mediante el portal de Azure, CLI de Azure, Azure PowerShell o una tecnología de infraestructura como código (IaC), como Bicep o Terraform. La mayoría de estas herramientas reintentan automáticamente si se produce un error transitorio. Si usa las API de Azure Resource Manager, asegúrese de que la aplicación vuelva a intentarlo automáticamente después de un breve retraso.

Use Administrador de redes virtuales para configurar y administrar los recursos de red. No participa en el procesamiento del tráfico de la carga de trabajo, por lo que los errores transitorios en el servicio no afectan directamente a la carga de trabajo.

Resistencia a errores de zona de disponibilidad

Las zonas de disponibilidad son grupos físicamente independientes de centros de datos dentro de una región de Azure. Cuando una zona falla, los servicios pueden transferirse a una de las zonas restantes.

Administrador de redes virtuales tiene redundancia de zona automáticamente al implementarlo en una región de Azure que admite zonas de disponibilidad.



Requirements

Compatibilidad con regiones: Administrador de redes virtuales redundancia de zona está disponible en todas las regiones de Azure que admiten Administrador de redes virtuales y proporcionan zonas de disponibilidad. Para obtener una lista de las regiones que admiten Administrador de redes virtuales, consulte Productos disponibles por región. Para obtener una lista completa de las regiones que admiten zonas de disponibilidad, consulte Regiones de Azure que admiten zonas de disponibilidad.

Comportamiento cuando todas las zonas están en buen estado

En esta sección se describe qué esperar al implementar una instancia de Administrador de redes virtuales en una región con zonas de disponibilidad y todas las zonas de disponibilidad están operativas.

- **Operación entre zonas:** Las operaciones de red de Azure enrutan el tráfico entre zonas de forma transparente.
- **Replicación de datos entre zonas:** Azure Virtual Network Manager replica sincrónicamente su estado de configuración entre zonas.

Comportamiento durante un fallo de zona

En esta sección se describe qué esperar al implementar una instancia de Administrador de redes virtuales en una región con zonas de disponibilidad y hay una interrupción en una de las zonas de disponibilidad.

- **Detección y respuesta:** Microsoft detecta errores de zona de disponibilidad y administra todas las acciones de respuesta. No es necesario realizar ninguna acción para iniciar una conmutación por error de zona.
- **Notificación:** Microsoft no le notifica automáticamente cuando una zona está inactiva. Sin embargo, puede usar Azure Service Health para comprender el estado general del servicio, incluidos los errores de zona, y puede configurar alertas de Service Health para notificarle problemas.
- **Implementaciones activas:** Las implementaciones activas pueden pausarse si usan la infraestructura afectada. Una vez recuperada la zona, se reanudan automáticamente.
- **Pérdida de datos esperada:** No se espera ninguna pérdida de datos o configuración durante una interrupción de zona.

- **Tiempo de inactividad esperado:** Una pequeña cantidad de tiempo de inactividad, normalmente unos segundos, puede producirse mientras el servicio redirige a una infraestructura correcta.
- **Redistribución:** Microsoft redirige automáticamente las solicitudes y las operaciones de administración hacia las zonas en buen estado.

Recuperación de zona

Cuando se recupera una zona de disponibilidad con errores, Administrador de redes virtuales restaura automáticamente las operaciones normales sin su intervención.

Prueba de fallos de zona

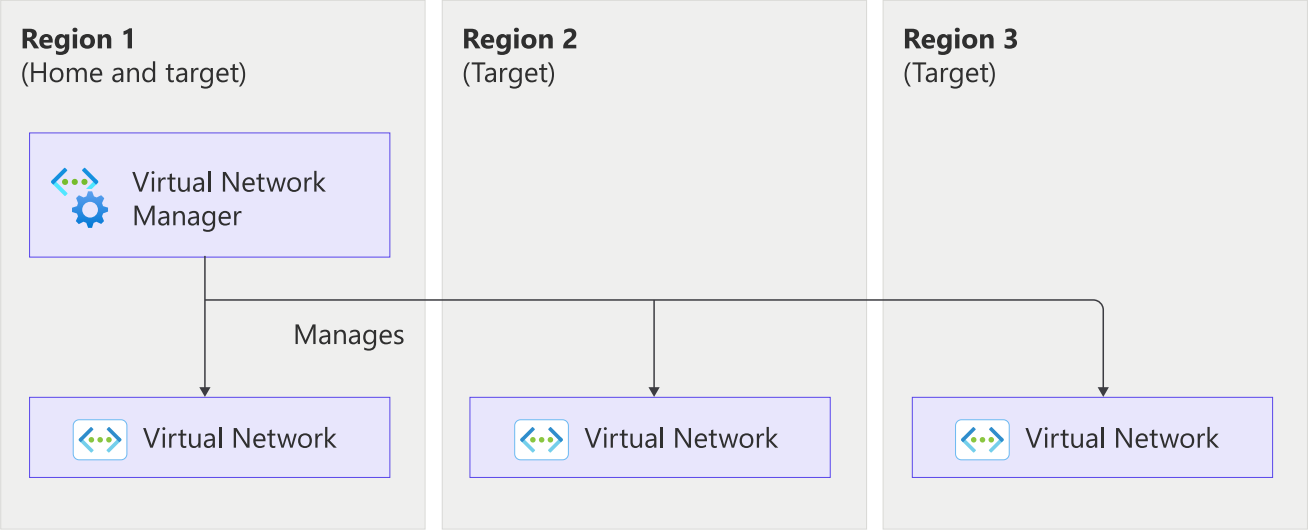
Administrador de redes virtuales es un servicio con redundancia de zona totalmente administrado por Microsoft. Dado que Microsoft gestiona la redundancia de las zonas de disponibilidad, no es necesario probar escenarios de conmutación por error de zonas.

Resistencia a errores en toda la región

Administrador de redes virtuales es un recurso de una sola región. Si la región que hospeda la instancia de Administrador de redes virtuales deja de estar disponible, la instancia tampoco está disponible. Sin embargo, una instancia de Administrador de redes virtuales puede administrar redes virtuales en otras regiones de Azure.

Administración de redes virtuales entre regiones

Administrador de redes virtuales puede administrar redes virtuales que se distribuyen entre varias regiones de Azure. Use este enfoque para administrar e implementar implementaciones en redes virtuales distribuidas globalmente. La región en la que implementa la instancia de Administrador de redes virtuales en es su *región principal* y cada una de las regiones que contiene redes virtuales es una *región de destino*.



Requirements

Compatibilidad con regiones: La compatibilidad con regiones depende del tipo de región:

- *Región principal:* Para obtener una lista de las regiones que admiten la implementación de instancias de Administrador de redes virtuales, consulte Productos disponibles por región.
- *Región de destino:* Puede conectar los recursos de red virtual en cualquier región de Azure.

Configuración de la compatibilidad con varias regiones

Al crear una instancia de Administrador de redes virtuales, seleccione la región donde se hospeda la instancia. A continuación, agregue redes virtuales a los grupos de red que administra la instancia y esas redes virtuales pueden estar en cualquier región de Azure.

Comportamiento cuando todas las regiones están en buen estado

En esta sección se describe qué esperar cuando todas las regiones están operativas.

- **Operación entre regiones:** La instancia de Administrador de redes virtuales implementa configuraciones de conectividad, seguridad y enrutamiento en redes virtuales en las regiones de destino que especifique.
- **Replicación de datos entre regiones:** Administrador de redes virtuales almacena su configuración en la región principal y replica la configuración implementada en cada región de destino. Cada región de destino usa su copia local de la configuración para aplicar la configuración a las redes virtuales administradas de esa región.

Comportamiento durante una interrupción del servicio en la región de origen

En esta sección se describe qué esperar cuando se produce una interrupción en la región principal, que hospeda la instancia de Administrador de redes virtuales.

- **Detección y respuesta:** Microsoft detecta el error de la región y administra todas las acciones de respuesta del servicio Administrador de redes virtuales.
- **Notificación:** Microsoft no le notifica automáticamente cuando una región está inactiva. Sin embargo, puede usar Azure Service Health para comprender el estado general del servicio, incluidos los errores de región, y puede configurar alertas de Service Health para notificarle problemas.
- **Implementaciones activas:** Es posible que se produzca un error en las implementaciones en curso porque la instancia no puede distribuir la configuración a las regiones de destino. Necesita reiniciar estos despliegues después de que se recupere la región de origen.
- **Pérdida de datos esperada:** No se espera ninguna pérdida de datos de configuración, pero la instancia no está disponible hasta que se recupere la región principal. Las regiones de destino siguen funcionando en la última configuración correcta conocida que se ha replicado en ellas, por lo que las configuraciones ya implementadas en redes virtuales administradas en esas regiones permanecen intactas y siguen aplicándose durante la interrupción.
- **Tiempo de inactividad esperado:** La instancia de Administrador de redes virtuales no está disponible durante la interrupción de la región. Las solicitudes de administración a la instancia de Administrador de redes virtuales fallan hasta que la región se recupere.
- **Redistribución:** Administrador de redes virtuales no replica la instancia en otras regiones, por lo que las solicitudes de administración no se redirigen automáticamente a regiones correctas.

Comportamiento durante una interrupción de una región de destino

En esta sección se describe qué esperar cuando hay una interrupción en una región de destino, que contiene redes virtuales administradas por la instancia de Administrador de redes virtuales, pero cuando la región principal sigue funcionando.

- **Detección y respuesta:** Microsoft detecta el error de la región y administra todas las acciones de respuesta para los servicios de plataforma de Azure afectados. La instancia de

Administrador de redes virtuales permanece disponible en su región principal.

- **Notificación:** Microsoft no le notifica automáticamente cuando una región está inactiva. Sin embargo, puede usar Azure Service Health para comprender el estado general del servicio, incluidos los errores de región, y puede configurar alertas de Service Health para notificarle problemas.

- **Implementaciones activas:** Si una implementación está en curso cuando se produce un error en la región de destino y la configuración aún no se ha aplicado a los recursos de esa región de destino, se produce un error en la implementación de esa región de destino. Debe reiniciar la implementación después de que se recupere la región de destino.

  La región de destino afectada no puede recibir nuevas implementaciones ni actualizaciones de configuración durante la interrupción. Las implementaciones en otras regiones de destino no se ven afectadas. Una vez recuperada la región de destino, Administrador de redes virtuales sincroniza la configuración más reciente con ella.

  Si la implementación ya se había aplicado a los recursos de la región afectada antes de la interrupción, esas configuraciones vuelven a estar en línea con los recursos después de restaurar la región.

- **Pérdida de datos esperada:** No se espera ninguna pérdida de datos de configuración en Administrador de redes virtuales.

- **Tiempo de inactividad esperado:** La instancia de Administrador de redes virtuales permanece disponible. Las redes virtuales y las cargas de trabajo que las usan en la región afectada no están disponibles durante la interrupción de la región.

- **Redistribución:** Administrador de redes virtuales sigue administrando redes virtuales en regiones no afectadas.

Recuperación de regiones

Cuando se recupera una región con errores, Administrador de redes virtuales restaura automáticamente las operaciones normales. Es posible que tenga que reiniciar las implementaciones que no se pudieron realizar durante la interrupción.

Prueba de fallos de región

Administrador de redes virtuales es un servicio totalmente administrado Microsoft. Dado que Microsoft administra la compatibilidad con varias regiones, no es necesario probar escenarios de

conmutación por error de región.

Copias de seguridad y restauración

Administrador de redes virtuales almacena la configuración de red. No almacena otros tipos de datos.

Para proteger la configuración, defina los recursos de Administrador de redes virtuales mediante infraestructura como código, como Bicep o Terraform, y almacene esas definiciones en el control de versiones. Si necesita volver a crear una instancia, vuelva a implementarla desde la configuración almacenada.

Resistencia al mantenimiento del servicio

Microsoft aplica periódicamente actualizaciones de servicio y realiza otro mantenimiento. La plataforma Azure controla estas actividades automáticamente, lo que garantiza que el mantenimiento sea transparente y sin problemas. No se espera ningún tiempo de inactividad durante los eventos de mantenimiento a menos que se le haya informado a través del mantenimiento planeado de Azure Service Health.

Acuerdo de nivel de servicio

El acuerdo de nivel de servicio (SLA) para Azure servicios describe la disponibilidad esperada de cada servicio y las condiciones que la solución debe cumplir para lograr esa expectativa de disponibilidad. Para obtener más información, consulte Acuerdos de Nivel de Servicio para servicios en línea.

Contenido relacionado

- Introducción a Azure Virtual Network Manager
- Preguntas más frecuentes sobre Azure Virtual Network Manager
- Implementaciones de configuración en Azure Virtual Network Manager
- Problemas comunes con Azure Virtual Network Manager
- Confiabilidad en Azure Virtual Network

ⓘ**Nota:** El autor ha creado este artículo con ayuda de inteligencia artificial. Más información

---

Last updated on 25/06/2026

Preguntas más frecuentes sobre Virtual WAN

¿Qué es Azure Virtual WAN en GA?

Sí, Azure Virtual WAN está disponible con carácter general (GA). Sin embargo, Virtual WAN consta de varias características y escenarios. Hay características o escenarios dentro de Virtual WAN en los que Microsoft aplica la etiqueta de versión preliminar. En esos casos, la característica específica o el propio escenario se encuentra en versión preliminar. Si no usa una característica específica en vista previa, se aplica la compatibilidad GA normal. Para obtener más información sobre la compatibilidad con la versión preliminar, consulte Términos de uso complementarios para las versiones preliminares de Microsoft Azure.

¿Qué ubicaciones y regiones están disponibles?

Para ver las regiones disponibles para Virtual WAN, consulte Productos disponibles por región. Especifique **Virtual WAN** como nombre del producto.

¿El usuario necesita disponer de una topología radial con los dispositivos SD-WAN/VPN para usar Azure Virtual WAN?

Virtual WAN proporciona muchas funcionalidades integradas en un solo panel de vidrio. Esto incluye:

- Conectividad VPN de sitio a sitio
- Conectividad de usuario/P2S
- Conectividad de ExpressRoute
- Conectividad de red virtual
- Interconectividad de VPN ExpressRoute
- Conectividad transitiva de red virtual a red virtual
- Enrutamiento centralizado
- Seguridad de Azure Firewall y Firewall Manager
- Monitoring
- Cifrado de ExpressRoute y
- Muchas otras funcionalidades

No es necesario tener todos estos casos de uso para empezar a utilizar Virtual WAN. Puede empezar a trabajar con uno solo de estos.

La arquitectura de Virtual WAN es una arquitectura "hub and spoke" con escalado y rendimiento integrados, donde las ramas (dispositivos VPN/SD-WAN), los usuarios (clientes de VPN de Azure, openVPN o IKEv2), los circuitos ExpressRoute y las instancias de Virtual Network sirven como radios para los centros de conectividad virtuales. Todos los centros de conectividad están conectados en una malla completa en una red WAN virtual estándar, lo que facilita al usuario el uso de la red troncal de Microsoft para la conectividad de cualquier tipo (cualquier radio). En el caso de la topología radial de los dispositivos SD-WAN/VPN, los usuarios pueden configurarla manualmente en el portal de Azure Virtual WAN o usar el CPE del asociado de Virtual WAN (SD-WAN/VPN) para configurar la conectividad con Azure.

Los asociados de Virtual WAN proporcionan automatización para la conectividad. Es la capacidad de exportar la información del dispositivo a Azure, descargar la configuración de Azure y establecer la conectividad con el centro de Azure Virtual WAN. Para la conectividad VPN de punto a sitio o usuario, se admite el cliente VPN de Azure, OpenVPN o el cliente IKEv2.

¿Se pueden deshabilitar los centros de conectividad totalmente en malla en una instancia de Virtual WAN?

Virtual WAN viene en dos tipos: Básico y Estándar. En la instancia de Virtual WAN básica, los centros de conectividad no están en malla. En una instancia de Virtual WAN estándar, los centros de conectividad están en malla y se conectan automáticamente cuando la instancia se configura por primera vez. No es necesario que el usuario haga nada específico. El usuario tampoco tiene que deshabilitar ni habilitar la funcionalidad para obtener concentradores en malla. Virtual WAN ofrece muchas opciones de enrutamiento para dirigir el tráfico entre radios cualesquiera (VNet, VPN o ExpressRoute). Proporciona la facilidad de los centros de conectividad totalmente en malla, así como la flexibilidad de enrutar el tráfico según sus necesidades.

¿Por qué veo un error sobre el ámbito y la autorización no válidos para realizar operaciones en recursos de Virtual WAN?

Si ve un error en el formato siguiente, asegúrese de que tiene los permisos siguientes configurados: Roles y permisos de Virtual WAN.

Formato del mensaje de error: "El cliente con el identificador de objeto {} no tiene autorización para realizar la acción {} en el ámbito {} o el ámbito no es válido". Para más información sobre los permisos necesarios, visite {}. Si se concedió el acceso recientemente, actualice las credenciales".

¿Cómo se administran Availability Zones y la resistencia en Virtual WAN?

Virtual WAN es una colección de centros de conectividad y servicios que están disponibles en el centro de conectividad. El usuario puede tener tantas instancias de Virtual WAN como necesite. En un centro de Virtual WAN, hay varios servicios, como VPN, ExpressRoute, etc. Cada uno de estos servicios se implementa automáticamente en Availability Zones (excepto Azure Firewall), si la región admite Availability Zones. Si una región se convierte en una zona de disponibilidad después de la implementación inicial en el centro de conectividad, el usuario puede volver a crear las puertas de enlace, lo que desencadenará una implementación de Availability Zones. Todas las puertas de enlace se aprovisionan en un centro como activo-activo, lo que implica que hay resistencia integrada dentro de un centro. Los usuarios pueden conectarse a varios centros de conectividad si desean resistencia entre regiones.

Actualmente, Azure Firewall se puede implementar para admitir zonas de disponibilidad mediante el portal de Azure Firewall Manager, PowerShell o la CLI. Actualmente no hay ninguna manera de configurar un firewall existente para que se implemente en zonas de disponibilidad. Debe eliminar y volver a implementar Azure Firewall.

Aunque el concepto de Virtual WAN es global, el recurso de Virtual WAN real se basa en Resource Manager y se implementa de forma regional. Si la propia región de virtual WAN tiene un problema, todos los centros de esa WAN virtual siguen funcionando. Sin embargo, el usuario no puede crear nuevos centros hasta que la región de virtual WAN esté disponible.

¿Cómo se eliminan o limpian mis recursos de Virtual WAN?

Cuando ya no necesite los recursos que ha creado, elimínelos. Algunos de los recursos de Virtual WAN deben eliminarse en un orden determinado debido a las dependencias. La eliminación puede tardar unos 30 minutos.

1. Abra la WAN virtual que ha creado.

2. Seleccione un centro virtual asociado a la WAN virtual para abrir la página del centro.

3. Elimine todas las entidades de puerta de enlace siguiendo el orden siguiente para cada tipo de puerta de enlace. Esto puede tardar 30 minutos en completarse.

   VPN:

   - Desconectar sitios VPN
   - Eliminación de conexiones VPN

- Eliminación de puertas de enlace de VPN

ExpressRoute:

- Eliminación de conexiones de ExpressRoute
- Eliminación de puertas de enlace de ExpressRoute

4. Repita el proceso para todos los centros asociados a la WAN virtual.

5. Puede eliminar los centros en este momento o más adelante, cuando elimine el grupo de recursos.

6. Vaya al grupo de recursos en el portal de Azure.

7. Seleccione **Eliminar grupo de recursos**. Esto elimina los demás recursos del grupo de recursos, incluidos los centros de conectividad y la WAN virtual.

¿Es posible compartir el firewall en un centro protegido con otros centros?

No, cada instancia de centro virtual de Azure debe tener su propio firewall. Se producirá un error en la implementación de rutas personalizadas para que apunten al firewall de otro centro protegido y no se completará correctamente. Considere la posibilidad de convertir esos centros en centros protegidos con sus propios firewalls.

¿Qué cliente admite la red privada virtual de usuario (de punto a sitio) de Azure Virtual WAN?

Virtual WAN admite el cliente VPN de Azure, el cliente OpenVPN o cualquier cliente IKEv2. La autenticación de Microsoft Entra se admite con el cliente VPN de Azure. Se requiere un mínimo del sistema operativo cliente de Windows 10 versión 17763.0 o superior. Los clientes de OpenVPN pueden admitir la autenticación basada en certificados. Una vez que se haya seleccionado la autenticación basada en certificados en la puerta de enlace, verá el archivo .ovpn* que se va a descargar en el dispositivo. IKEv2 admite la autenticación basada en certificados y RADIUS.

Retirada del Cliente VPN de Azure para Linux (versión preliminar)

El cliente VPN de Azure para Linux (versión preliminar) se va a retirar el 31 de agosto de 2026. Para obtener más información, consulte el artículo Cliente VPN de Azure para Linux (versión preliminar) — Información general sobre la retirada.

En el caso de la red privada virtual de usuario (de punto a sitio), ¿por qué el grupo de clientes de P2S se divide en dos rutas?

Cada puerta de enlace tiene dos instancias. La división se produce para que cada instancia de la puerta de enlace pueda asignar de forma independiente las direcciones IP de cliente para los clientes conectados, y el tráfico de la red virtual se redirija a la instancia de puerta de enlace correcta para evitar el salto de instancias entre puertas de enlace.

¿Cómo puedo agregar servidores DNS para los clientes de P2S?

Hay dos opciones para agregar servidores DNS para los clientes de P2S. Se prefiere el primer método porque agrega los servidores DNS personalizados a la puerta de enlace en lugar del cliente.

1. Use el siguiente script de PowerShell para agregar los servidores DNS personalizados. Reemplace los valores por los de su entorno.

PowerShell

```
// Define variables
$rgName = "testRG1"
$virtualHubName = "virtualHub1"
$P2SvpnGatewayName = "testP2SVpnGateway1"
$vpnClientAddressSpaces =
$vpnServerConfiguration1Name = "vpnServerConfig1"
$vpnClientAddressSpaces = New-Object string[] 2
$vpnClientAddressSpaces[0] = "192.168.2.0/24"
$vpnClientAddressSpaces[1] = "192.168.3.0/24"
$customDnsServers = New-Object string[] 2
$customDnsServers[0] = "7.7.7.7"
$customDnsServers[1] = "8.8.8.8"
$virtualHub = $virtualHub = Get-AzVirtualHub -ResourceGroupName $rgName -Name
$virtualHubName
$vpnServerConfig1 = Get-AzVpnServerConfiguration -ResourceGroupName $rgName -Name
$vpnServerConfiguration1Name

// Specify custom dns servers for P2SVpnGateway VirtualHub while creating gateway
createdP2SVpnGateway = New-AzP2sVpnGateway -ResourceGroupName $rgname -Name
$P2SvpnGatewayName -VirtualHub $virtualHub -VpnGatewayScaleUnit 1 -
VpnClientAddressPool $vpnClientAddressSpaces -VpnServerConfiguration
$vpnServerConfig1 -CustomDnsServer $customDnsServers

// Specify custom dns servers for P2SVpnGateway VirtualHub while updating
```

```
existing gateway
$P2SVpnGateway = Get-AzP2sVpnGateway -ResourceGroupName $rgName -Name
$P2SvpnGatewayName
$updatedP2SVpnGateway = Update-AzP2sVpnGateway -ResourceGroupName $rgName -Name
$P2SvpnGatewayName -CustomDnsServer $customDnsServers

// Re-generate VPN profile either from PS/Portal for VPN clients to have the
specified dns servers
```

2. O bien, si usa el cliente VPN de Azure para Windows 10, puede modificar el archivo XML del perfil descargado y agregar las **<etiquetas dnsservers><dnsserver></dnsserver></dnsservers>** antes de importarlo.

PowerShell

```
<azvpnprofile>
<clientconfig>

    <dnsservers>
        <dnsserver>x.x.x.x</dnsserver>
        <dnsserver>y.y.y.y</dnsserver>
    </dnsservers>

</clientconfig>
</azvpnprofile>
```

En el caso de la red privada virtual de usuario (de punto a sitio), ¿cuántos clientes se admiten?

En la tabla siguiente se describe el número de conexiones simultáneas y el rendimiento agregado de la puerta de enlace VPN de punto a sitio compatible con diferentes unidades de escalado.

Expandir tabla

Unidad de escala	Instancias de puerta de enlace	Conexiones simultáneas que se admiten	Rendimiento agregado
1	2	500	0,5 Gbps
2	2	500	1 Gbps
3	2	500	1,5 Gbps
4	2	1000	2 Gbps
5	2	1000	2,5 Gbps
6	2	1000	3 Gbps

Unidad de escala	Instancias de puerta de enlace	Conexiones simultáneas que se admiten	Rendimiento agregado
7	2	5000	3,5 Gbps
8	2	5000	4 Gbps
9	2	5000	4,5 Gbps
10	2	5000	5 Gbps
11	2	10000	5,5 Gbps
12	2	10000	6 Gbps
13	2	10000	6,5 Gbps
14	2	10000	7 Gbps
15	2	10000	7,5 Gbps
16	2	10000	8 Gbps
17	2	10000	8,5 Gbps
18	2	10000	9 Gbps
19	2	10000	9,5 Gbps
20	2	10000	10 Gbps
40	4	20000	20 Gbps
60	6	30000	30 Gbps
80	8	40000	40 Gbps
100	10	50000	50 Gbps
120	12	60000	60 Gbps
140	14	70000	70 Gbps
160	16	80000	80 Gbps
180	18	90000	90 Gbps
200	20	100000	100 Gbps

Como ejemplo, supongamos que el usuario elige 1 unidad de escalado. Cada unidad de escalado implica una puerta de enlace activo-activo implementada, y cada una de las instancias (2, en este caso) admitiría hasta 500 conexiones. Dado que puede obtener 500 conexiones x 2 por puerta de

enlace, no significa que tenga previstas 1000 en lugar de 500 para esta unidad de escalado. Es posible que se deban atender las instancias en que la conectividad de las 500 conexiones adicionales pueda sufrir interrupciones si supera el recuento de conexiones recomendado.

En el caso de las puertas de enlace con unidades de escalado superiores a 20, se implementan pares adicionales de instancias de puerta de enlace con gran disponibilidad para proporcionar capacidad adicional para conectar a los usuarios. Cada uno de estos pares de instancias admite hasta 10 000 usuarios adicionales. Por ejemplo, si implementa una puerta de enlace con 100 unidades de escalado, se implementan 5 pares de puertas de enlace (10 instancias en total) y se pueden conectar hasta 50 000 usuarios (10 000 x 5 pares de puertas de enlace) a la vez.

Además, asegúrese de planear el tiempo de inactividad en caso de que decida escalar o reducir verticalmente en la unidad de escalado o cambiar la configuración de punto a sitio en VPN Gateway.

Para VPN de usuario (punto a sitio), ¿se admite la aplicación registrada de Microsoft en Entra Id Authentication?

Sí, la aplicación registrada por Microsoft se admite en Virtual WAN. Puede migrar la VPN de usuario de una aplicación registrada manualmente a una aplicación registrada por Microsoft para una conectividad más segura.

¿Qué son las unidades de escalado de puerta de enlace de Virtual WAN?

Una unidad de escalado es una unidad definida para obtener un rendimiento agregado de una puerta de enlace de un centro de conectividad virtual. 1 unidad de escalado de VPN = 500 Mbps. 1 unidad de escalado de ExpressRoute = 2 Gbps. Ejemplo: 10 unidades de escalado de VPN significarían 500 Mbps * 10 = 5 Gbps.

Escale o reduzca verticalmente la unidad de escalado de la puerta de enlace según sus necesidades. Consulte Acerca de la configuración de puerta de enlace de Virtual WAN para más información sobre la configuración de la puerta de enlace.

Para obtener más información sobre los límites de soporte para las unidades de escala del gateway, consulte:

- En Virtual WAN, ¿cuáles son los rendimientos estimados de la SKU de puerta de enlace de ExpressRoute? para las puertas de enlace de ExpressRoute

- ¿Cuál es el algoritmo recomendado y los paquetes por segundo por instancia de sitio a sitio en el centro de conectividad de Virtual WAN? ¿Cuántos túneles se admiten por instancia? ¿Cuál es el rendimiento máximo admitido en un solo túnel? para puertas de enlace de VPN de sitio a sitio
- Para VPN de usuario (punto a sitio): ¿cuántos clientes se admiten? para pasarelas de VPN de usuario (punto a sitio)

¿Cuál es el impacto de modificar las unidades de escala de la puerta de enlace de red para una puerta de enlace ExpressRoute de Virtual WAN?

Al cambiar las unidades de escala del gateway, tenga en cuenta las siguientes consideraciones:

- **Capacidad de tráfico**: Asegúrese de que las unidades de escala aprovisionadas sean capaces de admitir los requisitos de tráfico.

- **Reconexiones TCP**: Puede experimentar reconexiones TCP durante el cambio de unidad de escalamiento, lo que puede provocar interrupciones menores.

- **Impacto en los puntos de conexión privados**: los cambios de unidad de escalado pueden afectar a la conectividad de los puntos de conexión privados. Revise la implementación y siga los procedimientos recomendados descritos en Uso de Private Link en Virtual WAN.

Para más información sobre las unidades de escala de puerta de enlace y la planificación de la capacidad, consulte Acerca de la configuración de puerta de enlace de Virtual WAN.

¿Cuál es la diferencia entre una puerta de enlace de Azure Virtual Network (VPN Gateway) y una instancia de VPN Gateway de Azure Virtual WAN?

Virtual WAN proporciona conectividad de sitio a sitio y a gran escala, y se ha creado para mejorar el rendimiento, la escalabilidad y la facilidad de uso. Cuando se conecta un sitio a una instancia de VPN Gateway de Virtual WAN, no es lo mismo que una puerta de enlace de red virtual normal que usa un tipo de puerta de enlace "VPN de sitio a sitio". Si quiere conectar usuarios remotos a Virtual WAN, use un tipo de puerta de enlace "VPN de punto a sitio". Las puertas de enlace de VPN de punto a sitio y de sitio a sitio son entidades independientes en el centro de Virtual WAN y deben implementarse individualmente. Del mismo modo, cuando se conecta un circuito ExpressRoute a un centro de conectividad de Virtual WAN, se usa un recurso para la puerta de

enlace de ExpressRoute que es distinto al de la puerta de enlace de red virtual normal que usa el tipo de puerta de enlace "ExpressRoute".

Virtual WAN admite un rendimiento agregado de hasta 20 Gbps para VPN y ExpressRoute. Virtual WAN también tiene automatización para la conectividad con un ecosistema de asociados de dispositivos de la rama CPE. Los dispositivos de la rama CPE tienen una automatización integrada que se aprovisiona automáticamente y se conecta a Azure Virtual WAN. Estos dispositivos están disponibles en un creciente ecosistema de asociados de VPN y SD-WAN. Consulta la lista de partners preferidos.

¿En qué difiere Virtual WAN de una puerta de enlace de Azure Virtual Network?

Una VPN de puerta de enlace de red virtual se limita a 100 túneles. En las conexiones, debe usar Virtual WAN para VPN a gran escala. Puede conectar hasta 1000 conexiones de rama por centro virtual con agregados de 20 Gbps por centro. Una conexión es un túnel de activo a activo del dispositivo VPN local al concentrador virtual. También puede tener varios centros virtuales por región. Esto significa que puede conectar más de 1000 ramas a una sola región de Azure mediante la implementación de varios centros de Virtual WAN en esa región de Azure, cada uno con su propia puerta de enlace de VPN de sitio a sitio.

¿Cuál es el algoritmo recomendado y los paquetes por segundo por instancia de sitio a sitio en el centro de conectividad de Virtual WAN? ¿Cuántos túneles se admiten por instancia? ¿Cuál es el rendimiento máximo admitido en un solo túnel?

Virtual WAN admite dos instancias de VPN Gateway activas de sitio a sitio en un centro virtual. Esto significa que hay dos instancias de VPN Gateway establecidas en modo activo-activo en un centro virtual. Durante las operaciones de mantenimiento, cada instancia se actualiza de una en una, por lo que un usuario puede experimentar una breve disminución en el rendimiento agregado de una instancia de puerta de enlace de VPN.

Aunque la VPN de Virtual WAN admite muchos algoritmos, nuestra recomendación es GCMAES256 para el cifrado IPSEC y la integridad para un rendimiento óptimo. AES256 y SHA256 se consideran menos avanzados y, por tanto, se puede esperar una degradación del rendimiento como la latencia y la caída de paquetes para tipos de algoritmo similares.

Los paquetes por segundo (o PPS) son un factor del número total de paquetes y el rendimiento admitido por instancia. Esto se explica mejor con un ejemplo. Supongamos que se implementa

en un centro de conectividad de Virtual WAN una instancia de VPN Gateway de sitio a sitio de 500 Mbps de una unidad de escalado. Suponiendo un tamaño de paquete de 1400, los PPS esperados para esa instancia de puerta de enlace VPN *en un máximo* = [(500 Mbps * 1024 * 1024) / 8 / 1400] ~ 47000.

Virtual WAN tiene conceptos de conexión VPN, conexión de enlace y túneles. Una única conexión VPN consta de conexiones de vínculo. Virtual WAN admite hasta cuatro conexiones de vínculo en una conexión VPN. Cada conexión de vínculo consta de dos túneles IPsec que finalizan en dos instancias de una puerta de enlace VPN activo-activo implementada en un centro virtual. El número total de túneles que pueden finalizar en una sola instancia activa es de 1000. Esto también implica que el rendimiento de una instancia está disponible agregado en todos los túneles que se conectan a esa instancia. Cada túnel tiene también ciertos valores de rendimiento. En los casos de varios túneles conectados a una puerta de enlace de unidad de escalado de menor valor, es mejor evaluar la necesidad por túnel y planear una puerta de enlace de VPN que sea un valor agregado para el rendimiento en todos los túneles que terminan en la instancia de VPN.

Valores de varias unidades de escalado admitidas en Virtual WAN

Expandir tabla

Unidad de escalado	Rendimiento máximo por túnel (Mbps)	Número máximo de PPS por túnel	Rendimiento máximo por instancia (Mbps)	Número máximo de PPS por instancia
1	500	47K	500	47K
2	1000	94K	1000	94K
3	1500	140K	1500	140K
4	1500	140K	2000	187K
5	1500	140K	2500	234K
6	1500	140K	3000	281K
7	2300	215K	3500	328K
8	2300	215K	4000	374K
9	2300	215K	4500	421K
10	2300	215K	5000	468K
11	2300	215K	5500	515K

Unidad de escalado	Rendimiento máximo por túnel (Mbps)	Número máximo de PPS por túnel	Rendimiento máximo por instancia (Mbps)	Número máximo de PPS por instancia
12	2300	215K	6000	562K
13	2300	215K	6500	609K
14	2300	215K	7000	655K
15	2300	215K	7500	702K
16	2300	215K	8000	749K
17	2300	215K	8500	796K
18	2300	215K	9000	843K
19	2300	215K	9500	889K
20	2300	215K	10000	936K

> ⓘ **Nota:**
>
> Todos los números se basan en el algoritmo de GCM.

¿Qué proveedores de dispositivos (asociados de Virtual WAN) se admiten?

En este momento, muchos asociados admiten la experiencia de Virtual WAN totalmente automatizada. Para más información, consulte Asociados de Virtual WAN.

¿Cuáles son los pasos de automatización de asociados de Virtual WAN?

Para conocer los pasos de automatización de asociados, consulte Automatización de asociados de Virtual WAN.

¿Debo usar un dispositivo de asociado preferido?

No. Puede usar cualquier dispositivo compatible con VPN que cumpla los requisitos de Azure para la compatibilidad con IPsec de IKEv1 o IKEv2. Virtual WAN también tiene soluciones de asociado de CPE que automatizan la conectividad con Azure Virtual WAN, lo que facilita la configuración de conexiones VPN de IPsec a escala.

¿Cómo automatizan la conectividad con Azure Virtual WAN los asociados de Virtual WAN?

Las soluciones de conectividad definidas por software suelen administrar sus dispositivos de rama con un controlador o un centro de aprovisionamiento de dispositivos. El controlador puede usar las API de Azure para automatizar la conectividad a Azure Virtual WAN. La automatización incluye la carga de información de la rama, la descarga de la configuración de Azure, la configuración de túneles IPsec en los centros de conectividad virtuales de Azure y la configuración automática de la conectividad desde el dispositivo de rama a Azure Virtual WAN. Cuando tiene cientos de ramas, la conexión mediante asociados de CPE de Virtual WAN es fácil ya que la experiencia de incorporación elimina la necesidad de configurar y administrar la conectividad de IPsec a gran escala. Para más información, consulte Automatización de asociados de Virtual WAN.

¿Qué ocurre si el dispositivo que uso no está en la lista de asociados de Virtual WAN? ¿Lo puedo usar a pesar de todo para conectarme a una VPN de Azure Virtual WAN?

Sí, siempre que el dispositivo admita IPsec IKEv1 o IKEv2. Los asociados de Virtual WAN automatizan la conectividad desde el dispositivo hasta los puntos de conexión de VPN de Azure. Esto supone automatizar pasos como "carga de información de la rama", "IPsec y configuración" y "conectividad". Dado que el dispositivo no procede de un ecosistema de asociados de Virtual WAN, debe realizar el trabajo pesado de tomar manualmente la configuración de Azure y actualizar el dispositivo para configurar la conectividad de IPsec.

¿Cómo consiguen los nuevos asociados que no aparecen en la lista de asociados de partida ser incluidos en ella?

Todas las API de WAN virtual son OpenAPI. Puede consultar la documentación automatización de asociados de Virtual WAN para evaluar la viabilidad técnica. Un asociado ideal es aquel que tiene un dispositivo que se puede aprovisionar para la conectividad IPsec de IKEv1 o IKEv2. Una vez que la empresa complete el trabajo de automatización para su dispositivo CPE en función de las directrices de automatización compartidas anteriores proporcionadas, puede ponerse en contacto con azurevirtualwan@microsoft.com para que se muestre aquí Conectividad a través de asociados. Si es un cliente que desea que una determinada solución de la empresa aparezca como un asociado de Virtual WAN, haga que la empresa se ponga en contacto con Virtual WAN mediante el envío de un correo electrónico a azurevirtualwan@microsoft.com.

¿Cómo es que Virtual WAN admite dispositivos SD-WAN?

Los asociados de Virtual WAN automatizan la conectividad IPsec a los puntos de conexión de VPN de Azure. Si el asociado de Virtual WAN es un proveedor de SD-WAN, se supone que el controlador de SD-WAN administra la automatización y la conectividad IPsec a los puntos de conexión de VPN de Azure. Si el dispositivo SD-WAN necesita su propio punto de conexión en lugar de la VPN de Azure para obtener funcionalidad de SD-WAN propietaria, puede implementar el punto de conexión de SD-WAN en una red virtual de Azure y coexistir con Azure Virtual WAN.

Virtual WAN admite el emparejamiento BGP y también tiene la capacidad de implementar NVAs en un hub de la Virtual WAN.

¿Cuántos dispositivos VPN pueden conectarse a un único centro?

Se admiten hasta 1000 conexiones por concentrador virtual. Cada conexión consta de cuatro vínculos y cada conexión de vínculo admite dos túneles que están en una configuración activo-activo. Los túneles terminan en un elemento VPN Gateway del centro de conectividad virtual de Azure. Los vínculos representan el vínculo del ISP físico en la sucursal o el dispositivo VPN.

¿Qué es una conexión de rama a Azure Virtual WAN?

Una conexión desde una rama o un dispositivo VPN a Azure Virtual WAN es una conexión VPN que se conecta de forma virtual al sitio VPN y a Azure VPN Gateway en un centro de conectividad virtual.

¿Qué ocurre si el dispositivo VPN local solo tiene un túnel a una puerta de enlace de VPN de Azure Virtual WAN?

Una conexión de Azure Virtual WAN se compone de dos túneles. Una puerta de enlace de VPN de Virtual WAN se implementa en un centro de conectividad virtual en modo activo-activo, lo que significa que hay distintos túneles desde dispositivos locales que terminan en distintas instancias. Esta es la recomendación para todos los usuarios. Sin embargo, si el usuario decide tener solo 1 túnel a una de las instancias de puerta de enlace de VPN de Virtual WAN, si por algún motivo (mantenimiento, revisiones, etc.), la instancia de puerta de enlace se desconecta, el túnel se mueve a la instancia activa secundaria y el usuario podría experimentar una reconexión. Las sesiones de BGP no se mueven entre instancias.

¿Qué ocurre durante un restablecimiento de puerta de enlace en una instancia de VPN Gateway de Virtual WAN?

El botón Restablecimiento de puerta de enlace debe usarse si todos los dispositivos locales funcionan según lo previsto, pero la conexión VPN de sitio a sitio en Azure está en estado desconectado. VPN Gateway de Virtual WAN siempre se implementa en un estado Activo-Activo de alta disponibilidad. Esto significa que siempre hay más de una instancia implementada en una instancia de VPN Gateway en cualquier momento. Cuando se usa el botón Restablecimiento de puerta de enlace, reinicia las instancias de VPN Gateway de forma secuencial, por lo que las conexiones no se interrumpen. Habrá un breve intervalo a medida que las conexiones se muevan de una instancia a otra, pero este intervalo debe ser inferior a un minuto. Además, restablecer las puertas de enlace no cambia las direcciones IP públicas.

Este escenario solo se aplica a las conexiones S2S.

¿Se puede conectar el dispositivo VPN local a varios centros de conectividad?

Yes. Al principio, el flujo de tráfico sería del dispositivo local al perímetro de red de Microsoft más cercano y, luego, al centro de conectividad virtual.

¿Hay disponibles nuevos recursos de Resource Manager para Virtual WAN?

Sí, Virtual WAN presenta nuevos recursos de Resource Manager. Para obtener más información, consulte La información general.

¿Puedo implementar y usar mi aplicación virtual de red favorita (en una red virtual de NVA) con Azure Virtual WAN?

Sí, puede conectar la red virtual de su aplicación virtual de red (NVA) favorita a Azure Virtual WAN.

¿Puedo crear una aplicación virtual de red dentro del centro de conectividad virtual?

Las aplicaciones virtuales de red (NVA) se pueden implementar en un centro de conectividad virtual. Para conocer los pasos necesarios para hacerlo, consulte Acerca de las aplicaciones virtuales de red en un centro de Virtual WAN.

¿Puedo conectar una red virtual a un centro virtual en un inquilino diferente?

Yes. Siga la guía en Conectar redes virtuales entre inquilinos a un hub de Virtual WAN.

¿Puede una red virtual de radio tener una puerta de enlace de red virtual?

No. La red virtual de radio no puede tener una puerta de enlace de red virtual si está conectada al centro de conectividad virtual.

¿Puede una red virtual radial tener un servidor de Azure Route Server?

No. La red virtual radial no puede tener un servidor de rutas si está conectado al centro de conectividad de virtual WAN.

¿Existe compatibilidad con BGP en la conectividad VPN?

Sí, BGP se admite. Al crear una VPN de sitio, puede proporcionar en ella los parámetros de BGP. Esto implica que todas las conexiones creadas en Azure para ese sitio están habilitadas para BGP.

¿Hay información de precios o licencias para Virtual WAN?

Yes. Consulte la página Precios .

¿Es posible construir Azure Virtual WAN con una plantilla de Resource Manager?

Se puede crear una configuración sencilla de una instancia de Virtual WAN con un centro de conectividad y un sitio vpn mediante una plantilla de inicio rápido . Virtual WAN es principalmente un servicio REST o basado en el portal.

¿Pueden las redes virtuales radiales conectadas a un centro virtual comunicarse entre sí (tránsito V2V)?

Yes. Virtual WAN estándar admite la conectividad transitiva entre redes virtuales mediante el centro de conectividad de Virtual WAN al que están conectadas las redes virtuales. En la terminología de Virtual WAN, nos referimos a estas rutas de acceso como "tránsito de red virtual

de Virtual WAN local" para las redes virtuales conectadas a un centro de Virtual WAN dentro de una sola región. El "Tránsito global de red virtual de Virtual WAN" es para redes virtuales conectadas a través de varios centros de Virtual WAN en dos o más regiones.

En algunos escenarios, las redes virtuales de radio también pueden conectarse directamente entre sí utilizando el emparejamiento de redes virtuales, además del tránsito de redes virtuales de Azure Virtual WAN locales o globales. En este caso, el emparejamiento de red virtual tiene prioridad sobre la conexión transitiva del centro de conectividad de Virtual WAN.

¿Se permite la conectividad de rama a rama en Virtual WAN?

Sí, este tipo de conectividad está disponible en Virtual WAN. La rama se aplica conceptualmente a sitios VPN, circuitos ExpressRoute o usuarios de VPN de punto a sitio o de usuario. La habilitación de la rama a rama está habilitada de forma predeterminada y se puede encontrar **en la configuración** de WAN. De esta forma, los usuarios o ramas VPN se pueden conectar a otras ramas VPN y la conectividad de tránsito también se puede habilitar entre los usuarios de VPN y ExpressRoute.

¿Atraviesa el tráfico de rama a rama Azure Virtual WAN?

Yes. El tráfico de rama a rama atraviesa Azure Virtual WAN.

¿La instancia de Virtual WAN requiere ExpressRoute desde cada sitio?

No. Virtual WAN no requiere ExpressRoute desde cada sitio. Los sitios pueden estar conectados a una red de proveedor mediante un circuito de ExpressRoute. En el caso de los sitios conectados mediante ExpressRoute a un centro de conectividad virtual, así como a la VPN de IPsec en el mismo centro de conectividad, el centro de conectividad virtual proporciona conectividad de tránsito entre el usuario de VPN y el de ExpressRoute.

¿Hay un límite en el rendimiento o la conexión de la red al utilizar Azure Virtual WAN?

El rendimiento de red es por servicio en un centro de conectividad de Virtual WAN. En cada centro de conectividad, el rendimiento agregado de VPN es de hasta 20 Gbps, el rendimiento agregado de ExpressRoute es de hasta 20 Gbps y el rendimiento agregado de VPN de punto a sitio o VPN de usuario es de hasta 200 Gbps. El enrutador del centro de conectividad virtual admite hasta 50 Gbps con los flujos de tráfico de red virtual a red virtual y supone una carga de

trabajo total de 2000 máquinas virtuales en todas las redes virtuales conectadas a un solo centro de conectividad virtual.

Para proteger la capacidad inicial sin esperar a que el centro virtual se escale horizontalmente cuando se necesite más rendimiento, puede establecer la capacidad mínima o modificar según sea necesario. Consulte Acerca de la configuración del centro virtual: capacidad del concentrador. Para obtener información sobre el costo, consulte el *costo de la unidad de infraestructura de enrutamiento* en la página Precios de Azure Virtual WAN.

Cuando los sitios VPN se conectan a un centro de conectividad, lo hacen con conexiones. Virtual WAN admite hasta 1000 conexiones o 2000 túneles IPsec por centro virtual. Cuando los usuarios remotos se conectan a un centro de conectividad virtual, se conectan a la instancia de VPN Gateway P2S, que admite hasta 100 000 usuarios, en función de la unidad de escalado (ancho de banda) elegida para la instancia de VPN Gateway P2S en el centro de conectividad virtual.

¿Puedo usar NAT-T en las conexiones VPN?

Sí, se admite NAT traversal (NAT-T). VPN Gateway de Virtual WAN NO llevará a cabo ninguna funcionalidad similar a la de NAT en los paquetes internos hacia y desde los túneles de IPsec. En esta configuración, asegúrese de que el dispositivo local inicie el túnel IPsec.

¿Cómo se puede configurar una unidad de escalado en un valor concreto, como por ejemplo 20 Gbps?

Vaya a la puerta de enlace de VPN dentro de un nodo central en el portal y, a continuación, seleccione la unidad de escalamiento para ajustarla a la configuración adecuada.

¿Azure Virtual WAN permite al dispositivo local usar varias ISP en paralelo o es siempre un solo túnel VPN?

Las soluciones en dispositivos locales pueden aplicar directivas de tráfico para dirigir el tráfico entre varios túneles en el centro de conectividad de Azure Virtual WAN (VPN Gateway en el centro de conectividad virtual).

¿Qué es la arquitectura de tránsito global?

Para más información, consulte Arquitectura de red de tránsito global y Virtual WAN.

¿Cómo se enruta el tráfico en la red troncal de Azure?

El tráfico sigue el patrón: dispositivo de rama ->ISP->perímetro de red de Microsoft->controlador de dominio de Microsoft (red virtual de centro)->perímetro de red de Microsoft->ISP->dispositivo de rama.

En este modelo, ¿qué es necesario en cada sitio? ¿Solo una conexión a Internet?

Yes. Una conexión a Internet y un dispositivo físico que admita IPsec, preferiblemente de nuestros asociados integrados de Virtual WAN. De forma opcional, puede administrar manualmente la configuración y la conectividad a Azure desde su dispositivo preferido.

¿Cómo habilito la ruta predeterminada (0.0.0.0/0) de una conexión (VPN, ExpressRoute o de red virtual)?

Un centro de conectividad virtual puede propagar una ruta predeterminada aprendida en una red virtual, una VPN de sitio a sitio y una conexión ExpressRoute si la marca está "habilitada" en la conexión. Esta marca está visible cuando el usuario edita una conexión de red virtual, una conexión VPN o una conexión ExpressRoute. De forma predeterminada, esta marca está deshabilitada cuando un sitio o circuito de ExpressRoute se conecta a un centro. Está habilitada de forma predeterminada cuando se agrega una conexión de red virtual para conectar una red virtual a un centro de conectividad virtual.

La ruta predeterminada no se origina en el centro de conectividad de Virtual WAN. La ruta predeterminada se propaga si el hub de Virtual WAN ya lo ha aprendido como resultado de la implementación de un firewall en el hub, o si otro sitio conectado tiene el túnel forzado habilitado.

Las rutas predeterminadas no se propagan entre los centros (entre centros).

¿Es posible crear varios centros de conectividad de Virtual WAN en la misma región?

Yes. Los clientes ahora pueden crear más de un centro de conectividad en la misma región para la misma instancia de Azure Virtual WAN.

¿Cómo selecciona el centro de conectividad virtual de una instancia de Virtual WAN la mejor ruta desde varios centros?

Para obtener información, consulte la página preferencias de enrutamiento del centro de conectividad virtual .

¿El centro de conectividad de Virtual WAN permite la conectividad entre circuitos ExpressRoute?

El transporte público entre urgencias y urgencias está disponible a través de Global Reach. Las puertas de enlace de centro de conectividad virtual se implementan en las regiones de Azure o del controlador de dominio. Cuando dos circuitos ExpressRoute se conectan a través de Global Reach, no es necesario que el tráfico llegue desde los enrutadores perimetrales hasta el controlador de dominio del centro de conectividad.

Routing Intent también se puede usar con directivas de enrutamiento de tráfico privado para habilitar la conectividad de tránsito de ExpressRoute a través de una aplicación de seguridad implementada en el centro virtual.

Para más información, consulte Acerca de Global Reach de Virtual WAN.

¿Existe un concepto de peso en los circuitos ExpressRoute de Azure Virtual WAN o en las conexiones VPN?

Cuando varios circuitos ExpressRoute están conectados a un centro de conectividad virtual, el peso del enrutamiento en la conexión proporciona un mecanismo para que ExpressRoute en el centro de conectividad virtual prefiera un circuito antes que el otro. No hay ningún mecanismo para establecer un peso en una conexión VPN. Azure siempre prefiere una conexión ExpressRoute antes que una conexión VPN en un solo centro de conectividad.

¿Virtual WAN prefiere ExpressRoute antes que una VPN para el tráfico que sale de Azure?

Yes. ¿Virtual WAN prefiere ExpressRoute antes que una VPN para el tráfico que sale de Azure? Sin embargo, se puede configurar la preferencia de enrutamiento del centro virtual para cambiar la preferencia predeterminada. Para conocer los pasos, consulte Configuración de la preferencia de enrutamiento del centro virtual.

Cuando un centro de conectividad de Virtual WAN tiene un circuito ExpressRoute y un sitio VPN conectados a él, ¿qué haría que una ruta de conexión VPN se prefiriera por encima de ExpressRoute?

Cuando un circuito ExpressRoute se conecta a un centro de conectividad virtual, los enrutadores de Microsoft Edge son el primer nodo para la comunicación entre el entorno local y Azure. Estos enrutadores perimetrales se comunican con las puertas de enlace de ExpressRoute de Virtual WAN que, a su vez, aprenden las rutas del enrutador del centro de conectividad virtual que controla todas las rutas entre todas las puertas de enlace de Virtual WAN. Los enrutadores de Microsoft Edge procesan las rutas de ExpressRoute del centro de conectividad virtual con una preferencia más alta que las rutas aprendidas del entorno local.

Por cualquier motivo, si la conexión VPN se convierte en el medio principal del centro de conectividad virtual para aprender rutas de (por ejemplo, escenarios de conmutación por error entre ExpressRoute y VPN), a menos que el sitio VPN tenga una longitud de ruta de acceso AS más larga, el centro virtual seguirá compartiendo las rutas aprendidas de VPN con la puerta de enlace de ExpressRoute. Esto hace que los enrutadores de Microsoft Edge prefieran las rutas VPN antes que las rutas locales.

¿ExpressRoute admite el Enrutamiento multidireccional de igual costo (ECMP) en Virtual WAN?

Cuando 1 o más circuitos de ExpressRoute están conectados a un centro de Virtual WAN, ECMP permite que el tráfico de las redes virtuales radiales a un entorno local a través de ExpressRoute se distribuya a través de 2 circuitos de ExpressRoute (correspondientes a hasta 4 vínculos de ExpressRoute) que anuncian las mismas rutas locales. ECMP no está habilitado actualmente de forma predeterminada para centros de Virtual WAN. Para habilitar ECMP para su entorno, puede crear un mapa de rutas para el centro virtual. Al crear un mapa de rutas, el centro virtual se actualiza automáticamente a la versión de software más reciente que admite ECMP, independientemente de si este mapa de rutas se aplica en cualquier conexión. Como resultado, solo tiene que seguir los pasos del 1 al 7 en Configuración de mapas de rutas. Si no tiene previsto usar el mapa de rutas, puede eliminar el mapa de rutas una vez completado el paso 7, ya que los centros con un mapa de rutas incurrirán en costos adicionales. Se recomienda crear un mapa de rutas en el entorno de prueba y validar el enrutamiento y la conectividad antes de crear un mapa de rutas en el entorno de producción.

Cuando dos centros de conectividad (1 y 2) están conectados y hay un circuito ExpressRoute conectado como un lazo para ambos centros de conectividad, ¿cuál es la ruta de acceso de una red virtual conectada al centro de conectividad

1 para llegar a una red virtual conectada en el centro de conectividad 2?

El comportamiento actual es preferir la ruta de acceso del circuito ExpressRoute frente a una conexión entre centros para la conectividad de red virtual a red virtual. Sin embargo, esto no se recomienda en una configuración de Virtual WAN. Para resolver esto, puede realizar una de estas dos acciones:

- Configure varios circuitos ExpressRoute (proveedores distintos) para que se conecten a un centro de conectividad y usen la conectividad entre centros de conectividad que proporciona Virtual WAN para los flujos de tráfico entre regiones.
- Configure la ruta de acceso AS como preferencia de enrutamiento del centro virtual. Esto garantiza que el tráfico entre los dos centros atraviese el enrutador de centro virtual en cada centro y use la ruta de acceso de centro a centro en lugar de la ruta de acceso de ExpressRoute (que atraviesa los enrutadores de Microsoft Edge). Para obtener más información, consulte Configuración de la preferencia de enrutamiento del centro de conectividad virtual.

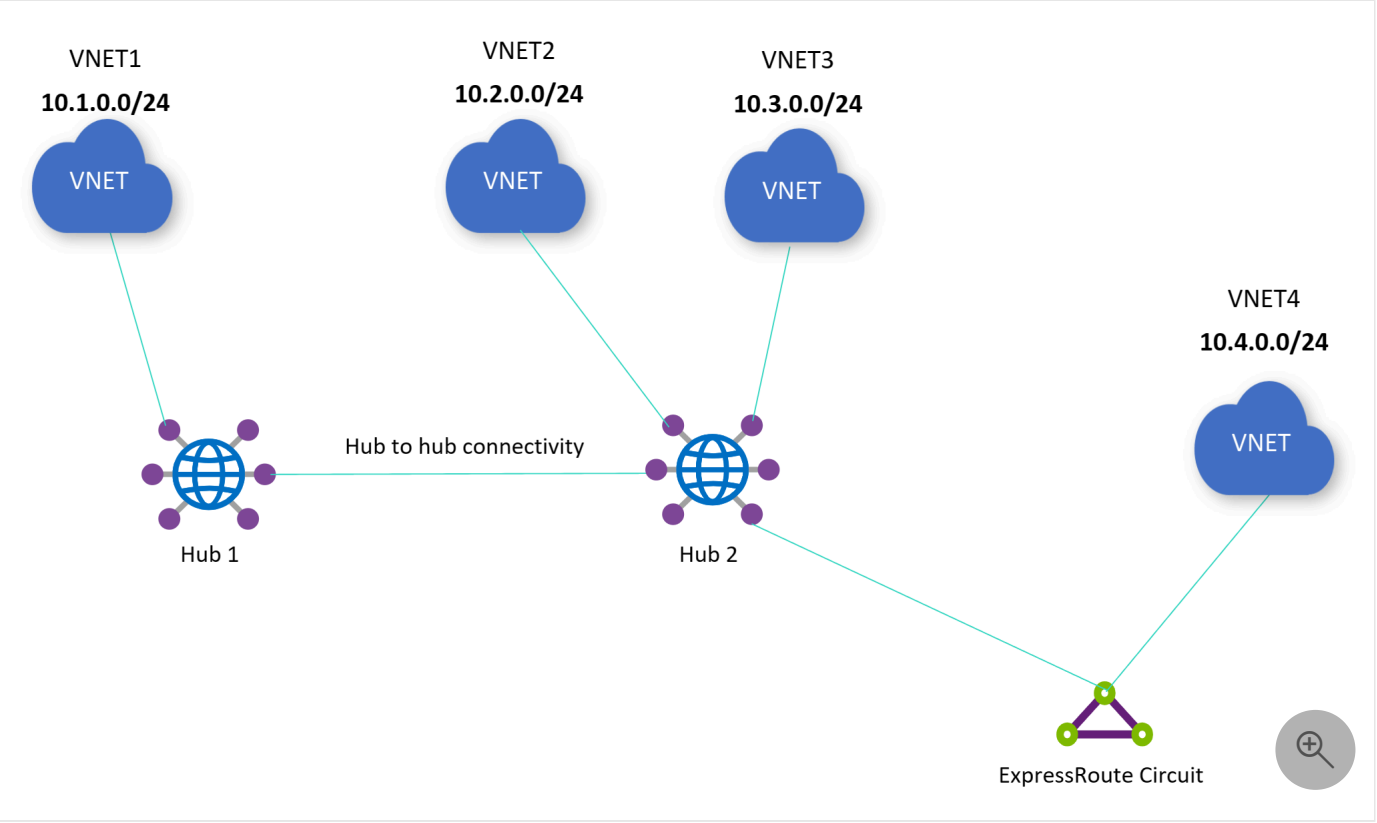
Cuando hay un circuito ExpressRoute conectado como lazo a un centro de Virtual WAN y a una red virtual independiente, ¿cuál es la ruta de acceso para que la red virtual autónoma llegue al centro de conectividad de Virtual WAN?

En el caso de las nuevas implementaciones, esta conectividad está bloqueada de forma predeterminada. Para permitir esta conectividad, puede habilitar estos conmutadores de puerta de enlace de ExpressRoute en el panel "Editar centro virtual" y en el panel "Puerta de enlace de red virtual" del Portal. Sin embargo, se recomienda mantener deshabilitadas estas alternancias y, en su lugar, crear una conexión de Virtual Network para conectar directamente redes virtuales independientes a un centro de Virtual WAN. Después, el tráfico de red virtual a red virtual atravesará el enrutador del concentrador de Virtual WAN, que ofrece un mejor rendimiento que la ruta de acceso de ExpressRoute. La ruta de acceso de ExpressRoute incluye:

- La puerta de enlace de ExpressRoute, que tiene límites de ancho de banda inferiores a los del enrutador del concentrador
- y los enrutadores de Microsoft Enterprise Edge/MSEE, que es un salto adicional en la ruta de acceso de datos.

En el diagrama siguiente, ambas alternancias deben habilitarse para permitir la conectividad entre la red virtual independiente 4 y las redes virtuales conectadas directamente al centro 2 (VNet 2 y VNet 3): **permitir el tráfico desde redes de Virtual WAN remotas** para la puerta de enlace de red virtual y **Permitir el tráfico desde redes que no son de Virtual WAN** para la puerta de enlace de ExpressRoute del centro virtual. Si se implementa Azure Route Server en la red virtual independiente 4 y el Route Server tiene habilitada la conexión de sitio a sitio, la conectividad se bloqueará entre la red virtual 1 y la red virtual independiente 4.

Habilitar o deshabilitar el botón de alternancia solo afectará al siguiente flujo de tráfico: el tráfico que fluye entre el centro de Virtual WAN y las redes virtuales independientes por el circuito ExpressRoute. Habilitar o deshabilitar el botón de alternancia **no** incurrirá en tiempo de inactividad para los demás flujos de tráfico (por ejemplo: el sitio local a la red virtual de radio 2 no se verá afectado, la red virtual 2 a la red virtual 3 no se verá afectada, etc.).



Cuando actualizo una conexión de ExpressRoute, ¿esto afectará a la conectividad de mis otras conexiones de ExpressRoute?

Al realizar una operación de creación, actualización o eliminación en una conexión de ExpressRoute, las demás conexiones de ExpressRoute aparecerán en un estado de "actualización" en el portal. Sin embargo, la conectividad a través de estas conexiones expressRoute no se verá afectada.

¿Por qué la conectividad no funciona cuando anuncio rutas con un ASN de 0 en la ruta de acceso del sistema autónomo (AS)?

El centro de conectividad de Virtual WAN quita las rutas con un ASN de 0 en la ruta de acceso del sistema autónomo (AS). Para asegurarse de que estas rutas se anuncian correctamente en Azure, la ruta de acceso del sistema autónomo (AS) no debe incluir 0.

¿Se pueden crear concentradores en grupos de recursos diferente en Virtual WAN?

Yes. Actualmente, esta opción solo está disponible a través de PowerShell. En el portal de Virtual WAN es necesario que los centros de conectividad estén en el mismo grupo de recursos que el propio recurso de Virtual WAN.

¿Cuál es el espacio de direcciones de centro de conectividad recomendado durante la creación del centro?

El espacio de direcciones del hub de Virtual WAN **no se puede modificar después de haber sido creado el hub**. Use la siguiente información para seleccionar el tamaño de dirección del centro adecuado para la implementación:

- Para dar cabida a las necesidades futuras de escalabilidad y arquitectura, mientras que el espacio de direcciones mínimo para un centro de Virtual WAN es **/24**, se recomienda especificar un **espacio de direcciones /23** o mayor durante la creación del centro.
- Si usa una instancia de Azure Firewall en Virtual WAN, se requiere un espacio mínimo de direcciones del centro **/22** para asegurarse de que Azure Firewall pueda asignar suficientes direcciones IP para escalar al máximo rendimiento.
- Si usa aplicaciones virtuales de red en el centro de Virtual WAN, el tamaño del centro de Virtual WAN determina el número de direcciones IP utilizables asignadas a las aplicaciones virtuales de red. Consulte la documentación sobre NVA para conocer la correspondencia entre el espacio de direcciones del concentrador y las direcciones IP asignables a las NVA.

¿Hay compatibilidad con IPv6 en un Virtual WAN?

IPv6 no se admite en el centro de conectividad de Virtual WAN y sus puertas de enlace. Si conecta una red virtual radial con un intervalo de direcciones IPv6 al centro de conectividad de Virtual WAN, solo funcionará la conectividad IPv4 con esta red virtual radial. No es compatible la conectividad IPv6 con esta red virtual spoke.

Si anuncia prefijos IPv6 desde el entorno local, esto interrumpe la conectividad IPv4 para los recursos de Azure.

En el escenario de VPN de usuario de punto a sitio con salida de Internet a través de Azure Firewall, probablemente tendrá que desactivar la conectividad IPv6 en el dispositivo cliente para forzar el tráfico al centro de conectividad de Virtual WAN. Esto se debe a que, de forma predeterminada, los dispositivos modernos usan direcciones IPv6.

¿Cuál es la versión de API recomendada para utilizar en los scripts que automatizan varias funcionalidades de Virtual WAN?

Se requiere una versión mínima de 05-01-2022 (1 de mayo de 2022).

¿Hay algún límite de Virtual WAN?

Consulte la sección Límites de Virtual WAN en la página Límites de suscripción y servicio.

¿Cuáles son las diferencias entre los tipos de instancias de Virtual WAN (básico y estándar)?

Consulte WAN virtuales básicas y estándar. Para obtener los precios, consulte la página Precios .

¿Virtual WAN almacena datos de clientes?

No. Virtual WAN no almacena datos de clientes.

¿Hay proveedores de servicios administrados que pueden administrar Virtual WAN para usuarios como servicio?

Yes. Para obtener una lista de las soluciones del proveedor de servicios administrados (MSP) habilitadas a través de Azure Marketplace, consulte Ofertas de Azure Marketplace por asociados de MSP de redes de Azure.

¿En qué se diferencia el enrutamiento del centro de conectividad de Virtual WAN de Azure Route Server en una red virtual?

Tanto el centro de conectividad de Azure Virtual WAN como Azure Route Server proporcionan funcionalidades de emparejamiento de Protocolo de puerta de enlace de borde (BGP) que pueden usar las aplicaciones virtuales de red (NVA) para anunciar direcciones IP desde la NVA a las redes virtuales de Azure del usuario. Las opciones de implementación difieren en el sentido de que Azure Route Server suele ser implementado por una red virtual de concentrador autoadministrada por el cliente, mientras que Azure Virtual WAN ofrece un servicio de concentrador completamente mallado y sin intervención, al cual los clientes conectan sus diversos puntos terminales tipo spoke (redes virtuales de Azure, sucursales locales mediante VPN sitio a sitio o SD-WAN, usuarios remotos mediante VPN punto a sitio/usuario remoto y conexiones privadas con ExpressRoute). Además, permite el emparejamiento BGP para NVAs implementadas en redes virtuales tipo spoke, junto con otras funcionalidades de Virtual WAN como conectividad de tránsito entre redes virtuales, conectividad de tránsito entre VPN y ExpressRoute, enrutamiento personalizado/avanzado, asociación y propagación de rutas personalizadas, intención/directivas de enrutamiento para una seguridad entre regiones sin complicaciones, concentrador seguro, firewall de Azure, entre otras. Para más detalles sobre el emparejamiento BGP de la WAN virtual, consulte Cómo emparejar BGP con un concentrador virtual.

Si uso un proveedor de seguridad de terceros (Zscaler, iBoss o Checkpoint) para proteger el tráfico de Internet, ¿por qué no veo el sitio de VPN asociado a él en Azure Portal?

Cuando decide implementar un proveedor de seguridad asociado para proteger el acceso a Internet para sus usuarios, el proveedor de seguridad de terceros crea un sitio VPN en su nombre. Este sitio de VPN no aparece en Azure Portal porque el proveedor crea automáticamente el proveedor de seguridad de terceros y no es un sitio VPN creado por el usuario.

Para obtener más información sobre las opciones disponibles de proveedores de seguridad de terceros y cómo configurarlo, consulte Despliegue de un proveedor de socios de seguridad.

¿Se conservarán las comunidades BGP generadas por el entorno local en Virtual WAN?

Sí, las comunidades BGP generadas por el entorno local se conservarán en Virtual WAN.

¿Se conservarán las comunidades de BGP generadas por pares de BGP (en una red virtual adjunta) en Virtual WAN?

Sí, las comunidades de BGP generadas por pares de BGP se conservarán en Virtual WAN. Las comunidades se conservan en el mismo centro y en las conexiones entre centros. Esto también se aplica a escenarios de Virtual WAN mediante directivas de intención de enrutamiento.

¿Qué números de ASN se admiten para redes locales adjuntas de forma remota que ejecuten el Protocolo de puerta de enlace de borde?

Se pueden usar los ASN públicos propios o los ASN privados para las redes locales. No se pueden usar los rangos reservados por Azure o IANA.

- ASN reservados por Azure:
  - ASN públicos: 8074, 8075, 12076
  - ASN privados: 65515, 65517, 65518, 65519, 65520
  - ASN reservados por IANA: 23456, 64496-64511, 65535-65551

¿Hay alguna manera de cambiar los ASN en Virtual WAN?

No. Virtual WAN no admite cambios de ASN para centros virtuales ni puertas de enlace.

En Virtual WAN, ¿cuáles son los rendimientos estimados de la SKU de puerta de enlace de ExpressRoute?

Expandir tabla

Unidad de escalado	Conexiones por segundo	Mega bits por segundo	Paquetes por segundo	Total de flujos
1	14,000	2,000	200,000	200,000
2	28,000	4,000	400,000	400,000
3	42,000	6,000	600,000	600,000
4	56,000	8,000	800,000	800,000
5	70,000	10,000	1,000,000	1,000,000
6	84,000	12,000	1,200,000	1,200,000
7	98,000	14,000	1,400,000	1,400,000
8	112,000	16,000	1,600,000	1,600,000
9	126,000	18,000	1,800,000	1,800,000

Unidad de escalado	Conexiones por segundo	Mega bits por segundo	Paquetes por segundo	Total de flujos
10	140,000	20,000	2,000,000	2,000,000

Es importante tener en cuenta los siguientes puntos:

- Durante las operaciones de mantenimiento, las unidades de escalado de 2 a 10 mantienen su rendimiento agregado. Sin embargo, la unidad de escalado 1 puede experimentar pequeñas variaciones en el rendimiento durante estas operaciones.
- Independientemente del número de unidades de escalado implementadas, si un único flujo TCP supera los 1,5 Gbps, el rendimiento del tráfico puede degradarse.

Si conecto un circuito local de ExpressRoute a un centro de conectividad de Virtual WAN, ¿podré acceder únicamente a las regiones en la misma área metropolitana que el circuito local?

Los circuitos locales solo se pueden conectar a puertas de enlace de ExpressRoute en su región de Azure correspondiente. Sin embargo, no hay ninguna limitación para enrutar el tráfico a redes virtuales radiales de otras regiones.

¿Por qué el enrutador del centro virtual requiere una dirección IP pública con puertos abiertos?

Estos puntos de conexión públicos son necesarios para que la plataforma de administración y SDN subyacente de Azure se comuniquen con el enrutador del centro virtual. Dado que el enrutador del centro virtual se considera parte de la red privada del cliente, la plataforma subyacente de Azure no puede acceder directamente al enrutador del centro ni administrarlo a través de sus puntos de conexión privados debido a los requisitos de cumplimiento. La conectividad a los puntos de conexión públicos del enrutador del centro se autentica a través de certificados y Azure realiza auditorías de seguridad rutinarias de estos puntos de conexión públicos. Como resultado, no constituyen una exposición de seguridad del centro virtual.

¿Hay un límite de ruta para los clientes de OpenVPN que se conectan a una puerta de enlace de VPN P2S de Azure?

El límite de ruta para los clientes de OpenVPN es 1000.

¿Cómo se calcula el contrato de nivel de servicio de Virtual WAN ?

Virtual WAN es una plataforma de red como servicio que tiene un contrato de nivel de servicio del 99,95 %. Sin embargo, Virtual WAN combina muchos componentes diferentes, como Azure Firewall, VPN de sitio a sitio, ExpressRoute, VPN de punto a sitio y aplicaciones virtuales de red integrada o de concentrador de Virtual WAN.

El contrato de nivel de servicio para cada componente se calcula individualmente. Por ejemplo, si ExpressRoute tiene un tiempo de inactividad de 10 minutos, la disponibilidad de ExpressRoute se calcularía como (Máximo de minutos disponibles - tiempo de inactividad) / Máximo de minutos disponibles * 100.

¿Puede cambiar el espacio de direcciones de la red virtual en una red virtual radial conectada al centro?

Sí, esto se puede hacer automáticamente sin necesidad de actualizar o restablecer la conexión de emparejamiento. Tenga en cuenta lo siguiente:

- No es necesario hacer clic en el botón "Sincronizar" debajo de la hoja Emparejamiento. Una vez que se cambia el espacio de direcciones de la red virtual, el emparejamiento de red virtual se sincroniza automáticamente con la red virtual del centro virtual.
- Asegúrese de que el espacio de direcciones actualizado no se superpone con el espacio de direcciones de las redes virtuales de radio existentes en la red virtual WAN.

Puede encontrar más información sobre cómo cambiar el espacio de direcciones de la red virtual en Crear, cambiar o eliminar una red virtual.

¿Cuál es el número máximo de direcciones de red virtual de radio admitidas para centros configurados con intención de enrutamiento?

El número máximo de espacios de direcciones en todas las redes virtuales conectadas directamente a un único centro de Virtual WAN es 600. Este límite se aplica individualmente a cada centro de Virtual WAN en una implementación de Virtual WAN. Los espacios de direcciones de red virtual conectados a centros remotos (otros centros de conectividad de Virtual WAN en la misma instancia de Virtual WAN) no se cuentan hacia este límite.

Para obtener más información sobre el límite y scripts de ejemplo para determinar el número de espacios de direcciones entre redes virtuales conectadas a un centro de Virtual WAN, consulte Enrutamiento de límites de espacio de direcciones de red virtual.

¿Puedo usar Azure Bastion con Virtual WAN?

Sí, pero hay limitaciones. Consulte las preguntas más frecuentes sobre Azure Bastion para más información.

Mantenimiento de puerta de enlace controlada por el cliente de Virtual WAN

¿Qué servicios se incluyen en el ámbito de configuración de mantenimiento de las puertas de enlace de red?

Para Virtual WAN, puede configurar ventanas de mantenimiento para puertas de enlace de VPN de sitio a sitio, puertas de enlace de VPN de punto a sitio y puertas de enlace de ExpressRoute.

Esta característica también se admite para Azure Firewalls en centros seguros de Virtual WAN.

¿Qué mantenimiento es compatible o no compatible con el mantenimiento controlado por el cliente?

Los servicios de Azure pasan por actualizaciones de mantenimiento periódicas para mejorar su funcionalidad, la confiabilidad, el rendimiento y la seguridad. Una vez configurada una ventana de mantenimiento para los recursos, el mantenimiento del sistema operativo invitado y el servicio se realizan durante esa ventana. Estas actualizaciones representan la mayoría de los elementos de mantenimiento que preocupan a clientes.

Las actualizaciones subyacentes de hardware de host y infraestructura del centro de datos no están cubiertas por el mantenimiento controlado por el cliente. Además, si hay un problema de seguridad de alta gravedad que podría poner en peligro a nuestros clientes, Azure podría necesitar invalidar el control del cliente de la ventana de mantenimiento e implementar el cambio. Estas son raras apariciones que solo se usarían en casos extremos.

¿Puedo recibir una notificación avanzada de mantenimiento?

En este momento, no se puede habilitar la notificación avanzada para el mantenimiento de los recursos de puerta de enlace de red.

¿Puedo configurar una ventana de mantenimiento inferior a cinco horas?

Actualmente, debe configurar una ventana de cinco horas como mínimo en la zona horaria preferida.

¿Puedo configurar una programación de mantenimiento que no se repita diariamente?

Actualmente, debe configurar una ventana de mantenimiento diaria.

¿Los recursos de configuración de mantenimiento deben estar en la misma región que el recurso de puerta de enlace?

Yes.

¿Es necesario implementar una unidad de escalado de puerta de enlace mínima para que sea apta para el mantenimiento controlado por el cliente?

No.

¿Cuánto tiempo tarda la directiva de configuración de mantenimiento en ser efectiva después de asignarla al recurso de puerta de enlace?

Las puertas de enlace de red pueden tardar hasta 24 horas en seguir la programación de mantenimiento después de que la directiva de mantenimiento esté asociada al recurso de puerta de enlace.

¿Cómo debo planear las ventanas de mantenimiento al usar VPN y ExpressRoute en un escenario de coexistencia?

Al trabajar con VPN y ExpressRoute en un escenario de coexistencia o siempre que tenga recursos que actúen como copias de seguridad, se recomienda configurar ventanas de mantenimiento independientes. Este enfoque garantiza que el mantenimiento no afecte a los recursos de copia de seguridad al mismo tiempo.

He programado una ventana de mantenimiento para uno de mis recursos en una fecha futura. ¿Se pausarán las actividades de mantenimiento en este recurso hasta entonces?

No, las actividades de mantenimiento no se pausarán en el recurso durante el período anterior a la ventana de mantenimiento programada. Durante los días no incluidos en la programación de mantenimiento, el mantenimiento continúa como de costumbre en el recurso.

¿Hay límites en el número de rutas que puedo anunciar?

Sí, hay límites. ExpressRoute admite hasta 4000 prefijos para el emparejamiento privado y 200 prefijos para el emparejamiento de Microsoft. Con ExpressRoute Premium, puede aumentar el límite a 10 000 rutas para el emparejamiento privado. El número máximo de rutas anunciadas desde el emparejamiento privado de Azure a través de una puerta de enlace de ExpressRoute a través de un circuito ExpressRoute es de 1000, que es la misma para los circuitos ExpressRoute estándar y premium. Para más información, puede revisar los límites de ruta de los circuitos ExpressRoute en la página Límites de rutas y límites de suscripción de Azure . Tenga en cuenta que los anuncios de rutas IPv6 no se admiten actualmente con Virtual WAN.

¿Existen restricciones en los intervalos IP que puedo anunciar durante la sesión BGP?

Sí, hay restricciones. Los prefijos privados (RFC1918) no se aceptan para la sesión BGP de emparejamiento de Microsoft. Sin embargo, cualquier tamaño de prefijo hasta un prefijo /32 se acepta en el emparejamiento privado y de Microsoft.

¿Qué ocurre si se supera el límite de rutas de BGP?

Si se supera el límite de rutas BGP, las sesiones de BGP se desconectarán. Las sesiones se restauran una vez que el número de prefijos se reduce por debajo del límite. Para obtener más información, consulte los límites de rutas de circuitos de ExpressRoute en la página de límites y cuotas de suscripción de Azure.

¿Puedo supervisar el número de rutas anunciadas o recibidas a través de un circuito ExpressRoute?

Sí, puede hacerlo. Para conocer los procedimientos recomendados y la configuración de la supervisión de alertas basadas en métricas, consulte los procedimientos recomendados de supervisión de Azure.

¿Cuál es la recomendación para reducir el número de prefijos IP?

Se recomienda agregar los prefijos antes de anunciarlos a través de ExpressRoute o la puerta de enlace VPN. Además, puede usar Route-Maps para resumir las rutas anunciadas desde o hacia Virtual WAN.

¿Puedo usar tablas de rutas definidas por el usuario en redes virtuales radiales conectadas al centro de Virtual WAN?

Yes. Las rutas que el centro de conectividad de Virtual WAN anuncia a los recursos implementados en redes virtuales radiales conectadas son rutas de tipo Protocolo de puerta de enlace de borde (BGP). Si una tabla de rutas definida por el usuario está asociada a una subred conectada a Virtual WAN, la configuración "Propagar rutas de puerta de enlace" **debe** establecerse en "Sí" para que Virtual WAN anuncie los recursos implementados en esa subred. La plataforma de red definida por software subyacente de Azure usa el siguiente algoritmo para seleccionar rutas en función del algoritmo de selección de rutas de Azure.

¿Por qué tengo problemas de conectividad después de volver a anunciar las rutas de Azure en Azure?

Si tiene previsto quitar las comunidades de Azure BGP de las rutas de red virtual y UDR, no anuncie estas rutas de nuevo en Azure, ya que esto provoca problemas de enrutamiento. No se recomienda volver a anunciar rutas de Azure en el entorno de Azure.

Pasos siguientes

Para más información sobre Virtual WAN, consulte Acerca de Virtual WAN.

Last updated on 27/05/2026

Diseño para la recuperación ante desastres con el emparejamiento privado de ExpressRoute

ExpressRoute está diseñado para ofrecer una alta disponibilidad y proporcionar conectividad de red privada de clase operador a los recursos de Microsoft. En otras palabras, no hay un único punto de error en la ruta de acceso de ExpressRoute dentro de la red de Microsoft. Para conocer las consideraciones de diseño para maximizar la disponibilidad de un circuito ExpressRoute, consulte Diseño de alta disponibilidad con ExpressRoute y Well-Architected Framework.

Sin embargo, teniendo en cuenta el popular adage de Murphy (*si algo puede ir mal, lo hará*), este artículo se centra en las soluciones que van más allá de los errores que un único circuito ExpressRoute puede abordar. Examina las consideraciones de arquitectura de red para crear una sólida conectividad de red back-end para la recuperación ante desastres mediante circuitos ExpressRoute con redundancia geográfica.

> ⓘ **Nota:**
>
> Los conceptos descritos en este artículo se aplican de la misma forma cuando se crea un circuito ExpressRoute en una red WAN virtual o fuera de él.

Necesidad de una solución de conectividad redundante

Las ubicaciones de emparejamiento de ExpressRoute o un servicio regional completo pueden degradarse. Una interrupción regional del servicio puede incluir calamidad natural. Planee la recuperación ante desastres para la continuidad empresarial y las aplicaciones críticas.

> ⓘ **Nota:**
>
> Cuando necesite implementar un diseño de recuperación ante desastres en una situación sujeta a limitaciones temporales, como mantener la continuidad empresarial durante un desastre natural, debe tener en cuenta los siguientes factores:
>
> - En este documento, se proporcionan instrucciones sobre cómo implementar un diseño sólido de recuperación ante desastres para varios circuitos ExpressRoute configurados

a través de diferentes ubicaciones de emparejamiento. En este escenario, se supone que tiene tiempo suficiente y recursos para configurar los circuitos ExpressRoute.

- Si necesita configurar rápidamente un diseño de recuperación ante desastres para un único circuito ExpressRoute que no sea con redundancia geográfica, puede usar las siguientes alternativas:
  - Use una **VPN de sitio a sitio como copia de seguridad** para el tráfico de emparejamiento privado.
  - Use la conectividad a Internet como copia de seguridad del tráfico de emparejamiento de Microsoft.

Independientemente de si ejecuta sus aplicaciones críticas en una región de Azure, en una región local o en cualquier otro lugar, puede usar otra región de Azure como sitio de conmutación por error. En los artículos siguientes se aborda la recuperación ante desastres desde las aplicaciones y las perspectivas de acceso de front-end:

- Recuperación ante desastres de escala empresarial
- Recuperación ante desastres SMB con Azure Site Recovery

Si confía en la conectividad de ExpressRoute entre una red local y Microsoft, debe tener en cuenta lo siguiente para planear la recuperación ante desastres a través de ExpressRoute:

- debe usar circuitos de ExpressRoute con redundancia geográfica.
- usando diversas redes de proveedores de servicios para diferentes circuitos de ExpressRoute
- debe diseñar cada uno de los circuitos de ExpressRoute para la alta disponibilidad.
- debe terminar el circuito de ExpressRoute diferente en una ubicación diferente que se encuentre en la red del cliente.
- debe usar puertas de enlace de red virtual de ExpressRoute compatibles con la zona disponibilidad

Desafíos de usar varios circuitos de ExpressRoute

Cuando interconecta el mismo conjunto de redes con más de una conexión, debe introducir rutas de acceso paralelas entre las redes. Las rutas de acceso paralelas, cuando no están correctamente diseñadas, podrían conducir a un enrutamiento asimétrico. Si tiene entidades con estado, por ejemplo, NAT o firewall en la ruta de acceso, el enrutamiento asimétrico podría bloquear el flujo de tráfico. Por lo general, si usa la ruta de emparejamiento privado de ExpressRoute no se encontrará con entidades con estado como NAT o los firewall. Por lo tanto, el enrutamiento

asimétrico sobre el emparejamiento privado de ExpressRoute no bloquea necesariamente el flujo de tráfico.

Pero si equilibra la carga del tráfico a través de rutas paralelas con redundancia geográfica, independientemente de si tiene entidades con estado o no, experimentará un rendimiento de red incoherente. Estas rutas de acceso paralelas con redundancia geográfica pueden encontrarse en el mismo metro o en un metro diferente que se encuentre en la página de proveedores por ubicación.

Redundancia con circuitos ExpressRoute en el mismo metro

La opción Varios metros tiene dos ubicaciones de ExpressRoute. Un ejemplo sería *Amsterdam* y *Amsterdam2*. Al diseñar la redundancia, puede crear dos rutas de acceso paralelas a Azure con ambas ubicaciones en el mismo metro. Puede lograr este objetivo con el mismo proveedor de servicios, o bien elegir otro para mejorar la resistencia. Otra ventaja de este diseño es cuando se produce la conmutación por error de la aplicación, la latencia de un extremo a otro entre las aplicaciones locales y Microsoft se mantiene aproximadamente igual. Sin embargo, si hay un desastre natural, como un terremoto, es posible que la conectividad de ambas rutas de acceso deje de estar disponible.

Redundancia con circuitos ExpressRoute en diferentes metros

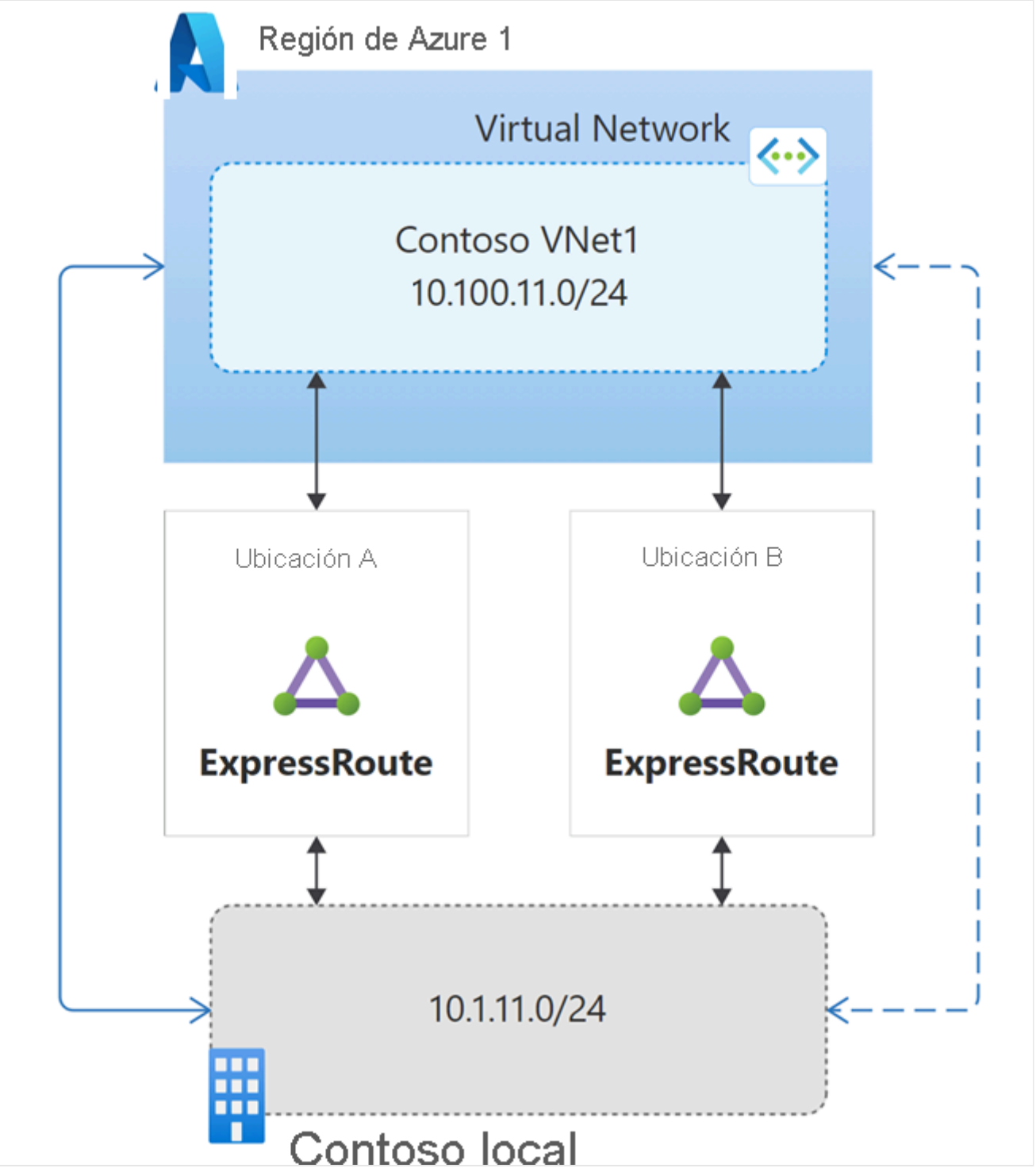
Al usar metros diferentes para la redundancia, debe seleccionar la ubicación secundaria en la misma región geográfica. Para elegir una ubicación fuera de la región geográfica, debe usar la SKU Premium para ambos circuitos en las rutas de acceso paralelas. La ventaja de esta configuración es que la posibilidad de que se produzca un desastre natural que provoque una interrupción de ambos vínculos es menor, pero a costa de una mayor latencia de un extremo a otro.

> **Nota:**
>
> La habilitación de BFD en los circuitos ExpressRoute será útil, ya que facilita una detección más rápida de errores en los vínculos entre los dispositivos Microsoft Enterprise Edge (MSEE) y los enrutadores perimetrales de clientes/partners. Sin embargo, las operaciones de conmutación por error general y convergencia a un sitio redundante pueden tardar hasta 180 segundos cuando se dan ciertas condiciones y es posible que experimente una mayor degradación del rendimiento o posterior durante este tiempo.

En las secciones siguientes se abordan los desafíos a los que podría enfrentarse al configurar rutas de acceso con redundancia geográfica.

Opciones de la red local pequeña a mediana

Considere la red de ejemplo que se muestra en el diagrama siguiente. En el ejemplo, la conectividad expressRoute con redundancia geográfica se establece entre la ubicación local de Contoso y la red virtual de Contoso en una región de Azure. En el diagrama, la línea azul continua indica la ruta preferida (a través de ExpressRoute 1) y la línea punteada representa la ruta de respaldo (a través de ExpressRoute 2).



De forma predeterminada, si anuncia rutas idénticas a través de todos los caminos de ExpressRoute, Azure distribuye el tráfico dirigido a locales entre un máximo de 4 circuitos de ExpressRoute mediante el enrutamiento de múltiples rutas de igual costo (ECMP).

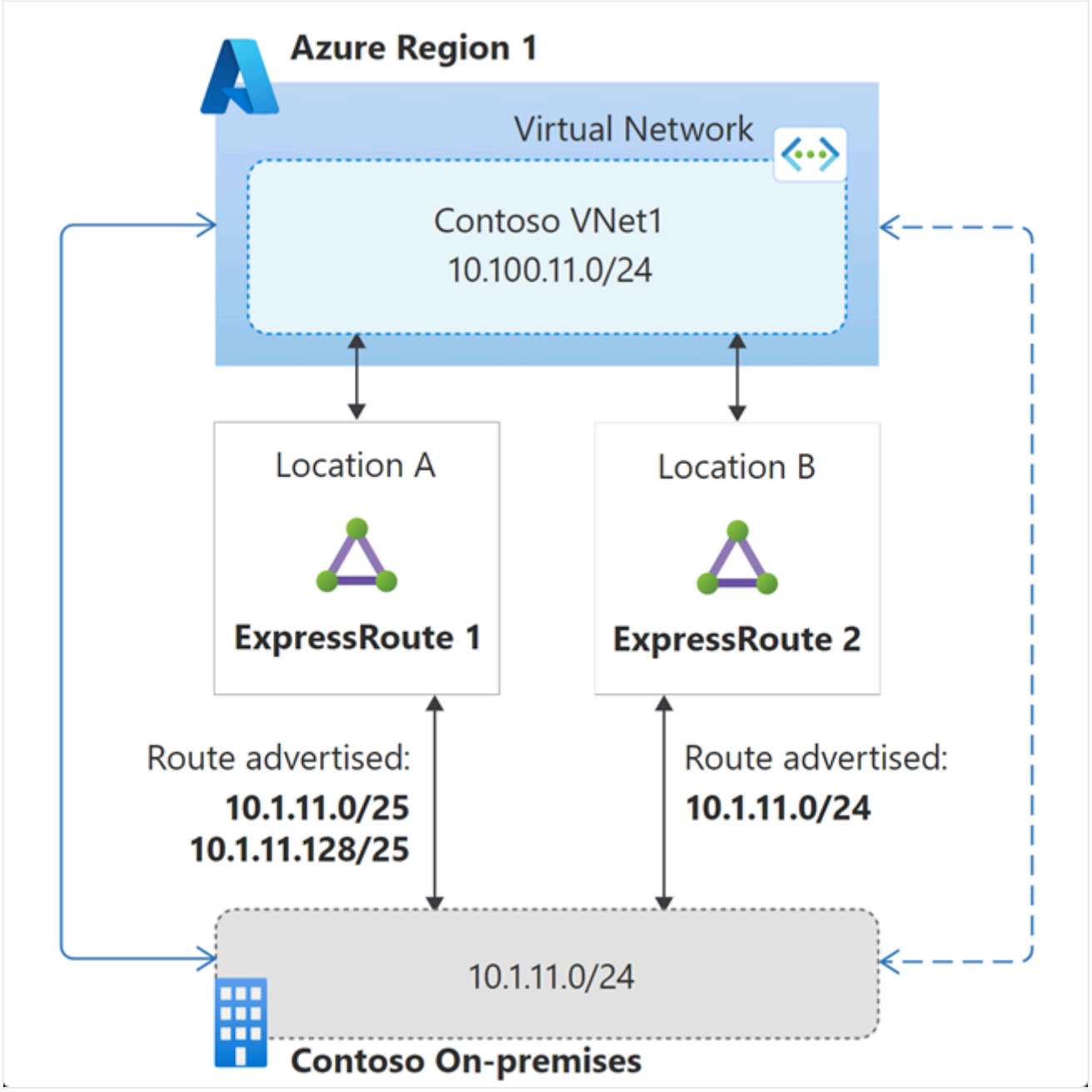
Sin embargo, con circuitos ExpressRoute con redundancia geográfica, hay que tener en cuenta las diferencias de rendimiento de la red entre distintas rutas de red, en particular la latencia. Para obtener un rendimiento de red más coherente durante el funcionamiento normal, es posible que desee preferir el circuito ExpressRoute que ofrece la latencia mínima.

Asimismo, puede influir en Azure para que use un circuito de ExpressRoute en vez de otro, mediante una de las siguientes técnicas (enumeradas en el orden de efectividad):

- indicar una ruta de acceso más específica a través del circuito de ExpressRoute que prefiera, en comparación con otros circuitos de ExpressRoute
- configurar el peso más alto en aquella conexión que conecta la red virtual con el circuito de ExpressRoute preferido.
- indicar las rutas a través del circuito de ExpressRoute que menos prefiera con una ruta de acceso AS más larga (anteponiendo la ruta de acceso de AS).

Una ruta más especifica

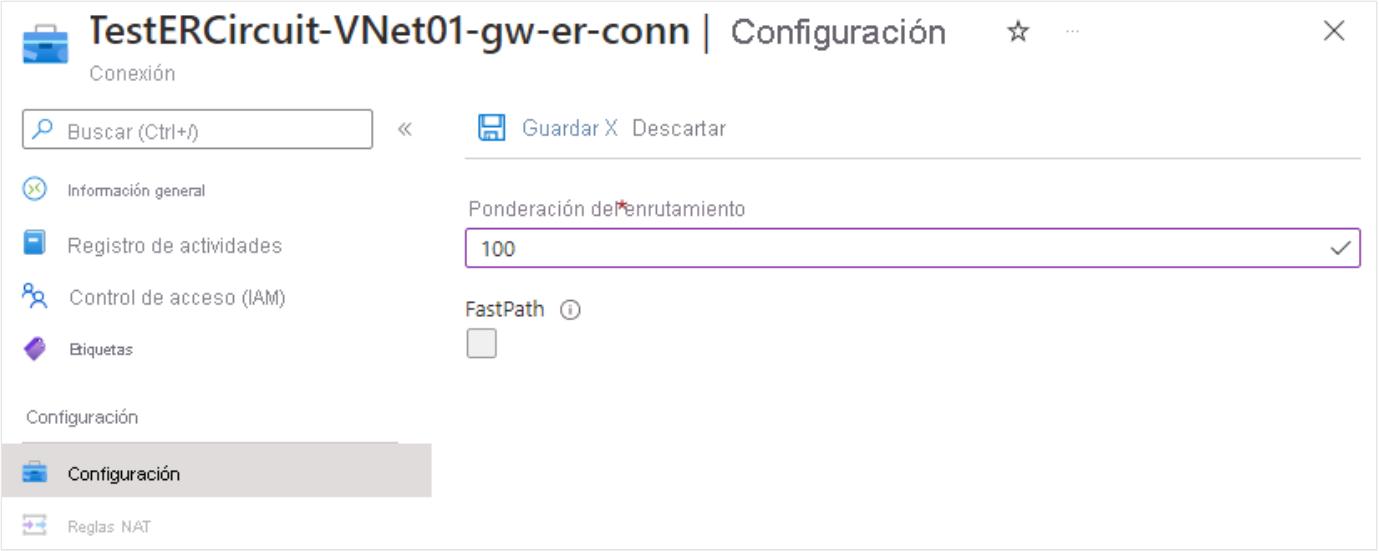
En el siguiente diagrama se indica cómo influir en la selección de rutas de acceso de ExpressRoute con un anuncio de ruta más específico. En el ejemplo ilustrado, el rango de IP local /24 de Contoso se indica como rangos de dos direcciones /25 a través de la ruta de acceso preferida (ExpressRoute 1) y como /24 a través de la ruta de acceso en espera (ExpressRoute 2).



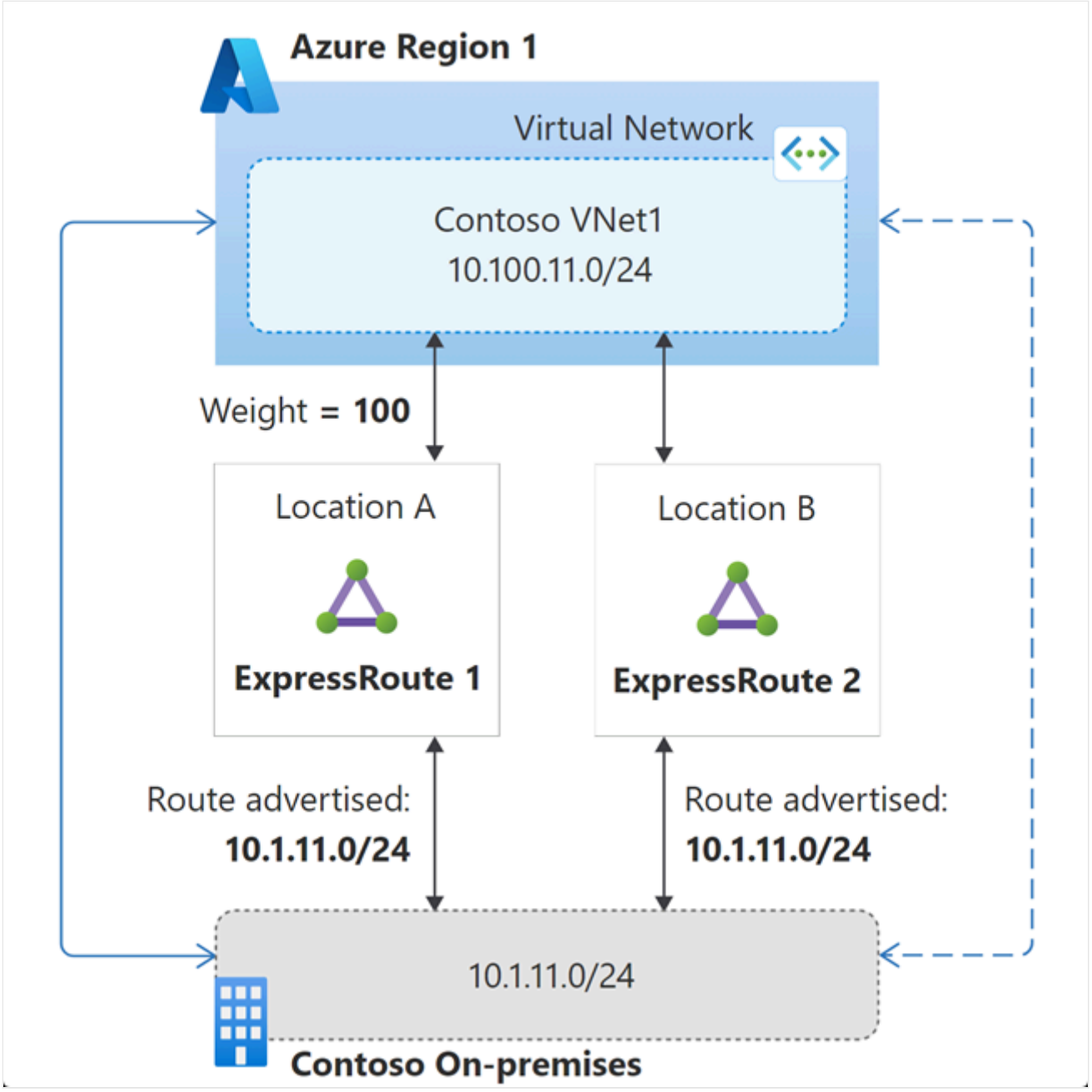
Como /25 es más específico en comparación con /24, Azure enviará el tráfico destinado a 10.1.11.0/24 a través de ExpressRoute 1 en un estado normal. Si ambas conexiones de ExpressRoute 1 se interrumpen, la red virtual verá el anuncio de la ruta 10.1.11.0/24 solo a través de ExpressRoute 2; por lo tanto, el circuito de espera se usará en este estado de error.

Peso de la conexión

En la siguiente captura de pantalla se indica la configuración del peso de una conexión de ExpressRoute a través de Azure Portal.



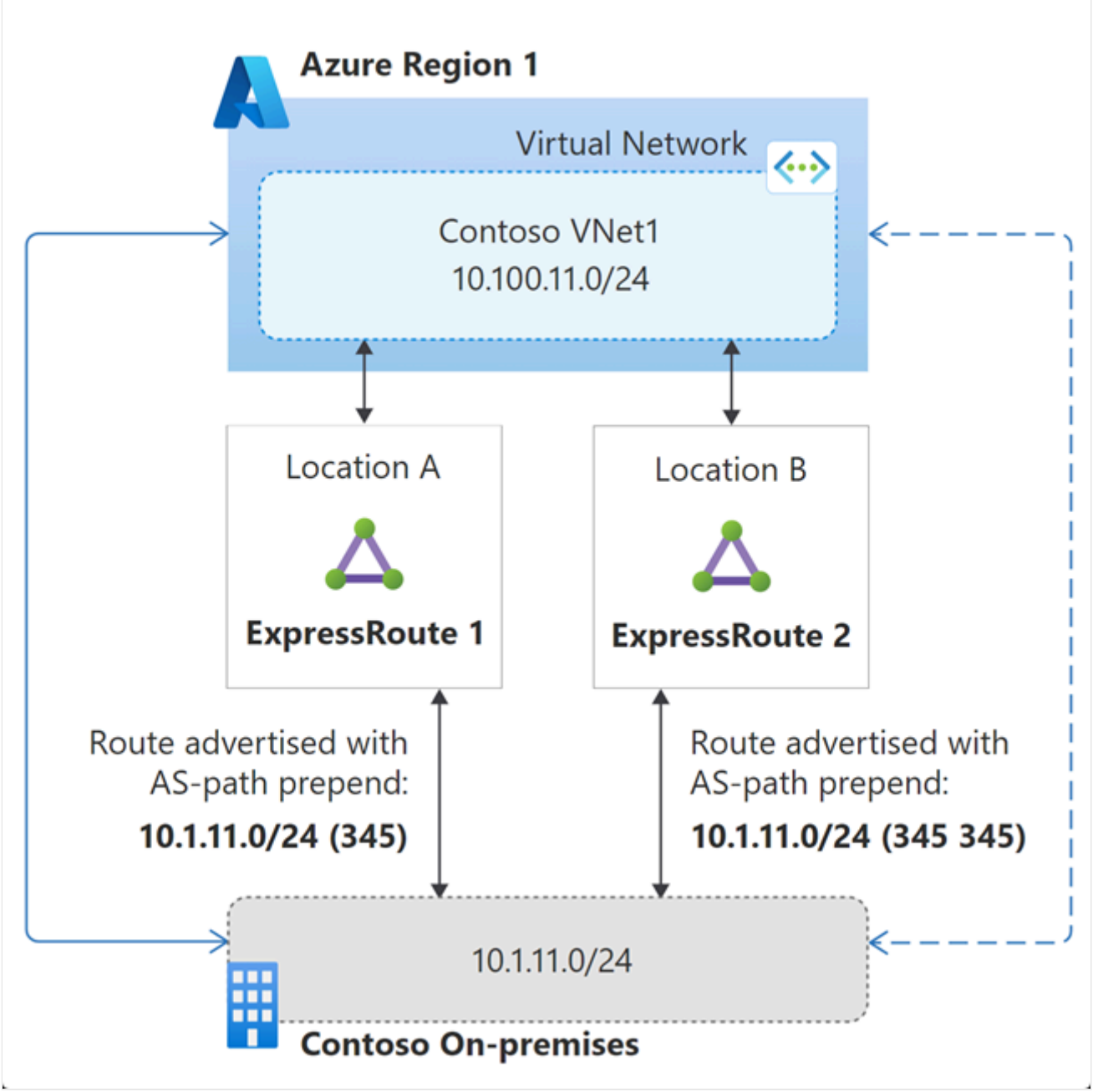
En el siguiente diagrama se ilustra cómo influir en la selección de la ruta de acceso de ExpressRoute con el peso de la conexión. El peso de conexión predeterminado es 0. En el siguiente ejemplo, el peso de la conexión para ExpressRoute 1 se configura como 100. Cuando una red virtual recibe un prefijo de ruta anunciado a través de más de un circuito de ExpressRoute, la red virtual usa la conexión con el mayor peso.



Si ambas conexiones de ExpressRoute 1 se interrumpen, la red virtual verá el anuncio de la ruta 10.1.11.0/24 solo a través de ExpressRoute 2; por lo tanto, el circuito de espera se usará en este estado de error.

Anteponer la ruta de acceso AS

En el siguiente diagrama se ilustra cómo influir en la selección de la ruta de acceso de ExpressRoute anteponiendo la ruta de acceso AS. En el diagrama, el anuncio de ruta sobre ExpressRoute 1 indica el comportamiento predeterminado de eBGP. En el anuncio de ruta sobre ExpressRoute 2, el ASN de la red local se antepone adicionalmente a la ruta de acceso AS de la ruta en cuestión. Cuando se recibe la misma ruta a través de varios circuitos de ExpressRoute, según el proceso de selección de ruta de eBGP, la red virtual usará la ruta con la ruta de acceso AS más corta.



Si las conexiones de ExpressRoute 1 se interrumpen, entonces la red virtual verá el anuncio de ruta 10.1.11.0/24 solo a través de ExpressRoute 2. Por lo tanto, la ruta de acceso as más larga sería irrelevante. Así pues, el circuito de espera se usa en este estado de error.

Al usar cualquiera de esas técnicas, si influye en Azure para que prefiera una de sus rutas de ExpressRoute sobre otras, también debe asegurarse de que la red local también prefiera la misma ruta de ExpressRoute para el tráfico con destino a Azure y así evitar flujos asimétricos. Normalmente, el valor de preferencia local se usa para influir en la red local y que así esta prefiera un circuito de ExpressRoute sobre otros. La preferencia local es una métrica interna de BGP (iBGP). Se prefiere la ruta de BGP que tenga el valor de preferencia local más alto.

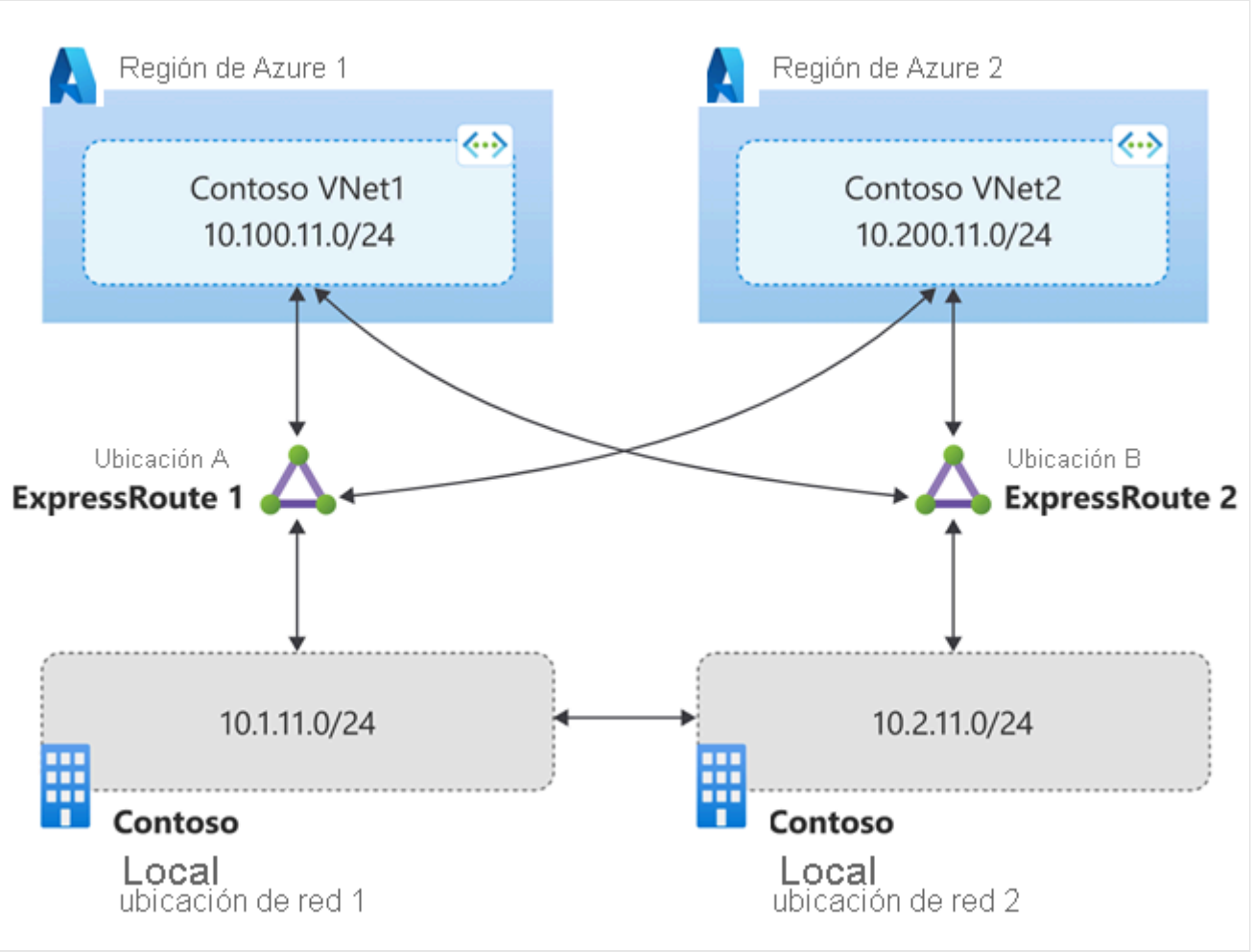
> ⓘ **Importante**

Cuando use ciertos circuitos de ExpressRoute como modo de espera, debe administrarlos activamente y probar periódicamente la operación de conmutación por error.

Red de empresa distribuida de gran tamaño

Cuando tiene una red de empresa distribuida de gran tamaño, es probable que tenga varios circuitos de ExpressRoute. En esta sección se muestra cómo diseñar la recuperación ante desastres mediante circuitos ExpressRoute activos y activos, sin necesidad de otro circuito independiente.

Considere el ejemplo que se muestra en el diagrama siguiente. En el ejemplo, Contoso tiene dos ubicaciones locales conectadas a dos implementaciones de IaaS de Contoso en dos regiones de Azure diferentes a través de circuitos ExpressRoute en dos ubicaciones de emparejamiento diferentes.

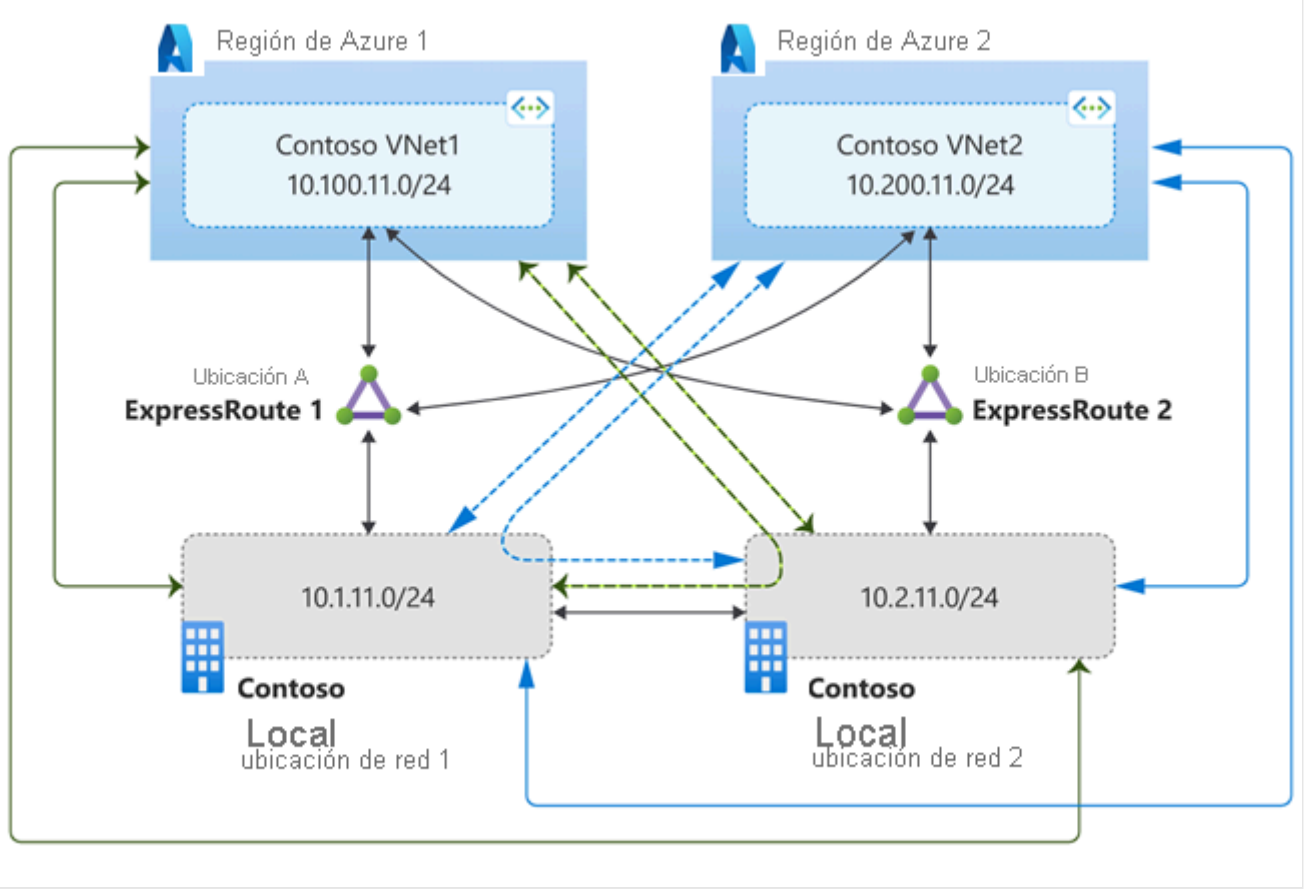


Cómo diseñar la recuperación ante desastres afecta a cómo se enruta el tráfico entre regiones y entre ubicaciones (region1/region2 a location2/location1). Considere dos arquitecturas de recuperación ante desastres que encaminan de manera diferente el tráfico entre ubicaciones de distintas regiones.

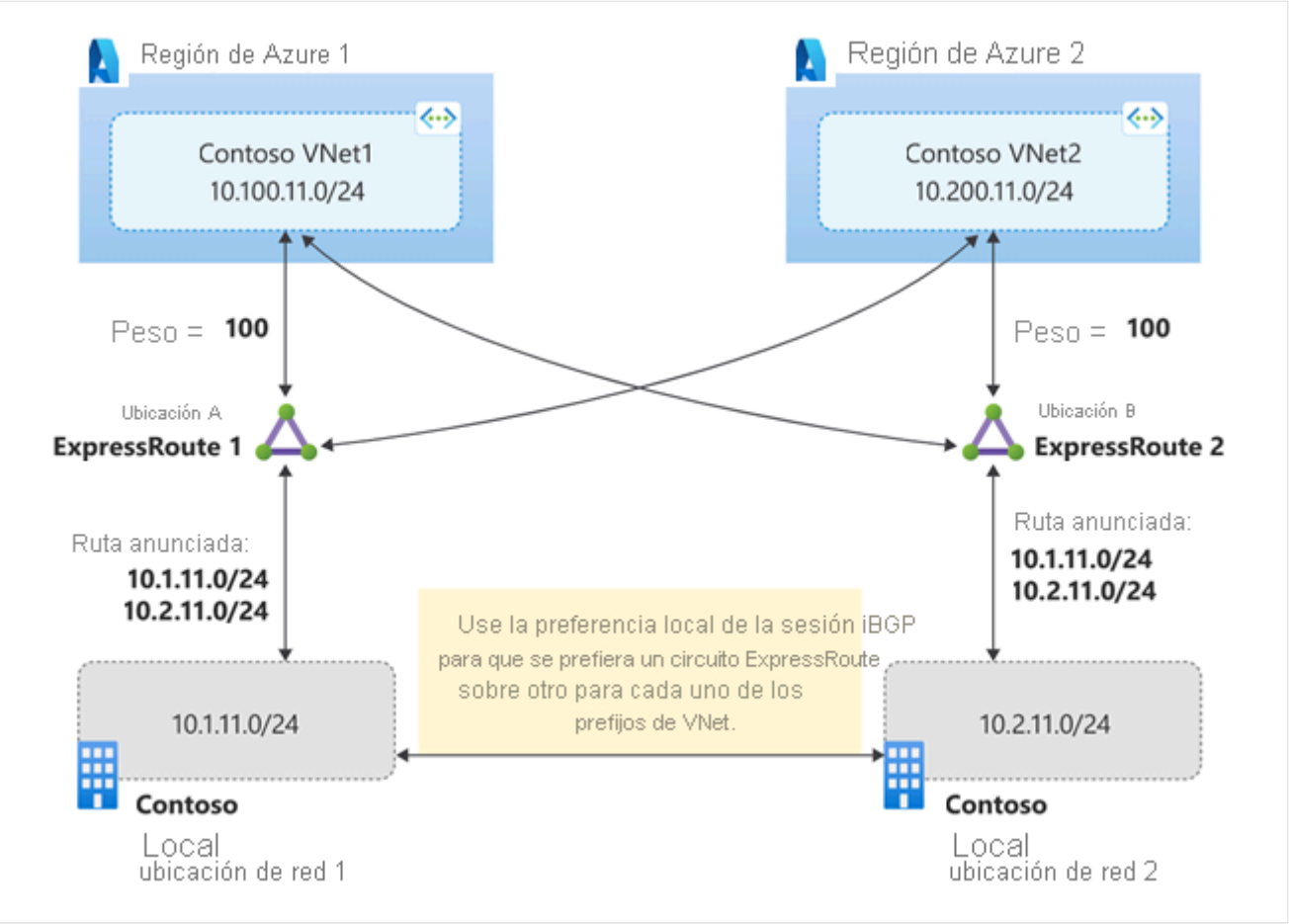
Escenario 1

En el primer escenario, diseñará la recuperación ante desastres para que todo el tráfico entre una región de Azure y la red local fluya a través del circuito ExpressRoute local en estado estable. Si se produce un error en el circuito ExpressRoute local, el circuito ExpressRoute remoto controla todos los flujos de tráfico entre Azure y la red local.

El escenario 1 se ilustra en el diagrama siguiente. En el diagrama, las líneas verdes indican las rutas de acceso del flujo de tráfico entre VNet1 y las redes locales. Las líneas azules indican las rutas de acceso del flujo de tráfico entre VNet2 y las redes locales. Las líneas continuas indican la ruta de acceso que se quiere usar y que tiene un estado estable; las líneas discontinuas, por su parte, indican la ruta de acceso del tráfico en el error del circuito de ExpressRoute correspondiente que transporta el flujo de tráfico en estado estable.

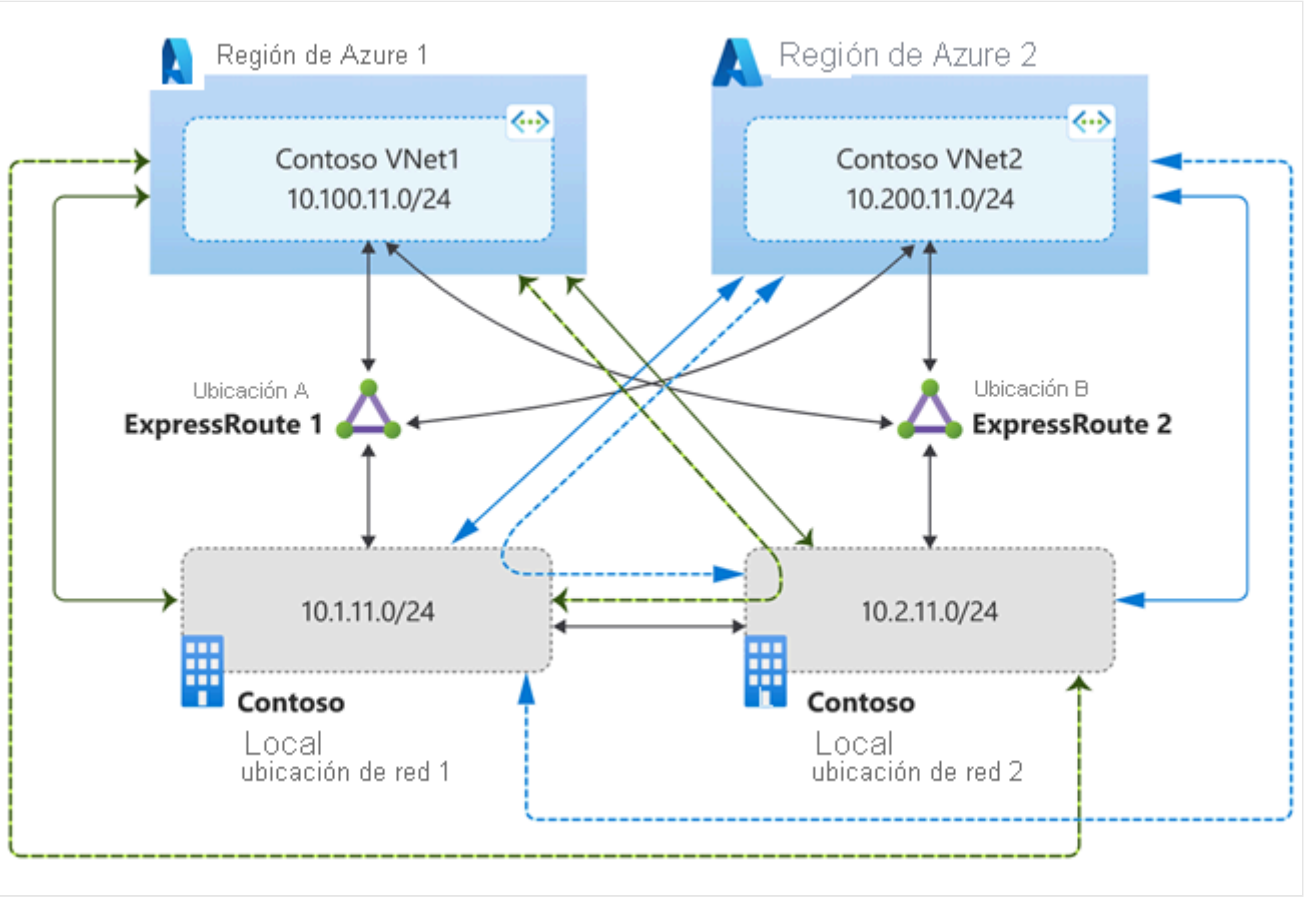


Puede configurar el escenario mediante la ponderación de la conexión para influir en las redes virtuales a fin de que den preferencia a la conexión con la ubicación de emparejamiento local de ExpressRoute para el tráfico destinado a la red local. Para completar la solución, debe garantizar un flujo de tráfico inverso simétrico. Puede usar la preferencia local en la sesión iBGP entre los enrutadores BGP (donde terminan los circuitos ExpressRoute en el lado local) para dar preferencia a un circuito ExpressRoute. La solución se ilustra en el siguiente diagrama.

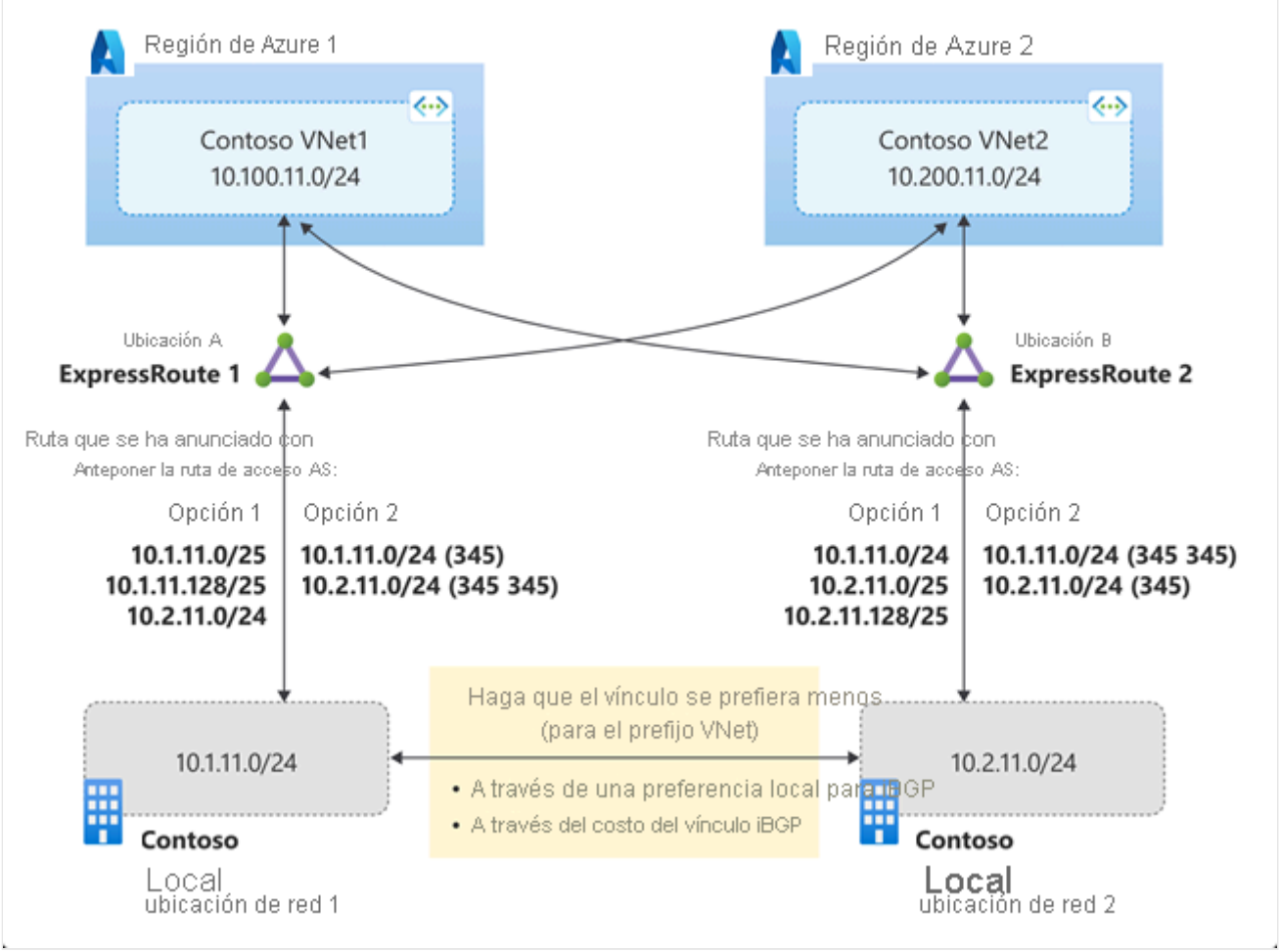


Escenario 2

El escenario 2 se ilustra en el diagrama siguiente. En el diagrama, las líneas verdes indican las rutas de acceso del flujo de tráfico entre VNet1 y las redes locales. Las líneas azules indican las rutas de acceso del flujo de tráfico entre VNet2 y las redes locales. En el estado estable (líneas continuas en el diagrama), todo el tráfico entre redes virtuales y ubicaciones locales fluye a través de la red troncal de Microsoft normalmente; además, también fluye a través de la interconexión entre ubicaciones locales solo en el estado de error, líneas punteadas en el diagrama, de una ruta de ExpressRoute.



La solución se ilustra en el siguiente diagrama. Como se muestra, puede diseñar el escenario usando una ruta más específica (opción 1) o la anteposición de AS-path (opción 2) para influir en la selección de ruta de la red virtual. Para influir en la selección de rutas de red locales para el tráfico destinado a Azure, debe configurar la interconexión con el entorno local para que sea menos preferible. La configuración del vínculo de interconexión depende del protocolo de enrutamiento utilizado en la red local. Puede usar la preferencia local con iBGP o la métrica con IGP (OSPF o IS-IS).



> **Importante**
>
> Cuando uno o varios circuitos de ExpressRoute están conectados a varias redes virtuales, el tráfico de red virtual a red virtual se puede enrutar a través de ExpressRoute. Sin embargo, esto no se recomienda. Para habilitar la conectividad de red virtual a red virtual, configure el emparejamiento de redes virtuales.

Contenido relacionado

En este artículo se describe cómo diseñar la recuperación ante desastres de la conectividad de emparejamiento privado del circuito ExpressRoute. En los artículos siguientes se aborda la recuperación ante desastres desde las aplicaciones y las perspectivas de acceso de front-end:

- Recuperación ante desastres de escala empresarial
- Recuperación ante desastres SMB con Azure Site Recovery

Last updated on 24/06/2026

Confiabilidad en puertas de enlace de red virtual de Azure

Una puerta de enlace de red virtual de Azure proporciona conectividad segura entre una red virtual de Azure y otras redes, ya sea una red local u otra red virtual en Azure.

Azure proporciona dos tipos de puertas de enlace de red virtual:

- Las puertas de enlace de Azure ExpressRoute usan conexiones privadas que no atraviesan la red pública de Internet.
- Las puertas de enlace de red privada virtual (VPN) de Azure usan túneles cifrados a través de Internet.

Como componente de Azure, una puerta de enlace de red virtual proporciona diversas funcionalidades para admitir los requisitos de confiabilidad.

Cuando se usa Azure, la confiabilidad es una responsabilidad compartida. Microsoft proporciona una variedad de funcionalidades para admitir resistencia y recuperación. Es responsable de comprender cómo funcionan esas funcionalidades dentro de todos los servicios que usa y de seleccionar las funcionalidades que necesita para cumplir los objetivos empresariales y los objetivos de tiempo de actividad.

En este artículo se describe cómo hacer que una puerta de enlace de red virtual sea resistente a varias posibles interrupciones y problemas, incluidos errores transitorios, interrupciones de zona de disponibilidad, interrupciones de regiones y mantenimiento planeado del servicio. También resalta información clave sobre el acuerdo de nivel de servicio (SLA) de la puerta de enlace de red virtual.

Para ver información sobre las puertas de enlace de ExpressRoute, seleccione el tipo de puerta de enlace de red virtual adecuado al principio de esta página.

> **ⓘ Importante**
>
> En este artículo se describe la confiabilidad de las puertas de enlace de red virtual, que son los componentes basados en Azure del servicio Azure VPN Gateway.
>
> Pero cuando se usan las VPNs, debe diseñar la arquitectura de toda la *red* (no solo la puerta de enlace) para cumplir con sus requisitos de resiliencia. Es responsable de administrar la confiabilidad de la conexión VPN, incluidos los dispositivos cliente para configuraciones de punto a sitio y dispositivos VPN remotos para configuraciones de sitio a sitio. Para obtener más información sobre cómo configurar la infraestructura para lograr

una alta disponibilidad, consulte [Diseño de la conectividad de puerta de enlace de alta disponibilidad para conexiones entre locales y de red virtual a red virtual](#).

Para garantizar una alta confiabilidad para las puertas de enlace de red virtual de producción, aplique los procedimientos siguientes:

- ✓ **Habilite la redundancia de zona** si los recursos de VPN Gateway residen en una región admitida. Implemente VPN Gateway mediante SKU admitidas (VpnGw1AZ o superior) para garantizar el acceso a las características de redundancia de zona.
- ✓ **Utilice direcciones IP públicas de SKU estándar.**
- ✓ **Configure el modo activo-activo** para una mayor disponibilidad cuando los dispositivos VPN remotos admitan este modo.
- ✓ **Implemente la supervisión adecuada** mediante Azure Monitor para recopilar y ver las métricas de VPN Gateway.

Introducción a la arquitectura de confiabilidad

Una VPN requiere que implemente componentes en el entorno local y en Azure:

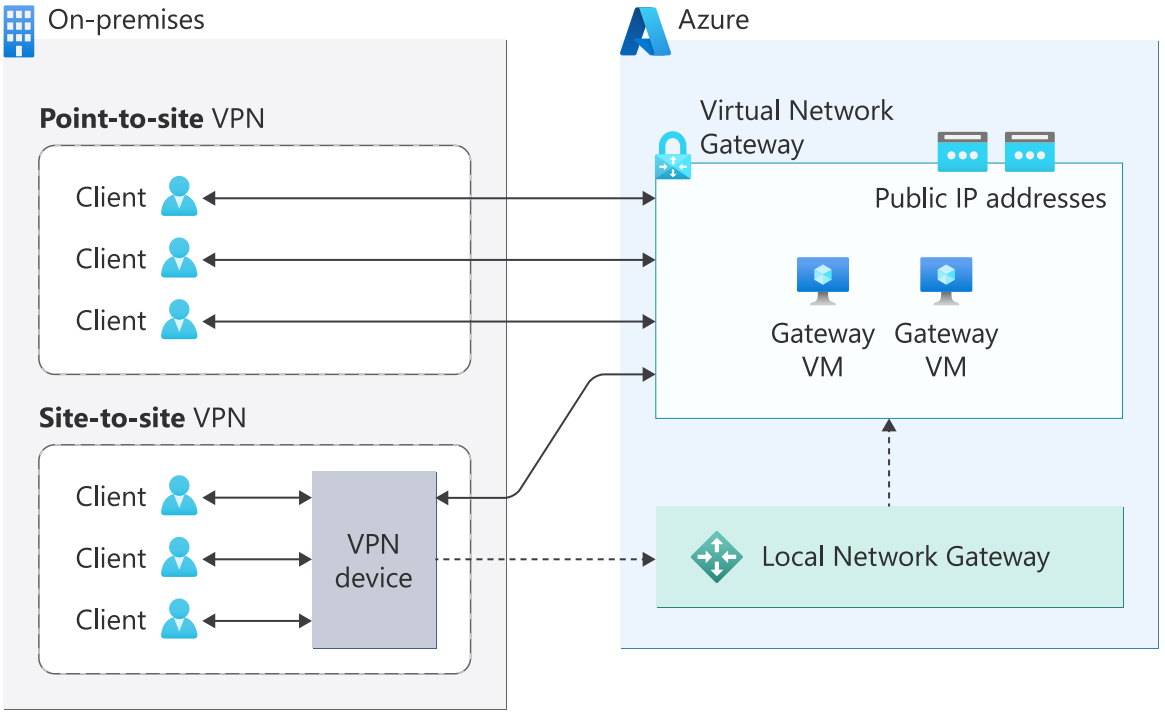
- **Componentes locales:** Los componentes que implemente dependen de si usa una configuración de punto a sitio o de sitio a sitio.
  - Las configuraciones de *sitio a sitio* requieren un dispositivo VPN local que sea responsable de implementar, configurar y administrar.
  - Las configuraciones de *punto a sitio* requieren que implemente una aplicación cliente VPN en un dispositivo remoto, como un equipo portátil o un escritorio, e importe el perfil de usuario en el cliente VPN. Cada conexión de punto a sitio tiene su propio perfil de usuario. Es responsable de implementar y configurar los dispositivos cliente.

  Para más información sobre las diferencias, consulte Topología y diseño de VPN Gateway.

- **Puerta de enlace de red virtual de Azure:** En Azure, creará una puerta de enlace de red virtual, también denominada *puerta de enlace de VPN*, que actúa como punto de terminación para las conexiones VPN.
- **Puerta de enlace de red local:** Una configuración vpn de sitio a sitio también requiere una puerta de enlace de red local, que representa el dispositivo VPN remoto. La puerta de enlace de red local almacena la siguiente información:

- La dirección IP pública del dispositivo VPN local para establecer las conexiones de intercambio de claves de Internet (IKE) fase 1 y fase 2
- Las redes IP locales, para el enrutamiento estático
- Dirección IP del Protocolo de Puerta de Enlace de Frontera (BGP) del par remoto para el enrutamiento dinámico

En el diagrama siguiente se muestran los componentes clave de una VPN que se conecta desde un entorno local a Azure.



Puerta de enlace de red virtual

Una puerta de enlace de red virtual vpn contiene exactamente dos *máquinas virtuales (VM) de puerta de enlace*, que son las máquinas virtuales subyacentes que usa la puerta de enlace para procesar el tráfico VPN.

No ves ni gestionas directamente las máquinas virtuales de la puerta de enlace. La plataforma gestiona automáticamente la creación de máquinas virtuales de puerta de enlace, el monitoreo del estado y la sustitución de máquinas virtuales de puerta de enlace no saludables. Para protegerse frente a fallos de servidor y de bastidor de servidores, Azure distribuye automáticamente las máquinas virtuales de puerta de enlace en varios dominios de fallo dentro de una región. Si ocurre una falla en un rack de servidores, la plataforma Azure migra automáticamente las máquinas virtuales de puerta de enlace de ese clúster a otro clúster.

Usted configura la SKU de la puerta de enlace. Cada SKU admite un nivel de rendimiento diferente y un número diferente de conexiones VPN. Consulte Gateway SKUs para obtener más

información.

En función de los requisitos de alta disponibilidad, puede configurar la puerta de enlace en uno de estos dos modos:

- *Activo-en espera:* Una máquina virtual de puerta de enlace procesa el tráfico y la otra máquina virtual de puerta de enlace permanece en espera.
- *Activo-activo:* Ambas máquinas virtuales de puerta de enlace procesan el tráfico al mismo tiempo. Este modo no siempre es posible porque los flujos de conexión pueden ser asimétricos.

Para más información, consulte Diseño de conectividad del gateway de alta disponibilidad para conexiones entre locales y entre redes virtuales.

Puede protegerse contra fallos en zonas de disponibilidad distribuyendo las máquinas virtuales de la puerta de enlace entre varias zonas. Esta distribución proporciona conmutación automática dentro de la región y mantiene la conectividad durante el mantenimiento o las interrupciones en la zona. Para obtener más información, consulte Resilience to availability zone failures (Resistencia a errores de zona de disponibilidad).

Resistencia a errores transitorios

Los errores transitorios son errores breves e intermitentes en los componentes. Se producen con frecuencia en un entorno distribuido como la nube y son una parte normal de las operaciones. Los errores transitorios se corrigen después de un breve período de tiempo. Es importante que las aplicaciones puedan controlar errores transitorios, normalmente mediante el reintento de solicitudes afectadas.

Todas las aplicaciones hospedadas en la nube deben seguir las instrucciones de control de errores transitorios de Azure cuando se comunican con cualquier API, bases de datos y otros componentes hospedados en la nube. Para obtener más información, consulte Recomendaciones para controlar errores transitorios.

En el caso de las aplicaciones que se conectan a través de una puerta de enlace de red virtual, implemente la lógica de reintento con retroceso exponencial para controlar posibles problemas de conexión transitorios. La naturaleza con estado de las puertas de enlace de red virtual garantiza que las conexiones legítimas se mantengan durante breves interrupciones de red.

En un entorno de red distribuido, los errores transitorios pueden producirse en varias capas, incluidas las siguientes ubicaciones:

- Su entorno local

- Internet
- Azur

Si configura correctamente el enrutamiento de direcciones IP en el dispositivo local, el tráfico de datos como el Protocolo de control de transmisión (TCP) fluye automáticamente a través de túneles activos de Seguridad del protocolo de Internet (IPsec) cuando se produce una desconexión.

A veces, los errores transitorios pueden afectar a los túneles IPsec o a los flujos de datos TCP. Cuando se produce una desconexión, IKE renegocia las asociaciones de seguridad (SAs) para la fase 1 y la fase 2 para restablecer el túnel IPsec.

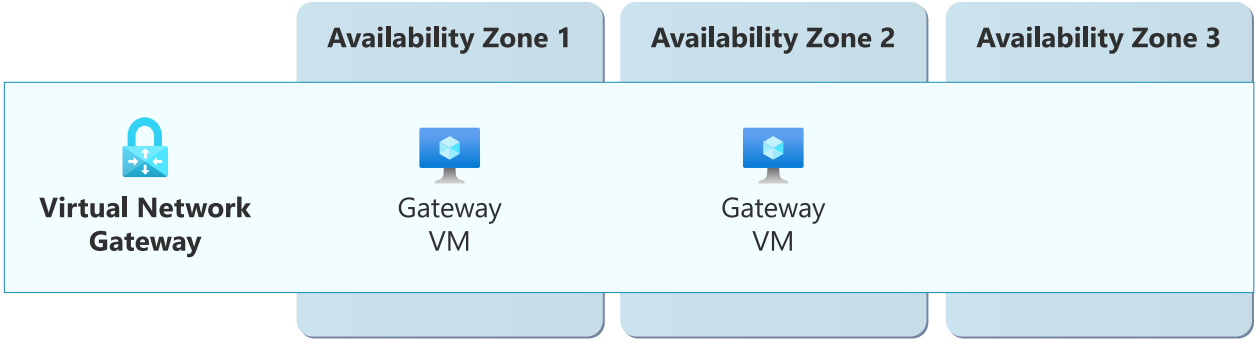
Resistencia a errores de zona de disponibilidad

Las zonas de disponibilidad son grupos físicamente independientes de centros de datos dentro de una región de Azure. Cuando una zona falla, los servicios pueden transferirse a una de las zonas restantes.

Las puertas de enlace de red virtual tienen automáticamente *redundancia de zona* cuando cumplen los requisitos. La redundancia de zona elimina cualquier zona única como punto de error y proporciona el nivel más alto de resistencia de zona. Las puertas de enlace con redundancia de zona proporcionan conmutación automática por error dentro de la región y mantienen la conectividad durante el mantenimiento o las interrupciones de zona.

En VPN Gateway, la redundancia de zona significa que las máquinas virtuales de puerta de enlace se distribuyen automáticamente entre varias zonas de disponibilidad.

En el diagrama siguiente se muestra una puerta de enlace de red virtual con redundancia de zona con dos máquinas virtuales de puerta de enlace distribuidas en distintas zonas de disponibilidad.



> ⓘ **Nota:**

Las puertas de enlace de red local no requieren configuración de zona de disponibilidad porque son automáticamente resistentes a la zona.

Al usar una SKU compatible, las puertas de enlace recién creadas tienen automáticamente redundancia de zona. Se recomienda la redundancia de zona para todas las cargas de trabajo de producción.

Requisitos

- **Compatibilidad con regiones:** Las puertas de enlace de red virtual con redundancia de zona están disponibles en todas las regiones que admiten zonas de disponibilidad.
- **SKU:** Para que una puerta de enlace de red virtual tenga redundancia de zona, debe usar una SKU que admita redundancia de zona. Todos los niveles de VPN Gateway admiten redundancia de zona, excepto la SKU básica, que solo es para entornos de desarrollo. Para más información sobre las opciones de SKU, consulte Gateway SKUs.
- **Direcciones IP públicas:** Debe usar direcciones IP públicas de SKU estándar y configurarlas para que sean redundantes de zona.

Cost

No hay ningún costo adicional para una puerta de enlace implementada en varias zonas de disponibilidad si usa una SKU compatible. Para más información, consulte Precios de VPN Gateway.

Configurar soporte de zonas de disponibilidad

En esta sección se explica cómo configurar la redundancia de zona para las puertas de enlace de red virtual.

- **Cree una nueva puerta de enlace de red virtual que admita zonas de disponibilidad.** Las nuevas puertas de enlace de red virtual son automáticamente redundantes de zona si cumplen los requisitos enumerados anteriormente. Para más información, consulte Creación de una puerta de enlace de red virtual con redundancia de zona en zonas de disponibilidad.
- **Cambie la configuración de zona de disponibilidad de una puerta de enlace de red virtual existente.** Es posible que las puertas de enlace de red virtual que ya haya creado no sean redundantes de zona. Puede migrar una puerta de enlace no zonal a una puerta

de enlace con redundancia de zona con un tiempo de inactividad mínimo. Para obtener más información, consulte Consolidación y migración de SKU.

Comportamiento cuando todas las zonas están en buen estado

En la sección siguiente se describe qué esperar cuando la puerta de enlace de red virtual está configurada para la redundancia de zona y todas las zonas de disponibilidad están operativas.

- **Enrutamiento de tráfico entre zonas:** La redundancia de zona no afecta a cómo se enruta el tráfico. El tráfico se enruta entre las máquinas virtuales de puerta de enlace de tu puerta de enlace en función de la configuración de tus clientes. Si tu gateway utiliza una configuración activa-activa y emplea dos direcciones IP públicas, ambas máquinas virtuales del gateway podrían recibir tráfico. Para la configuración en espera activa, el tráfico se enruta a una sola máquina virtual de puerta de enlace principal que Azure selecciona.

- **Replicación de datos entre zonas:** VPN Gateway no necesita sincronizar el estado de conexión entre zonas de disponibilidad. En el modo activo-activo, la máquina virtual de puerta de enlace que procesa la VPN es responsable de gestionar el estado de la conexión.

- **Administración de máquinas virtuales de puerta de enlace:** La plataforma selecciona automáticamente las zonas de las máquinas virtuales de puerta de enlace y administra la ubicación en todas las zonas. La supervisión del estado garantiza que solo las máquinas virtuales de puerta de enlace correctas reciban tráfico.

Comportamiento durante un fallo de zona

En la sección siguiente se describe qué esperar cuando la puerta de enlace de red virtual está configurada para zona de redundancia y hay una interrupción en la zona de disponibilidad.

- **Detección y respuesta:** La plataforma Azure detecta y responde a un error en una zona de disponibilidad. No es necesario iniciar una conmutación por error de zona.

- **Notificación**: Microsoft no le notifica automáticamente cuando una zona está inactiva. Sin embargo, puede usar Azure Resource Health para supervisar el estado de un recurso individual y puede configurar alertas de Resource Health para notificarle problemas. También puede usar Azure Service Health para comprender el estado general del servicio, incluidos los errores de zona, y puede configurar alertas de Service Health para notificarle problemas.

- **Solicitudes activas:** Se finalizan las solicitudes activas conectadas a través de máquinas virtuales de puerta de enlace en la zona que falla. Las aplicaciones cliente deben reintentar las solicitudes siguiendo las instrucciones para controlar errores transitorios.
- **Pérdida de datos esperada:** Los errores de zona no deben provocar pérdida de datos porque las puertas de enlace de red virtual no almacenan datos persistentes de los clientes.
- **Tiempo de inactividad esperado:** Durante las interrupciones de zona, las conexiones pueden experimentar interrupciones breves que suelen durar hasta un minuto cuando se redistribuye el tráfico. Las aplicaciones cliente deben reintentar las solicitudes siguiendo las instrucciones para controlar errores transitorios.
- **Reenrutamiento del tráfico:** El tráfico se vuelve a enrutar automáticamente a la otra máquina virtual de puerta de enlace en una zona de disponibilidad diferente.

Recuperación de zona

Cuando se recupera la zona de disponibilidad afectada, Azure restaura automáticamente las máquinas virtuales de puerta de enlace en la zona recuperada y vuelve a la distribución de tráfico normal en todas las zonas que usa la puerta de enlace.

Prueba de fallos de zona

La plataforma Azure gestiona el enrutamiento del tráfico, la conmutación por error y la conmutación por recuperación para las puertas de enlace de red virtual con redundancia de zona. Esta característica está totalmente administrada, por lo que no es necesario iniciar ni validar los procesos de error de zona de disponibilidad.

Resistencia a errores en toda la región

Una puerta de enlace de red virtual es un recurso de una sola región. Si la región deja de estar disponible, la puerta de enlace tampoco está disponible.

Soluciones personalizadas de varias regiones para la resistencia

Puede implementar puertas de enlace de VPN independientes en dos o más regiones diferentes. Cada puerta de enlace está conectada a una red virtual diferente y las puertas de enlace funcionan de forma independiente. No hay interacción ni replicación de la configuración o el estado entre ellos. También es responsable de configurar los clientes y

dispositivos remotos para conectarse a la VPN correcta o cambiar entre VPN cuando sea necesario.

Resistencia al mantenimiento del servicio

Azure realiza un mantenimiento normal en las puertas de enlace de red virtual para garantizar un rendimiento y seguridad óptimos. Durante estas ventanas de mantenimiento, se pueden producir algunas interrupciones del servicio, pero Azure diseña estas actividades para minimizar el efecto en la conectividad.

Durante las operaciones de mantenimiento planificado en las puertas de enlace de red virtual, el proceso se ejecuta secuencialmente en las máquinas virtuales de la puerta de enlace, no simultáneamente. Este proceso garantiza que una máquina virtual de puerta de enlace siempre permanezca activa durante el mantenimiento, lo que minimiza el impacto en las conexiones activas.

Para reducir la probabilidad de interrupciones inesperadas, puede configurar las ventanas de mantenimiento del gateway para que coincidan con los requisitos operativos.

Para más información, consulte Configuración de ventanas de mantenimiento para las puertas de enlace de red virtual.

Acuerdo de nivel de servicio

El contrato de nivel de servicio (SLA) para los servicios de Azure describe la disponibilidad esperada de cada servicio y las condiciones que la solución deberá cumplir para lograr esa expectativa de disponibilidad. Para obtener más información, consulte Acuerdos de Nivel de Servicio para servicios en línea.

Todas las SKU de VPN Gateway que no sean la SKU básica son aptas para un Acuerdo de Nivel de Servicio de mayor disponibilidad. La SKU básica proporciona un Acuerdo de Nivel de Servicio de menor disponibilidad y funcionalidades limitadas, y solo se debe usar para pruebas y desarrollo. Para obtener más información, consulte Gateway SKU: cargas de trabajo de producción y pruebas de desarrollo

Contenido relacionado

- Crear una puerta de enlace de red virtual con redundancia de zona
- Supervisión de la pasarela VPN
- Diseño de conectividad de alta disponibilidad de pasarela para conexiones entre ubicaciones y de red virtual a red virtual

ⓘ **Nota:** El autor ha creado este artículo con ayuda de inteligencia artificial. Más información

---

Last updated on 22/01/2026

Confiabilidad en Azure Firewall

Azure Firewall es un servicio de seguridad de red administrado basado en la nube que protege los recursos de Azure Virtual Network. Es un servicio de firewall con estado completo que incluye alta disponibilidad integrada y escalabilidad de nube sin restricciones.

Cuando se usa Azure, la confiabilidad es una responsabilidad compartida. Microsoft proporciona una variedad de funcionalidades para admitir resistencia y recuperación. Es responsable de comprender cómo funcionan esas funcionalidades dentro de todos los servicios que usa y de seleccionar las funcionalidades que necesita para cumplir los objetivos empresariales y los objetivos de tiempo de actividad.

En este artículo se describe cómo hacer que Azure Firewall sea resistente a una variedad de posibles interrupciones y problemas, como errores transitorios, interrupciones de zona de disponibilidad y interrupciones de regiones. También describe la resistencia durante el mantenimiento del servicio y resalta cierta información clave sobre el acuerdo de nivel de servicio (SLA) de Firewall.

Recomendaciones de implementación de producción

Para obtener información sobre cómo implementar Azure Firewall para admitir los requisitos de confiabilidad de la solución y cómo afecta la confiabilidad a otros aspectos de la arquitectura, consulte Procedimientos recomendados de arquitectura para Azure Firewall en Azure Well-Architected Framework.

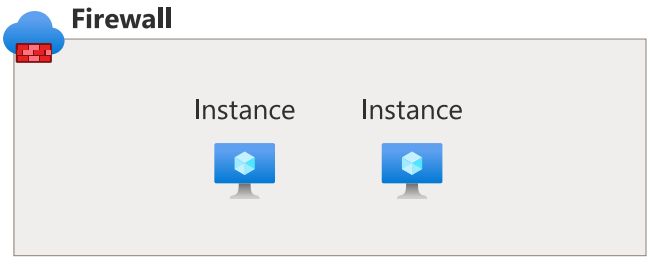
Introducción a la arquitectura de confiabilidad

Una *instancia* hace referencia a una unidad de nivel de máquina virtual (VM) del firewall. Cada instancia representa la infraestructura que controla el tráfico y realiza comprobaciones de firewall.

Para lograr una alta disponibilidad de un firewall, Azure Firewall proporciona automáticamente al menos dos instancias, sin necesidad de intervención o configuración. El firewall se escala horizontalmente automáticamente cuando el rendimiento promedio, el consumo de CPU y el uso de la conexión alcanzan umbrales predefinidos. Para más información, consulte Rendimiento de Azure Firewall. La plataforma administra automáticamente la creación de instancias, la supervisión del estado y el reemplazo de instancias incorrectas.

Para protegerse frente a errores de servidor y bastidor de servidores, Azure Firewall distribuye automáticamente las instancias entre varios dominios de error dentro de una región.

En el diagrama siguiente se muestra un firewall con dos instancias:



Para aumentar la redundancia y la disponibilidad durante los errores del centro de datos, Azure Firewall habilita automáticamente la redundancia de zona en regiones que admiten varias zonas de disponibilidad, distribuyendo instancias entre al menos dos zonas de disponibilidad.

Resistencia a errores transitorios

Los errores transitorios son errores breves e intermitentes en los componentes. Se producen con frecuencia en un entorno distribuido como la nube y son una parte normal de las operaciones. Los errores transitorios se corrigen después de un breve período de tiempo. Es importante que las aplicaciones puedan controlar errores transitorios, normalmente mediante el reintento de solicitudes afectadas.

Todas las aplicaciones hospedadas en la nube deben seguir las instrucciones de control de errores transitorios de Azure cuando se comunican con cualquier API, bases de datos y otros componentes hospedados en la nube. Para obtener más información, consulte Recomendaciones para controlar errores transitorios.

En el caso de las aplicaciones que se conectan a través de Azure Firewall, implemente la lógica de reintento con retroceso exponencial para controlar posibles problemas de conexión transitorios. La naturaleza con estado de Azure Firewall garantiza que las conexiones legítimas permanezcan activas durante breves interrupciones de red.

Durante las operaciones de escalado, que tardan cinco a siete minutos, el firewall conserva las conexiones existentes mientras agrega nuevas instancias de firewall para controlar la mayor carga.

Resistencia a errores de zona de disponibilidad

Las zonas de disponibilidad son grupos físicamente independientes de centros de datos dentro de una región de Azure. Cuando una zona falla, los servicios pueden transferirse a una de las zonas restantes.

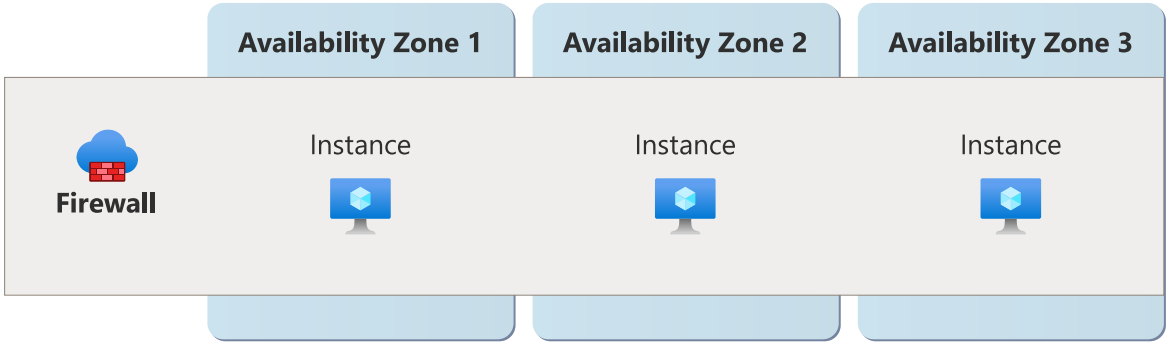
Azure Firewall se implementa automáticamente con redundancia de zona en regiones que admiten múltiples zonas de disponibilidad. Un firewall tiene redundancia de zona al implementarlo en al menos dos zonas de disponibilidad.

Azure Firewall admite modelos de implementación con redundancia de zona y zonales:

- **Con redundancia de zona:** En regiones que admiten zonas de disponibilidad, Azure distribuye automáticamente las instancias de firewall entre varias zonas de disponibilidad (al menos dos). Azure administra automáticamente el equilibrio de carga y la conmutación por error entre zonas. Este modelo de implementación es el valor predeterminado para todos los firewalls nuevos.

  Los firewalls con redundancia de zona logran el acuerdo de nivel de servicio (SLA) de tiempo de actividad más alto. Úselas para cargas de trabajo de producción que requieran la máxima disponibilidad.

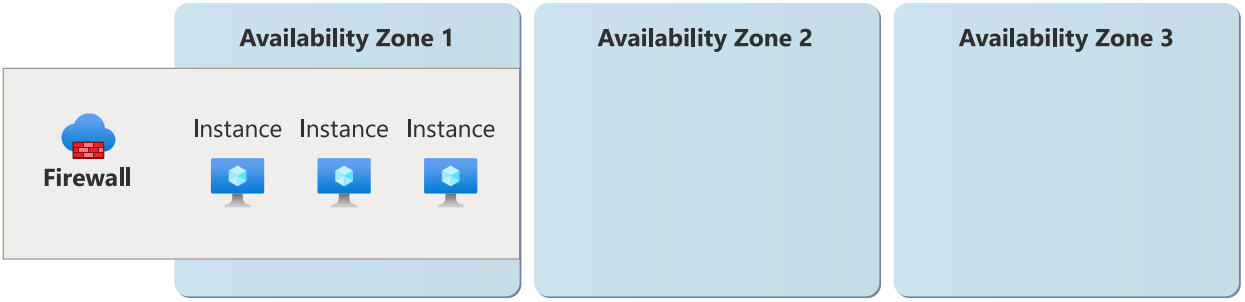
  En el diagrama siguiente se muestra un firewall con redundancia de zona con tres instancias distribuidas entre tres zonas de disponibilidad:



> ⓘ **Nota:**
>
> Todas las implementaciones de firewall en regiones con varias zonas de disponibilidad son automáticamente con redundancia de zona. Esta regla se aplica a las implementaciones a través de Azure Portal y implementaciones basadas en API (CLI de Azure, PowerShell, Bicep, plantillas de ARM, Terraform).

- **Zonal:** En escenarios específicos en los que existen restricciones de capacidad o los requisitos de latencia son críticos, puede implementar Azure Firewall en una zona de disponibilidad específica mediante herramientas basadas en API (CLI de Azure, PowerShell, Bicep, plantillas de ARM, Terraform). Implemente todas las instancias de un firewall zonal dentro de esa zona.

  En el diagrama siguiente se muestra un firewall zonal con tres instancias que se implementan en la misma zona de disponibilidad:



> **Importante**
>
> Solo puede crear implementaciones zonales a través de herramientas basadas en API. No se pueden configurar a través de Azure Portal. Las implementaciones de firewall zonales existentes se migrarán a implementaciones con redundancia de zona en el futuro. Use implementaciones con redundancia de zona siempre que sea posible para alcanzar el mayor nivel de disponibilidad según el SLA (Acuerdo de Nivel de Servicio). Un firewall zonal solo no proporciona resistencia a una interrupción de zona de disponibilidad.

Migración de implementaciones existentes

Anteriormente, las implementaciones de Azure Firewall que no están configuradas para que sean con redundancia de zona o zonales son *no zonales* o *regionales*. A lo largo del año natural 2026, Azure está migrando todas las implementaciones de firewall no zonales existentes a implementaciones con redundancia de zona en regiones que admiten varias zonas de disponibilidad.

Soporte para regiones

Azure Firewall admite zonas de disponibilidad en todas las regiones que admiten zonas de disponibilidad, donde el servicio Azure Firewall está disponible.

Requisitos

- Todos los niveles de Azure Firewall admiten zonas de disponibilidad.
- Los firewalls con redundancia de zona requieren direcciones IP públicas estándar configuradas para tener redundancia de zona.
- Los firewalls zonales (implementados mediante herramientas basadas en API) requieren direcciones IP públicas estándar y se pueden configurar para que sean con redundancia de zona o zonales en la misma zona que el firewall.

Cost

No hay ningún costo adicional para las implementaciones de firewall con redundancia de zona.

Configurar soporte de zonas de disponibilidad

En esta sección se explica la configuración de zona de disponibilidad para los firewalls.

- **Cree un nuevo firewall:** Todas las nuevas implementaciones de Azure Firewall en regiones con varias zonas de disponibilidad son automáticamente con redundancia de zona de forma predeterminada. Esta regla se aplica a las implementaciones basadas en el portal y basadas en API.

  - *Con redundancia de zona (valor predeterminado):* Al implementar un nuevo firewall en una región con varias zonas de disponibilidad, Azure distribuye automáticamente las instancias entre al menos dos zonas de disponibilidad. No se requiere ninguna configuración adicional. Para más información, consulte Implementar Azure Firewall mediante Azure Portal.
    - **Azure Portal:** Implementa automáticamente firewalls con redundancia de zona. No puede seleccionar una zona de disponibilidad específica a través del portal.
    - **Herramientas basadas en API (CLI de Azure, PowerShell, Bicep, plantillas de ARM, Terraform):** Implemente firewalls con redundancia de zona de forma predeterminada. Opcionalmente, puede especificar zonas para la implementación.

    Para más información sobre cómo implementar un firewall con redundancia de zona, consulte Implementación de Azure Firewall con zonas de disponibilidad.

  - *Zonal (solo herramientas basadas en API):* Para implementar un firewall en una zona de disponibilidad específica (por ejemplo, debido a restricciones de capacidad en una región), use herramientas basadas en API, como la CLI de Azure, PowerShell, Bicep, plantillas de ARM o Terraform. Especifique una sola zona en la configuración de implementación. Esta opción no está disponible a través de Azure Portal.

    > ⓘ **Nota:**
    >
    > Cuando selecciona qué zonas de disponibilidad usar, en realidad está seleccionando la *zona de disponibilidad lógica*. Si implementa otros componentes de la carga de trabajo en una suscripción de Azure diferente, podrían usar un número de zona de disponibilidad lógica distinto para acceder a la misma zona de disponibilidad física. Para obtener más información, consulte **zonas de disponibilidad física y lógica**.

- **Firewalls existentes:** Todas las implementaciones de firewall no zonales (regionales) existentes se migran automáticamente a implementaciones con redundancia de zona en regiones que admiten varias zonas de disponibilidad. Las implementaciones de firewall zonales existentes (ancladas a una zona específica) se migran a implementaciones con redundancia de zona en una fecha futura.

- **Restricciones de capacidad:** Si una región no tiene capacidad para una implementación con redundancia de zona (que requiere al menos dos zonas de disponibilidad), se produce un error en la implementación. En este escenario, puede implementar un firewall zonal en una zona de disponibilidad específica mediante herramientas basadas en API.

Comportamiento cuando todas las zonas están en buen estado

En esta sección se describe qué esperar cuando Azure Firewall está configurado con compatibilidad con la zona de disponibilidad y todas las zonas de disponibilidad están operativas.

- **Enrutamiento de tráfico entre zonas:** El comportamiento de enrutamiento del tráfico depende de la configuración de zona de disponibilidad que usa el firewall.

  - *Con redundancia de zona:* Azure Firewall distribuye automáticamente las solicitudes entrantes entre instancias de todas las zonas que usa el firewall. Esta configuración activa-activa garantiza un rendimiento y una distribución de carga óptimos en condiciones de funcionamiento normales.

  - *Zonal:* Si implementa varias instancias zonales en distintas zonas, debe configurar el enrutamiento de tráfico mediante soluciones de equilibrio de carga externas, como Azure Load Balancer o Azure Traffic Manager.

- **Administración de instancias:** La plataforma administra automáticamente la ubicación de la instancia en las zonas que usa el firewall. Reemplaza las instancias con errores y mantiene el recuento de instancias configurados. El seguimiento de estado garantiza que solo las instancias correctas reciban tráfico.

- **Replicación de datos entre zonas:** Azure Firewall no necesita sincronizar el estado de conexión entre zonas de disponibilidad. La instancia que procesa la solicitud mantiene el estado de cada conexión.

Comportamiento durante un fallo de zona

En esta sección se describe qué esperar cuando Azure Firewall está configurado con compatibilidad con la zona de disponibilidad y una o varias zonas de disponibilidad no están disponibles.

- **Detección y respuesta:** La responsabilidad de la detección y la respuesta depende de la configuración de la zona de disponibilidad que usa el firewall.
  - *Con redundancia de zona:* En el caso de las instancias configuradas para usar la redundancia de zona, la plataforma Azure Firewall detecta y responde a un error en una zona de disponibilidad. No es necesario iniciar una conmutación por error de zona.
  - *Zonal:* Para que los firewalls configurados sean zonales, debe detectar la pérdida de una zona de disponibilidad e iniciar una conmutación por error a un firewall secundario que cree en otra zona de disponibilidad.
- **Notificación:** Microsoft no le notifica automáticamente cuando una zona está inactiva. Sin embargo, puede usar Azure Service Health para comprender el estado general del servicio, incluidos los errores de zona, y puede configurar alertas de Service Health para notificarle problemas.
- **Conexiones activas:** Cuando una zona de disponibilidad no está disponible, las solicitudes en curso que se conectan a una instancia de firewall en la zona de disponibilidad errónea pueden finalizar y requerir reintentos.
- **Pérdida de datos esperada:** No se espera ninguna pérdida de datos durante la conmutación por error de zona porque Azure Firewall no almacena datos persistentes del cliente.
- **Tiempo de inactividad esperado:** El tiempo de inactividad depende de la configuración de zona de disponibilidad que use el firewall.
  - *Con redundancia de zona:* Espere un tiempo de inactividad mínimo (normalmente unos segundos) durante una interrupción de la zona de disponibilidad. Las aplicaciones cliente deben seguir los procedimientos para el control de errores transitorios, incluida la implementación de directivas de reintento con retroceso exponencial.
  - *Zonal:* Cuando una zona no está disponible, el firewall sigue sin estar disponible hasta que se recupere la zona de disponibilidad.
- **Redireccionamiento del tráfico:** El comportamiento de redireccionar el tráfico depende de la configuración de zona de disponibilidad que usa el firewall.
  - *Con redundancia de zona:* El tráfico se vuelve a enrutar automáticamente a zonas de disponibilidad correctas. Si es necesario, la plataforma crea nuevas instancias de

firewall en zonas correctas.

- *Zonal:* Cuando una zona no está disponible, el firewall zonal tampoco está disponible. Si tiene un firewall secundario en otra zona de disponibilidad, es responsable de redirigir el tráfico a ese firewall.

Failback

El comportamiento de la conmutación por recuperación depende de la configuración de zona de disponibilidad que usa el firewall.

- *Con redundancia de zona:* Una vez recuperada la zona de disponibilidad, Azure Firewall redistribuye automáticamente las instancias de todas las zonas que usa el firewall y restaura el equilibrio de carga normal entre zonas.
- *Zonal:* Una vez recuperada la zona de disponibilidad, es responsable de redirigir el tráfico al firewall en la zona de disponibilidad original.

Prueba de fallos de zona

Las opciones para las pruebas de errores de zona dependen de la configuración de la zona de disponibilidad del firewall.

- *Con redundancia de zona:* La plataforma Azure Firewall administra el enrutamiento del tráfico, la conmutación por error y la conmutación por recuperación para los recursos de firewall con redundancia de zona. Esta característica está totalmente administrada, por lo que no es necesario iniciar ni validar los procesos de error de zona de disponibilidad.
- *Zonal:* Puede simular aspectos del error de una zona de disponibilidad deteniendo un firewall. Use este enfoque para probar cómo otros sistemas y equilibradores de carga controlan una interrupción en el firewall. Para más información, consulte Detener e iniciar Azure Firewall.

Resistencia a errores en toda la región

Azure Firewall es un servicio de una sola región. Si la región no está disponible, el recurso firewall tampoco está disponible.

Soluciones personalizadas de varias regiones para la resistencia

Para implementar una arquitectura de varias regiones, use firewalls independientes. Este enfoque requiere implementar un firewall independiente en cada región, enrutar el tráfico al firewall regional adecuado e implementar la lógica de conmutación por error personalizada. Considere los siguientes puntos:

- **Use Azure Firewall Manager** para la administración centralizada de directivas en varios firewalls. Use el método de Directiva de firewall para la administración centralizada de reglas en varias instancias de firewall.
- **Implemente el enrutamiento de tráfico** mediante Traffic Manager o Azure Front Door.

Para obtener una arquitectura de ejemplo que muestra arquitecturas de seguridad de red de varias regiones, consulte Equilibrio de carga de varias regiones mediante Traffic Manager, Azure Firewall y Application Gateway.

Resistencia al mantenimiento del servicio

Azure Firewall realiza periódicamente actualizaciones de servicio y otras formas de mantenimiento.

Puede configurar ventanas de mantenimiento diarias para alinear las programaciones de actualización con sus necesidades operativas. Para más información, consulte Configuración del mantenimiento controlado por el cliente para Azure Firewall.

Acuerdo de nivel de servicio

El contrato de nivel de servicio (SLA) para los servicios de Azure describe la disponibilidad esperada de cada servicio y las condiciones que la solución deberá cumplir para lograr esa expectativa de disponibilidad. Para obtener más información, consulte Acuerdos de Nivel de Servicio para servicios en línea.

Azure Firewall proporciona un Acuerdo de Nivel de Servicio de mayor disponibilidad para firewalls con redundancia de zona implementados en dos o más zonas de disponibilidad.

Contenido relacionado

- Información general de Azure Firewall
- Características de Azure Firewall
- Procedimientos recomendados de Azure Firewall para el rendimiento
- Firewall Manager

ⓘ **Nota:** El autor ha creado este artículo con ayuda de inteligencia artificial. Más información

Last updated on 10/02/2026

Confiabilidad en Azure Key Vault

Azure Key Vault es un servicio en la nube que proporciona un almacén seguro para secretos, como claves, contraseñas, certificados y otra información confidencial. Key Vault proporciona una gama de características de confiabilidad integradas para ayudar a garantizar que los secretos permanezcan disponibles.

Al usar Azure, relibilidad es una responsabilidad compartida. Microsoft proporciona una variedad de funcionalidades para admitir resistencia y recuperación. Es responsable de comprender cómo funcionan esas funcionalidades dentro de todos los servicios que usa y de seleccionar las funcionalidades que necesita para cumplir los objetivos empresariales y los objetivos de tiempo de actividad.

Este artículo describe cómo Key Vault es resistente a una variedad de posibles interrupciones y problemas, incluidos errores transitorios, interrupciones de zonas de disponibilidad e interrupciones en regiones. También se describe cómo puede usar copias de seguridad para recuperarse de otros tipos de problemas, características de recuperación para evitar la eliminación accidental y resaltar cierta información clave sobre el contrato de nivel de servicio (SLA) de Key Vault.

Recomendaciones de implementación de producción para la confiabilidad

En el caso de las cargas de trabajo de producción, se recomienda que usted:

- ✓ Use almacenes de claves de nivel Estándar o Premium.
- ✓ Active el borrado suave y la protección contra purga para evitar la eliminación accidental o malintencionada.
- ✓ En el caso de las cargas de trabajo críticas, considere la posibilidad de implementar estrategias de varias regiones que se describen en esta guía.

Introducción a la arquitectura de confiabilidad

Para garantizar una alta durabilidad y disponibilidad de las claves, secretos y certificados si se produce un error de hardware o una interrupción de red, Key Vault proporciona varias capas de redundancia para mantener la disponibilidad durante los siguientes eventos:

- Errores de hardware
- Interrupciones de red
- Desastres localizados
- Actividades de mantenimiento

De forma predeterminada, Key Vault consigue redundancia mediante la replicación del key vault y su contenido dentro de la región.

Si la región tiene una región emparejada y esa región emparejada está en la misma geografía que la región primaria, el contenido también se replica en la región emparejada. Este enfoque garantiza una alta durabilidad de las claves y secretos, que protege frente a errores de hardware, interrupciones de red o desastres localizados.

Resistencia a errores transitorios

Los errores transitorios son errores breves e intermitentes en los componentes. Se producen con frecuencia en un entorno distribuido como la nube y son una parte normal de las operaciones. Los errores transitorios se corrigen después de un breve período de tiempo. Es importante que las aplicaciones puedan controlar errores transitorios, normalmente mediante el reintento de solicitudes afectadas.

Todas las aplicaciones hospedadas en la nube deben seguir las instrucciones de control de errores transitorios Azure cuando se comunican con cualquier API, bases de datos y otros componentes hospedados en la nube. Para más información, vea Recomendaciones para gestionar errores temporales.

Para controlar los errores transitorios que puedan producirse, las aplicaciones cliente deben implementar lógica de reintento cuando interactúan con Key Vault. Puede usar los siguientes procedimientos recomendados:

- Use el SDK de Azure, que normalmente incluye mecanismos de reintento integrados.
- Implemente políticas de reintento con retroceso exponencial si sus clientes se conectan directamente a Key Vault.
- Almacene en caché los secretos en memoria siempre que sea posible para reducir las solicitudes directas a Key Vault.
- Supervise si hay errores de limitación porque superar los límites del servicio Key Vault provoca la limitación.

Si usa Key Vault en escenarios de alto rendimiento, considere la posibilidad de distribuir las operaciones entre varios almacenes de claves para evitar límites. Tenga en cuenta las instrucciones específicas de Key Vault para los siguientes escenarios:

- Un escenario de alto rendimiento es aquel que se aproxima o supera los límites de service para las operaciones de Key Vault, como 200 operaciones por segundo para claves protegidas por software.
- Para cargas de trabajo de alto rendimiento, divida el tráfico de Key Vault entre varios vaults y regiones distintas.
- Un límite aplicado a toda la suscripción para todos los tipos de transacciones es cinco veces el límite individual de la bóveda de claves.
- Use un almacén independiente para cada dominio de seguridad o disponibilidad. Por ejemplo, si tiene cinco aplicaciones en dos regiones, considere la posibilidad de usar 10 almacenes.
- Para las operaciones de clave pública, como el cifrado, el ajuste y la comprobación, realice estas operaciones localmente almacenando en caché el material de clave pública.

Para obtener más información, consulte la guía de limitación de Key Vault.

Resistencia a errores de zona de disponibilidad

Availability zones son grupos de centros de datos físicamente independientes dentro de una región de Azure. Cuando una zona falla, los servicios pueden transferirse a una de las zonas restantes.

Key Vault proporciona automáticamente redundancia de zona en regiones que admiten zonas de disponibilidad. Esta redundancia proporciona alta disponibilidad dentro de una región sin necesidad de ninguna configuración específica.

Cuando una zona de disponibilidad deja de estar disponible, Key Vault redirige automáticamente las solicitudes a otras zonas de disponibilidad correctas para garantizar una alta disponibilidad.

Requisitos

- **Compatibilidad con regiones:** Key Vault permite la redundancia de zona de forma predeterminada en todas las regiones de Azure que admiten zonas de disponibilidad.

- Todas las SKU de Key Vault, Estándar y Premium, admiten el mismo nivel de disponibilidad y resistencia. No hay ningún requisito específico del nivel para lograr resiliencia de zona.

Cost

No hay ningún costo adicional asociado a la redundancia de zona en Key Vault. Los precios se basan en la SKU, estándar o Premium, y el número de operaciones realizadas.

Comportamiento cuando todas las zonas están en buen estado

En esta sección se describe qué esperar cuando los almacenes de claves están en una región que tiene zonas de disponibilidad y todas las zonas de disponibilidad están operativas:

- **Enrutamiento de tráfico entre zonas:** Key Vault administra automáticamente el enrutamiento del tráfico entre zonas de disponibilidad. Durante las operaciones normales, las solicitudes se distribuyen entre zonas de forma transparente.

- **Replicación de datos entre zonas:** Los datos de Key Vault se replican sincrónicamente entre zonas de disponibilidad en regiones que las admiten. Esta replicación garantiza que las claves, los secretos y los certificados sigan siendo coherentes y estén disponibles incluso si una zona deja de estar disponible.

Comportamiento durante un fallo de zona

En la sección siguiente se describe qué esperar cuando los almacenes de claves están en una región que tiene zonas de disponibilidad y una o varias zonas de disponibilidad no están disponibles:

- **Detección y respuesta:** El servicio Key Vault es responsable de detectar errores de zona y responder automáticamente a ellos. No es necesario realizar ninguna acción ante un fallo de zona.

- **Notificación**: Microsoft no le notifica automáticamente cuando una zona está inactiva. Sin embargo, puede usar Azure Resource Health para supervisar el estado de un recurso individual y puede configurar Resource Health alertas para notificarle problemas. También puede usar Azure Service Health para comprender el estado general del servicio, incluidos los errores de zona, y puede configurar alertas de Service Health para notificarle problemas.

- **Solicitudes activas:** durante un error de zona, es posible que la zona afectada no controle las solicitudes en curso, lo que requiere que las aplicaciones cliente vuelvan a intentarlas.

Las aplicaciones cliente deben seguir los procedimientos transitorios de control de errores para asegurarse de que pueden reintentar solicitudes si se produce un error de zona.

- **Pérdida de datos esperada:** No se espera ninguna pérdida de datos durante un error de zona debido a la replicación sincrónica entre zonas.
- **Tiempo de inactividad esperado:** En el caso de las operaciones de lectura, debe haber un tiempo de inactividad mínimo o nulo durante un error de zona. Las operaciones de escritura pueden experimentar una falta de disponibilidad temporal mientras el servicio se ajusta al error de zona. Se espera que las operaciones de lectura permanezcan disponibles durante fallos de zona.
- **Traffic rerouting:** Key Vault vuelve a enrutar automáticamente el tráfico de la zona afectada a zonas correctas sin necesidad de intervención del cliente.

Recuperación de zona

Cuando se recupera la zona de disponibilidad afectada, Key Vault restaura automáticamente las operaciones en esa zona. La plataforma Azure administra completamente este proceso y no requiere ninguna intervención del cliente.

Resistencia a errores en toda la región

Los recursos de Key Vault se despliegan en una única región de Azure. Si la región deja de estar disponible, el almacén de claves tampoco está disponible. Sin embargo, hay enfoques que puede usar para ayudar a garantizar la resistencia a las interrupciones de la región. Estos enfoques dependen de si el almacén de claves está en una región emparejada o no emparejada y en sus requisitos y configuración específicos.

Conmutación por error administrada por Microsoft en una región emparejada

Key Vault admite la replicación administrada por Microsoft y el failover para los almacenes de claves en la mayoría de las regiones emparejadas. El contenido del almacén de claves se replica automáticamente dentro de la región y, de forma asincrónica, en la región emparejada. Este enfoque garantiza una alta durabilidad de las claves y los secretos. En el caso poco probable de un error prolongado en la región, Microsoft podría iniciar una conmutación por error regional del almacén de claves.

Las regiones siguientes no admiten la replicación administrada por Microsoft ni la conmutación por error entre regiones:

- Sur de Brasil
- Sudeste de Brasil
- Oeste de EE. UU. 3
- Cualquier región que no tenga una región emparejada

> **Importante**
>
> Microsoft inicia la conmutación por error administrada por Microsoft. Es probable que se produzca después de un retraso significativo y se realice con el mejor esfuerzo. También hay algunas excepciones a este proceso. La conmutación por error de almacenes de claves puede producirse en un momento diferente del tiempo de conmutación por error de otros servicios de Azure.
>
> Si necesita ser resistente a las interrupciones de regiones, considere la posibilidad de usar una de las soluciones personalizadas de varias regiones para lograr resistencia.

También puede usar la función de copia de seguridad y restauración para replicar el contenido de la bóveda en otra región de su elección.

Consideraciones

- **Tiempo de inactividad:** aunque la conmutación por error está en curso, es posible que el almacén de claves no esté disponible durante unos minutos.

- **Solo lectura después de la conmutación por error:** después de la conmutación por error, el almacén de claves se convierte en de solo lectura y solo admite acciones limitadas. No se pueden cambiar las propiedades del almacén de claves mientras se trabaja en la región secundaria y no se pueden modificar las configuraciones de directiva de acceso y firewall mientras se trabaja en la región secundaria.

  Cuando el almacén de claves está en modo de solo lectura, solo se admiten las siguientes operaciones:
  - Enumerar certificados
  - Obtención de certificados
  - Enumerar secretos
  - Obtención de secretos

- Enumerar claves
- Obtener claves y sus propiedades
- Cifrar
- Descifrar
- Envoltorio
- Desempaquetar
- Verify
- Firma
- Backup

Cost

No hay ningún costo adicional para las funcionalidades de replicación multiregión integradas de Key Vault.

Comportamiento cuando todas las regiones están en buen estado

En la sección siguiente se describe qué esperar cuando un almacén de claves se encuentra en una región que admite la replicación administrada por Microsoft y la conmutación por error y la región primaria está operativa:

- **Enrutamiento de tráfico entre regiones:** durante las operaciones normales, todas las solicitudes se enrutan a la región primaria donde se implementa el almacén de claves.

- **Replicación de datos entre regiones:** Key Vault replica los datos de forma asincrónica en la región emparejada. Al realizar cambios en el contenido de la bóveda de claves, esos cambios se confirman primero en la región primaria y, a continuación, se replican en la región secundaria.

Comportamiento durante una falla de región

En la sección siguiente se describe qué esperar cuando un almacén de claves se encuentra en una región que admite la replicación administrada por Microsoft y la conmutación por error y hay una interrupción en la región primaria:

- **Detección y respuesta:** Microsoft puede decidir realizar una conmutación por error si se pierde la región primaria. Este proceso puede tardar varias horas después de la pérdida de la región primaria o más tiempo en algunos escenarios. Es posible que la conmutación por error de almacenes de claves no se produzca al mismo tiempo que otros servicios de Azure.

- **Notificación**: Microsoft no le notifica automáticamente cuando una zona está inactiva. Sin embargo, puede usar Azure Resource Health para supervisar el estado de un recurso individual y puede configurar Resource Health alertas para notificarle problemas. También puede usar Azure Service Health para comprender el estado general del servicio, incluidos los errores de zona, y puede configurar alertas de Service Health para notificarle problemas.

- **Solicitudes activas:** Durante una conmutación por error de región, las solicitudes activas pueden producir errores y las aplicaciones cliente deben reintentarlas una vez completada la conmutación por error.

- **Pérdida de datos esperada:** Puede haber alguna pérdida de datos si los cambios no se replican en la región secundaria antes de que se produzca un error en la región primaria.

- **Tiempo de inactividad esperado:** durante una interrupción importante de la región primaria, es posible que el almacén de claves no esté disponible durante varias horas o hasta que Microsoft inicie la conmutación por error a la región secundaria.

  Si usa Private Link para conectarse al almacén de claves, la conexión puede tardar hasta 20 minutos en restablecerse después de la conmutación por error de la región.

- **Reenrutamiento del tráfico:** Una vez completada la conmutación por error de una región, las solicitudes se enrutan automáticamente a la región emparejada sin necesidad de intervención del cliente.

Soluciones de varias regiones personalizadas para la resistencia

Hay escenarios en los que las funcionalidades de conmutación por error entre regiones administradas por Microsoft de Key Vault no son adecuadas:

- El almacén de claves está en una región no emparejada.

- El almacén de claves está en una región emparejada que no admite la replicación y conmutación por error administradas por Microsoft en el Sur de Brasil, el Sudeste de Brasil y el Oeste de EE. UU. 3.

- Los objetivos de tiempo de actividad empresarial no están satisfechos por el tiempo de recuperación o la pérdida de datos que proporciona la conmutación por error entre regiones administrada por Microsoft.

- Debe conmutar por error a una región que no esté emparejada con la región principal.

Puede diseñar una solución de conmutación por error personalizada entre regiones siguiendo estos pasos:

1. Cree almacenes de claves independientes en diferentes regiones.
2. Use la funcionalidad de copia de seguridad y restauración para mantener secretos coherentes entre regiones.
3. Implemente la lógica de nivel de aplicación para conmutar por error entre almacenes de claves.

Copias de seguridad y restauración

Key Vault puede realizar copias de seguridad y restaurar secretos, claves y certificados individuales. Las copias de seguridad están pensadas para proporcionarle una copia offline de sus secretos en el improbable caso de perder acceso al almacén de claves.

Tenga en cuenta los siguientes factores clave relacionados con la funcionalidad de copia de seguridad:

- Las copias de seguridad crean blobs cifrados que no se pueden descifrar fuera de Azure.
- Las copias de seguridad solo se pueden restaurar en un almacén de claves dentro de la misma suscripción de Azure y geografía de Azure.
- Hay una limitación de realizar una copia de seguridad de no más de 500 versiones anteriores de una clave, un secreto o un objeto de certificado.
- Las copias de seguridad son instantáneas de un momento específico y no se actualizan automáticamente cuando cambian las claves.

Para la mayoría de las soluciones, no debe confiar exclusivamente en copias de seguridad. En su lugar, utilice las otras capacidades descritas en esta guía para apoyar los requisitos de resiliencia. Sin embargo, las copias de seguridad protegen contra algunos riesgos que otros métodos no cubren, como el borrado accidental de secretos específicos. Para obtener más información, consulte Key Vault backup.

Características de recuperación

Key Vault proporciona dos características clave de recuperación para evitar la eliminación accidental o malintencionada:

- **Eliminación temporal:** cuando está habilitada, la eliminación temporal permite recuperar almacenes y objetos eliminados durante un período de retención configurable. Este período es un valor predeterminado de 90 días. Piense en la eliminación temporal como una papelera de reciclaje para los recursos del almacén de claves.
- **Protección de purga:** cuando está habilitada, la protección de purga evita la eliminación permanente del almacén de claves y sus objetos hasta que transcurre el período de retención. Esta protección evita que los actores malintencionados destruyan permanentemente los secretos.

Se recomienda encarecidamente ambas características para entornos de producción. Para más información, consulte eliminación suave y protección contra purga en la documentación de gestión de recuperación de Key Vault.

Acuerdo de nivel de servicio

El acuerdo de nivel de servicio (SLA) para Azure servicios describe la disponibilidad esperada de cada servicio y las condiciones que la solución debe cumplir para lograr esa expectativa de disponibilidad. Para obtener más información, vea SLAs for servicios en línea.

Contenido relacionado

- Copia de seguridad de Key Vault
- Administración de recuperación de Key Vault
- Confiabilidad en Azure

ⓘ**Nota:** El autor ha creado este artículo con ayuda de inteligencia artificial. Más información

---

Last updated on 13/08/2026

Confiabilidad en HSM administrado de Azure Key Vault

Azure Key Vault HSM administrado es un servicio en la nube totalmente administrado, de alta disponibilidad, de un solo inquilino y compatible con estándares que le permite proteger las claves criptográficas de las aplicaciones en la nube mediante módulos de seguridad de hardware (HSM) que se validan en FIPS 140-3 nivel 3. Managed HSM proporciona una gama de características de confiabilidad integradas para ayudar a garantizar que las claves permanezcan disponibles.

Cuando se usa Azure, la confiabilidad es una responsabilidad compartida. Microsoft proporciona una variedad de capacidades para apoyar la resiliencia y la recuperación. Es responsable de comprender cómo funcionan esas funcionalidades dentro de todos los servicios que usa y de seleccionar las funcionalidades que necesita para cumplir los objetivos empresariales y los objetivos de tiempo de actividad.

En este artículo se describe cómo HSM administrado es resistente a varias posibles interrupciones y problemas, incluidos errores transitorios, errores de partición y interrupciones de región. También se describe cómo usar copias de seguridad y el dominio de seguridad para la recuperación ante desastres, cómo las características de recuperación protegen contra la eliminación accidental e información clave sobre el contrato de nivel de servicio (SLA) de HSM administrado.

Recomendaciones de implementación de producción

En el caso de las cargas de trabajo de producción, se recomienda que usted:

- ✓ Descargue y almacene de forma segura el dominio de seguridad inmediatamente después de aprovisionar el HSM administrado. Necesita el dominio de seguridad para la recuperación ante desastres.
- ✓ Establezca un cuórum de varias personas para el dominio de seguridad con al menos tres poseedores de claves.
- ✓ Habilite la protección de purga para evitar la eliminación accidental o malintencionada.
- ✓ Implemente copias de seguridad periódicas en una cuenta de Azure Storage y use el almacenamiento con redundancia geográfica en regiones admitidas.

- ✓ Habilite la replicación en varias regiones para cargas de trabajo críticas que requieran un Acuerdo de Nivel de Servicio superior.

Introducción a la arquitectura de confiabilidad

Cuando se usa HSM administrado, se implementa una *instancia*, que a veces se *denomina grupo*.

La arquitectura de HSM administrado está diseñada para lograr alta disponibilidad y durabilidad.

- **Aislamiento de un solo inquilino:** Cada instancia de HSM administrada se dedica a un único cliente y consta de un clúster de varias particiones HSM que están aisladas criptográficamente.
- **Particiones con redundancia triple:** Un grupo de HSM administrado consta de tres particiones HSM con equilibrio de carga distribuidas entre bastidores independientes dentro de un centro de datos. Esta distribución proporciona redundancia frente a errores de hardware y garantiza que la pérdida de un único componente, como la fuente de alimentación o el conmutador de red de un bastidor, no afecta a todas las particiones.
- **Computación confidencial:** Cada instancia de servicio se ejecuta en un entorno de ejecución de confianza (TEE) que usa enclaves de Intel SGX. El personal de Microsoft, incluidas las personas con acceso físico a los servidores, no puede acceder a su material criptográfico de claves.
- **Recuperación automática:** Si un error de hardware u otro problema afecta a una de las tres particiones, el servicio vuelve a generar automáticamente la partición afectada en hardware correcto sin intervención del cliente y sin exponer secretos.

Para obtener información sobre cómo Managed HSM implementa estas funcionalidades, consulte Soberanía clave, disponibilidad, rendimiento y escalabilidad en HSM administrado.

Dominio de seguridad

El dominio de seguridad es un componente crítico de HSM administrado con fines de recuperación ante desastres. Es un blob cifrado que contiene todas las credenciales necesarias para recompilar una instancia de HSM administrada desde cero, incluida la clave de propietario de la partición, las credenciales de partición, la clave de ajuste de datos y una copia de seguridad inicial del HSM.

> ⓘ **Importante**

Sin el dominio de seguridad, no es posible la recuperación ante desastres. Microsoft no tiene ninguna manera de recuperar el dominio de seguridad y no puede acceder a las claves sin ella.

Los dominios de seguridad son una parte fundamental de la seguridad y confiabilidad del HSM administrado. Se recomienda seguir estos procedimientos recomendados:

- **Genere claves de forma segura:** En entornos de producción, genere los pares de claves RSA que protegen el dominio de seguridad en un entorno con disponibilidad inalámbrica, como un HSM local o una estación de trabajo aislada.
- **Almacenar sin conexión:** Almacene claves de dominio de seguridad en unidades USB cifradas u otro almacenamiento sin conexión, con cada recurso compartido de claves en un dispositivo independiente en ubicaciones geográficas independientes.
- **Establecer un cuórum de varias personas:** Utilice al menos tres custodios de claves para evitar que una sola persona tenga acceso a todas las claves del cuórum y para evitar depender de una sola persona.

Para obtener más información, consulte Dominio de seguridad en la información general sobre el HSM administrado.

Resistencia a errores transitorios

Los errores transitorios son errores breves e intermitentes en los componentes. Se producen con frecuencia en un entorno distribuido como la nube y son una parte normal de las operaciones. Los errores transitorios se corrigen después de un breve período de tiempo. Es importante que las aplicaciones puedan controlar errores transitorios, normalmente mediante el reintento de solicitudes afectadas.

Todas las aplicaciones hospedadas en la nube deben seguir las instrucciones de control de errores transitorios de Azure cuando se comunican con cualquier API, bases de datos y otros componentes hospedados en la nube. Para obtener más información, consulte Recomendaciones para controlar errores transitorios.

Cuando se usan servicios de Azure que se integran con HSM administrado, esos servicios controlan los errores transitorios automáticamente.

Si compila aplicaciones personalizadas que se integran con HSM administrado, tenga en cuenta los procedimientos recomendados siguientes para controlar los errores transitorios que puedan

producirse:

- **Use los SDK proporcionados por Microsoft** para Azure Key Vault, que incluyen mecanismos de reintento integrados. Los SDK están disponibles para .NET, Python y JavaScript.
- **Implemente la lógica de reintento**, incluido el retroceso exponencial, para cualquier código que interactúe directamente con HSM administrado.
- **Reduzca el número de dependencias directas en el HSM administrado.** Almacene en caché los resultados de la operación criptográfica cuando sea posible para reducir las solicitudes directas a HSM administrado. Realice operaciones de clave pública, como el cifrado, el ajuste y la comprobación, localmente mediante el almacenamiento en caché del material de clave pública. La realización de las operaciones localmente reduce la dependencia del HSM administrado y reduce la probabilidad de errores transitorios que interrumpen estas operaciones.

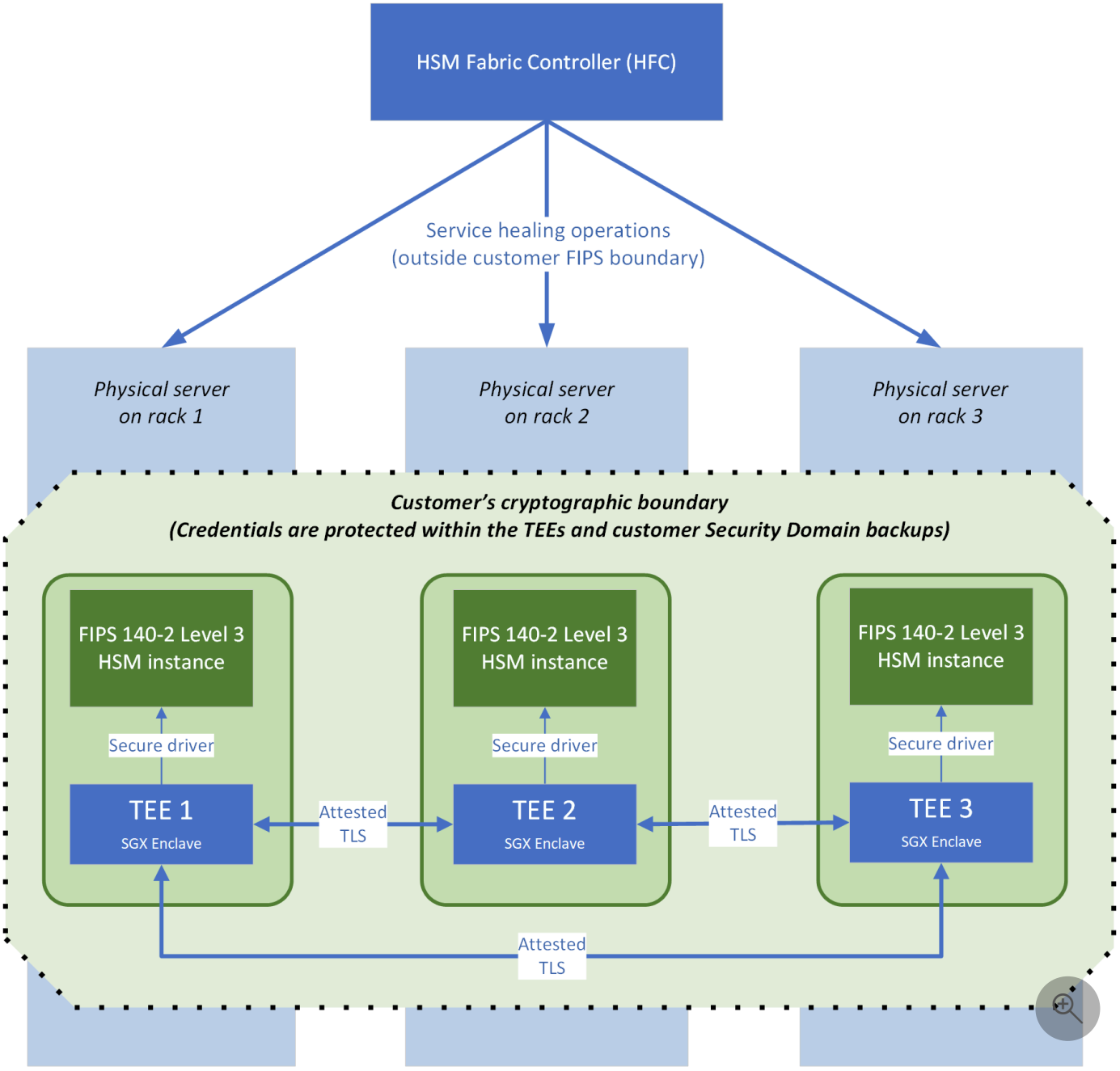
Si usa Managed HSM en escenarios de alto rendimiento, Managed HSM no limita las operaciones criptográficas. Utiliza su hardware HSM al máximo de su capacidad. Cada instancia de HSM administrada tiene tres particiones. Durante las operaciones de mantenimiento o recuperación, es posible que una partición no esté disponible. Para la planificación de la capacidad, suponga que hay dos particiones disponibles. Si necesita un rendimiento garantizado, planee como si solo una partición estuviera disponible. Supervise la métrica Disponibilidad de HSM administrada para comprender el estado del servicio.

Para escalar el cifrado de grandes volúmenes de datos, use una jerarquía de claves. Almacene solo la clave de cifrado de claves (KEK) en HSM administrado y úsela para encapsular las claves de cifrado de datos de nivel inferior almacenadas en otro lugar del almacenamiento seguro de claves.

Para obtener más información sobre las pruebas comparativas de rendimiento y el planeamiento de la capacidad, consulte Azure guía de escalado de HSM administrado.

Resistencia a errores de partición

HSM administrado logra una alta disponibilidad a través de su arquitectura triple redundante, donde cada grupo de HSM consta de tres particiones HSM distribuidas entre bastidores de servidores independientes dentro de un centro de datos. Esta distribución de nivel de bastidor proporciona redundancia frente a errores de hardware localizados.



Cuando se producen errores de hardware o interrupciones localizadas, el HSM administrado redirige automáticamente las solicitudes a particiones correctas y vuelve a generar particiones afectadas a través de un proceso denominado *recuperación del servicio confidencial*. Las particiones fallidas se reconstruyen automáticamente en hardware en buen estado mediante TLS con atestación y enclaves Intel SGX para proteger los secretos durante la recuperación.

Cost

La alta disponibilidad integrada en HSM administrado no agrega ningún costo adicional. Los precios se basan en el número de grupos de HSM y en el número de operaciones que se realizan. Para obtener más información, consulte los precios de Azure Managed HSM.

Comportamiento cuando todas las particiones están en buen estado

En esta sección se describe qué esperar cuando los grupos de HSM administrados están operativos y no hay particiones disponibles.

- **Enrutamiento del tráfico:** HSM administrado administra automáticamente el enrutamiento de tráfico entre sus tres particiones. Durante las operaciones normales, distribuye de forma transparente las solicitudes entre particiones.
- **Replicación de datos:** HSM administrado replica de forma sincrónica todos los datos, incluidas las claves, las asignaciones de roles y las directivas de control de acceso, en las tres particiones. Este enfoque garantiza la coherencia y la disponibilidad incluso si una partición deja de estar disponible.

Comportamiento durante un error de partición

En esta sección se describe qué esperar cuando una o varias particiones no están disponibles.

- **Detección y respuesta:** El servicio HSM administrado detecta errores de partición y responde automáticamente a ellos. No es necesario realizar ninguna acción durante un fallo de partición.
- **Solicitudes activas:** Durante un error de partición, las solicitudes en curso a la partición afectada pueden producir un error y requerir que las aplicaciones cliente vuelvan a intentarlas. Para minimizar los efectos de las interrupciones de particiones, las aplicaciones cliente deben seguir los procedimientos transitorios de control de errores.
- **Pérdida de datos esperada:** No se espera ninguna pérdida de datos durante un error de partición debido a la replicación sincrónica entre particiones.
- **Tiempo de inactividad esperado:** Para las operaciones de lectura y la mayoría de las operaciones criptográficas, debe haber un tiempo de inactividad mínimo o sin tiempo de inactividad durante un error de partición. Las particiones sanas restantes siguen atendiendo solicitudes.
- **Reenrutamiento del tráfico:** HSM administrado vuelve a enrutar automáticamente el tráfico de la partición afectada a particiones correctas sin necesidad de intervención del cliente.

Recuperación de particiones

Cuando se recupera la partición afectada, Managed HSM restaura automáticamente las operaciones mediante el restablecimiento confidencial del servicio. Este proceso:

1. Crea una nueva instancia de servicio en hardware en buen estado.
2. Establece una conexión TLS atestada con la partición principal.
3. Intercambia de forma segura las credenciales y el material criptográfico.
4. Asegura los datos de servicio a una nueva CPU.

La plataforma Azure administra completamente este proceso y no requiere ninguna intervención del cliente.

Resistencia a errores de zona de disponibilidad

La alta disponibilidad en Managed HSM se basa en la distribución a nivel de bastidor dentro de un centro de datos, no en el despliegue explícito en zonas de disponibilidad. Cada partición se ejecuta en un servidor independiente en un bastidor diferente, que protege frente a errores de nivel de bastidor, como problemas de fuente de alimentación o conmutador de red.

Para protegerse frente a interrupciones en todo el centro de datos o en toda la zona de disponibilidad, use uno de los enfoques descritos en Resistencia a errores de toda la región.

Resistencia a errores en toda la región

Los recursos de HSM administrados se implementan en una sola región de Azure. Si la región deja de estar disponible, el HSM administrado tampoco está disponible. Sin embargo, hay enfoques que puede usar para ayudar a garantizar la resistencia a las interrupciones de la región.

Replicación en varias regiones

Managed HSM admite la replicación opcional de varias regiones, que puede usar para ampliar un grupo de HSM administrado desde una región de Azure (región *primaria*) a una segunda región Azure (la *región extendida*). Al configurar esta característica:

- Ambas regiones están activas y pueden atender solicitudes.
- El material clave, los roles y los permisos se replican automáticamente entre regiones.
- Azure Traffic Manager enruta las solicitudes a la región disponible más cercana.
- El Acuerdo de Nivel de Servicio combinado aumenta.

Requirements

- **Compatibilidad con regiones:** Todas las regiones que admiten HSM administrado se pueden usar como regiones principales. No hay ninguna dependencia en los emparejamientos de regiones de Azure.

  HSM administrado no admite todas las regiones como regiones extendidas. Para más información, consulte Compatibilidad con regiones de Azure.

- **Número máximo de regiones:** Puede agregar una región extendida, para un máximo de dos regiones en total.

Cost

La replicación en varias regiones incurre en facturación adicional porque la región extendida consume un segundo grupo de HSM. Para obtener más información, consulte los precios de Azure Managed HSM.

Configuración de la replicación en varias regiones

- **Agregue una región extendida:** Para obtener más información sobre cómo agregar una región extendida a una región primaria existente, consulte Extensión de un HSM principal a una región extendida.

  La extensión de un HSM administrado a otra región puede tardar hasta 30 minutos.

- **Quite una región extendida:** Para más información sobre cómo quitar una región extendida de una región primaria existente, consulte Quitar una región extendida del HSM principal.

Comportamiento cuando todas las regiones están en buen estado

En esta sección se describe qué esperar al configurar la replicación en varias regiones y ambas regiones están operativas.

- **Enrutamiento del tráfico:** Todas las regiones pueden atender solicitudes. Azure Traffic Manager enruta las solicitudes a la región con la proximidad geográfica más cercana o la latencia más baja.

  Si usa Azure Private Link para acceder a HSM administrado, configure puntos de conexión privados en ambas regiones para un enrutamiento óptimo durante la conmutación por error. Para obtener más información, consulte Comportamiento de Private Link con replicación en varias regiones.

- **Replicación de datos:** Todos los cambios en las claves, las definiciones de roles y las asignaciones de roles se replican de forma asincrónica en la región extendida en un plazo de seis minutos. Espere seis minutos después de crear o actualizar una clave antes de usarla en la región extendida.

Comportamiento durante una falla de región

En esta sección se describe qué esperar al configurar la replicación en varias regiones y se produce una interrupción en una de las regiones de réplica.

- **Detección y respuesta:** Azure Traffic Manager detecta la región incorrecta y enruta las solicitudes futuras a la región correcta. Los registros DNS tienen un tiempo de vida (TTL) de cinco segundos, aunque los clientes que almacenan en caché las búsquedas de DNS podrían experimentar tiempos de conmutación por error ligeramente más prolongados.

- **Notificación:** Microsoft no le notifica automáticamente cuando una región está inactiva. Sin embargo, puede usar Azure Service Health para comprender el estado general del servicio, incluidos los errores de región, y puede configurar alertas de Service Health para notificarle problemas.

- **Peticiones activas:** Las peticiones en curso a la región afectada pueden fallar y requerir que se vuelvan a intentar.

- **Pérdida de datos esperada:** Es posible que los cambios realizados en un plazo de seis minutos antes del error de la región no se repliquen en la región extendida. Esos cambios podrían perderse si la región primaria es irrecuperable.

- **Tiempo de inactividad esperado:** Las operaciones de lectura y escritura permanecen disponibles en la región correcta durante la conmutación por error.

  Las aplicaciones cliente que están cerca de la región incorrecta pueden seguir siendo dirigidas a esa región hasta que se actualicen los registros DNS, pero esta actualización tiene lugar en aproximadamente cinco segundos. Para minimizar el tiempo de conmutación por error, los clientes deben evitar el almacenamiento en caché de las búsquedas DNS por más tiempo del TTL del registro DNS.

- **Reenrutar:** Azure Traffic Manager vuelve a enrutar automáticamente las solicitudes a la región correcta.

Recuperación de regiones

Cuando se recupera la región afectada, HSM administrado reanuda automáticamente las operaciones. Azure Traffic Manager comienza a enrutar las solicitudes a ambas regiones de nuevo en función de la proximidad.

Prueba de fallos de región

HSM administrado gestiona completamente el enrutamiento del tráfico, la conmutación por error y la recuperación tras el fallo para los fallos regionales, por lo que no es necesario validar los procesos de fallos regionales ni proporcionar ninguna intervención adicional.

Soluciones personalizadas de varias regiones para la resistencia

Si la replicación en varias regiones no es adecuada para sus necesidades, puede implementar la recuperación ante desastres manual. Este enfoque requiere:

- **Dominio de seguridad** del HSM de origen.
- **Las claves privadas** (al menos el número de cuórum) que cifran el dominio de seguridad.
- **Una copia de seguridad completa de HSM reciente** del HSM de origen.

Para realizar la recuperación ante desastres:

1. Cree una nueva instancia de HSM administrada en otra región.
2. Active el modo de recuperación de dominio de seguridad y cargue el dominio de seguridad.
3. Realice una copia de seguridad del nuevo HSM (necesario antes de la restauración).
4. Restaure la copia de seguridad desde el HSM de origen.

> **Importante**
>
> El nuevo HSM tiene un nombre y un URI de punto de conexión de servicio diferentes. Debe actualizar la configuración de la aplicación para usar la nueva ubicación.

Para procedimientos detallados de recuperación ante desastres, consulte Recuperación ante desastres de HSM administrado.

Copias de seguridad y restauración

Managed HSM admite copias de seguridad y restauración completas de todas las claves, versiones, atributos, etiquetas y asignaciones de roles. Las copias de seguridad se almacenan en una cuenta de Azure Storage. Si la región lo admite, se recomienda realizar una copia de

seguridad de HSM administrado en una cuenta de Azure Storage que tenga habilitado el almacenamiento con redundancia geográfica (GRS).

El HSM cifra las copias de seguridad mediante claves criptográficas asociadas al dominio de seguridad de HSM. Solo puede restaurar copias de seguridad en un HSM con el mismo dominio de seguridad.

Managed HSM no admite la programación de copias de seguridad, pero puede crear su propio programador usando un servicio como Azure Functions o Azure Automation.

Mientras una copia de seguridad está en curso, es posible que el HSM no funcione en un rendimiento completo porque algunas particiones están ocupadas realizando la operación de copia de seguridad.

Para obtener procedimientos detallados de copia de seguridad y restauración, consulte Copia de seguridad y restauración completas.

Resistencia a la eliminación accidental

Managed HSM proporciona dos características clave de recuperación para evitar la eliminación accidental o malintencionada.

- **Eliminación temporal:** Los HSM y las claves eliminados no se eliminan definitivamente de inmediato. Permanecen recuperables durante un período de retención configurable de 7 a 90 días (valor predeterminado: 90 días). La eliminación temporal siempre está habilitada y no se puede deshabilitar.

  > ⓘ **Nota:**
  >
  > Los recursos HSM administrados eliminados temporalmente continúan incurr en facturación hasta que se purgan.

- **Protección de purga:** Cuando está habilitada, evita la eliminación permanente del HSM administrado y sus claves hasta que transcurre el período de retención. Nadie puede deshabilitar ni invalidar la protección de purga, incluido Microsoft.

Se recomienda encarecidamente habilitar la protección de purga para entornos de producción. Para más información, consulte Managed HSM soft-delete and purge protection (Eliminación temporal de HSM administrado y protección de purga).

Resistencia al mantenimiento del servicio

HSM administrado controla el mantenimiento del servicio, incluidas las actualizaciones de firmware, la aplicación de revisiones y la recuperación de hardware, sin intervención del cliente. Durante el mantenimiento:

- Es posible que el servicio deje de estar disponible temporalmente las particiones mientras aplica las actualizaciones.
- Al menos dos de tres particiones permanecen disponibles durante el mantenimiento rutinario.
- Las aplicaciones cliente deben implementar lógica de reintento para controlar breves interrupciones.

El proceso de autorreparación del servicio confidencial garantiza que el servicio nunca exponga secretos durante las operaciones de mantenimiento.

Acuerdo de nivel de servicio

El acuerdo de nivel de servicio (SLA) para Azure servicios describe la disponibilidad esperada de cada servicio y las condiciones que la solución debe cumplir para lograr esa expectativa de disponibilidad. Para obtener más información, consulte Acuerdos de Nivel de Servicio para servicios en línea.

HSM administrado proporciona un Acuerdo de Nivel de Servicio de disponibilidad estándar para implementaciones de una sola región. Habilitar la replicación multirregional aumenta el tiempo de actividad general previsto, porque las peticiones se pueden atender desde cualquiera de las dos regiones si una deja de estar disponible.

Contenido relacionado

- ¿Qué es HSM administrado de Azure Key Vault?
- Habilitación de la replicación en varias regiones
- Recuperación ante desastres de HSM administrado
- Copia de seguridad y restauración completas
- Descripción general del dominio de seguridad en HSM administrado
- Eliminación temporal y protección contra la purga de HSM administrado
- Guía de escalado de HSM administrado de Azure
- ¿Qué es Azure documentación de confiabilidad?

ⓘ**Nota:** El autor ha creado este artículo con ayuda de inteligencia artificial. Más información

---

Last updated on 24/06/2026

Confiabilidad en Azure Backup

Azure Backup es un servicio de Azure integrado que protege de forma segura las cargas de trabajo locales y en la nube. La copia de seguridad puede escalar su protección en múltiples cargas de trabajo y proporciona integración nativa con las cargas de trabajo de Azure, incluidas las máquinas virtuales (VM) en Azure, SAP HANA en máquinas virtuales en Azure, SQL en máquinas virtuales en Azure, Azure Files, Azure Blob Storage, Azure Data Lake Storage, discos administrados de Azure, volúmenes de Azure Elastic SAN y Azure Kubernetes Service (AKS). No es necesario administrar la automatización ni la infraestructura, escribir scripts ni aprovisionar el almacenamiento.

Al usar Azure, relibilidad es una responsabilidad compartida. Microsoft proporciona una variedad de capacidades para admitir resiliencia y recuperación. Es responsable de comprender cómo funcionan esas funcionalidades dentro de todos los servicios que usa y de seleccionar las funcionalidades que necesita para cumplir los objetivos empresariales y los objetivos de tiempo de actividad.

En este artículo se describe cómo la copia de seguridad puede ser resistente a una variedad de posibles interrupciones y problemas, incluidos errores transitorios, interrupciones de zona de disponibilidad y interrupciones de regiones. También resalta cierta información clave sobre el acuerdo de nivel de servicio (SLA) de copia de seguridad.

> **Nota:**
>
> En este artículo se describe cómo el propio servicio Backup es resistente a varios problemas y cómo puede hacer que sea más resistente. No explica cómo usar backup para proteger las máquinas virtuales, los datos u otros recursos. Para obtener información sobre cómo usar Backup, consulte **Introducción a la copia de seguridad**.

Recomendaciones de implementación de producción para la confiabilidad

Para realizar copias de seguridad de las cargas de trabajo de producción, se recomienda configurar la bóveda de las siguientes maneras:

- ✔ Use el almacenamiento con redundancia de zona (ZRS) como nivel de redundancia mínimo para las copias de seguridad. ZRS replica las copias de seguridad en varias zonas de disponibilidad para poder restaurar las copias de seguridad durante una interrupción de la zona de disponibilidad.
- ✔ Si usa almacenamiento con redundancia geográfica (GRS) para replicar las copias de seguridad en una región de Azure emparejada, habilite la restauración entre regiones (CRR) para orígenes de datos admitidos. CRR le permite restaurar las copias de seguridad en la región emparejada en cualquier momento.

En las secciones siguientes de este artículo se proporcionan más detalles sobre estas configuraciones.

> ⓘ **Nota:**
>
> Estas recomendaciones de redundancia de almacenamiento se aplican a ubicaciones en las que se replican las copias de seguridad, no al servicio Backup o a los recursos de los que realiza una copia de seguridad. La protección de copia de seguridad y la redundancia de almacenamiento se complementan entre sí. Las copias de seguridad protegen contra la pérdida de datos y la redundancia protege frente a errores de infraestructura.

Para obtener una lista de otras recomendaciones para copia de seguridad, incluidas las recomendaciones centradas en la confiabilidad, consulte Copia de seguridad de cargas de trabajo en la nube y cargas de trabajo locales a la nube.

Introducción a la arquitectura de confiabilidad

En esta sección se describen algunos de los aspectos importantes de cómo funciona el servicio que es más relevante desde una perspectiva de confiabilidad. En la sección se presenta la arquitectura lógica, que incluye algunos de los recursos y características que se implementan y usan. También se describe la arquitectura física, que proporciona detalles sobre cómo funciona el servicio en segundo plano.

Arquitectura lógica

La copia de seguridad puede realizar copias de seguridad y restaurar una variedad de *orígenes de datos*. Las copias de seguridad se configuran de forma diferente en función del origen de datos con el que trabaje. Los orígenes de datos siguientes son comunes:

- máquinas virtuales de Azure
- Varias bases de datos
- cuentas de Blob Storage
- Clústeres de AKS
- Servidores locales a través del agente de Microsoft Azure Recovery Services (MARS)

La copia de seguridad almacena los datos de copia de seguridad en *almacenes*. Las bóvedas son entidades de almacenamiento en línea en Azure que contienen datos, como copias de seguridad, puntos de recuperación y políticas de copia de seguridad. Los almacenes de Recovery Services y los almacenes de Backup son dos tipos de almacenes. Puede usar uno o ambos tipos en función de lo que necesite proteger. Para obtener una lista de los orígenes de datos que admite cada tipo de almacén, consulte Preguntas más frecuentes sobre los almacenes admitidos para la copia de seguridad y restauración.

Los trabajos representan la actividad de copia de seguridad o restauración de los datos. Los trabajos de copia de seguridad incluyen operaciones programadas o bajo demanda que copian sus datos desde el origen a la bóveda. Los trabajos de restauración incluyen operaciones que recuperan los datos del almacenamiento de copia de seguridad en una ubicación de destino. Cada trabajo tiene un identificador único y un seguimiento de estado para que pueda supervisar el progreso y solucionar problemas que se producen durante las operaciones de copia de seguridad y restauración. Tú también creas *directivas de copia de seguridad* asociadas a trabajos. Las directivas especifican la configuración como la programación de copia de seguridad y cuánto tiempo desea conservar los datos.

Bóvedas almacenan las directivas de copia de seguridad y la configuración junto con los metadatos de los trabajos, permitiéndole realizar un seguimiento de los trabajos y solucionar problemas.

Arquitectura física

Microsoft administra la infraestructura principal del servicio Backup. Esta infraestructura es responsable de la administración y el funcionamiento del servicio, incluidos los trabajos de desencadenamiento y supervisión.

Los respaldos se almacenan en la bóveda. Las bóvedas se construyen sobre Azure Storage. Las bóvedas replican automáticamente los datos de copias de seguridad, y la durabilidad y resistencia de la copia de seguridad dependen de la redundancia de almacenamiento de la bóveda.

- Almacenamiento con redundancia de ubicación (LRS) replica los datos dentro de su almacén en una o varias zonas de disponibilidad de Azure ubicadas en la región primaria de su elección. No puede elegir la zona de disponibilidad preferida, pero Azure puede mover o expandir cuentas de LRS entre zonas para mejorar el equilibrio de carga. No se garantiza que los datos se repartan entre zonas. Para obtener más información, consulte Introducción a las zonas de disponibilidad.

- ZRS y GRS proporcionan protecciones adicionales. En este artículo se describen estas opciones al detalle.

> ⓘ **Nota:**
>
> Algunas fuentes de datos admiten copias de seguridad de *nivel operativo*, que almacenan datos en otra ubicación en lugar del almacén. Por ejemplo, la **copia de seguridad de discos administrados de Azure** y las **copias de seguridad de AKS** admiten copias de seguridad de nivel operativo, que se almacenan en instantáneas de disco. En este artículo no se describe el almacenamiento de copia de seguridad de nivel operativo, pero puede aplicar las instrucciones de resistencia de este artículo a operaciones y flujos de trabajo de copia de seguridad para estos tipos de copia de seguridad.

Resistencia a errores transitorios

Los errores transitorios son errores breves e intermitentes en los componentes. Se producen con frecuencia en un entorno distribuido como la nube y son una parte normal de las operaciones. Los errores transitorios se corrigen después de un breve período de tiempo. Es importante que las aplicaciones puedan controlar errores transitorios, normalmente mediante el reintento de solicitudes afectadas.

Todas las aplicaciones hospedadas en la nube deben seguir las instrucciones de control de errores transitorios Azure cuando se comunican con cualquier API, bases de datos y otros componentes hospedados en la nube. Para obtener más información, consulte Recomendaciones para controlar errores transitorios.

Cuando se usa Backup, los flujos de trabajo de copia de seguridad y restauración son resistentes a errores intermitentes. El servicio vuelve a intentarlo automáticamente cuando encuentra errores transitorios de red o interrupciones temporales del servicio. Tú no configuras ninguna lógica de reintento. Si experimenta errores repetidos, consulte Solución de problemas de las operaciones de administración de la bóveda de copias de seguridad.

Resistencia a errores de zona de disponibilidad

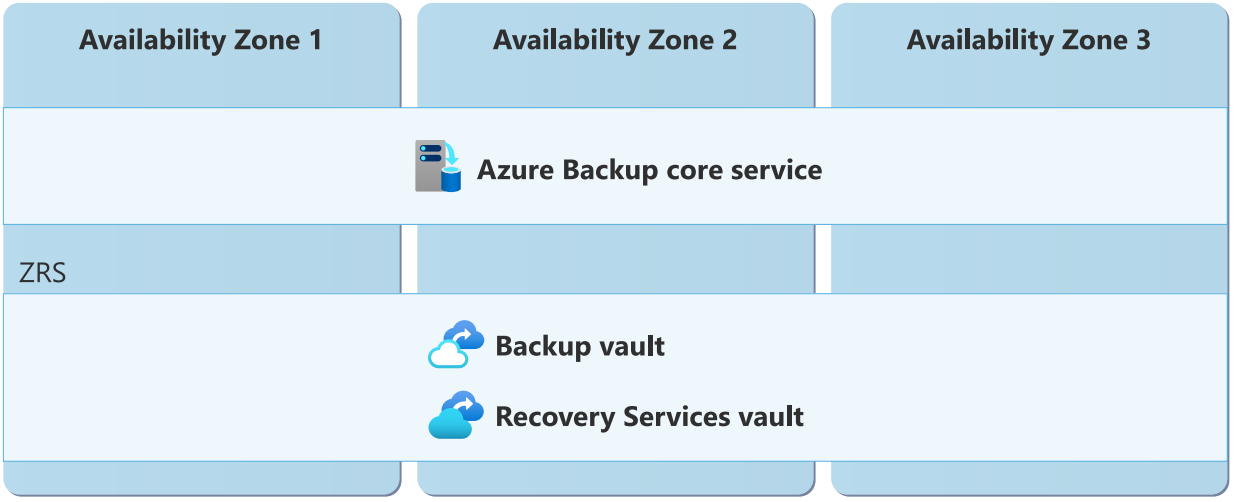
Availability zones son grupos de centros de datos físicamente independientes dentro de una región de Azure. Cuando una zona falla, los servicios pueden transferirse a una de las zonas restantes.

La copia de seguridad administra por separado la configuración de zona de disponibilidad del servicio y de los datos.

- **Servicio:** El servicio de Backup es automáticamente resistente a las zonas en regiones admitidas. Sin embargo, esta resistencia de zona integrada no se aplica a los datos de copia de seguridad.

- **Redundancia de almacenamiento de copia de seguridad:** Seleccione el nivel de redundancia que desea para los datos de copia de seguridad mediante la configuración del almacén de Recovery Services o el almacén de Backup. Si selecciona ZRS, las copias de los datos de copia de seguridad se almacenan automáticamente en varias zonas de disponibilidad en la región de Azure que use.

  Si no usa ZRS, los datos de copia de seguridad se consideran *no zonales* y pueden almacenarse en cualquier zona. Si alguna zona de la región tiene un problema, es posible que los datos de copia de seguridad no zonales no estén disponibles.



Requisitos

- **Compatibilidad con regiones:** El servicio es automáticamente resistente a zonas en todas las regiones que tienen zonas de disponibilidad. Los almacenes de ZRS se admiten en las mismas regiones.

- **Solo bóvedas nuevas:** Configure ZRS en su bóveda antes de la primera copia de seguridad.

Cost

Al habilitar ZRS para las copias de seguridad, se le cobra a una tarifa diferente de LRS debido a la sobrecarga adicional de replicación y almacenamiento. Para obtener más información, consulte Precios de copia de seguridad .

Configurar soporte de zonas de disponibilidad

- **Cree un nuevo almacén que use ZRS:** Configure la redundancia de almacenamiento al crear un almacén. Siga los diferentes pasos según el tipo de bóveda. Para obtener más información, consulte los artículos siguientes:
  - Crear y eliminar almacenes de copias de seguridad
  - Creación y configuración de un almacén de Recovery Services

- **Configuración de ZRS en almacenes existentes:** En el caso de los almacenes de Backup, configure la redundancia de almacenamiento al crear el almacén. Después de crear un almacén de Backup, la configuración está bloqueada y no se puede cambiar.

  En el caso de las bóvedas de Recovery Services, debe configurar la redundancia de almacenamiento antes de proteger cualquier carga de trabajo. Después de proteger una carga de trabajo, la configuración está bloqueada y no se puede cambiar.

  Puede crear una nueva bóveda configurada para usar ZRS y reasignar las cargas de trabajo a la nueva bóveda. Sin embargo, este enfoque requiere tiempo de inactividad. Para obtener más información, consulte Modificación de la configuración predeterminada. También es responsable de eliminar manualmente los puntos de recuperación existentes y otros datos porque las directivas de retención del almacén antiguo ya no se aplican. Para obtener más información, consulte Eliminación de un almacén de Backup o Eliminación de un almacén de Recovery Services.

Comportamiento cuando todas las zonas están en buen estado

En esta sección se describe qué esperar al configurar bóvedas para ZRS cuando todas las zonas están operativas.

- **Operación entre zonas:** Los trabajos de copia de seguridad se ejecutan en la infraestructura replicada entre zonas. Azure administra los trabajos de la infraestructura en cualquier zona.

- **Replicación de datos entre zonas:** ZRS replica los datos de copia de seguridad entre zonas. La replicación se produce de forma sincrónica, lo que significa que varias zonas reconocen cada operación de escritura antes de que se complete.

Comportamiento durante un fallo de zona

En esta sección se describe qué esperar al configurar bóvedas para ZRS cuando se produce una interrupción en una de las zonas.

- **Detección y respuesta:** Para el servicio de copia de seguridad, Microsoft es responsable de detectar fallos en las zonas de disponibilidad y responder. No es necesario hacer nada para iniciar una conmutación por error de la zona.

  > ⓘ **Importante**
  >
  > En el caso de los datos o recursos que no están disponibles debido a una falla de la zona, usted es responsable de detectar la falla y tomar medidas de recuperación, incluida la restauración de copias de seguridad en una zona operativa.

- **Notification**: Microsoft no le notifica automáticamente cuando una zona está inactiva. Sin embargo, puede usar Azure Resource Health para supervisar el estado de un recurso individual y puede configurar Resource Health alertas para notificarle problemas. También puede usar Azure Service Health para comprender el estado general del servicio, incluidos los errores de zona, y puede configurar alertas de Service Health para notificarle problemas.

- **Solicitudes activas:** El comportamiento de las tareas activas depende de qué zona falla.

  - Para cualquier origen de datos en la zona de disponibilidad fallida, el fallo de la zona hace que los orígenes de datos no estén disponibles. Los trabajos activos podrían pausarse o fallar.

  - En el caso de los orígenes de datos en zonas de disponibilidad correctas que ejecutan trabajos activos, una pequeña cantidad de tiempo de inactividad, normalmente unos segundos, puede producirse mientras la plataforma cambia a zonas de disponibilidad correctas para el servicio Backup.

- **Pérdida de datos esperada:** La cantidad esperada de pérdida de datos también se conoce como *objetivo de punto de recuperación (RPO)*. El RPO de los datos de copia de seguridad depende de varios factores, incluida la programación de las copias de seguridad. En general, para una interrupción de zona, no se espera ninguna pérdida de datos de copia de seguridad porque todos los datos se replican de manera sincrónica entre las zonas.

- **Tiempo de inactividad esperado:** La cantidad esperada de tiempo de inactividad también se conoce como *objetivo de tiempo de recuperación (RTO)*. El RTO es diferente para cada uno de los escenarios siguientes:

  - En el caso de los orígenes de datos de la zona de disponibilidad fallida, es posible que estos no estén disponibles hasta que la zona se recupere. Es posible que los trabajos de copia de seguridad no se ejecuten hasta que el origen de datos esté disponible de nuevo. El RTO no está definido.

  - En el caso de los orígenes de datos en zonas de disponibilidad correctas, una pequeña cantidad de tiempo de inactividad, normalmente unos segundos, puede producirse mientras la plataforma cambia a zonas de disponibilidad correctas para el servicio Backup.

- **Redistribución:** Las ejecuciones posteriores de trabajos usan automáticamente la infraestructura en zonas saludables siempre y cuando los orígenes de datos estén disponibles.

  Es responsable de restaurar la copia de seguridad en la infraestructura de una zona saludable y reconfigurar los balanceadores de carga, los clientes y otros sistemas para redirigir el tráfico a una infraestructura saludable en la nueva zona.

Recuperación de zona

Cuando se recupera la zona de disponibilidad, Backup restaura automáticamente las operaciones en la zona de disponibilidad y vuelve a enrutar el tráfico entre las zonas como es normal. Los trabajos siguen ejecutándose y los datos permanecen disponibles.

Prueba de fallos de zona

La plataforma de respaldo gestiona el enrutamiento del tráfico, la replicación de datos, la conmutación por error y la recuperación después de fallos. Esta característica está totalmente administrada, por lo que no es necesario iniciar ni validar los procesos de error de zona de disponibilidad.

Resistencia a errores en toda la región

Backup admite la redundancia geográfica y la conmutación por error a través de GRS y CRR.

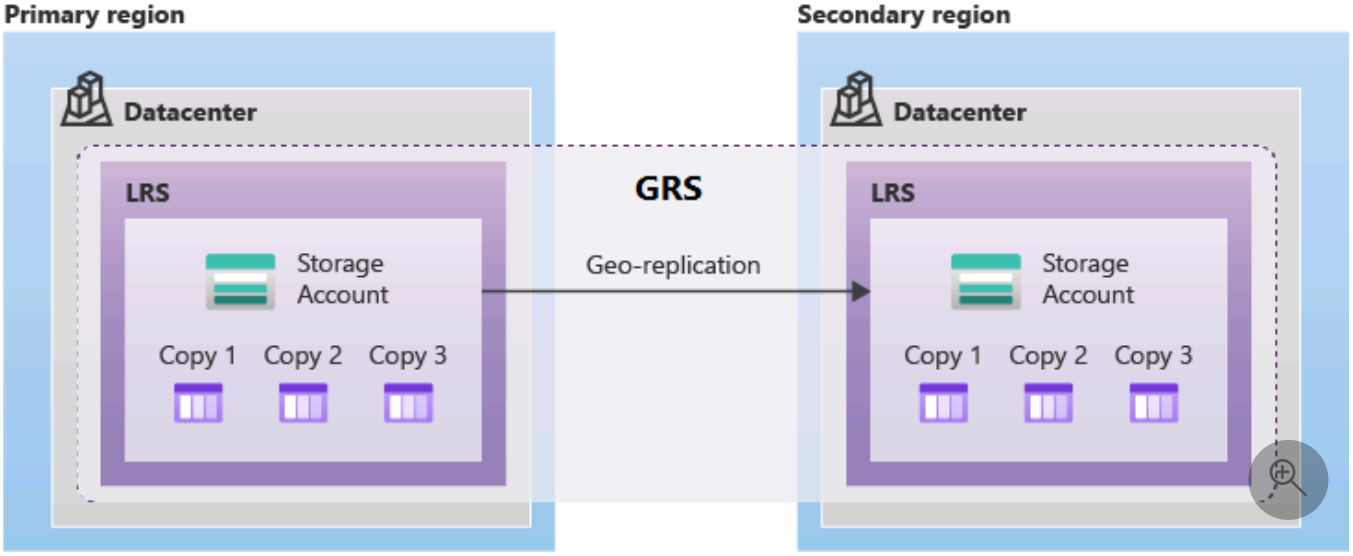
> ⓘ **Importante**
>
> GRS for Backup solo funciona en regiones Azure emparejadas.

Almacenamiento con redundancia geográfica y restauración entre regiones

Para lograr la redundancia regional de los datos de copia de seguridad, use Backup para replicar las copias de seguridad en una región emparejada Azure mediante GRS. GRS protege las copias de seguridad frente a interrupciones regionales.

La región en la que se implementa el almacén se llama *región primaria*. Los orígenes de datos deben encontrarse en la región primaria. No se pueden configurar copias de seguridad en una bóveda de otra región.

La región emparejada también se conoce como *región secundaria*.



Si no configura GRS y se produce una interrupción en la región de la bóveda, es posible que pueda acceder a la bóveda y ver los elementos de copia de seguridad. Sin embargo, sin redundancia regional, los datos de copia de seguridad subyacentes permanecen no disponibles para las operaciones de restauración.

Restauración cruzada de regiones

Al configurar GRS en un almacén, Microsoft pone a disposición las copias de seguridad en la región emparejada después de que se produzca una interrupción en la región primaria. Si el origen de datos admite CRR, puede restaurar desde puntos de recuperación de regiones secundarias incluso cuando no se produzca ninguna interrupción en la región primaria. CRR también le permite ejecutar simulacros para evaluar la resistencia frente a interrupciones regionales. Al activar CRR, Microsoft actualiza el almacenamiento de copia de seguridad de GRS al almacenamiento con redundancia geográfica con acceso de lectura (RA-GRS).

Requisitos

- **Soporte de región:** GRS para Backup solo funciona dentro de regiones emparejadas de Azure.
- **Solo bóvedas nuevas:** Debe configurar GRS en su bóveda antes de realizar la primera copia de seguridad.

Consideraciones

- **CRR:** Después de activar CRR, los elementos de copia de seguridad pueden tardar hasta 48 horas en estar disponibles en la región secundaria.

Cost

Las bóvedas GRS incurren en costos adicionales para la replicación entre regiones y para el almacenamiento en la región secundaria. La transferencia de datos entre regiones de Azure se cobra en función de las tasas estándar de ancho de banda entre regiones. La CRR se cobra a una tarifa distinta porque Microsoft actualiza el almacenamiento del cofre de GRS a RA-GRS. Para obtener más información, consulte Precios de copia de seguridad .

Configuración de la compatibilidad con varias regiones

- **Cree un nuevo almacén que use GRS y CRR:** Al crear un almacén, también debe configurar la redundancia de almacenamiento. Después de seleccionar GRS, opcionalmente puede habilitar CRR en el almacén. Los pasos que siga dependen del tipo de bóveda. Para obtener más información, consulte los artículos siguientes:
  - Crear y eliminar almacenes de copias de seguridad
  - Creación y configuración de un almacén de Recovery Services
- **Configuración de GRS y CRR en almacenes existentes:** En el caso de los almacenes de Backup, debe configurar la redundancia de almacenamiento al crear el almacén.

En el caso de las bóvedas de Recovery Services, debe configurar la redundancia de almacenamiento antes de proteger cualquier carga de trabajo. Una vez protegida una carga de trabajo, la configuración está bloqueada y no se puede cambiar.

Puede habilitar CRR en almacenes GRS existentes. Después de habilitar CRR, no se puede deshabilitar.

Comportamiento cuando todas las regiones están en buen estado

En esta sección se describe qué esperar al configurar bóvedas para el uso de GRS cuando todas las regiones están operativas.

- **Operación entre regiones:** Las copias de seguridad siempre se completan en la región primaria, que es la región donde están desplegados el almacén y el origen de datos.
- **Replicación de datos entre regiones:** Cuando configuras el vault para usar GRS, las copias de seguridad se confirman primero en la región primaria mediante LRS. Después de finalizar correctamente en la región primaria, los datos se replican de forma asincrónica en la región secundaria. La región secundaria usa LRS para almacenar datos. Los datos de copia de seguridad pueden tardar hasta 12 horas en replicarse desde la región primaria a la región secundaria.

Comportamiento durante una falla de región

En esta sección se describe qué esperar al configurar bóvedas para usar GRS en caso de que ocurra una interrupción en la región primaria.

- **Detección y respuesta:** Para los orígenes de datos que admiten CRR y donde CRR está habilitado en la bóveda, puede iniciar su propio CRR en la región emparejada en cualquier momento, incluido durante una caída de la región o un desastre. Es responsable de detectar la interrupción y realizar acciones de recuperación, incluida la restauración de copias de seguridad en una región correcta.

  En todos los demás escenarios, los datos replicados en la región secundaria están disponibles para restaurarse en la región secundaria solo si Azure declara un desastre en la región primaria. Microsoft es responsable de declarar un desastre. La cantidad de tiempo que se tarda en declarar un desastre depende de la gravedad del incidente y del tiempo necesario para evaluar la situación. Microsoft normalmente declara un desastre solo después de un período de tiempo prolongado.

- **Notification**: Microsoft no le notifica automáticamente cuando una región está inactiva. Sin embargo:
  - Puede usar Azure Resource Health para supervisar el estado de un recurso individual y puede configurar Resource Health alertas para notificarle problemas.
  - Puede usar Azure Service Health para comprender el estado general del servicio, incluidos los errores de región, y puede configurar alertas de Service Health para notificarle los problemas.
- **Pérdida de datos esperada:** El RPO de los datos de copia de seguridad depende de varios factores, incluida la programación de copia de seguridad. En términos generales, para una interrupción regional, se puede esperar hasta 36 horas de pérdida de datos porque el RPO en la región primaria es de 24 horas, y puede tardar hasta 12 horas en replicar los datos de copia de seguridad de la región primaria a la región secundaria.
- **Tiempo de inactividad esperado:** El RTO es diferente para cada uno de los escenarios siguientes:
  - Es posible que los orígenes de datos y otros recursos de la región con errores no estén disponibles hasta que se recupere la región, por lo que el RTO no está definido.
  - Es posible que la copia de seguridad no pueda realizar operaciones de copia de seguridad o restauración en la región afectada hasta que la región se recupere, por lo que el RTO no está definido.
  - Si usa CRR, el RTO para iniciar la restauración de las copias de seguridad ya replicadas en la región emparejada es cero. Si no usa CRR, el RTO depende del tiempo que tarda Microsoft en declarar un desastre en la región afectada.
- **Redistribución:** No se pueden ejecutar trabajos de copia de seguridad mientras la región primaria está sin conexión. Puede restaurar datos en la bóveda, pero no puede agregar nuevos.

  Es responsable de restaurar la copia de seguridad en la infraestructura de la región emparejada y de volver a configurar equilibradores de carga, clientes y otros sistemas para redirigir el tráfico a una infraestructura correcta en la región emparejada.

Recuperación de regiones

Cuando se recupera la región primaria, Backup restaura automáticamente las operaciones en la región. Los trabajos se reanudan y los datos permanecen disponibles.

Prueba de fallos de región

Puede usar CRR para realizar una operación de restauración en la región emparejada. Puede usar este enfoque para comprobar la restauración y otros procesos de recuperación.

Resistencia a la pérdida de datos de copia de seguridad

La copia de seguridad proporciona dos características clave de recuperación para evitar la eliminación accidental o malintencionada de los datos de copia de seguridad:

- **La eliminación temporal** permite recuperar los objetos y almacenes eliminados durante un período de retención configurable. De forma predeterminada, este período es de 14 días, pero puede editarlo. Piensa en la eliminación suave como una papelera de reciclaje para tus copias de seguridad y bóvedas. Para obtener más información, consulte Seguro por defecto con la eliminación suave de Backup.

- **Los almacenes inmutables** pueden ayudarle a proteger los datos de copia de seguridad bloqueando las operaciones que podrían provocar la pérdida de puntos de recuperación. Puede bloquear la configuración del almacén inmutable para que sea irreversible. También puede usar el almacenamiento de escribir una vez, leer muchas veces (WORM) para los respaldos, para evitar que usuarios malintencionados deshabiliten la inmutabilidad y eliminen copias de seguridad. Para obtener más información, consulte almacén inmutable para copia de seguridad.

Acuerdo de nivel de servicio

El acuerdo de nivel de servicio (SLA) para Azure servicios describe la disponibilidad esperada de cada servicio y las condiciones que la solución debe cumplir para lograr esa expectativa de disponibilidad. Para obtener más información, vea SLAs for servicios en línea.

El Acuerdo de Nivel de Servicio de copia de seguridad cubre la disponibilidad del servicio para las operaciones de copia de seguridad y restauración. Para estar cubierto por el Acuerdo de Nivel de Servicio, debe reintentar los trabajos de copia de seguridad o restauración con errores al menos una vez cada 30 minutos.

Contenido relacionado

- Confiabilidad en Azure
- Reliability in Azure Virtual Machines
- Reliability in Azure Disk Storage
- Fiabilidad en Azure Site Recovery

ⓘ **Nota:** El autor ha creado este artículo con ayuda de inteligencia artificial. Más información

---

Last updated on 14/04/2026

Confiabilidad en Azure Blob Storage

Azure Blob Storage es una solución de almacenamiento de objetos para la nube desde Microsoft. Se diseñó para almacenar grandes cantidades de datos no estructurados, como texto, datos binarios, documentos, archivos multimedia y copias de seguridad de aplicaciones. Como servicio básico de Azure Storage, Blob Storage proporciona varias características de confiabilidad para asegurarse de que los datos permanezcan disponibles y duren durante eventos planeados y no planeados.

Cuando se usa Azure, la confiabilidad es una responsabilidad compartida. Microsoft proporciona una variedad de funcionalidades para admitir resistencia y recuperación. Es responsable de comprender cómo funcionan esas funcionalidades dentro de todos los servicios que usa y de seleccionar las funcionalidades que necesita para cumplir los objetivos empresariales y los objetivos de tiempo de actividad.

En este artículo se describe cómo hacer que Blob Storage sea resistente a una variedad de posibles interrupciones y problemas, incluidos errores transitorios, interrupciones de zona de disponibilidad y interrupciones de regiones. También se describe cómo puede usar copias de seguridad para recuperarse de otros tipos de problemas y se resalta cierta información clave sobre el acuerdo de nivel de servicio (SLA) de Blob Storage.

> ⓘ **Nota:**
>
> Blob Storage forma parte de la plataforma Azure Storage. Algunas de las capacidades de Blob Storage son comunes en muchos servicios de Azure Storage. En este artículo, se usa *Azure Storage* para hacer referencia a estas características.

Recomendaciones de implementación de producción

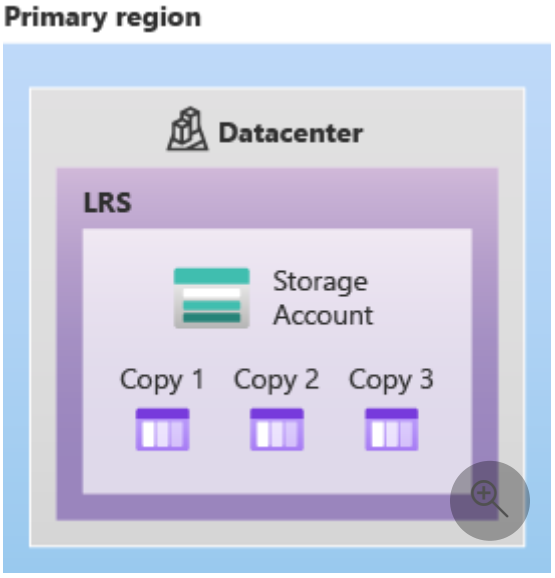
Para obtener información sobre cómo implementar Blob Storage para admitir los requisitos de confiabilidad de la solución y cómo afecta la confiabilidad a otros aspectos de la arquitectura, consulte Procedimientos recomendados de arquitectura para Blob Storage en el marco de arquitectura recomendada de Azure.

Introducción a la arquitectura de confiabilidad

Azure Storage proporciona varias opciones de redundancia para ayudarle a proteger los datos frente a distintos tipos de errores. Cada opción proporciona un nivel específico de redundancia

de datos, por lo que es posible elegir el nivel que mejor se adapte a los requisitos de la aplicación.

El almacenamiento con redundancia local (LRS) replica los datos de las cuentas de almacenamiento en una o varias zonas de disponibilidad de Azure ubicadas en la región primaria de su elección. Aunque no hay ninguna opción para elegir la zona de disponibilidad preferida, Azure puede mover o expandir cuentas LRS entre zonas para mejorar el equilibrio de carga. No hay ninguna garantía de que los datos se distribuirán entre zonas. Para obtener más información sobre las zonas de disponibilidad, consulte ¿Qué son las zonas de disponibilidad?.



El almacenamiento con redundancia de zona (ZRS), el almacenamiento con redundancia geográfica (GRS) y el almacenamiento con redundancia de zona geográfica (GZRS) proporcionan protecciones adicionales. En este artículo se describen estas opciones al detalle.

Resistencia a errores transitorios

Los errores transitorios son errores breves e intermitentes en los componentes. Se producen con frecuencia en un entorno distribuido como la nube y son una parte normal de las operaciones. Los errores transitorios se corrigen después de un breve período de tiempo. Es importante que las aplicaciones puedan controlar errores transitorios, normalmente mediante el reintento de solicitudes afectadas.

Todas las aplicaciones hospedadas en la nube deben seguir las instrucciones de control de errores transitorios de Azure cuando se comunican con cualquier API, bases de datos y otros componentes hospedados en la nube. Para obtener más información, consulte Recomendaciones para controlar errores transitorios.

Para administrar de forma eficaz los errores transitorios al usar Blob Storage, implemente las siguientes recomendaciones:

- **Utilice las bibliotecas cliente de Azure Storage**, que incluyen directivas de reintento integradas con retroceso exponencial y fluctuación. Los SDK de .NET, Java, Python y JavaScript controlan automáticamente los reintentos en caso de fallos transitorios. Para obtener más información sobre las opciones de configuración de reintentos, consulte Implementación de directivas de reintentos con .NET.

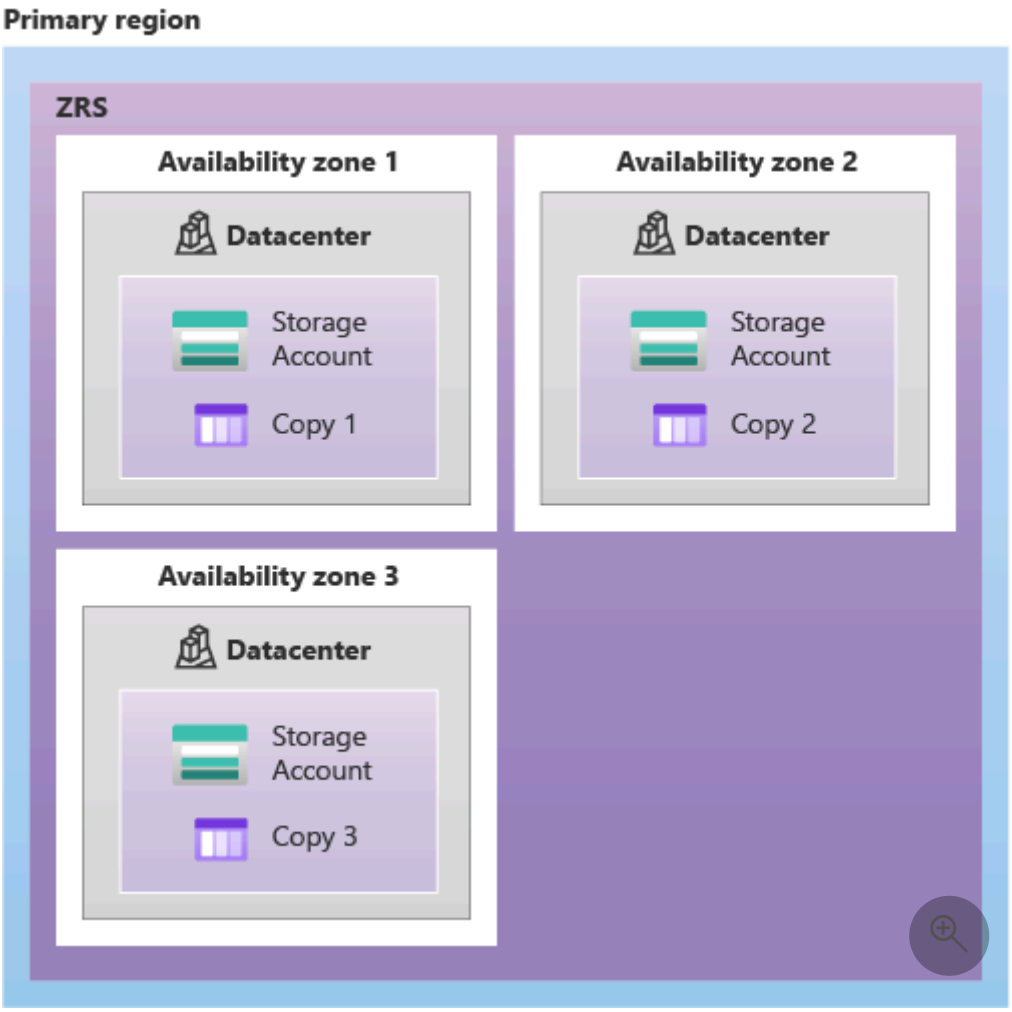
- **Configure los valores de tiempo de espera adecuados** para sus operaciones con blobs en función del tamaño de los blobs y las condiciones de la red. Los blobs más grandes requieren tiempos de espera más largos, pero las operaciones más pequeñas pueden usar valores más cortos para detectar errores rápidamente.

Resistencia a errores de zona de disponibilidad

Las zonas de disponibilidad son grupos físicamente independientes de centros de datos dentro de una región de Azure. Cuando una zona falla, los servicios pueden transferirse a una de las zonas restantes.

Blob Storage proporciona una sólida compatibilidad con zonas de disponibilidad a través de configuraciones de ZRS que distribuyen automáticamente los datos entre varias zonas de disponibilidad dentro de una región. A diferencia del almacenamiento con redundancia local (LRS), ZRS garantiza que Azure replique de forma sincrónica los datos del blob en varias zonas de disponibilidad. ZRS garantiza que sus datos sigan estando accesibles incluso si se produce un corte en una zona.

La redundancia de zona se habilita en el nivel de la cuenta de almacenamiento y se aplica a todos los contenedores de blobs dentro de esa cuenta. No se pueden establecer distintos niveles de redundancia para contenedores individuales. La configuración de redundancia se aplica a toda la cuenta de almacenamiento. Cuando una zona de disponibilidad experimente una interrupción, Azure Storage enrutará automáticamente las solicitudes a zonas correctas sin necesidad de intervención de usted o de la aplicación.



Requisitos

- **Compatibilidad con regiones:** Puede implementar cuentas de Azure Storage con redundancia de zona en cualquier región que admita zonas de disponibilidad.
- **Tipos de cuenta de almacenamiento:** La redundancia de zona está disponible tanto para los tipos de cuentas de almacenamiento v2 de uso general estándar como para las cuentas de almacenamiento Blob en bloques Premium. Los blobs en bloques, los blobs en anexos y los blobs en páginas admiten configuraciones con redundancia de zona, pero el tipo de cuenta de almacenamiento que se use determinará qué funcionalidades estarán disponibles. Para obtener más información, consulte Tipos de cuentas de almacenamiento compatibles.

Cost

Cuando se habilita el almacenamiento con redundancia de zona (ZRS), se cobra una tarifa diferente a la del almacenamiento redundante local (LRS) debido a la replicación adicional y los gastos generales de almacenamiento.

Para más información, consulte Precios de Blob Storage.

Configurar soporte de zonas de disponibilidad

- **Crear una cuenta de Blob Storage con redundancia de zona.** Para crear una cuenta de almacenamiento con ZRS, consulte Creación de cuentas de almacenamiento y seleccione **ZRS**, **almacenamiento con redundancia de zona geográfica (GZRS)** o **almacenamiento con redundancia geográfica con acceso de lectura (RA-GZRS)** como opción de redundancia durante la creación de la cuenta.

- **Cambiar el tipo de replicación.** Para obtener información sobre cómo cambiar una cuenta de almacenamiento existente a almacenamiento con redundancia de zona (ZRS) y sobre las opciones y requisitos de configuración, consulte Cambiar la forma en que se replica una cuenta de almacenamiento.

- **Desactivar la redundancia de zona.** Convierta las cuentas de ZRS de nuevo a una configuración no zonal, como el almacenamiento con redundancia local (LRS), utilizando el mismo proceso de cambio de configuración de redundancia.

Comportamiento cuando todas las zonas están en buen estado

En esta sección se describe qué se puede esperar cuando se configura una cuenta de almacenamiento de blobs para la redundancia de zona y todas las zonas de disponibilidad están operativas.

- **Enrutamiento del tráfico entre zonas:** Azure Storage con almacenamiento con redundancia de zona (ZRS) distribuye automáticamente las solicitudes entre los clústeres de almacenamiento en varias zonas de disponibilidad. La distribución del tráfico es transparente para las aplicaciones y no requiere ninguna configuración del lado cliente.

- **Replicación de datos entre zonas:** todas las operaciones de escritura en ZRS se replican de forma sincrónica en todas las zonas de disponibilidad dentro de la región. Al cargar o modificar datos, la operación no se considera completa hasta que los datos se hayan replicado correctamente en todas las zonas de disponibilidad. Esta replicación sincrónica garantiza una fuerte coherencia y cero pérdida de datos durante los errores de zona.

Comportamiento durante un fallo de zona

En esta sección se describe qué puede ocurrir cuando se configura una cuenta de almacenamiento de blobs para ZRS y se produce una interrupción en la zona de disponibilidad.

- **Detección y respuesta:** Microsoft detecta automáticamente errores de zona e inicia procesos de recuperación. No se requiere ninguna acción del cliente para las cuentas de

almacenamiento con redundancia de zona (ZRS). Si una zona deja de estar disponible, Azure lleva a cabo actualizaciones de red como el repunto del sistema de nombres de dominio (DNS).

- **Notificación**: Microsoft no le notifica automáticamente cuando una zona está inactiva. Sin embargo, puede usar Azure Resource Health para supervisar el estado de un recurso individual y puede configurar alertas de Resource Health para notificarle problemas. También puede usar Azure Service Health para comprender el estado general del servicio, incluidos los errores de zona, y puede configurar alertas de Service Health para notificarle problemas.
- **Solicitudes activas:** Es posible que las solicitudes en curso se descarten durante el proceso de recuperación y se deban reintentar. Las aplicaciones deben implementar lógica de reintento para controlar estas interrupciones temporales.
- **Pérdida de datos esperada:** No se produce ninguna pérdida de datos durante los errores de zona porque los datos se replican sincrónicamente en varias zonas antes de completar las operaciones de escritura.
- **Tiempo de inactividad previsto:** es posible que se produzca un breve periodo de inactividad, normalmente de unos segundos, durante la recuperación automática, ya que el tráfico se redirige a zonas correctas. Cuando diseñe aplicaciones para ZRS, siga las prácticas de manejo de fallos transitorios, incluyendo la implementación de directivas de reintento con retroceso exponencial.
- **Reenrutamiento del tráfico:** si una zona de disponibilidad se quedase sin conexión, Azure iniciará los cambios de red, como la reconfiguración del punto de conexión del sistema de nombres de dominio (DNS). Estas actualizaciones garantizan que el tráfico se redireccione a las zonas de disponibilidad correctas restantes. El servicio mantiene toda la funcionalidad mediante las zonas supervivientes y no requiere intervención del cliente.

Recuperación de zona

Cuando se recupera la zona de disponibilidad con errores, Azure Storage restaura automáticamente las operaciones normales en todas las zonas de disponibilidad. El servicio garantiza automáticamente la coherencia de los datos mediante la sincronización de las operaciones que se produjeron durante el período de interrupción.

Prueba de fallos de zona

El usar el almacenamiento con redundancia de zona (ZRS), Azure Storage administrará automáticamente la replicación, el enrutamiento del tráfico y las respuestas ante caídas de

zonas. Dado que esta característica está totalmente administrada, no es necesario iniciar ni validar los procesos de error de zona de disponibilidad.

Resistencia a errores en toda la región

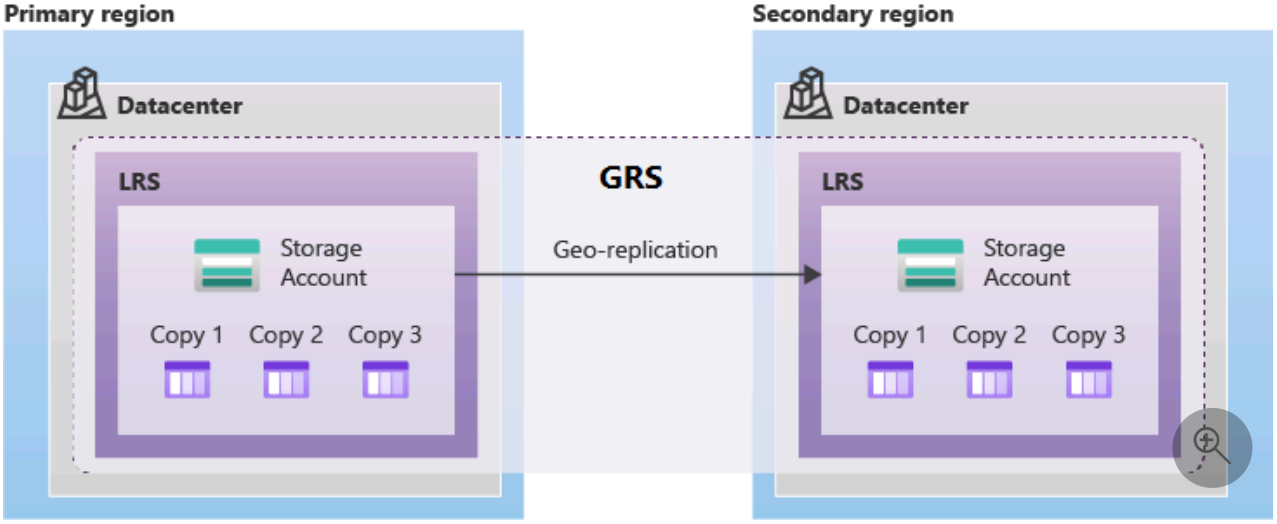
Azure Storage, que incluye Azure Blob Storage, Azure Files, Azure Table Storage y Azure Queue Storage, ofrece una amplia gama de capacidades de redundancia geográfica y conmutación por error para adaptarse a diferentes requisitos.

> ⓘ **Importante**
>
> El GRS solo funciona dentro de regiones emparejadas de Azure. Si la región de la cuenta de almacenamiento no está emparejada, considere la posibilidad de usar las soluciones personalizadas de varias regiones para lograr resistencia.

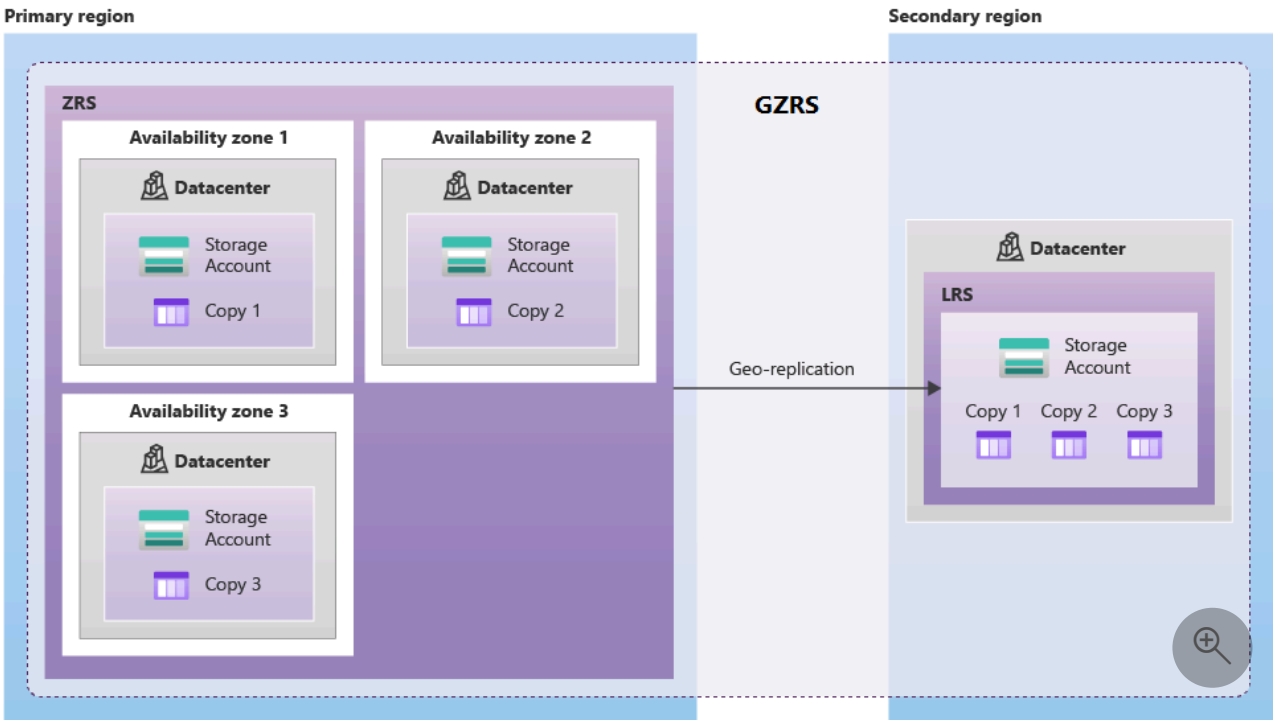
Almacenamiento con redundancia geográfica para regiones emparejadas

Azure Storage ofrece varios tipos de GRS en regiones emparejadas. Independientemente del tipo de GRS que utilice, los datos de la región secundaria siempre se replicarán mediante almacenamiento con redundancia local (LRS). Este enfoque proporciona protección contra fallos de hardware dentro de la región secundaria.

- GRS proporciona soporte para conmutaciones por error planificadas y no planificadas a la región emparejada de Azure cuando se produce una interrupción en la región principal. GRS replica de forma asincrónica los datos de la región primaria a la región emparejada.



- El almacenamiento con redundancia de zona geográfica (GZRS) replica los datos en varias zonas de disponibilidad de la región principal y en la región emparejada.



- El almacenamiento con redundancia geográfica con acceso de lectura (RA-GRS) y el almacenamiento con redundancia de zona geográfica con acceso de lectura (RA-GZRS) amplían el almacenamiento con redundancia geográfica (GRS) y el almacenamiento con redundancia de zona geográfica (GZRS), con la ventaja adicional del acceso de lectura al punto de conexión secundario. Estas opciones son idóneas y están diseñadas para obtener alta disponibilidad en aplicaciones críticas para la empresa. En el improbable caso de que el punto de conexión principal experimente una interrupción, las aplicaciones configuradas para el acceso de lectura a la región secundaria pueden seguir funcionando.

Tipos de conmutación por error

Azure Storage admite tres tipos de conmutación por error para distintos escenarios.

- **Conmutación por error no planeada administrada por el cliente:** es responsable de iniciar la recuperación si hubiera un error de almacenamiento en toda la región principal.

- **Conmutación por error planeada administrada por el cliente:** Usted es responsable de iniciar la recuperación si otra parte de su solución sufre un fallo en la región primaria y necesita trasladar toda su solución a la región secundaria. Use una conmutación por error planeada cuando el almacenamiento permanezca operativo en la región primaria, pero debe realizar una conmutación por error de toda la solución a una región secundaria, como para los simulacros de recuperación ante desastres diseñados para garantizar el cumplimiento de los requisitos de auditoría.

- **Conmutación por error administrada por Microsoft:** en circunstancias excepcionales, Microsoft podría iniciar la conmutación por error para todas las cuentas de almacenamiento con redundancia geográfica (GRS) de una región. Sin embargo, la

conmutación por error administrada por Microsoft es un último recurso y se espera que solo se realice después de un período prolongado de interrupción. No se debe confiar en el failover administrado por Microsoft.

Las cuentas de GRS pueden utilizar cualquiera de estos tipos de conmutación por error. No es necesario preconfigurar una cuenta de almacenamiento para usar ninguno de los tipos de conmutación por error con antelación.

Requisitos

- **Compatibilidad con regiones:** Las configuraciones con redundancia geográfica de Azure Storage usan regiones emparejadas de Azure para la replicación de regiones secundarias. La región secundaria se determina automáticamente en función de la selección de la región principal y no se puede personalizar. Para obtener una lista completa de las regiones emparejadas de Azure, consulte Lista de regiones de Azure.

  Si la región de la cuenta de almacenamiento no está emparejada, considere la posibilidad de usar las soluciones personalizadas de varias regiones para lograr resistencia.

- **Tipos de cuenta de almacenamiento:** El almacenamiento con redundancia geográfica (GRS) y la conmutación por error y la conmutación por recuperación iniciadas por el cliente están disponibles en todas las regiones emparejadas de Azure que admiten cuentas de Azure Storage de uso general v2.

Consideraciones

Al implementar Blob Storage en varias regiones, tenga en cuenta los siguientes factores clave:

- **Latencia de replicación asincrónica:** la replicación de datos en la región secundaria es asincrónica, lo que significa que hay un retraso entre cuando los datos se escriben en la región primaria y cuando estarán disponibles en la región secundaria. Este retraso podría provocar una posible pérdida de datos si se produjese un fallo en la región principal antes de que se repliquen los datos recientes. La pérdida de datos se mide mediante el objetivo de punto de recuperación (RPO). Espere que el retraso de replicación sea inferior a 15 minutos, pero este tiempo es una estimación y no se garantiza.

  Compruebe la propiedad Última hora de sincronización para saber cuántos datos se podrían perder si su cuenta de almacenamiento sufriese una conmutación por error no planificada.

- **Acceso a la región secundaria:** con las configuraciones de almacenamiento con redundancia geográfica (GRS) y almacenamiento con redundancia de zona geográfica

(GZRS), la región secundaria no será accesible a las lecturas hasta que se produzca una conmutación por error.

Las configuraciones de almacenamiento con redundancia geográfica con acceso de lectura (RA-GRS) y almacenamiento con redundancia de zona geográfica con acceso de lectura (RA-GZRS) proporcionan acceso de lectura a la región secundaria durante las operaciones normales, pero debido a la latencia de replicación asincrónica, podrían devolver datos ligeramente obsoletos.

- **Limitaciones de las características:** algunas características de Azure Storage no son compatibles o tienen limitaciones cuando se utiliza almacenamiento con redundancia geográfica (GRS) o conmutación por error administrada por el cliente. Revise la compatibilidad de las característica antes de implementar la redundancia geográfica.

Cost

Las configuraciones de cuentas de Azure Storage multirregionales conllevan costes adicionales por la replicación entre regiones y el almacenamiento en la región secundaria. La transferencia de datos entre regiones de Azure se cobra en función de las tarifas estándar de ancho de banda entre regiones.

Para más información, consulte Precios de Blob Storage.

Configuración de la compatibilidad con varias regiones

- **Cree una nueva cuenta de almacenamiento con redundancia geográfica (GRS).** Para crear una cuenta de GRS, consulte Creación de una cuenta de almacenamiento y seleccione GRS, almacenamiento con redundancia geográfica con acceso de lectura (RA-GRS), almacenamiento con redundancia de zona geográfica (GZRS) o almacenamiento con redundancia de zona geográfica con acceso de lectura (RA-GZRS) durante la creación de la cuenta.

- **Habilite la redundancia geográfica en una cuenta de almacenamiento existente.** Para convertir una cuenta de almacenamiento existente en almacenamiento con redundancia geográfica (GRS), consulte Cambio de cómo se replica una cuenta de almacenamiento.

> ⚠ **Advertencia**
>
> Una vez que su cuenta se haya reconfigurado para la redundancia geográfica, es posible que transcurra un tiempo considerable antes de que los datos existentes en la nueva región principal se copien por completo a la nueva región secundaria.

Para evitar una pérdida importante de datos, compruebe el valor de la **propiedad Última hora de sincronización** antes de iniciar una conmutación por error no planificada. Para evaluar la posible pérdida de datos, compare la última hora de sincronización con la última hora en que se escribieron los datos en la nueva región principal.

- **Deshabilite la redundancia geográfica.** Convierta las cuentas de GRS en configuraciones de una sola región, como el almacenamiento con redundancia local (LRS) o el almacenamiento con redundancia local (ZRS) con el mismo proceso de cambio de configuración de redundancia.

Comportamiento cuando todas las regiones están en buen estado

En esta sección se describe qué esperar cuando una cuenta de almacenamiento está configurada para redundancia geográfica y todas las regiones están operativas.

- **Enrutamiento del tráfico entre regiones:** Azure Storage utiliza un enfoque activo-pasivo en el que todas las operaciones de escritura y la mayoría de las operaciones de lectura se dirigen a la región principal.

  En las configuraciones de almacenamiento con redundancia geográfica con acceso de lectura (RA-GRS) y almacenamiento con redundancia de zona geográfica con acceso de lectura (RA-GZRS), las aplicaciones pueden leer opcionalmente desde la región secundaria accediendo al punto de conexión secundario. Este enfoque requiere una configuración explícita de la aplicación y no es automática. Además, debido al retraso de replicación asincrónica, los datos de la región secundaria podrían estar ligeramente obsoletos.

- **Replicación de datos entre regiones:** las operaciones de escritura se confirman primero en la región principal utilizando los siguientes tipos de redundancia configurados:
  - Almacenamiento con redundancia local (LRS) para almacenamiento con redundancia geográfica (GRS) y RA-GRS
  - Almacenamiento con redundancia de zona (ZRS) para el almacenamiento con redundancia de zona geográfica (GZRS) y RA-GZRS

  Tras completarse con éxito en la región principal, los datos se replican de forma asíncrona en la región secundaria, donde se almacenan mediante LRS.

  La naturaleza asíncrona de la replicación entre regiones significa que suele haber un retraso entre el momento en que los datos se escriben en la región principal y el momento en que están disponibles en la región secundaria. Supervise el tiempo de replicación utilizando la propiedad Última hora de sincronización.

Comportamiento durante una falla de región

En esta sección se describe qué esperar cuando una cuenta de almacenamiento está configurada para la redundancia geográfica y hay una interrupción en la región primaria.

- **Conmutación por error administrada por el cliente (no planificada):** utilice una conmutación por error no planificada cuando el almacenamiento en la región principal no esté disponible.
  - **Detección y respuesta:** En el caso poco probable de que su cuenta de almacenamiento no esté disponible en la región primaria, puede considerar iniciar un failover no planeado administrado por el cliente. Para tomar esta decisión, tenga en cuenta los siguientes factores:
    - Si Azure Resource Health muestra problemas al acceder a la cuenta de almacenamiento en su región principal
    - Si Microsoft le recomienda realizar una conmutación por error a otra región

    > ⚠ **Advertencia**
    >
    > Una conmutación por error no planeada puede **provocar la pérdida de datos**. Antes de iniciar una conmutación por error administrada por el cliente, decida si la restauración del servicio justifica el riesgo de pérdida de datos.

  - **Notificación**: Microsoft no le notifica automáticamente cuando una región está inactiva. Sin embargo:
    - Puede usar Azure Resource Health para supervisar el estado de un recurso individual y puede configurar alertas de Resource Health para notificarle problemas.
    - Puede usar Azure Service Health para comprender el estado general del servicio, incluidos los errores de región, y puede configurar alertas de Service Health para notificarle problemas.
  - **Solicitudes activas:** Durante el proceso de conmutación por error, los puntos de conexión de la cuenta de almacenamiento principal y secundaria dejan de estar disponibles temporalmente para lecturas y escrituras. Las solicitudes activas podrían fallar, y las aplicaciones cliente deben intentar nuevamente después de que se complete la conmutación por error.
  - **Pérdida de datos prevista:** la pérdida de datos es habitual durante una conmutación por error no planificada debido al retraso en la replicación asíncrona, lo que significa que es posible que las escrituras recientes no se repliquen. Consulte la propiedad

Última hora de sincronización para saber cuántos datos se podrían perder durante una conmutación por error no planificada. La pérdida de datos esperada se conoce a menudo como objetivo de punto de recuperación (RPO). Normalmente, puede esperar que el RPO sea inferior a 15 minutos, pero ese tiempo no está garantizado.

  - **Tiempo de inactividad esperado:** La cantidad de tiempo de inactividad esperado se conoce a menudo como objetivo de tiempo de recuperación (RTO). La conmutación por error administrada por el cliente se completa normalmente en un plazo de hasta 60 minutos, dependiendo del tamaño y la complejidad de la cuenta.

  - **Reenrutamiento del tráfico:** A medida que se completa la conmutación por error, Azure actualiza automáticamente los puntos de conexión de la cuenta de almacenamiento para que no sea necesario volver a configurar las aplicaciones. Si su aplicación mantiene entradas del Sistema de nombres de dominio (DNS) almacenadas en caché, puede que sea necesario borrar la caché para asegurarse de que la aplicación envíe el tráfico a la nueva región principal.

  - **Configuración posterior a la conmutación por error:** una vez completada una conmutación por error no planificada, su cuenta de almacenamiento en la región de destino utilizará el nivel de almacenamiento con redundancia local (LRS). Si necesitase volver a replicar geográficamente los datos, deberá volver a habilitar el almacenamiento con redundancia geográfica (GRS) y esperar a que los datos se repliquen en la nueva región secundaria.

  Para obtener más información sobre cómo iniciar una conmutación por error administrada por el cliente, consulte Cómo funciona la conmutación por error administrada por el cliente (no planificada) e Iniciar una conmutación por error de una cuenta de almacenamiento.

- **Conmutación por error administrada por el cliente (planificada):** utilice una conmutación por error planificada cuando el almacenamiento siga operativo en la región principal, pero necesite conmutar por error toda su solución a una región secundaria por otro motivo. Por ejemplo, otro servicio de Azure podría estar experimentando un problema y debe cambiar al uso de una región secundaria para toda la solución. O bien, puede usar una conmutación por error planeada para realizar un simulacro de recuperación ante desastres (DR) con fines de cumplimiento y auditoría.

  - **Detección y respuesta:** Usted es responsable de decidir conmutar por error. Normalmente, debe tomar esta decisión si necesita realizar una conmutación entre regiones, aunque su cuenta de almacenamiento esté en buen estado. Por ejemplo, puede desencadenar una conmutación por error cuando se produce una interrupción importante en otro componente de la aplicación que no se puede recuperar en la región principal.

- **Notificación**: Microsoft no le notifica automáticamente cuando una región está inactiva. Sin embargo:
  - Puede usar Azure Resource Health para supervisar el estado de un recurso individual y puede configurar alertas de Resource Health para notificarle problemas.
  - Puede usar Azure Service Health para comprender el estado general del servicio, incluidos los errores de región, y puede configurar alertas de Service Health para notificarle problemas.

- **Solicitudes activas:** Durante el proceso de conmutación por error, los puntos de conexión de la cuenta de almacenamiento principal y secundaria dejan de estar disponibles temporalmente para lecturas y escrituras. Las solicitudes activas podrían fallar, y las aplicaciones cliente deben intentar nuevamente después de que se complete la conmutación por error.

- **Pérdida de datos esperada:** No se espera ninguna pérdida de datos porque el proceso de conmutación por error se completa solo después de sincronizar todos los datos, lo que da como resultado un RPO de cero.

- **Tiempo de inactividad esperado:** La conmutación por error normalmente se completa en un plazo de 60 minutos, lo que significa que el tiempo de recuperación objetivo (RTO) esperado es de 60 minutos, en función del tamaño y la complejidad de la cuenta. Durante el proceso de conmutación por error, los puntos de conexión de la cuenta de almacenamiento principal y secundaria dejan de estar disponibles temporalmente para lecturas y escrituras.

- **Reenrutamiento del tráfico:** A medida que se completa la conmutación por error, Azure actualiza automáticamente los puntos de conexión de la cuenta de almacenamiento para que no sea necesario volver a configurar las aplicaciones. Si su aplicación mantiene entradas DNS almacenadas en caché, puede que sea necesario borrar la caché para asegurarse de que la aplicación envíe tráfico a la nueva región principal.

- **Configuración posterior a la conmutación por error:** Una vez completada una conmutación por error planeada, la cuenta de almacenamiento de la región de destino sigue siendo replicada geográficamente y permanece en el nivel GRS.

Para obtener más información sobre cómo iniciar una conmutación por error administrada por el cliente, consulte Cómo funciona la conmutación por error administrada por el cliente (planificada) e Iniciar una conmutación por error de una cuenta de almacenamiento.

- **Conmutación por error administrada por Microsoft:** En el caso excepcional de un desastre importante en el que Microsoft determina que la región primaria es irrecuperable permanentemente, es posible que se inicie una conmutación automática por error a la región secundaria. Microsoft controla todo el proceso y no se requiere ninguna acción del cliente. El tiempo que transcurre antes de que ocurra un failover depende de la gravedad del desastre y del tiempo necesario para evaluar la situación.
  - **Notificación**: Microsoft no le notifica automáticamente cuando una región está inactiva. Sin embargo:
    - Puede usar Azure Resource Health para supervisar el estado de un recurso individual y puede configurar alertas de Resource Health para notificarle problemas.
    - Puede usar Azure Service Health para comprender el estado general del servicio, incluidos los errores de región, y puede configurar alertas de Service Health para notificarle problemas.

> ⓘ **Importante**
>
> Utilice las opciones de conmutación por error administradas por el cliente para desarrollar, probar e implementar tus planes de recuperación ante desastres. **No confíe en la conmutación por error administrada por Microsoft**, que solo se puede utilizar en circunstancias extremas. Es probable que se haya iniciado una conmutación por error administrada por Microsoft para toda una región. No se puede iniciar para cuentas de almacenamiento individuales, suscripciones o clientes. La conmutación por error puede producirse en momentos diferentes para distintos servicios de Azure. Recomendamos que utilice la conmutación por error administrada por el cliente.

Recuperación de regiones

El proceso de conmutación por recuperación difiere significativamente entre los escenarios de conmutación por error administrados por Microsoft y los administrados por el cliente.

- **Conmutación por error administrada por el cliente (no planeada):** después de una conmutación por error no planeada, la cuenta de almacenamiento se configura con almacenamiento con redundancia local (LRS). Para realizar una conmutación por error, es necesario restablecer la relación de almacenamiento con redundancia geográfica (GRS) y esperar a que se repliquen los datos.
- **Conmutación por error administrada por el cliente (planificada):** después de una conmutación por error planificada, la cuenta de almacenamiento seguirá estando

replicada geográficamente. Inicie otra conmutación por error administrada por el cliente para volver a la región principal original. Se aplican las mismas consideraciones de conmutación por error.

- **Conmutación por error administrada por Microsoft:** si Microsoft iniciase una conmutación por error, es probable que se haya producido un desastre importante en la región principal y que esta no se pueda recuperar. Cualquier calendario o plan de recuperación dependerá del alcance del desastre regional y de los esfuerzos de recuperación. Debe supervisar las comunicaciones de Azure Service Health para más información.

Prueba de fallos de región

Simule fallos regionales para probar sus procedimientos de recuperación ante desastres.

- **Pruebas de conmutación por error planificadas:** en el caso de las cuentas de almacenamiento con redundancia geográfica (GRS), es posible realizar operaciones de conmutación por error planificadas durante las ventanas de mantenimiento para probar el proceso completo de conmutación por error y conmutación por recuperación. La conmutación por error planificada no implica pérdida de datos, pero sí conlleva tiempo de inactividad tanto durante la conmutación por error como durante la conmutación por recuperación.

- **Pruebas de puntos de conexión secundarios:** para el almacenamiento con redundancia geográfica con acceso de lectura (RA-GRS) y las configuraciones de almacenamiento con redundancia de zona geográfica con acceso de lectura (RA-GZRS), pruebe periódicamente las operaciones de lectura en el punto de conexión secundario para asegurarse de que la aplicación pueda leer correctamente los datos de la región secundaria.

Soluciones personalizadas de varias regiones para la resistencia

Las funcionalidades de conmutación por error entre regiones de Azure Storage podrían no ser adecuadas por los siguientes motivos:

- La cuenta de almacenamiento está en una región no emparejada.

- Los objetivos de tiempo de actividad empresarial no se satisfacen por el tiempo de recuperación o la pérdida de datos que proporciona las opciones de conmutación por error integradas.

- Debe conmutar por error a una región que no esté emparejada con la región principal.
- Necesita una configuración activa o activa entre regiones.

En esta sección se proporciona información general de alto nivel de algunos enfoques que se deben tener en cuenta. Una introducción completa de las topologías de implementación de varias regiones para Azure Storage está fuera del ámbito de este artículo.

Implemente Azure Storage en varias regiones utilizando cuentas de almacenamiento independientes en cada región. Este enfoque ofrece flexibilidad en la selección de regiones, la posibilidad de utilizar regiones no emparejadas y un control más granular sobre el momento de la replicación y la coherencia de los datos. Al implementar varias cuentas de almacenamiento entre regiones, es necesario configurar la replicación de datos entre regiones, implementar directivas de equilibrio de carga y conmutación por error y garantizar la coherencia de los datos entre regiones.

La replicación de objetos proporciona una opción adicional para la replicación de datos entre regiones que proporciona copia asincrónica de blobs en bloques entre cuentas de almacenamiento. A diferencia de las opciones de GRS integradas que usan regiones emparejadas fijas, la replicación de objetos permite replicar datos entre cuentas de almacenamiento en cualquier región de Azure, incluyendo las regiones no emparejadas. Este enfoque le ofrece un control total sobre las regiones de origen y destino, las directivas de replicación y los contenedores y prefijos de blobs específicos que se van a replicar.

Configure la replicación de objetos para replicar todos los blobs de un contenedor o subconjuntos específicos en función de prefijos y etiquetas de blobs. La replicación es asincrónica y se produce en segundo plano. Puede configurar varias directivas de replicación e incluso encadenar la replicación entre varias cuentas de almacenamiento para crear topologías multirregionales sofisticadas.

La replicación de objetos no es compatible con todas las cuentas de almacenamiento. Por ejemplo, no funciona con cuentas de almacenamiento que usen espacios de nombres jerárquicos (también conocidos como *cuentas de Azure Data Lake Storage Gen2*).

Para obtener más información, consulte Replicación de objetos para blobs en bloques y Configuración de la replicación de objetos.

Copia de seguridad y recuperación

Blob Storage proporciona varios mecanismos de protección de datos que complementan la redundancia para estrategias de copia de seguridad completas. La redundancia integrada del servicio protege frente a errores de infraestructura, mientras que las funcionalidades de copia

de seguridad adicionales protegen contra actividades malintencionadas, daños y eliminación accidentales.

La restauración a un momento dado (PITR) permite restaurar los datos de blobs en bloques a un estado anterior dentro de un período de retención configurado de hasta 365 días. Microsoft administra completamente esta característica. También proporciona funcionalidades de recuperación pormenorizadas en el nivel de contenedor o blob. Los datos PITR se almacenan en la misma región que la cuenta de origen y son accesibles incluso durante interrupciones regionales si se usan configuraciones con redundancia geográfica.

El **control de versiones de blobs** mantiene automáticamente las versiones anteriores de blobs cuando se modifiquen o eliminen. Cada versión se almacena como un objeto independiente y se puede acceder a ella de forma independiente. Las versiones se almacenan en la misma región que el blob actual y siguen la misma configuración de redundancia que la cuenta de almacenamiento.

La eliminación temporal proporciona una red de seguridad para los blobs y contenedores eliminados accidentalmente, ya que conserva los datos eliminados durante un período configurable. Los datos eliminados temporalmente permanecen en la misma cuenta de almacenamiento y región, lo que hace que estén disponibles inmediatamente para la recuperación. En el caso de las cuentas con redundancia geográfica, los datos eliminados de forma temporal también se replican en la región secundaria.

Las **instantáneas de blobs** crean copias de solo lectura y a un momento dado de blobs que se pueden usar para escenarios de copia de seguridad y recuperación. Las instantáneas se almacenan en la misma cuenta de almacenamiento y siguen la misma configuración de redundancia y replicación geográfica que el blob base.

En el caso de los requisitos de copia de seguridad entre regiones, considere la posibilidad de usar **Azure Backup para blobs**, que proporciona administración centralizada de copias de seguridad y almacena datos de copia de seguridad en regiones diferentes de los datos de origen. Este servicio proporciona opciones de copia de seguridad operativas y de almacén que tienen directivas de retención configurables, así como funcionalidades de restauración. Para más información, consulte Introducción a Backup para blobs.

Para la mayoría de las soluciones, no debe confiar exclusivamente en copias de seguridad. En su lugar, utilice las otras capacidades descritas en esta guía para apoyar los requisitos de resiliencia. Sin embargo, las copias de seguridad protegen contra algunos riesgos que otros enfoques no. Para más información, consulte ¿Qué son la redundancia, la replicación y la copia de seguridad?.

Acuerdo de nivel de servicio

El acuerdo de nivel de servicio (SLA) para Azure Storage describe la disponibilidad esperada del servicio y las condiciones que se deben cumplir para lograr esa expectativa de disponibilidad. El contrato de nivel de servicio de disponibilidad al que tiene derecho dependerá del nivel de almacenamiento y del tipo de replicación que utilice. Para obtener más información, consulte los SLAs for Online Services.

Contenido relacionado

- Introducción a Blob Storage
- Redundancia de Azure Storage
- Recuperación ante desastres y failover de la cuenta de almacenamiento
- Uso de redundancia geográfica para diseñar aplicaciones de alta disponibilidad
- Cambio de la replicación de una cuenta de almacenamiento

ⓘ **Nota:** El autor ha creado este artículo con ayuda de inteligencia artificial. Más información

---

Last updated on 22/01/2026

Azure Data Box Disk: Preguntas más frecuentes

La solución en la nube Microsoft Azure Data Box Disk permite enviar terabytes de datos a Azure de forma rápida, económica y confiable. Estas preguntas frecuentes contienen preguntas y respuestas que puede tener al usar Data Box Disks en Azure Portal.

Acerca del servicio

¿Qué es el servicio Azure Data Box?

El servicio Azure Data Box está diseñado para la ingesta de datos sin conexión. Este servicio administra una matriz de productos adaptados para el transporte de datos para diferentes capacidades de almacenamiento.

¿Qué son los discos de Azure Data Box?

Los discos de Azure Data Box permiten transferir terabytes de datos de forma rápida, barata y segura a y desde Azure. Microsoft envía entre 1 y 5, con una capacidad de almacenamiento máxima de 35 TB. Dichos discos se pueden configurar, conectar y desbloquear fácilmente con el servicio Data Box en Azure Portal.

Los discos cuentan con cifrado de unidad BitLocker de Microsoft, mientras que las claves de cifrado se administran desde Azure Portal. Luego se copian los datos de los servidores del cliente. En el centro de datos, Microsoft migra los datos de la unidad a la nube con un vínculo de carga en la red rápido y privado, y los carga en Azure.

¿Cuándo deben usarse los discos de Data Box?

Si desea transferir 40 TB de datos (o menos) a Azure, es aconsejable usar de los discos de Data Box.

¿Cuánto cuestan los discos de Data Box?

Para más información sobre el precio de Data Box Disks, vaya a la página de precios.

¿Cómo se obtienen los discos de Data Box?

Para obtener Azure Data Box Disks, inicie sesión en Azure Portal y cree un pedido de Data Box para discos. Proporcione su información de contacto y los datos de notificación. Una vez que realiza el pedido, y en función de la disponibilidad, los discos se envían en un plazo máximo de 10 días.

¿Cuál es la cantidad máxima de datos que se puede transferir con los discos de Data Box en una instancia?

Para cinco discos de 8 TB cada uno (7 TB de capacidad usable), la capacidad máxima usable es 35 TB. Por tanto, puede transferir 35 TB de datos en una instancia. Para transferir más datos, es preciso solicitar más discos.

¿Cómo se puede comprobar si los discos de Data Box están disponibles en una región concreta?

Para ver dónde están disponibles actualmente las instancias de Data Box Disk, vaya a Disponibilidad en regiones.

¿En qué regiones puedo almacenar datos con los discos de Data Box?

Data Box Disk es admitido con todas las regiones de EE. UU., Canadá, Europa, Reino Unido, Australia, Singapur, India, China, Región administrativa especial de Hong Kong, Japón, Corea y Sudáfrica. Solo se admiten las regiones con nube pública de Azure. No se admiten Azure Government u otras nubes soberanas.

¿Cómo puedo importar datos de origen presentes en mi ubicación de un país o región a una región de Azure en un país o región diferente?

Data Box Disk solo admite la ingesta de datos dentro del mismo país o región que su destino y no cruzará ninguna frontera internacional. La única excepción es para los pedidos de la Unión Europea (EU), donde Data Box Disks se puede enviar a cualquier país o región de la UE y desde ellas.

Por ejemplo, si quiere mover datos de su ubicación en Canadá a una cuenta de almacenamiento del Oeste de EE. UU. de Azure, podría conseguirlo de la siguiente manera:

Opción 1:

Envíe un disco compatible que contenga datos mediante el servicio Azure Import/Export desde la ubicación de origen en Canadá hasta el centro de datos del Oeste de EE. UU. de Azure.

Opción 2:

1. Pida un dispositivo Data Box Disk en Canadá al seleccionar una cuenta de almacenamiento del Centro de Canadá. Los discos SSD se envían desde el centro de datos de Azure en Centro de Canadá a la dirección de envío (en Canadá) proporcionada durante la creación del pedido.

2. Después de copiar los datos del servidor local en los discos, devuélvalos al centro de datos de Azure en Canadá con las etiquetas de devolución proporcionadas por Microsoft. Los datos presentes en Data Box Disks se cargan en la cuenta de almacenamiento de destino en la región de Azure de Canadá elegida durante la creación del pedido.

3. A continuación, puede usar una herramienta como AzCopy para copiar los datos en una cuenta de almacenamiento en la región del Oeste de EE. UU. Este paso incurre en almacenamiento estándar y cargos de ancho de banda que no se incluyen en la facturación de Data Box Disk.

¿Data Box Disk almacena los datos de clientes fuera de la región de servicio?

No. Data Box Disk no almacena datos de clientes fuera de la región de servicio. El cliente tiene la propiedad total de sus datos y puede guardarlos en una ubicación determinada en función de la cuenta de almacenamiento que seleccione durante la creación del pedido.

Además de los datos del cliente, hay datos de Data Box Disk que incluyen metadatos y registros de supervisión. En todas las regiones (excepto el Sur de Brasil, el Sudeste Asiático y el Este de Asia), los datos de Data Box Disk se almacenan y replican en una región emparejada a través de una cuenta de almacenamiento con redundancia geográfica como protección contra la pérdida de datos.

Debido a requisitos de residencia de datos en El Sur de Brasil, Sudeste asiático y Este de Asia, los datos de Data Box Disk se almacenan en una cuenta de almacenamiento con redundancia de

zona (ZRS) para que esté contenida en una sola región. Para el Sudeste Asiático, todos los datos de Data Box Disk se almacenan en Singapur. En el caso del Sur de Brasil, se almacenan en Brasil. En cuanto al Este de Asia, los datos se almacenan en Hong Kong.

Si hay una interrupción del servicio en el Sur de Brasil, Sudeste asiático o Este de Asia, los clientes pueden crear nuevos pedidos desde otra región. Los nuevos pedidos se sirven desde la región en la que se crean y los clientes son responsables del envío de ida y vuelta del Disco Data Box.

¿Debe estar mi región de Data Box Disk en la misma región de Azure que mi cuenta de almacenamiento de destino?

No. El recurso Data Box Disk no necesita ser el mismo que la región de la cuenta de almacenamiento en la que desea transferir datos. Puede elegir una cuenta de almacenamiento en cualquier región dentro del país o región o límite comercial seleccionados en el pedido. La región de Data Box Disk solo se usa para implementar el recurso al crear el pedido. Esto se fija para cada país o región o límite comercial y no se puede cambiar.

¿Cómo puedo recuperar mis datos si se produce un error en toda una región?

En circunstancias extremas en las que se pierde una región debido a un desastre significativo, Microsoft podría iniciar una conmutación por error regional. En este caso, no se requieren acciones por su parte. El pedido se cumple a través de la región de conmutación por error si está dentro del mismo país o región o límite comercial. Sin embargo, algunas regiones de Azure no tienen una región emparejada en el mismo límite geográfico o de comercio. Si se produce un desastre en cualquiera de esas regiones, tendrá que volver a crear el pedido de Data Box desde otra región que esté disponible y copiar los datos en Azure en la región nueva. Para obtener más información, consulte Replicación entre regiones en Azure: continuidad empresarial y recuperación ante desastres.

¿Con quién debo ponerme en contacto si surge algún problema con los discos de Data Box?

Si surge algún problema con los dispositivos de Data Box Disk, póngase en contacto con el Soporte técnico de Microsoft.

Pedir el dispositivo

¿Cómo se obtiene un dispositivo Data Box Disk?

Para obtener Azure Data Box Disk, inicie sesión en Azure Portal y cree un pedido de Data Box Disk. Proporcione su información de contacto y los datos de notificación. Una vez realizado el pedido, y en función de la disponibilidad, Data Box Disk se envía en un plazo máximo de 10 días. Para obtener más información, vaya a Order a Data Box (Pedir Data Box).

No pude hacer un pedido de Data Box Disk en Azure Portal. ¿Por qué?

Si no pudo pedir un dispositivo Data Box Disk, es que hay un problema con su tipo de suscripción o con el acceso.

Compruebe su suscripción. Data Box Disk solo está disponible para las ofertas de suscripción de Contratos Enterprise (EA) y del Proveedor de soluciones en la nube (CSP). Si su suscripción no se encuentra entre ninguno de los tipos anteriores, póngase en contacto con el Soporte técnico de Microsoft para actualizar la suscripción.

Si tiene un tipo de oferta compatible para la suscripción, compruebe el nivel de acceso de la suscripción. Debe ser un colaborador o propietario de la suscripción para poder realizar un pedido.

¿Cuánto tardará mi pedido desde su creación hasta que se carguen los datos en Azure?

Los siguientes plazos estimados para cada fase del procesamiento de pedidos le proporcionarán una idea precisa de lo que puede esperar.

Recuerde que estos plazos son *estimaciones*. El tiempo de cada fase de procesamiento de pedidos se ve afectado por la carga del centro de recursos, los pedidos simultáneos y otras condiciones del entorno.

Plazos estimados para una solicitud de Data Box Disk:

1. Pedido de Data Box Disk: unos minutos desde el portal
2. Asignación y preparación de discos: hasta 5 días laborables, según la disponibilidad del inventario y el número de pedidos pendientes que se van a procesar

3. Envío: 2 a 3 días laborables
4. Copia de datos en el sitio del cliente: depende de la naturaleza de los datos, el tamaño y el número de archivos.
5. Envío de devolución: 2 a 3 días laborables
6. Procesamiento en el centro de datos y carga en Azure: la carga de datos comienza en el centro de datos en cuanto se completa el procesamiento operativo y el disco está conectado. El tiempo de carga depende de la naturaleza de los datos, su tamaño y el número de archivos.

¿Data Box Disk admite la exportación?

No, solo se admite la importación para Data Box Disk. No hay ninguna funcionalidad de exportación disponible. Si quiere exportar datos desde Azure, puede usar Azure Data Box.

Configuración y conexión

¿Se puede especificar el número de discos de Data Box en el pedido?

No. Obtiene discos de 8 TB (un máximo de cinco discos) en función del tamaño de los datos y de la disponibilidad de los discos.

¿Cómo se desbloquean los discos de Data Box?

En Azure Portal, vaya a su pedido de discos de Data Box y, después, a **Detalles del dispositivo**. Copie la clave de paso. Descargue y extraiga la herramienta de desbloqueo Data Box Disk para su sistema operativo desde Azure Portal. Ejecute la herramienta en el equipo que contiene los datos que desea copiar en los discos. Escriba la clave de paso para desbloquear los discos. La misma clave desbloquea todos los discos.

Para obtener instrucciones paso a paso, vaya a Desbloquear discos.

¿Se puede usar un equipo host Linux para conectarse y copiar los datos de los discos de Data Box?

Sí. Tanto los clientes de Linux como los de Windows pueden utilizarse para conectar y copiar datos en los discos de Data Box Disk. Para más información, vaya a la lista de sistemas operativos compatibles del equipo host.

Quiero cancelar un pedido una vez que los discos se han enviado. ¿Por qué no está disponible el botón de cancelación?

El pedido solo se puede cancelar durante el periodo que va entre la realización del pedido y el momento en que se realiza el envío. Una vez que los discos se han enviado, no de puede cancelar el pedido. Sin embargo, puede devolver los discos a un precio.

¿Se pueden conectar varios discos de Data Box al mismo al equipo host para transferir datos?

Sí. Se pueden conectar varios discos de Data Box al mismo equipo host para transferir datos y se pueden ejecutar varios trabajos de copia en paralelo.

Seguimiento del estado

¿Cómo se realiza el seguimiento de los discos desde el momento en que se realizó el pedido a la devolución de los discos?

El seguimiento del estado del pedido de Data Box Disk se puede realizar desde Azure Portal. Al crear el pedido, también se le solicita que indique una dirección de correo electrónico para cualquier notificación. Si ya la ha especificado, los cambios de estado del pedido se le notificarán por correo electrónico. Para más información acerca de cómo configurar correos electrónicos de notificación.

¿Cómo se devuelven los discos?

Microsoft incluye una etiqueta de envío con los discos de Data Box en el paquete en que estos se envían. Pegue dicha etiqueta en la caja de envío y lleve el paquete precintado a las instalaciones de la empresa que va a realizar el envío. Si la etiqueta se daña o se pierde, vaya a **Información general** > **Descargar la etiqueta de envío** y descargue una nueva etiqueta de envío de devolución.

¿Puedo seleccionar mi pedido de Data Box Disk? ¿Puedo usar el transportista que prefiera para

devolver los discos?

Sí. Microsoft también ofrece el envío autoadministrado, pero solo en la región US Gov. Al realizar el pedido de Data Box Disk, puede elegir la opción de envío autoadministrado. Para recoger el pedido de Data Box Disk, siga estos pasos:

1. Después de haber realizado el pedido, este se procesa y se preparan los discos. Se le enviará por correo electrónico una notificación cuando el pedido se pueda recoger.
2. Una vez que el pedido esté listo para su recogida, vaya a él en Azure Portal y. después, a la hoja **Información general**.
3. Verá una notificación con un código en Azure Portal. Envíe un correo electrónico al Equipo de operaciones de Azure Data Box y proporcione el código. El equipo proporcionará la ubicación y programará una fecha y hora de recogida. Debe llamar al equipo en un plazo máximo de cinco días laborables después de recibir la notificación por correo electrónico.

Una vez que completan la copia y la validación de los datos, siga estos pasos siguientes para devolver el disco:

1. Una vez completada la validación de datos, desconecte los discos. Quite los cables de conexión.

2. Envuelva todos los discos y los cables de conexión en un envoltorio de burbujas y colóquelos en la caja de envío. La falta de accesorios puede acarrear un costo.

    - Utilice el empaquetado que se usó en el envío inicial. Es aconsejable usar un envoltorio de burbujas bien protegido para empaquetar los discos.
    - Asegúrese de que todo encaja perfectamente para reducir los movimientos dentro de la caja.

3. Vaya a la **hoja Información general** del pedido en Azure Portal. Debería ver una notificación con un código.

4. Use ese código y envíe un correo electrónico al Equipo de operaciones de Azure Data Box y proporcione el código. Le proporcionarían información sobre dónde y cuándo soltar los discos.

Migración de los datos

¿Cuál es el tamaño máximo de los datos que se puede usar con los discos de Data Box?

La solución Data Box Disks puede tener hasta cinco discos, con una capacidad máxima utilizable de 35 TB. Los discos en sí tienen 8 TB (7 TB utilizables).

¿Qué tamaño máximo de blobs en bloques y blobs en páginas admiten los discos de Data Box?

El tamaño máximo lo determinan los límites de Azure Storage. El blob en bloques máximo es, aproximadamente, de 4,768 TiB, mientras que el tamaño de blob en páginas máximo es de 8 TiB. Para más información, consulte Objetivos de escalabilidad y rendimiento de Blob Storage.

¿Cuál es la velocidad de transferencia de datos de los discos de Data Box?

Cuando se prueba con discos conectados mediante USB 3.0, el rendimiento del disco era de hasta 430 MB/s. Los números reales varían según el tamaño de archivo usado. Con archivos más pequeños, es posible que observe un rendimiento menor.

¿Cómo se sabe si los datos están protegidos durante el tránsito?

Con los discos de Data Box se emplea el cifrado BitLocker AES de 128 bits y la clave de paso solo está disponible en Azure Portal. Inicie sesión en Azure Portal con sus credenciales de cuenta para obtener la clave de acceso. Escriba dicha clave al ejecutar la herramienta de desbloqueo de Data Box Disk.

¿Cómo se copian los datos a los discos de Data Box?

Use una herramienta de copia de SMB, como `Robocopy`, `Diskboss` o incluso el Explorador de archivos de Windows, para arrastrar y colocar los datos en los discos, y así copiarlos.

¿Algunas sugerencia para agilizar la copia de datos?

Para acelerar el proceso de copia:

- Use varias secuencias de copia de datos. Por ejemplo, si usa `Robocopy`, emplee la opción de varios procesos. Para más información acerca del comando exacto que se debe usar, vaya a Tutorial: Copia de datos a Azure Data Box Disk y comprobación de los mismos.
- Utilice varias sesiones.
- En lugar de realizar la copia a través de un recurso compartido de red (en el que puede tener la limitación de la velocidad de la red), asegúrese de que los datos se encuentran localmente en el equipo al que se conectan los discos.
- Asegúrese de que usa USB 3.0 o una versión posterior durante todo el proceso de copia. Descargue la herramienta USBView y úsela para identificar los controladores y dispositivos USB conectados al equipo.
- Evalúe el rendimiento del equipo usado para copiar los datos. Descargue la herramienta BluestopFIO y úsela para evaluar el rendimiento del hardware del servidor. Seleccione la compilación x86 o x64 más reciente, seleccione la pestaña **Artefactos** y descargue el MSI.

¿Cómo acelerar los datos si los datos de origen tienen archivos pequeños (de KB o pocos MB)?

Para acelerar el proceso de copia:

- Cree un VHDx local en el almacenamiento rápido, o bien cree disco duro virtual vacío en la HDD o SSD (más lenta).
- Móntelo en una máquina virtual.
- Copie los archivos en el disco de la máquina virtual.

¿Se pueden usar varias cuentas de almacenamiento con los discos de Data Box?

No. Actualmente, los discos de Data Box solo admiten una cuenta de almacenamiento, general o clásica. Se admiten blobs frecuentes y esporádicos.

> **Nota:**
>
> Las cuentas de almacenamiento clásicas no se admitirán a partir del **1 de agosto de 2023** .

¿Qué contiene el conjunto de herramientas disponible para mis datos con Data Box Disk?

El conjunto de herramientas disponible con Data Box Disk contiene tres herramientas:

- **Herramienta de desbloqueo**de Data Box Disk: use esta herramienta para desbloquear los discos cifrados que se envían desde Microsoft. Cuando desbloquee los discos mediante la herramienta, deberá proporcionar una clave de paso disponible en el pedido de Data Box Disk en Azure Portal.
- **Herramienta de validación de Data Box Disk**: use esta herramienta para validar el tamaño, el formato y los nombres de blob según las convenciones de nomenclatura de Azure. La herramienta también genera sumas de comprobación para los datos copiados que, posteriormente, se usan para comprobar los datos cargados en Azure.
- **Herramienta de copia dividida de Data Box Disk**: use esta herramienta cuando use varios discos y tenga un conjunto de datos grande que deba dividirse y copiarse en todos los discos. Esta herramienta está actualmente disponible para Windows. Esta herramienta no se admite con discos administrados. Esta herramienta valida los datos a medida que los copia, de modo que puede omitir el paso de validación al usarla.

El conjunto de herramientas está disponible para Windows y Linux. Puede descargar el conjunto de herramientas desde aquí:

- Descargar conjunto de herramientas de Data Box Disk para Windows
- Descargar conjunto de herramientas de Data Box Disk para Linux

¿Puedo usar Data Box Disk para transferir datos a Azure Files y, a continuación, usar los datos con Azure File Sync?

Sí, ahora se admite el uso de Azure File Sync con Data Box Disks.

Comprobación y carga

¿Cuándo se puede acceder a los datos en Azure una vez que se han devuelto los discos?

Una vez que el estado del pedido de Copia de datos indica que se ha completado, debería poder acceder a los datos de inmediato.

¿Dónde se encuentran los datos en Azure después de la carga?

Cuando se copian los datos de las carpetas *BlockBlob* y *PageBlob* del disco, se crea un contenedor en la cuenta de Azure Storage para cada una de las subcarpetas de las carpetas *BlockBlob* y *PageBlob*. Si ha copiado los archivos de las carpetas *BlockBlob* y *PageBlob* directamente, estos se encontrarán en el contenedor predeterminado *$root* de la cuenta de Azure Storage. Al copiar los datos en una carpeta situada bajo la carpeta *AzureFile*, se crea un recurso compartido de archivos.

Acabo de darme cuenta de que no seguí los requisitos de nomenclatura de Azure en los contenedores. ¿Pueden aparecer errores al cargar los datos en Azure?

Las letras mayúsculas de los nombres de los contenedores se convierten automáticamente a minúsculas. Si los nombres no son compatibles de otras maneras (por ejemplo, contienen caracteres especiales u otros idiomas), se producirá un error en la carga. Para más información, vaya a Convenciones de nomenclatura de Azure.

¿Cómo se comprueban los datos que se han copiado en varios discos de Data Box?

Una vez completada la copia de datos, puede ejecutar `DataBoxDiskValidation.cmd`, que se encuentra en la carpeta *DataBoxDiskImport*, para generar sumas de comprobación de validación. Si tiene varios discos, es preciso que abra una ventana Comandos por disco y que ejecute este comando. Tenga en cuenta que esta operación puede tardar mucho tiempo (horas) en función del tamaño de los datos.

¿Qué ocurre con los datos después de que se devuelven los discos?

Una vez que se completa la copia de datos en Azure, los datos de los discos se borran de forma segura según las directrices de la revisión 1 de NIST SP 800-88.

¿Cómo se protegen los datos durante el tránsito?

Los dispositivos Data Box Disk se cifran con el cifrado de Microsoft BitLocker AES-128 y se requiere una sola clave de paso para desbloquear todos los discos y obtener acceso a los datos.

¿Es preciso volver a ejecutar la validación de suma de comprobación si se agregan más datos a los discos de Data Box?

Sí. Si decide validar los datos (es aconsejable que lo haga), tendrá que volver a ejecutar la validación si agrega más datos a los discos.

He usado todos los discos para transferir datos y necesito pedir más discos. ¿Hay alguna manera rápida de realizar el pedido?

Puede clonar el pedido anterior. La clonación crea el mismo pedido que antes y permite editar los detalles del pedido solo sin necesidad de escribir los datos de dirección, contacto y notificación.

He copiado datos en la carpeta ManagedDisk. No veo ningún disco administrado con el grupo de recursos especificado para los discos administrados. ¿Los datos se cargaron en Azure? ¿Cómo puedo encontrarlos?

Sí. Los datos se cargaron en Azure, pero si no ve ningún disco administrado con los grupos de recursos especificados, es probable que se deba a que los datos no eran válidos. Si los blob en páginas, los blob en bloques, Azure Files o los discos administrados no son válidos, irán a las siguientes carpetas:

- Los blob en páginas van a un contenedor de blob en bloques a partir de *databoxdisk-invalid-pb-*.
- Azure Files irá a un contenedor de blobs en bloques que empieza por *databoxdisk-invalid-af-*.
- Los discos administrados van a un contenedor de blob en bloques a partir de *databoxdisk-invalid-md-*.

Pasos siguientes

- Consulte los requisitos del sistema de Data Box Edge.
- Información acerca de los límites de Data Box Disk.
- Implemente rápidamente Azure Data Box Disk en Azure Portal.

Last updated on 18/08/2026

Confiabilidad en Azure Disk Storage

Azure Disk Storage proporciona discos administrados para máquinas virtuales (VM) de Azure. Creado para cargas de trabajo críticas, garantiza la confiabilidad y disponibilidad de nivel empresarial. Los datos se replican automáticamente para protegerse frente a errores de hardware, con varias opciones de redundancia para satisfacer los requisitos de durabilidad.

Al usar Azure, relibilidad es una responsabilidad compartida. Microsoft proporciona una variedad de capacidades para admitir resiliencia y recuperación. Es responsable de comprender cómo funcionan esas funcionalidades dentro de todos los servicios que usa y de seleccionar las funcionalidades que necesita para cumplir los objetivos empresariales y los objetivos de tiempo de actividad.

En este artículo se describe cómo hacer que Azure Disk Storage sea resistente a una variedad de posibles interrupciones y problemas, incluidos fallos transitorios, fallas en la zona de disponibilidad y fallos en toda la región. También describe las opciones de copia de seguridad y recuperación, y resalta la información clave sobre el acuerdo de nivel de servicio (SLA) de Azure Disk Storage.

> **Importante**
>
> Al considerar la confiabilidad de un disco, también debe tener en cuenta la confiabilidad de las máquinas virtuales, la infraestructura de red y las aplicaciones que se ejecutan en las máquinas virtuales. Mejorar la resistencia del disco solo podría tener un impacto limitado si los otros componentes no son igualmente resistentes. En función de los requisitos de resistencia, es posible que tenga que realizar cambios de configuración en varias áreas.

Recomendaciones de implementación de producción

El marco de Azure Well-Architected proporciona recomendaciones sobre confiabilidad, rendimiento, seguridad, costo y operaciones. Para comprender cómo estas áreas influyen entre sí y contribuir a una solución de Azure Disk Storage confiable, consulte Architecture best practices for Azure Disk Storage.

Introducción a la arquitectura de confiabilidad

Cada máquina virtual (VM) usa discos para diferentes propósitos:

- *Disco del sistema* operativo: un único disco del sistema operativo ejecuta el sistema operativo. De forma predeterminada, es un disco administrado que conserva los datos. También puede usar discos de sistema operativo efímeros, que no se administran. Evite usar el disco del sistema operativo para almacenar aplicaciones o datos.
- *Discos de datos*: cero o más discos administrados para almacenar aplicaciones y datos.
- *Disco temporal*: un disco no administrado no persistente incluido con cada máquina virtual.

Esta guía se centra específicamente en discos administrados, que conservan datos de forma confiable. Para más información sobre los distintos roles de disco, consulte Roles de disco.

Los discos administrados están diseñados para la disponibilidad de máquinas virtuales de 99,999% y proporcionan al menos 99,999999999% (11 9) de durabilidad. Con los discos administrados, los datos se replican tres veces. Si una de las tres copias deja de estar disponible, Azure genera automáticamente una nueva copia de los datos en segundo plano. Esto garantiza la persistencia de los datos y la alta tolerancia a errores.

De forma predeterminada, los discos administrados usan almacenamiento con redundancia local (LRS). LRS mantiene tres copias de los datos del disco dentro de un único centro de datos, protegiendo frente a errores de hardware, como problemas de unidad o bastidor del servidor.

Aunque LRS protege tus discos frente a fallos en el bastidor del servidor y las unidades de disco, no tiene en cuenta desastres como incendios o inundaciones dentro del centro de datos. Para obtener mayores niveles de protección, use el almacenamiento con redundancia de zona (ZRS), que replica los discos en varias zonas de disponibilidad.

En el caso de las aplicaciones que se ejecutan en varias máquinas virtuales, varias máquinas virtuales tienen el Acuerdo de Nivel de Servicio de mayor disponibilidad cuando se distribuyen entre varias zonas de disponibilidad. En el caso de las máquinas virtuales y los discos distribuidos en varias zonas de disponibilidad, los discos y sus máquinas virtuales primarias se colocan respectivamente en la misma zona, lo que evita que varias máquinas virtuales dejen de funcionar si se produce una interrupción en una zona entera.

Cuando las zonas no están disponibles o la carga de trabajo es sensible a la latencia entre máquinas virtuales, implemente máquinas virtuales y discos en varios dominios de error. Los dominios de error no proporcionan redundancia de zona, pero reducen el impacto de errores de hardware, interrupciones de red o interrupciones de energía. Esto impide que varias máquinas virtuales produzcan errores si un dominio de error de almacenamiento deja de funcionar.

Resistencia a errores transitorios

Los errores transitorios son errores breves e intermitentes en los componentes. Se producen con frecuencia en un entorno distribuido como la nube y son una parte normal de las operaciones. Los errores transitorios se corrigen después de un breve período de tiempo. Es importante que las aplicaciones puedan controlar errores transitorios, normalmente mediante el reintento de solicitudes afectadas.

Todas las aplicaciones hospedadas en la nube deben seguir las instrucciones de control de errores transitorios Azure cuando se comunican con cualquier API, bases de datos y otros componentes hospedados en la nube. Para obtener más información, consulte Recomendaciones para controlar errores transitorios.

Los discos administrados se recuperan automáticamente de errores transitorios en la infraestructura de Azure.

Resistencia a errores de zona de disponibilidad

Availability zones son grupos de centros de datos físicamente independientes dentro de una región de Azure. Cuando una zona falla, los servicios pueden transferirse a una de las zonas restantes.

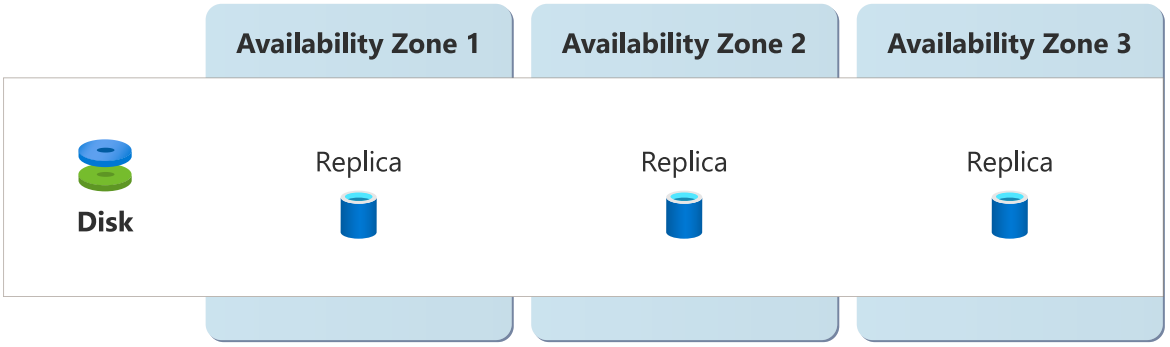
Hay dos maneras de usar zonas de disponibilidad con discos administrados:

- Puede implementar un disco con redundancia de zona (ZRS), que se encuentra en tres zonas de disponibilidad de una región. Para obtener la mejor confiabilidad, se recomienda usar discos ZRS porque los discos ZRS proporcionan resistencia de zona automática.
- Puede implementar un disco LRS zonal, que solo se encuentra en una sola zona. Con los discos LRS zonales, usted es responsable de configurar la carga de trabajo para que sea resiliente a las interrupciones de la zona. Para ello, se implementan varias máquinas virtuales y discos y se distribuyen entre zonas de disponibilidad.

Si no configura la compatibilidad con zonas de disponibilidad, el disco es *no zonal* o *regional* y podría colocarse en cualquier zona de disponibilidad de la región. Estos discos se consideran LRS porque se replican dentro de la región.

Discos redundantes por zonas

El almacenamiento con redundancia de zona (ZRS) replica los datos de forma sincrónica en tres zonas de disponibilidad dentro de una región. Al habilitar la redundancia de zona para un disco administrado, Azure garantiza que un error en cualquier zona única no afecte a la disponibilidad de los datos.



Los discos ZRS se pueden compartir entre máquinas virtuales para mejorar la disponibilidad de aplicaciones en clúster o distribuidas, como FCI de SQL, ASCS/SCS de SAP o GFS2. Puede conectar un disco ZRS compartido a máquinas virtuales principales y secundarias en distintas zonas, aprovechando tanto los discos ZRS como las máquinas virtuales distribuidas entre varias zonas de disponibilidad. Si se produce un error en la zona principal, puede conmutar por error rápidamente a la máquina virtual secundaria mediante la reserva persistente SCSI.

Si un disco ZRS está conectado como un disco de datos a una sola máquina virtual en una zona que falla, puede forzar la desasociación del disco de la máquina virtual fallida y conectarlo a otra máquina virtual.

Requisitos

- **Compatibilidad con regiones:** Para obtener una lista de las regiones que admiten discos administrados de ZRS, consulte Opciones de redundancia para discos administrados.
- **Tipos de disco:** Los discos con redundancia de zona se admiten con discos administrados SSD Premium y SSD estándar. ZRS no se admite con discos SSD Premium v2, Discos Ultra ni discos HDD estándar administrados.

Costo

ZRS incurre en un costo mayor que LRS debido a la sobrecarga de replicación adicional y la infraestructura necesarias para mantener los datos en varias zonas. La diferencia exacta de costos varía según la región y el tipo de disco. Para obtener información detallada sobre los precios, consulte Precios de discos administrados de Azure.

Configurar soporte de zonas de disponibilidad

- **Crear un nuevo disco ZRS:** Para crear un disco administrado ZRS, consulte Tutorial: administración de discos Azure con el CLI de Azure para máquinas virtuales Linux o Tutorial: Administrar discos con Azure PowerShell para máquinas virtuales Windows. Seleccione un nivel de disco ZRS durante la creación del disco.

Es responsable de conectar el disco a las máquinas virtuales, incluida la configuración de discos compartidos en varias máquinas virtuales en distintas zonas si es adecuado para la carga de trabajo.

- **Cambie un disco existente para usar ZRS:** Puede convertir un disco nozonal (regional) existente en ZRS.

  Aunque no se puede convertir un disco LRS zonal a ZRS, puede crear un nuevo disco ZRS a partir de una instantánea. Consulte Conversión de un disco de LRS a ZRS para conocer los requisitos y procedimientos de migración paso a paso.

- **Deshabilite la compatibilidad con zonas de disponibilidad:** No es posible cambiar la configuración de zona de disponibilidad de un disco ZRS existente. En su lugar, debe crear un nuevo disco con la nueva configuración mediante una instantánea del disco anterior y eliminar el anterior.

Comportamiento cuando todas las zonas están en buen estado

En esta sección se describe qué esperar cuando los discos administrados están configurados para usar ZRS y todas las zonas de disponibilidad están operativas.

- **Operación entre zonas:** Azure administra automáticamente el enrutamiento del tráfico entre zonas de disponibilidad cuando se utiliza una máquina virtual con un disco con redundancia de zona. Durante las operaciones normales, las solicitudes se distribuyen entre zonas de forma transparente.

- **Replicación de datos entre zonas:** Los discos ZRS replican cada escritura de forma sincrónica en varias zonas de disponibilidad de la región. Una operación de escritura solo se completa después de que los datos se almacenen en clústeres en varias zonas. Este enfoque proporciona una alta coherencia y alta disponibilidad, pero puede introducir una latencia de escritura ligeramente mayor en comparación con los discos LRS.

Comportamiento durante un fallo de zona

En esta sección se describe qué esperar cuando un disco administrado está configurado para usar ZRS y hay una interrupción de la zona de disponibilidad.

- **Detección y respuesta:** Algunas interrupciones de zona pueden afectar solo a los discos, solo las máquinas virtuales o ambas. El comportamiento que observará depende de si la interrupción de zona afecta a la máquina virtual conectada al disco.

  Si la máquina virtual sigue en buen estado, pero el disco se ve afectado por la interrupción, la máquina virtual continúa funcionando. Microsoft redirige

automáticamente las operaciones de disco para trabajar con los datos en zonas de disponibilidad saludables, y no necesitas hacer nada.

Si la máquina virtual está inactiva, debe cambiar la carga de trabajo a otra máquina virtual en una zona de disponibilidad diferente.

- *Discos compartidos:* Si ya ha creado la máquina virtual secundaria en una zona diferente y ha configurado discos compartidos, el disco está disponible para que la máquina virtual secundaria se use sin cambios de configuración necesarios.
- *Discos que no se comparten:* Puede *forzar la desasociación* del disco de la máquina virtual con errores y luego conectarlo a una máquina virtual en una zona saludable. Para realizar una desconexión forzada:
  - CLI de Azure: Utiliza el comando az vm disk detach con el argumento `--force-detach`.
  - Azure PowerShell: Use el cmdlet Remove-AzVMDataDisk junto con el argumento `-ForceDetach`.

- **Notification**: Microsoft no le notifica automáticamente cuando una zona está inactiva. Sin embargo, puede usar Azure Resource Health para supervisar el estado de un recurso individual y puede configurar Resource Health alertas para notificarle problemas. También puede usar Azure Service Health para comprender el estado general del servicio, incluidos los errores de zona, y puede configurar alertas de Service Health para notificarle problemas.
- **Pérdida de datos esperada:** No se produce ninguna pérdida de datos durante los errores de zona.
- **Tiempo de inactividad esperado:** Cuando el disco se comparte entre varias máquinas virtuales, no se espera ningún tiempo de inactividad.
- **Redistribution:** Azure vuelve a enrutar automáticamente el tráfico a otra copia de su disco en una zona saludable.

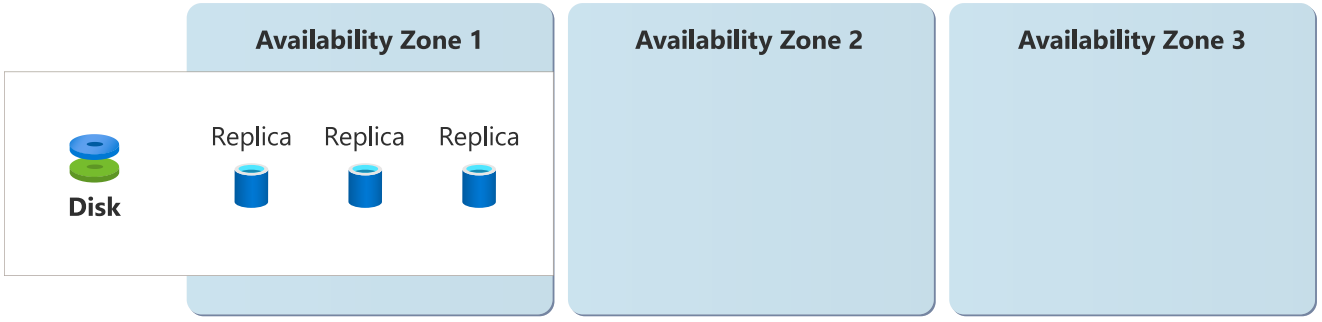
Recuperación de zona

Azure detecta automáticamente cuando la zona que falló anteriormente está saludable y restablece la sincronización de datos en la zona recuperada.

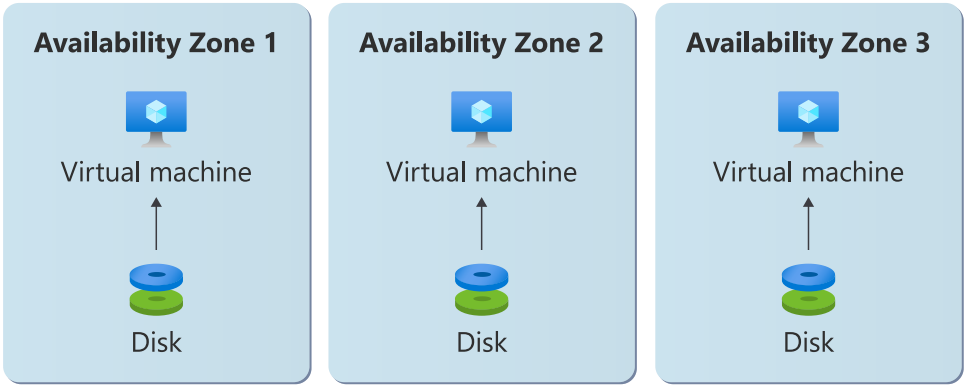
Discos LRS zonales

Los discos LRS zonales residen en una zona de disponibilidad específica y solo se conectan a las máquinas virtuales de esa zona. Todas las copias de los datos del disco están en la misma

zona. Un único disco LRS zonal y una máquina virtual no proporcionan resistencia de zona. Si la zona que contiene el disco experimenta una interrupción, el disco puede dejar de estar disponible.



En el caso de las cargas de trabajo de varias máquinas virtuales, puede lograr resistencia de zona mediante la implementación de varias máquinas virtuales y sus discos LRS zonales en diferentes zonas de disponibilidad. Este enfoque es la manera más común de lograr alta disponibilidad para cargas de trabajo como servidores web, niveles de aplicación y clústeres de bases de datos. Si se produce un error en una zona, puede configurar la carga de trabajo para seguir funcionando mediante las máquinas virtuales en zonas correctas.



Este patrón de distribución de varias zonas funciona con todos los tipos de disco, incluidos discos SSD Premium v2 y Ultra, que solo admiten LRS. Para más información sobre este enfoque, consulte Distribución de máquinas virtuales y discos entre zonas de disponibilidad.

Requisitos

- **Compatibilidad con regiones:** Los discos administrados LRS zonales se admiten en todas las regiones con zonas de disponibilidad.
- **Tipos de disco:** Todos los tipos de disco administrado admiten implementaciones de LRS zonales.

Costo

Los discos LRS zonales se cobran a la misma tarifa que los discos no zonales. Para obtener información detallada sobre los precios, consulte Precios de discos administrados de Azure.

Configurar soporte de zonas de disponibilidad

- **Crear un nuevo disco con compatibilidad con zona de disponibilidad:** Para crear un nuevo disco administrado con redundancia de LRS zonal, consulte Tutorial: Administración de discos Azure con el CLI de Azure para máquinas virtuales Linux o Tutorial: Administrar discos con Azure PowerShell para máquinas virtuales Windows.

  Seleccione la zona de disponibilidad durante la creación del disco.

  > **Importante**
  >
  > La asignación a una sola zona de disponibilidad solo se recomienda cuando la latencia entre zonas es demasiado alta para sus necesidades y después de comprobar que la latencia no cumple sus requisitos. Por sí mismo, un recurso zonal no proporciona resistencia a una interrupción de zona de disponibilidad. Para mejorar la resiliencia de un recurso zonal, debe desplegar explícitamente recursos independientes en múltiples zonas de disponibilidad y configurar el enrutamiento del tráfico y la conmutación por error. Para obtener más información, consulte Recursos zonales y resistencia de zona.

- **Cambie la configuración de zona de disponibilidad de un disco existente:** No es posible cambiar la configuración de zona de disponibilidad de un disco LRS zonal existente. En su lugar, debe crear un nuevo disco con la nueva configuración mediante una instantánea del disco anterior y eliminar el anterior.

Comportamiento cuando todas las zonas están en buen estado

En esta sección se describe qué esperar cuando se configura un disco administrado para usar LRS zonal y todas las zonas de disponibilidad están operativas.

- **Operación entre zonas:** El tráfico entre una máquina virtual zonal y un disco LRS zonal de la misma zona permanece dentro de la zona de disponibilidad.

  Al desplegar varias máquinas virtuales a través de zonas, es responsable de distribuir las solicitudes entrantes entre las máquinas virtuales. Cada máquina virtual lee y escribe en su propio disco zonal.

- **Replicación de datos entre zonas:** Todas las operaciones de escritura en discos LRS zonales se replican sincrónicamente dentro de la zona de disponibilidad.

Al implementar varias máquinas virtuales entre zonas, si la carga de trabajo requiere coherencia de datos entre máquinas virtuales, es responsable de sincronizar los datos, como mediante la replicación de bases de datos o la replicación de capa de aplicación.

Comportamiento durante un fallo de zona

En esta sección se describe qué esperar cuando un disco administrado está configurado para usar LRS zonal y hay una interrupción de zona de disponibilidad.

- **Detección y respuesta:** Si tienes una sola máquina virtual con un disco LRS zonal (almacenamiento con redundancia local), eres responsable de detectar una interrupción de zona y desencadenar una conmutación por error u otra respuesta.

  Cuando tienes máquinas virtuales distribuidas entre varias zonas, eres responsable de configurar tu tarea para detectar el fallo de zona y seguir ejecutándose en las máquinas virtuales en zonas saludables.

- **Notification**: Microsoft no le notifica automáticamente cuando una zona está inactiva. Sin embargo, puede usar Azure Resource Health para supervisar el estado de un recurso individual y puede configurar Resource Health alertas para notificarle problemas. También puede usar Azure Service Health para comprender el estado general del servicio, incluidos los errores de zona, y puede configurar alertas de Service Health para notificarle problemas.

- **Pérdida de datos esperada:** La replicación LRS proporciona al menos 99,999999999% (11 9) de durabilidad, lo que significa que el disco conserva sus datos y los datos se pueden recuperar después de que se recupere la zona.

  Cuando tiene máquinas virtuales distribuidas entre zonas, los datos que solo estaban en los discos de la zona con errores no están disponibles temporalmente. Si la aplicación sincroniza los datos entre máquinas virtuales, las máquinas virtuales de zonas correctas siguen atendiendo solicitudes mediante sus propios datos.

- **Tiempo de inactividad esperado:** Un único disco LRS zonal no está disponible hasta que se recupere la zona de disponibilidad.

  Cuando tiene máquinas virtuales y discos distribuidos entre zonas, la carga puede continuar operando en las máquinas virtuales en zonas saludables.

- **Redistribución:** Si tiene una sola máquina virtual con un disco LRS zonal, es responsable de redirigir el tráfico a otra máquina virtual si tiene una disponible.

  Cuando tenga máquinas virtuales distribuidas entre zonas, puede configurar la carga de trabajo para redistribuir automáticamente el tráfico a las máquinas virtuales en zonas

correctas.

Recuperación de zona

Cuando la zona de disponibilidad con errores se recupera, los discos administrados se recuperan automáticamente. Si la máquina virtual conectada al disco se ha visto afectada por la interrupción, se reinicia. Usted es responsable de resincronizar los datos de la aplicación con otras máquinas virtuales y discos en otras zonas de disponibilidad, si los usa.

Prueba de fallos de zona

No puede simular directamente errores de zona en el nivel de disco, pero puede usar la compatibilidad de Azure Chaos Studio para simular eventos de zona descendente en conjuntos de escalado de máquinas virtuales y simular la pérdida de una máquina virtual individual.

Debe probar la resistencia de la aplicación a errores de zona y el comportamiento del disco administrado durante las interrupciones. Supervise el rendimiento del disco durante las interrupciones de zona simulada y valide que las aplicaciones controlen una mayor latencia adecuadamente. Implemente escenarios de pruebas automatizadas que comprueben que las aplicaciones pueden controlar retrasos temporales de E/S y forzar las operaciones de desasociación de discos compartidos.

Resistencia a errores en toda la región

Azure Disk Storage es un servicio de una sola región que funciona dentro de los límites de una región de Azure específica. El servicio no incluye funciones nativas multirregionales ni recuperación automática frente a errores entre regiones. Si una región deja de estar disponible, los recursos de disco administrado de esa región tampoco están disponibles.

Soluciones personalizadas de varias regiones para la resistencia

Puede crear una solución de varias regiones mediante la implementación de máquinas virtuales y discos en cada región, replicación o copia de seguridad de datos entre regiones y conmutación por error o restauración a partir de copias de seguridad cuando sea necesario. Es responsable de gestionar recursos en cada región, coordinar y sincronizar datos, y gestionar la conmutación por error o la restauración. Algunos enfoques comunes son:

- Azure Site Recovery, que proporciona replicación entre regiones de las máquinas virtuales y discos.

- Azure Backup proporciona servicios de copia de seguridad administrados, incluidos los discos. Puede usar la restauración entre regiones para restaurar máquinas virtuales en otra región.
- Puede crear su propia solución basada en instantáneas copiando instantáneas de sus discos entre regiones.
- Algunas bases de datos y aplicaciones proporcionan enfoques de replicación que funcionan entre regiones mediante la replicación de cambios y la administración de clústeres. Por ejemplo, grupos de disponibilidad Siempre activas de SQL Server proporciona protección de datos entre regiones, con coherencia personalizable y un comportamiento de conmutación por error adaptado a las aplicaciones.

Copias de seguridad y restauración

Azure discos administrados admiten varios enfoques de copia de seguridad para protegerse contra la pérdida de datos y los daños. Azure Disk Backup es una solución nativa basada en la nube que automatiza la administración del ciclo de vida de las instantáneas. Proporciona copias de seguridad incrementales consistentes ante fallos con políticas de retención configurables. Este enfoque sin agente admite varias copias de seguridad al día sin afectar al rendimiento de la aplicación y se integra con Azure Backup Center para la administración centralizada. Puede usar instantáneas incrementales para reducir los costos de almacenamiento y los tiempos de copia de seguridad.

Para protección a nivel de MÁQUINA VIRTUAL, Azure Backup ofrece copias de seguridad coherentes con la aplicación para toda la máquina virtual, incluidos todos los discos conectados. Este enfoque es ideal cuando se necesita una copia de seguridad coordinada de varios discos o copias de seguridad compatibles con aplicaciones. En el caso de las cargas de trabajo de base de datos, considere las soluciones de copia de seguridad específicas de la aplicación que proporcionan una protección coherente con las transacciones y opciones de recuperación más rápidas.

En el caso de las cargas de trabajo críticas, implemente una estrategia de copia de seguridad superpuesta que combine Azure Disk Backup, la replicación de instantáneas entre regiones y las copias de seguridad de nivel de aplicación para la coherencia de las transacciones. Configure las directivas de copia de seguridad en función de los requisitos de recuperación, las necesidades de cumplimiento y las consideraciones de costos.

Acuerdo de nivel de servicio

El acuerdo de nivel de servicio (SLA) para Azure servicios describe la disponibilidad esperada de cada servicio y las condiciones que la solución debe cumplir para lograr esa expectativa de disponibilidad. Para obtener más información, vea SLAs for servicios en línea.

Azure Disk Storage no ofrece su propio SLA de disponibilidad, sino que está incluido en el SLA para máquinas virtuales. La configuración del disco puede afectar al Acuerdo de Nivel de Servicio de disponibilidad de la máquina virtual.

Contenido relacionado

- tipos de discos administrados de Azure
- Backup y recuperación ante desastres para discos administrados de Azure
- Mejores prácticas para lograr una alta disponibilidad con máquinas virtuales de Azure y discos administrados.
- Opciones de redundancia para Managed Disks
- confiabilidad Azure

ⓘ **Nota:** El autor ha creado este artículo con ayuda de inteligencia artificial. Más información

---

Last updated on 07/04/2026

Confiabilidad en Azure Elastic SAN

Azure Elastic SAN es un servicio de red de área de almacenamiento nativo en la nube (SAN) que proporciona una solución de almacenamiento escalable, rentable, de alto rendimiento y completa para una variedad de opciones de proceso. Elastic SAN permite crear y administrar volúmenes, que son discos virtuales que se pueden conectar a las máquinas virtuales, contenedores u otros servicios de Azure a través del protocolo iSCSI.

Cuando se usa Azure, la confiabilidad es una responsabilidad compartida. Microsoft proporciona una variedad de capacidades para apoyar la resiliencia y la recuperación. Es responsable de comprender cómo funcionan esas funcionalidades dentro de todos los servicios que usa y de seleccionar las funcionalidades que necesita para cumplir los objetivos empresariales y los objetivos de tiempo de actividad.

En este artículo se describe cómo hacer que Azure Elastic SAN sea resiliente a una variedad de posibles interrupciones y problemas, incluidos fallos transitorios, fallos de zona de disponibilidad y fallos a nivel regional. También describe las opciones de copia de seguridad y recuperación, y resalta la información clave sobre el acuerdo de nivel de servicio (SLA) de Azure Elastic SAN.

Recomendaciones de implementación de producción para la confiabilidad

En el caso de las cargas de trabajo de producción, se recomienda que usted:

- ✓ **Uso del almacenamiento con redundancia de zona (ZRS):** ZRS distribuye copias de los datos en tres zonas de disponibilidad.
- ✓ **Configure el acceso a la red para reducir la intervención manual durante los errores de zona:** Los puntos de conexión privados permiten la conmutación automática por error de zona sin intervención manual, mientras que los puntos de conexión de servicio pueden requerir intervención manual.
- ✓ **Para cargas de trabajo que requieren recuperación ante desastres entre regiones:** Cree instantáneas de los volúmenes con regularidad y expórtelas a instantáneas de disco administrado para la protección de datos. Copie las instantáneas en una región secundaria que esté geográficamente distante de la región primaria.

Introducción a la arquitectura de confiabilidad

Elastic SAN tiene una jerarquía de recursos de tres niveles:

- **ELASTIC SAN**: el recurso de nivel superior donde se configura la redundancia, se asigna la capacidad de almacenamiento y se establecen los límites de rendimiento. El número de unidades base asignadas a elastic SAN determina el total de IOPS y el rendimiento disponibles en toda la SAN.
- **Grupos de volúmenes**: construcciones de administración usadas para administrar volúmenes a escala. La configuración de acceso a la red, como puntos de conexión privados o puntos de conexión de servicio, se configura en el nivel de grupo de volúmenes y se hereda por todos los volúmenes del grupo.
- **Volúmenes**: Volúmenes de almacenamiento individuales divididos de la capacidad total del SAN. Los volúmenes están conectados a recursos de proceso a través del protocolo iSCSI.

Para más información sobre los recursos que implemente, consulte Recursos elásticos de SAN.

Internamente, Elastic SAN almacena los datos dentro de *los clústeres de almacenamiento*. Al configurar la SAN elástica para usar el almacenamiento con redundancia local (LRS), los datos se replican tres veces dentro de un único clúster de almacenamiento en un centro de datos. Puede configurar el almacenamiento con redundancia de zona (ZRS) para almacenar copias de los datos en tres zonas de disponibilidad. Para obtener más información, consulte Resilience to availability zone failures (Resistencia a errores de zona de disponibilidad).

Resistencia a errores transitorios

Los errores transitorios son errores breves e intermitentes en los componentes. Se producen con frecuencia en un entorno distribuido como la nube y son una parte normal de las operaciones. Los errores transitorios se corrigen después de un breve período de tiempo. Es importante que las aplicaciones puedan controlar errores transitorios, normalmente mediante el reintento de solicitudes afectadas.

Todas las aplicaciones hospedadas en la nube deben seguir las instrucciones de control de errores transitorios de Azure cuando se comunican con cualquier API, bases de datos y otros componentes hospedados en la nube. Para obtener más información, consulte Recomendaciones para controlar errores transitorios.

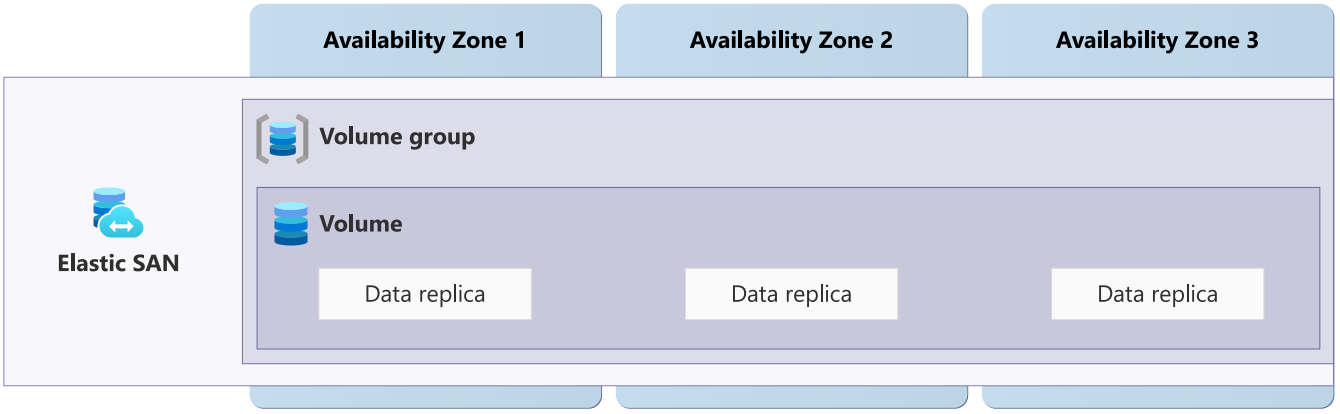
Si se interrumpe la conexión iSCSI a un volumen de Elastic SAN, el iniciador iSCSI en el cliente intenta volver a conectarse automáticamente. Es posible que experimente una breve pausa en las operaciones de E/S durante la reconexión. Configure el iniciador iSCSI con los valores de reintento y tiempo de espera adecuados para controlar las interrupciones transitorias. Para

obtener recomendaciones detalladas sobre la configuración de iSCSI y MPIO, consulte Optimización del rendimiento de elastic SAN.

Resistencia a errores de zona de disponibilidad

Las zonas de disponibilidad son grupos físicamente independientes de centros de datos dentro de una región de Azure. Cuando una zona falla, los servicios pueden transferirse a una de las zonas restantes.

Azure Elastic SAN se pueden configurar para usar el almacenamiento con redundancia de zona (ZRS), lo que significa que los datos se replican de forma sincrónica en tres zonas de disponibilidad de la región. La redundancia de zona le ayuda a lograr resistencia y confiabilidad para las cargas de trabajo de producción.



El enfoque de conectividad afecta a la capacidad de la carga de trabajo para conmutar de manera transparente durante una falla de zona. Se recomienda usar puntos de conexión privados para conectarse a los volúmenes. Los puntos de conexión privados admiten la conmutación automática por error. Si usa puntos de conexión de servicio, es posible que la conmutación por error requiera una intervención manual. Para obtener más información sobre los enfoques de conectividad, consulte Más información sobre las configuraciones de red para Elastic SAN.

El uso del almacenamiento con redundancia local (LRS) deja la SAN elástica desprotegida frente a errores de zona de disponibilidad. LRS hace que el SAN *elástico no es zonal* almacenando datos en una sola zona de disponibilidad en lugar de distribuirlos entre zonas con ZRS.

Requisitos

Compatibilidad con regiones: Los recursos de SAN elásticos con redundancia de zona se pueden implementar en un subconjunto de regiones. Para obtener una lista de las regiones, consulte Objetivos de escalado de Elastic SAN.

Cost

Al crear una SAN elástica con ZRS, el costo es mayor que LRS. Para obtener más información sobre los precios, consulte precios Azure Elastic SAN.

Configurar soporte de zonas de disponibilidad

- **Cree una nueva SAN elástica con ZRS:** Al crear una SAN elástica y seleccionar ZRS como opción de redundancia, la SAN elástica es automáticamente redundante por zona. No se puede cambiar la opción de redundancia después de crear la SAN elástica. Para obtener más información sobre cómo crear un nuevo recurso de Elastic SAN, consulte Implementación de una SAN elástica.

- **Habilite la redundancia de zona en una SAN elástica de LRS existente:** No se puede convertir una SAN elástica de LRS a ZRS directamente. Para migrar, realice una instantánea de los volúmenes de Elastic SAN, expórtelos a instantáneas de disco administrado, implemente una nueva SAN elástica en ZRS y, a continuación, cree volúmenes en el nuevo SAN elástico mediante esas instantáneas de disco. Para obtener más información, consulte Snapshot Azure Elastic SAN volúmenes.

Comportamiento cuando todas las zonas están en buen estado

En esta sección se describe qué esperar al configurar una SAN elástica para la redundancia de zona y todas las zonas están operativas.

- **Operación entre zonas:** Cuando se conecta a un volumen de Elastic SAN, la conexión iSCSI se enruta a un clúster en una de las zonas de disponibilidad. La plataforma enruta automáticamente el tráfico entre zonas.

- **Replicación de datos entre zonas:** Cuando un cliente escribe datos en un volumen de Elastic SAN, esos datos se escriben sincrónicamente en clústeres dentro de tres zonas de disponibilidad antes de que se confirme la operación de escritura. La replicación sincrónica garantiza un alto nivel de coherencia de datos y garantiza que no se pierdan datos durante un error de zona.

  La implementación de una Elastic SAN de ZRS proporciona más confiabilidad que una Elastic SAN de LRS, pero agrega más latencia de escritura. Haga pruebas comparativas de su Elastic SAN y simule la carga de trabajo de la aplicación para comparar la latencia entre LRS y ZRS, así podrá ver si afecta su carga de trabajo.

Comportamiento durante un fallo de zona

En esta sección se describe qué se puede anticipar al configurar una SAN elástica para la redundancia de zona, cuando ocurre una interrupción en una de las zonas.

- **Detección y respuesta:** La plataforma elastic SAN detecta los errores en una zona de disponibilidad. No es necesario hacer nada para iniciar una conmutación por error de zona para las SAN Elastic de ZRS.

- **Notificación:** Microsoft no le notifica automáticamente cuando una zona está inactiva. Sin embargo, puede usar Azure Service Health para comprender el estado general del servicio, incluidos los errores de zona, y puede configurar alertas de Service Health para notificarle problemas.

- **Solicitudes activas:** Cuando una zona de disponibilidad deja de estar disponible, el servicio puede finalizar las operaciones de E/S en curso que se conectan a las réplicas de la zona afectada y debe reintentar esas operaciones. Si usa puntos de conexión privados, el servicio conmuta automáticamente. Si usa puntos de conexión de servicio, es posible que deba reiniciar el iniciador iSCSI para cambiar a una zona sana.

- **Pérdida de datos esperada:** Elastic SAN evita la pérdida de datos durante un error de zona mediante la replicación sincrónica de datos entre tres zonas de disponibilidad.

- **Tiempo de inactividad esperado:** Cuando se usan puntos de conexión privados, la conmutación por error de zona se produce automáticamente. Es posible que experimente una degradación del rendimiento y la disponibilidad durante unos minutos después de una conmutación por error mientras la SAN se reequilibra por sí misma.

  Cuando usas puntos de conexión de servicio, Elastic SAN no cambia automáticamente a una zona saludable. Es posible que tenga que reiniciar el iniciador iSCSI para iniciar una conmutación por error a una zona correcta diferente.

- **Reenrutamiento del tráfico:** Cuando una zona no está disponible, la plataforma Elastic SAN detecta la pérdida de la zona y enruta el tráfico a las zonas correctas restantes.

  Para garantizar la conmutación automática por error de zona sin intervención manual, use puntos de conexión privados para conectarse a la SAN elástica. Al usar puntos de conexión de servicio en lugar de puntos de conexión privados, es posible que tenga que reiniciar el iniciador iSCSI manualmente para iniciar una conmutación por error a una zona en buen estado.

Recuperación de zona

Cuando se recupera la zona de disponibilidad, la plataforma Elastic SAN restaura automáticamente las operaciones normales y reanuda la replicación en tres zonas. No tiene que realizar ninguna acción.

Prueba de fallos de zona

La plataforma Azure Elastic SAN administra el enrutamiento del tráfico, la conmutación por error y la recuperación de zona para los recursos con redundancia de zona. Dado que esta característica está totalmente administrada, no es necesario validar los procesos de error de zona de disponibilidad.

Resistencia a errores en toda la región

Azure Elastic SAN es un servicio de una sola región. Si la región deja de estar disponible, el recurso de Elastic SAN tampoco está disponible. No hay replicación entre regiones ni conmutación por error integrada de una región a otra para Elastic SAN. Es responsable de diseñar su propia solución de recuperación ante desastres de varias regiones si la carga de trabajo requiere resistencia de nivel de región.

Soluciones personalizadas de varias regiones para la resistencia

Es responsable de implementar la recuperación ante desastres de varias regiones para los datos de Elastic SAN. El enfoque recomendado es usar instantáneas de volumen:

1. **Cree instantáneas periódicamente.** Utiliza instantáneas de volumen para capturar copias puntuales de tus volúmenes de Elastic SAN.

   El objetivo de punto de recuperación (RPO) depende de la frecuencia con la que cree y copie instantáneas en la región secundaria. Cuanto más frecuentemente cree instantáneas y cópielas, menor será la pérdida de datos potencial durante un desastre.

2. **Exportar instantáneas a instantáneas de disco administrado.**Exporte las instantáneas de volumen a instantáneas de disco administrado, que se pueden copiar en otras regiones.

3. **Copie instantáneas en una región secundaria.**Copie la instantánea incremental en una nueva región que esté geográficamente distante de la región primaria. Esto reduce el riesgo de que varias regiones se vean afectadas por un único desastre.

4. **Restaurar desde capturas.** En un escenario de recuperación ante desastres, cree nuevos volúmenes en la SAN elástica secundaria a partir de las instantáneas de disco administrado copiadas.

El objetivo de tiempo de recuperación (RTO) depende del tamaño de los datos, el tiempo necesario para copiar instantáneas entre regiones y el tiempo necesario para implementar y configurar una nueva SAN elástica en la región secundaria. Para reducir el tiempo de recuperación, considere la posibilidad de implementar una SAN elástica secundaria en la región de recuperación antes de que se produzca un desastre. Esto también ayuda a evitar restricciones de capacidad durante una interrupción.

Copias de seguridad y restauración

Para la mayoría de las soluciones, no debe confiar exclusivamente en copias de seguridad. En su lugar, utilice las otras capacidades descritas en esta guía para apoyar los requisitos de resiliencia. Sin embargo, las copias de seguridad protegen contra algunos riesgos que otros enfoques no. Para más información, consulte ¿Qué son la redundancia, la replicación y la copia de seguridad?.

Elastic SAN admite dos tipos de copia de seguridad:

- **Azure Elastic SAN backup:** Esta funcionalidad de Azure Backup ofrece una solución totalmente administrada para programar copias de seguridad, establecer escalas de tiempo de expiración para puntos de recuperación y datos de recuperación en un nuevo volumen. Ayuda a protegerse contra la pérdida de datos frente a eliminaciones accidentales, ransomware y actualizaciones de aplicaciones. Para obtener más información, consulte Azure Elastic SAN backup (versión preliminar).

  > **Importante**
  >
  > Como esta solución está cubierta por los términos complementarios de Microsoft para las versiones preliminares de Azure, úsela para pruebas y no para uso de producción.

- **Instantáneas:** Como alternativa, puede crear y administrar instantáneas en función de los requisitos de protección de datos.

  Azure Elastic SAN admite instantáneas de volumen para la protección de datos. Las instantáneas son copias incrementales en un momento específico de sus volúmenes que consumen espacio de la capacidad total de su Elastic SAN. Para proteger los datos, cree instantáneas periódicamente. La frecuencia depende de la cantidad de datos que puede

permitirse perder (su RPO). Puede crear instantáneas manualmente o crear su propia automatización para crearlas según una programación.

Las instantáneas se almacenan en la misma SAN elástica que tus volúmenes y usan la misma configuración de redundancia. Para protegerse frente a errores en toda la región, exporte las instantáneas a instantáneas de disco administrado y cópielas en otra región. Para obtener más información, consulte Exportación de instantáneas de volumen y Copiar una instantánea incremental a una nueva región.

Puede crear un nuevo volumen SAN elástico a partir de una instantánea o de una instantánea de un disco administrado. Para obtener más información, consulte Creación de un volumen a partir de una instantánea.

Resistencia al mantenimiento del servicio

Microsoft aplica periódicamente actualizaciones de servicio y realiza otro mantenimiento. La plataforma Azure controla estas actividades automáticamente, lo que garantiza que el mantenimiento sea transparente y sin problemas. No se espera ningún tiempo de inactividad durante los eventos de mantenimiento a menos que se le haya informado a través del mantenimiento planeado de Azure Service Health.

Acuerdo de nivel de servicio

El acuerdo de nivel de servicio (SLA) para Azure servicios describe la disponibilidad esperada de cada servicio y las condiciones que la solución debe cumplir para lograr esa expectativa de disponibilidad. Para obtener más información, consulte Acuerdos de Nivel de Servicio para servicios en línea.

Contenido relacionado

- ¿Qué es Azure Elastic SAN?
- Plan para implementar una instancia de Elastic SAN
- Confiabilidad en Azure

Last updated on 28/04/2026

Confiabilidad en Azure Files

En este artículo se describe el soporte para la confiabilidad en Azure Files. Azure Files proporciona recursos compartidos de archivos totalmente administrados en la nube a los que se puede acceder a través de protocolos estándar del bloque de mensajes del servidor (SMB) y del sistema de archivos de red (NFS) estándar del sector.

Cuando se usa Azure, la confiabilidad es una responsabilidad compartida. Microsoft proporciona una variedad de funcionalidades para admitir resistencia y recuperación. Es responsable de comprender cómo funcionan esas funcionalidades dentro de todos los servicios que usa y de seleccionar las funcionalidades que necesita para cumplir los objetivos empresariales y los objetivos de tiempo de actividad.

En este artículo se describe cómo hacer que Azure Files sea resistente a una variedad de posibles interrupciones y problemas, incluidos errores transitorios, interrupciones de zona de disponibilidad y interrupciones de regiones. También se describe cómo puede usar copias de seguridad para recuperarse de otros tipos de problemas y se resalta cierta información clave sobre el acuerdo de nivel de servicio (SLA) de Azure Files.

> **Nota:**
>
> Azure Files forma parte de la plataforma de Azure Storage. Algunas de las funcionalidades de Azure Files son comunes en muchos servicios de Azure Storage. En este artículo, se usa *Azure Storage* para hacer referencia a estas funcionalidades comunes.

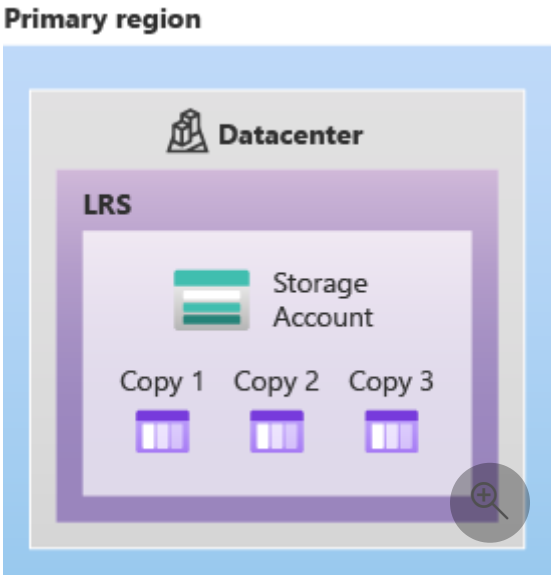
Recomendaciones de implementación de producción

Para obtener información sobre cómo implementar Azure Files para admitir los requisitos de confiabilidad de la solución y cómo afecta la confiabilidad a otros aspectos de la arquitectura, consulte Procedimientos recomendados de arquitectura para Azure Files en Azure Well-Architected Framework.

Introducción a la arquitectura de confiabilidad

El almacenamiento con redundancia local (LRS) replica los datos de las cuentas de almacenamiento en una o varias zonas de disponibilidad de Azure ubicadas en la región primaria de su elección. Aunque no hay ninguna opción para elegir la zona de disponibilidad preferida, Azure puede mover o expandir cuentas LRS entre zonas para mejorar el equilibrio de

carga. No hay ninguna garantía de que los datos se distribuirán entre zonas. Para obtener más información sobre las zonas de disponibilidad, consulte ¿Qué son las zonas de disponibilidad?.



El almacenamiento con redundancia de zona (ZRS), el almacenamiento con redundancia geográfica (GRS) y el almacenamiento con redundancia de zona geográfica (GZRS) proporcionan protecciones adicionales. En este artículo se describen estas opciones al detalle.

Azure Files está disponible en dos niveles multimedia:

- El **nivel Premium** usa unidades de estado sólido (SSD) para un alto rendimiento. Este nivel se recomienda para cargas de trabajo que requieren baja latencia.
- El **nivel Estándar** admite unidades de disco duro (HDD). Los recursos compartidos de archivos HDD proporcionan una opción de almacenamiento rentable para recursos compartidos de archivos de uso general.

Para más información, consulte Planeamiento de la implementación de Azure Files: niveles de almacenamiento.

Azure Files implementa redundancia en el nivel de cuenta de almacenamiento y los recursos compartidos de archivos heredan esa configuración de redundancia automáticamente. El servicio admite varios modelos de redundancia que difieren en su enfoque de protección de datos.

Resistencia a errores transitorios

Los errores transitorios son errores breves e intermitentes en los componentes. Se producen con frecuencia en un entorno distribuido como la nube y son una parte normal de las operaciones. Los errores transitorios se corrigen después de un breve período de tiempo. Es

importante que las aplicaciones puedan controlar errores transitorios, normalmente mediante el reintento de solicitudes afectadas.

Todas las aplicaciones hospedadas en la nube deben seguir las instrucciones de control de errores transitorios de Azure cuando se comunican con cualquier API, bases de datos y otros componentes hospedados en la nube. Para obtener más información, consulte Recomendaciones para controlar errores transitorios.

Para administrar de forma eficaz los errores transitorios al usar Azure Files, configure los valores de tiempo de espera adecuados para las operaciones de archivo en función del tamaño de archivo y las condiciones de red. Los archivos más grandes requieren tiempos de espera más largos, mientras que las operaciones más pequeñas pueden usar valores más cortos para detectar errores rápidamente.

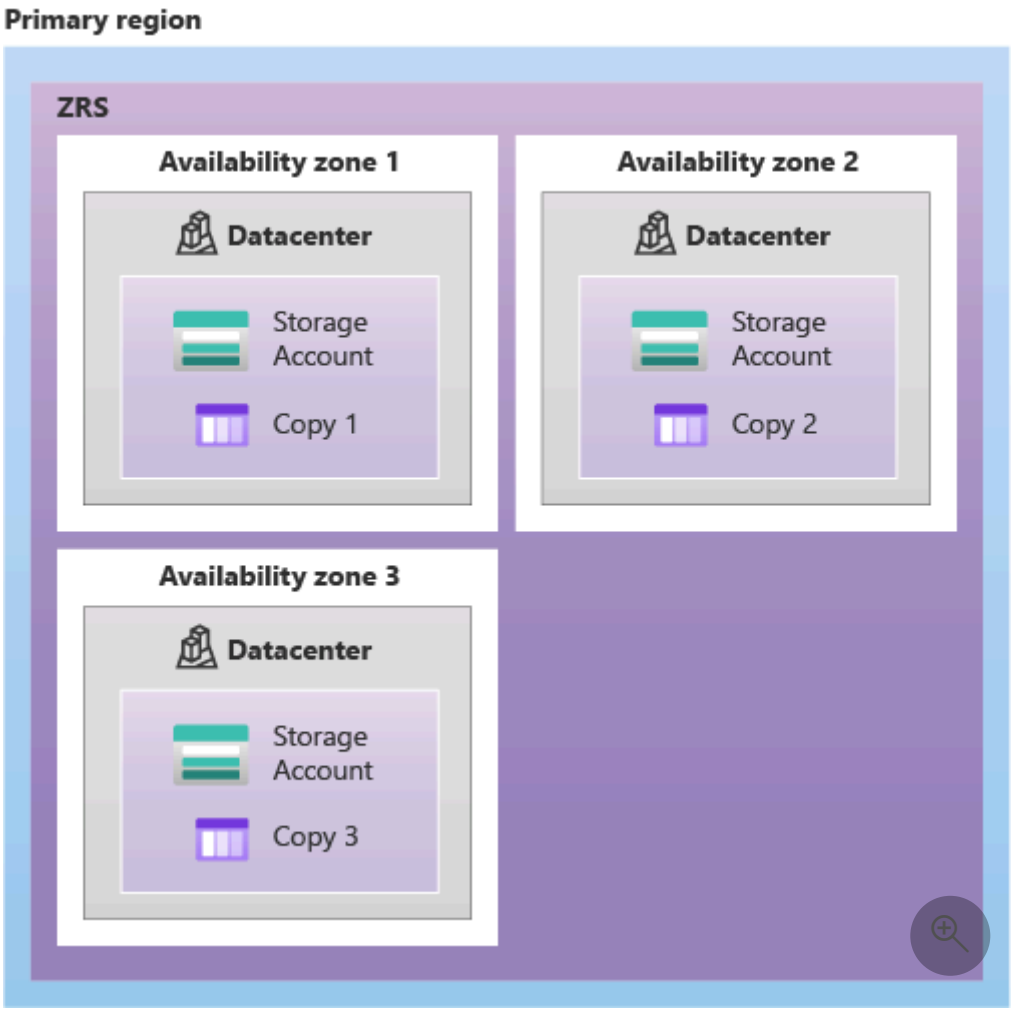
Para asegurarse de que solo se establecen conexiones seguras en el recurso compartido NFS, se recomienda configurar un punto de conexión privado para la cuenta de almacenamiento. Un punto de conexión privado usa Azure Private Link para asignar una dirección IP estática a la cuenta de almacenamiento desde el espacio de direcciones privadas de la red virtual. Un punto de conexión privado ayuda a evitar interrupciones de conectividad de los cambios dinámicos de direcciones IP. Para obtener más información sobre la seguridad de los recursos compartidos NFS, consulte Recursos compartidos de archivos NFS: seguridad y redes.

Resistencia a errores de zona de disponibilidad

Las zonas de disponibilidad son grupos físicamente independientes de centros de datos dentro de una región de Azure. Cuando una zona falla, los servicios pueden transferirse a una de las zonas restantes.

Azure Files proporciona dos tipos de compatibilidad con zonas de disponibilidad:

- **Almacenamiento con redundancia de zona (ZRS):** Las configuraciones de ZRS distribuyen automáticamente los datos entre varias zonas de disponibilidad dentro de una región. A diferencia de LRS, ZRS garantiza que Azure replica de forma sincrónica los datos de archivo en varias zonas de disponibilidad. ZRS garantiza que sus datos sigan estando accesibles incluso si se produce un corte en una zona.



- **Ubicación zonal con LRS:** En el caso de las cuentas de almacenamiento premium (nivel multimedia SSD), puede usar la ubicación zonal para seleccionar la zona de disponibilidad específica en la que reside la cuenta de almacenamiento de Azure Files. Puede usar la ubicación zonal si necesita colocar máquinas virtuales (VM) en la misma zona para reducir la latencia entre el proceso y el almacenamiento.

> ⓘ **Importante**
>
> La asignación a una sola zona de disponibilidad solo se recomienda cuando la latencia entre zonas es demasiado alta para sus necesidades y después de comprobar que la latencia no cumple sus requisitos. Por sí mismo, un recurso zonal no proporciona resistencia a una interrupción de zona de disponibilidad. Para mejorar la resiliencia de un recurso zonal, debe desplegar explícitamente recursos independientes en múltiples zonas de disponibilidad y configurar el enrutamiento del tráfico y la conmutación por error. Para obtener más información, consulte Recursos zonales y resistencia de zona.

Requisitos

- **Compatibilidad con regiones:**
  - *ZRS:* ZRS es compatible con:
    - *Recursos compartidos de archivos HDD (estándar)* en todas las regiones con zonas de disponibilidad.
    - *Recursos compartidos de archivos SSD (Premium)* a través del `FileStorage` tipo de cuenta de almacenamiento. Para obtener una lista de las regiones que admiten ZRS para cuentas de recursos compartidos de archivos SSD, consulte Compatibilidad con ZRS para recursos compartidos de archivos SSD.
  - *LRS con ubicación zonal:* LRS con ubicación zonal se admite para recursos compartidos de archivos SSD (Premium) en regiones admitidas.
- **Tipos de recurso compartido de archivos:**
  - *ZRS:* ZRS es compatible con todos los tipos de recursos compartidos de archivos.
  - *LRS con ubicación zonal:* LRS con ubicación zonal está disponible para las cuentas de almacenamiento que cumplen los siguientes requisitos:
    - Debe usar el nivel de almacenamiento Premium (nivel multimedia SSD).
    - Solo comparticiones de archivos clásicas de Azure (mediante el proveedor de recursos Microsoft.Storage). La ubicación zonal no es posible actualmente para los recursos compartidos de archivos creados con el proveedor de recursos Microsoft.FileShares (versión preliminar).

Cost

El impacto en el costo es diferente en función del tipo de soporte de zona de disponibilidad que use:

- *ZRS:* Al habilitar el almacenamiento con redundancia de zona (ZRS), se le cobra a una velocidad diferente del almacenamiento con redundancia local (LRS) debido a la sobrecarga adicional de replicación y almacenamiento.
- *LRS con ubicación zonal:* LRS con ubicación zonal se cobra a la misma tarifa que LRS.

Para obtener información detallada sobre los precios, consulte Precios de Azure Files.

Configurar soporte de zonas de disponibilidad

- **Cree un recurso compartido de archivos con soporte para zonas de disponibilidad:**

- *ZRS:* Para crear un nuevo recurso compartido de archivos con ZRS, consulte Creación de un recurso compartido de archivos de Azure y seleccione **ZRS** o **GZRS** como opción de redundancia durante la creación de la cuenta.
- *LRS con ubicación zonal:* Para crear una nueva cuenta de almacenamiento de archivos con ubicación zonal, consulte Creación de una nueva cuenta de almacenamiento zonal.

- **Cambiar el tipo de replicación:**
  - *ZRS:* Para convertir una cuenta de almacenamiento existente en ZRS y obtener información sobre las opciones y los requisitos de migración, consulte Cambio de la configuración de redundancia para Azure Files.
  - *LRS con ubicación zonal:* Para anclar una cuenta de almacenamiento existente a una zona seleccionada por Azure, consulte Anclar una cuenta de almacenamiento existente a una zona seleccionada por Azure.
- **Deshabilite la compatibilidad con zonas de disponibilidad:**
  - *ZRS:* Vuelva a convertir las cuentas de ZRS en una configuración no zonal, como LRS, a través del mismo proceso de cambio de configuración de redundancia.
  - *LRS con ubicación zonal:* Para desanclar una cuenta de almacenamiento de una zona y, a continuación, convertir la cuenta de almacenamiento zonal en una cuenta de almacenamiento regional, consulte Desanclar una cuenta de almacenamiento de una zona.

Comportamiento cuando todas las zonas están en buen estado

En esta sección se describe qué esperar cuando se configura una cuenta de almacenamiento de archivos para la compatibilidad con zonas de disponibilidad y todas las zonas de disponibilidad están operativas.

- **Enrutamiento de tráfico entre zonas:**
  - *ZRS:* Azure Storage con almacenamiento con redundancia de zona (ZRS) distribuye automáticamente las solicitudes entre clústeres de almacenamiento en varias zonas de disponibilidad. La distribución del tráfico es transparente para las aplicaciones y no requiere ninguna configuración del lado cliente.
  - *LRS con ubicación zonal:* Azure Storage con almacenamiento con redundancia local (LRS) distribuye automáticamente las solicitudes entre clústeres de almacenamiento en

la zona de disponibilidad seleccionada. La distribución del tráfico es transparente para las aplicaciones y no requiere ninguna configuración del lado cliente.

- **Replicación de datos entre zonas:**
  - *ZRS:* Todas las operaciones de escritura en ZRS se replican sincrónicamente en todas las zonas de disponibilidad de la región. Al cargar o modificar datos, la operación no se considera completa hasta que los datos se hayan replicado correctamente en todas las zonas de disponibilidad. Esta replicación sincrónica garantiza una fuerte coherencia y cero pérdida de datos durante los errores de zona.
  - *LRS con ubicación zonal:* Todas las operaciones de escritura en LRS se replican sincrónicamente en varias réplicas de almacenamiento dentro de la zona. Al cargar o modificar datos, la operación no se considera completa hasta que los datos se hayan replicado correctamente en todas las réplicas.

Comportamiento durante un fallo de zona

En esta sección se describe qué esperar cuando se configura una cuenta de almacenamiento de archivos para la compatibilidad con la zona de disponibilidad y se produce una interrupción de zona de disponibilidad.

- **Detección y respuesta:**
  - *ZRS:* Microsoft detecta automáticamente errores de zona e inicia procesos de recuperación. No se requiere ninguna acción del cliente para las cuentas de almacenamiento con redundancia de zona (ZRS). Si una zona deja de estar disponible, Azure lleva a cabo actualizaciones de red como el repunto del sistema de nombres de dominio (DNS).
  - *LRS con ubicación zonal:* Debe detectar la pérdida de una zona de disponibilidad. Si es necesario, puede iniciar una conmutación por error a un recurso compartido de archivos secundario que haya creado previamente en otra zona de disponibilidad.
- **Notificación**: Microsoft no le notifica automáticamente cuando una zona está inactiva. Sin embargo, puede usar Azure Resource Health para supervisar el estado de un recurso individual y puede configurar alertas de Resource Health para notificarle problemas. También puede usar Azure Service Health para comprender el estado general del servicio, incluidos los errores de zona, y puede configurar alertas de Service Health para notificarle problemas.
- **Solicitudes activas:**

- *ZRS:* Es posible que las solicitudes en curso se interrumpan durante el proceso de recuperación y se deben reintentar. Las aplicaciones deben implementar lógica de reintento para controlar estas interrupciones temporales.
- *LRS con colocación zonal:* Las solicitudes en vuelo se descartan y pueden reintentarse cuando la zona se recupere.

- **Pérdida de datos esperada:**
  - *ZRS*: no se produce ninguna pérdida de datos durante los errores de zona porque los datos se replican sincrónicamente en varias zonas antes de completar las operaciones de escritura.
  - *LRS con ubicación zonal:* Los datos de los recursos compartidos de archivos de la zona afectada no están disponibles hasta que se recupere la zona.
- **Tiempo de inactividad esperado:**
  - *ZRS:* Una pequeña cantidad de tiempo de inactividad, normalmente, unos segundos, puede producirse durante la recuperación automática a medida que el tráfico se redirige a zonas saludables. Cuando diseñe aplicaciones para ZRS, siga las prácticas de manejo de fallos transitorios, incluyendo la implementación de directivas de reintento con retroceso exponencial.
  - *LRS con ubicación zonal:* Los recursos compartidos de archivos de la zona afectada permanecen inactivos hasta que se recupere la zona de disponibilidad.
- **Reenrutamiento del tráfico:**
  - *ZRS:* Azure vuelve a enrutar automáticamente el tráfico a las zonas de disponibilidad correctas restantes. El servicio mantiene la funcionalidad completa mediante el uso de las zonas supervivientes sin que se requiera intervención del cliente. No es necesario volver a montar los recursos compartidos de archivos de Azure desde clientes conectados.
  - *LRS con ubicación zonal:* Usted es responsable de migrar a otras cuentas de almacenamiento de archivos en zonas saludables, si es necesario.

Recuperación de zona

El comportamiento de recuperación de zona depende del tipo de replicación que usa la cuenta de almacenamiento de archivos:

- *ZRS:* Cuando se recupera la zona de disponibilidad con errores, Azure Storage restaura automáticamente las operaciones normales en todas las zonas de disponibilidad. El

servicio garantiza automáticamente la coherencia de los datos mediante la sincronización de las operaciones que se produjeron durante el período de interrupción.

- *LRS con ubicación zonal:* Una vez que la zona está en buen estado, las comparticiones de archivos en la zona están disponibles de nuevo. Es responsable de cualquier procedimiento de recuperación de zona y sincronización de datos que requieran las cargas de trabajo.

Prueba de fallos de zona

Las opciones de prueba en una zona reducida dependen del tipo de replicación que utiliza la cuenta de almacenamiento de archivos.

- *ZRS:* Cuando se usa el almacenamiento con redundancia de zona (ZRS), Azure Storage administra automáticamente las respuestas de replicación, enrutamiento de tráfico y reducción de zona. Dado que esta característica está totalmente administrada, no es necesario iniciar ni validar los procesos de error de zona de disponibilidad.

- *LRS con ubicación zonal:* No hay forma de simular una interrupción de la zona de disponibilidad que contiene la cuenta de almacenamiento de archivos. Sin embargo, puede configurar manualmente aplicaciones ascendentes, firewalls, puertas de enlace o equilibradores de carga para redirigir el tráfico a una cuenta de almacenamiento de archivos diferente en una zona de disponibilidad diferente.

Resistencia a errores en toda la región

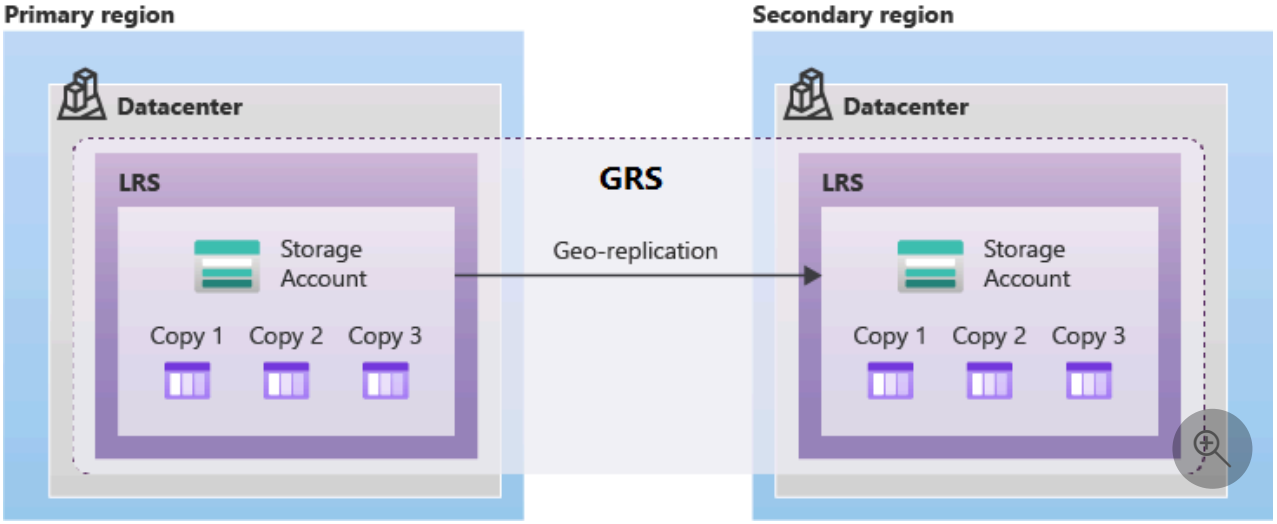
Azure Storage, que incluye Azure Blob Storage, Azure Files, Azure Table Storage y Azure Queue Storage, ofrece una amplia gama de capacidades de redundancia geográfica y conmutación por error para adaptarse a diferentes requisitos.

> **ⓘ Importante**
>
> El GRS solo funciona dentro de regiones emparejadas de Azure. Si la región de la cuenta de almacenamiento no está emparejada, considere la posibilidad de usar las soluciones personalizadas de varias regiones para lograr resistencia.

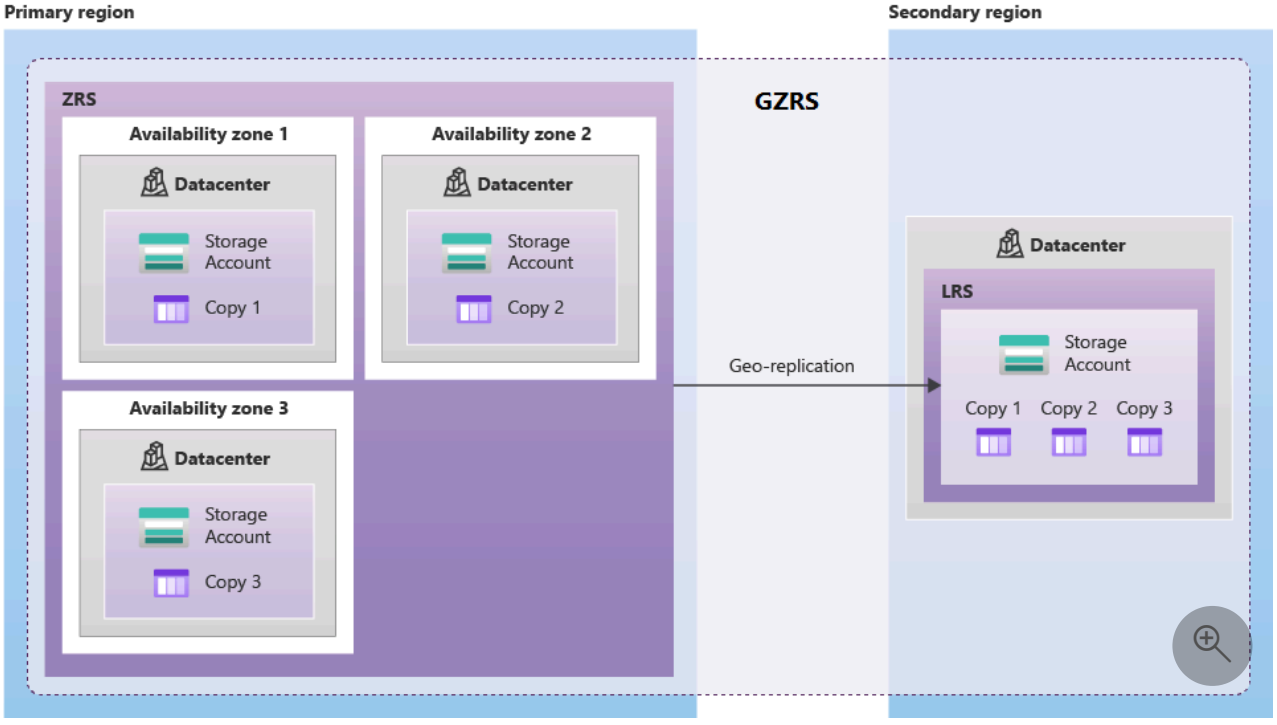
Almacenamiento con redundancia geográfica para regiones emparejadas

Azure Storage ofrece varios tipos de GRS en regiones emparejadas. Independientemente del tipo de GRS que utilice, los datos de la región secundaria siempre se replicarán mediante almacenamiento con redundancia local (LRS). Este enfoque proporciona protección contra fallos de hardware dentro de la región secundaria.

- GRS proporciona soporte para conmutaciones por error planificadas y no planificadas a la región emparejada de Azure cuando se produce una interrupción en la región principal. GRS replica de forma asincrónica los datos de la región primaria a la región emparejada.



- El almacenamiento con redundancia de zona geográfica (GZRS) replica los datos en varias zonas de disponibilidad de la región principal y en la región emparejada.



> **Importante**
>
> Azure Files solo admite redundancia geográfica (GRS o GZRS) para recursos compartidos de archivos estándar (HDD).

Azure Files no admite almacenamiento georeduntante de solo lectura (RA-GRS) ni almacenamiento geozonareduntante de solo lectura (RA-GZRS). Si una cuenta de almacenamiento está configurada para usar RA-GRS o RA-GZRS, los recursos compartidos de archivos estándar (HDD) se configuran y facturan como GRS o GZRS.

Tipos de conmutación por error

Azure Storage admite tres tipos de conmutación por error para distintos escenarios.

- **Conmutación por error no planeada administrada por el cliente:** es responsable de iniciar la recuperación si hubiera un error de almacenamiento en toda la región principal.
- **Conmutación por error planeada administrada por el cliente:** Usted es responsable de iniciar la recuperación si otra parte de su solución sufre un fallo en la región primaria y necesita trasladar toda su solución a la región secundaria. Use una conmutación por error planeada cuando el almacenamiento permanezca operativo en la región primaria, pero debe realizar una conmutación por error de toda la solución a una región secundaria, como para los simulacros de recuperación ante desastres diseñados para garantizar el cumplimiento de los requisitos de auditoría.
- **Conmutación por error administrada por Microsoft:** en circunstancias excepcionales, Microsoft podría iniciar la conmutación por error para todas las cuentas de almacenamiento con redundancia geográfica (GRS) de una región. Sin embargo, la conmutación por error administrada por Microsoft es un último recurso y se espera que solo se realice después de un período prolongado de interrupción. No se debe confiar en el failover administrado por Microsoft.

Las cuentas de GRS pueden utilizar cualquiera de estos tipos de conmutación por error. No es necesario preconfigurar una cuenta de almacenamiento para usar ninguno de los tipos de conmutación por error con antelación.

Requisitos

- **Compatibilidad con regiones:** Las configuraciones con redundancia geográfica de Azure Storage usan regiones emparejadas de Azure para la replicación de regiones secundarias. La región secundaria se determina automáticamente en función de la selección de la región principal y no se puede personalizar. Para obtener una lista completa de las regiones emparejadas de Azure, consulte Lista de regiones de Azure.

  Si la región de la cuenta de almacenamiento no está emparejada, considere la posibilidad de usar las soluciones personalizadas de varias regiones para lograr resistencia.

- **Solo recursos compartidos de archivos estándar:** Azure Files solo admite redundancia geográfica (GRS o GZRS) para recursos compartidos de archivos estándar (HDD). Los recursos compartidos de archivos Premium (SSD) deben utilizar LRS o ZRS. Si tiene archivos compartidos premium y quiere replicar los datos a través de regiones para una mayor resiliencia, consulte Soluciones personalizadas de varias regiones para resiliencia.

- **Solo GRS y GZRS:** Azure Files no admite el almacenamiento con redundancia geográfica con acceso de lectura (RA-GRS) ni el almacenamiento con redundancia de zona geográfica con acceso de lectura (RA-GZRS). Si una cuenta de almacenamiento está configurada para usar RA-GRS o RA-GZRS, los recursos compartidos de archivos estándar (HDD) se configuran y facturan como GRS o GZRS.

Consideraciones

Al implementar Azure Files de varias regiones, tenga en cuenta los siguientes factores importantes:

- **Latencia de replicación asincrónica:** la replicación de datos en la región secundaria es asincrónica, lo que significa que hay un retraso entre cuando los datos se escriben en la región primaria y cuando estarán disponibles en la región secundaria. Este retraso podría provocar una posible pérdida de datos si se produjese un fallo en la región principal antes de que se repliquen los datos recientes. La pérdida de datos se mide mediante el objetivo de punto de recuperación (RPO). Espere que el retraso de replicación sea inferior a 15 minutos, pero este tiempo es una estimación y no se garantiza.

  Compruebe la propiedad Última hora de sincronización para saber cuántos datos se podrían perder si su cuenta de almacenamiento sufriese una conmutación por error no planificada.

- **Hora de la última sincronización:** Para Azure Files, la hora de la última sincronización se basa en la instantánea del sistema más reciente de la región secundaria.

  El cálculo de la hora de la última sincronización puede producirse un tiempo de espera si hay más de 100 comparticiones de archivos en una cuenta de almacenamiento. Se recomienda implementar 100 o menos recursos compartidos de archivos para cada cuenta de almacenamiento para evitar tiempos de espera.

- **Acceso a la región secundaria:** No se accede a la región secundaria para lectura hasta que se produzca un failover.

- **Limitaciones de características:** Algunas características de Azure Files no se admiten o tienen limitaciones al usar GRS o conmutación por error administrada por el cliente. Estas limitaciones incluyen tipos de recursos compartidos de archivos específicos, niveles de

acceso y herramientas y operaciones de administración. Revise la documentación de compatibilidad de características antes de implementar la redundancia geográfica.

Cost

Las configuraciones de cuentas de Azure Storage multirregionales conllevan costes adicionales por la replicación entre regiones y el almacenamiento en la región secundaria. La transferencia de datos entre regiones de Azure se cobra en función de las tarifas estándar de ancho de banda entre regiones.

Para obtener información detallada sobre los precios, consulte Precios de Azure Files.

Configuración de la compatibilidad con varias regiones

- **Cree una nueva cuenta de almacenamiento con redundancia geográfica (GRS).** Para crear una cuenta de GRS, consulte Creación de una cuenta de almacenamiento y seleccione GRS, almacenamiento con redundancia geográfica con acceso de lectura (RA-GRS), almacenamiento con redundancia de zona geográfica (GZRS) o almacenamiento con redundancia de zona geográfica con acceso de lectura (RA-GZRS) durante la creación de la cuenta.

- **Habilite la redundancia geográfica en una cuenta de almacenamiento de archivos existente.** Para convertir una cuenta de almacenamiento de archivos existente en GRS, consulte Cambio de la configuración de redundancia para Azure Files.

  > ⚠ **Advertencia**
  >
  > Una vez que su cuenta se haya reconfigurado para la redundancia geográfica, es posible que transcurra un tiempo considerable antes de que los datos existentes en la nueva región principal se copien por completo a la nueva región secundaria.
  >
  > **Para evitar una pérdida importante de datos**, compruebe el valor de la **propiedad Última hora de sincronización** antes de iniciar una conmutación por error no planificada. Para evaluar la posible pérdida de datos, compare la hora de la última sincronización con la última vez en la que los datos se escribieron en la nueva región primaria.

- **Deshabilite la redundancia geográfica.** Vuelva a convertir las cuentas de GRS en configuraciones de una sola región (LRS o ZRS) a través del mismo proceso de cambio de configuración de redundancia.

Comportamiento cuando todas las regiones están en buen estado

En esta sección se describe qué esperar cuando una cuenta de almacenamiento está configurada para redundancia geográfica y todas las regiones están operativas.

- **Enrutamiento del tráfico entre regiones:** Azure Files usa un enfoque activo-pasivo en el que todas las operaciones de lectura y escritura se dirigen a la región primaria.
- **Replicación de datos entre regiones:** Las operaciones de escritura se confirman primero en la región primaria mediante el tipo de redundancia configurado (LRS para GRS o ZRS para GZRS). Después de finalizar correctamente en la región primaria, los datos se replican de forma asincrónica en la región secundaria, donde se almacenan mediante LRS.

  La naturaleza asíncrona de la replicación entre regiones significa que suele haber un retraso entre el momento en que los datos se escriben en la región principal y el momento en que están disponibles en la región secundaria. Supervise el tiempo de replicación utilizando la propiedad Última hora de sincronización.

Comportamiento durante una falla de región

En esta sección se describe qué esperar cuando una cuenta de almacenamiento está configurada para la redundancia geográfica y hay una interrupción en la región primaria.

- **Conmutación por error administrada por el cliente (no planificada):** utilice una conmutación por error no planificada cuando el almacenamiento en la región principal no esté disponible.
  - **Detección y respuesta:** En el caso poco probable de que su cuenta de almacenamiento no esté disponible en la región primaria, puede considerar iniciar un failover no planeado administrado por el cliente. Para tomar esta decisión, tenga en cuenta los siguientes factores:
    - Si Azure Resource Health muestra problemas al acceder a la cuenta de almacenamiento en su región principal
    - Si Microsoft le recomienda realizar una conmutación por error a otra región

    > ⚠ **Advertencia**
    >
    > Una conmutación por error no planeada puede provocar la pérdida de datos. Antes de iniciar una conmutación por error administrada por el cliente, decida si la restauración del servicio justifica el riesgo de pérdida de datos.

- **Notificación**: Microsoft no le notifica automáticamente cuando una región está inactiva. Sin embargo:
  - Puede usar Azure Resource Health para supervisar el estado de un recurso individual y puede configurar alertas de Resource Health para notificarle problemas.
  - Puede usar Azure Service Health para comprender el estado general del servicio, incluidos los errores de región, y puede configurar alertas de Service Health para notificarle problemas.
- **Solicitudes activas:** Durante el proceso de conmutación por error, los puntos de conexión de la cuenta de almacenamiento principal y secundaria dejan de estar disponibles temporalmente para lecturas y escrituras. Las solicitudes activas podrían fallar, y las aplicaciones cliente deben intentar nuevamente después de que se complete la conmutación por error.
- **Pérdida de datos prevista:** la pérdida de datos es habitual durante una conmutación por error no planificada debido al retraso en la replicación asíncrona, lo que significa que es posible que las escrituras recientes no se repliquen. Consulte la propiedad Última hora de sincronización para saber cuántos datos se podrían perder durante una conmutación por error no planificada. La pérdida de datos esperada se conoce a menudo como objetivo de punto de recuperación (RPO). Normalmente, puede esperar que el RPO sea inferior a 15 minutos, pero ese tiempo no está garantizado.
- **Tiempo de inactividad esperado:** La cantidad de tiempo de inactividad esperado se conoce a menudo como objetivo de tiempo de recuperación (RTO). La conmutación por error administrada por el cliente se completa normalmente en un plazo de hasta 60 minutos, dependiendo del tamaño y la complejidad de la cuenta.
- **Reenrutamiento del tráfico:** A medida que se completa la conmutación por error, Azure actualiza automáticamente los puntos de conexión de la cuenta de almacenamiento para que no sea necesario volver a configurar las aplicaciones. Si su aplicación mantiene entradas del Sistema de nombres de dominio (DNS) almacenadas en caché, puede que sea necesario borrar la caché para asegurarse de que la aplicación envíe el tráfico a la nueva región principal.
- **Configuración posterior a la conmutación por error:** una vez completada una conmutación por error no planificada, su cuenta de almacenamiento en la región de destino utilizará el nivel de almacenamiento con redundancia local (LRS). Si necesitase volver a replicar geográficamente los datos, deberá volver a habilitar el almacenamiento con redundancia geográfica (GRS) y esperar a que los datos se repliquen en la nueva región secundaria.

Para obtener más información sobre cómo iniciar una conmutación por error administrada por el cliente, consulte Cómo funciona la conmutación por error administrada por el cliente (no planificada) e Iniciar una conmutación por error de una cuenta de almacenamiento.

- **Conmutación por error administrada por el cliente (planificada):** utilice una conmutación por error planificada cuando el almacenamiento siga operativo en la región principal, pero necesite conmutar por error toda su solución a una región secundaria por otro motivo. Por ejemplo, otro servicio de Azure podría estar experimentando un problema y debe cambiar al uso de una región secundaria para toda la solución. O bien, puede usar una conmutación por error planeada para realizar un simulacro de recuperación ante desastres (DR) con fines de cumplimiento y auditoría.
  - **Detección y respuesta:** Usted es responsable de decidir conmutar por error. Normalmente, debe tomar esta decisión si necesita realizar una conmutación entre regiones, aunque su cuenta de almacenamiento esté en buen estado. Por ejemplo, puede desencadenar una conmutación por error cuando se produce una interrupción importante en otro componente de la aplicación que no se puede recuperar en la región principal.
    - **Notificación**: Microsoft no le notifica automáticamente cuando una región está inactiva. Sin embargo:
      - Puede usar Azure Resource Health para supervisar el estado de un recurso individual y puede configurar alertas de Resource Health para notificarle problemas.
      - Puede usar Azure Service Health para comprender el estado general del servicio, incluidos los errores de región, y puede configurar alertas de Service Health para notificarle problemas.
  - **Solicitudes activas:** Durante el proceso de conmutación por error, los puntos de conexión de la cuenta de almacenamiento principal y secundaria dejan de estar disponibles temporalmente para lecturas y escrituras. Las solicitudes activas podrían fallar, y las aplicaciones cliente deben intentar nuevamente después de que se complete la conmutación por error.
  - **Pérdida de datos esperada:** No se espera ninguna pérdida de datos porque el proceso de conmutación por error se completa solo después de sincronizar todos los datos, lo que da como resultado un RPO de cero.
  - **Tiempo de inactividad esperado:** La conmutación por error normalmente se completa en un plazo de 60 minutos, lo que significa que el tiempo de recuperación objetivo (RTO) esperado es de 60 minutos, en función del tamaño y la complejidad de la cuenta.

Durante el proceso de conmutación por error, los puntos de conexión de la cuenta de almacenamiento principal y secundaria dejan de estar disponibles temporalmente para lecturas y escrituras.

- **Reenrutamiento del tráfico:** A medida que se completa la conmutación por error, Azure actualiza automáticamente los puntos de conexión de la cuenta de almacenamiento para que no sea necesario volver a configurar las aplicaciones. Si su aplicación mantiene entradas DNS almacenadas en caché, puede que sea necesario borrar la caché para asegurarse de que la aplicación envíe tráfico a la nueva región principal.
- **Configuración posterior a la conmutación por error:** Una vez completada una conmutación por error planeada, la cuenta de almacenamiento de la región de destino sigue siendo replicada geográficamente y permanece en el nivel GRS.

Para obtener más información sobre cómo iniciar una conmutación por error administrada por el cliente, consulte Cómo funciona la conmutación por error administrada por el cliente (planificada) e Iniciar una conmutación por error de una cuenta de almacenamiento.

- **Conmutación por error administrada por Microsoft:** En el caso excepcional de un desastre importante en el que Microsoft determina que la región primaria es irrecuperable permanentemente, es posible que se inicie una conmutación automática por error a la región secundaria. Microsoft controla todo el proceso y no se requiere ninguna acción del cliente. El tiempo que transcurre antes de que ocurra un failover depende de la gravedad del desastre y del tiempo necesario para evaluar la situación.
  - **Notificación**: Microsoft no le notifica automáticamente cuando una región está inactiva. Sin embargo:
    - Puede usar Azure Resource Health para supervisar el estado de un recurso individual y puede configurar alertas de Resource Health para notificarle problemas.
    - Puede usar Azure Service Health para comprender el estado general del servicio, incluidos los errores de región, y puede configurar alertas de Service Health para notificarle problemas.

> **ⓘ Importante**
>
> Utilice las opciones de conmutación por error administradas por el cliente para desarrollar, probar e implementar tus planes de recuperación ante desastres. **No confíe en la conmutación por error administrada por Microsoft**, que solo se puede utilizar en circunstancias extremas. Es probable que se haya iniciado una

conmutación por error administrada por Microsoft para toda una región. No se puede iniciar para cuentas de almacenamiento individuales, suscripciones o clientes. La conmutación por error puede producirse en momentos diferentes para distintos servicios de Azure. Recomendamos que utilice la conmutación por error administrada por el cliente.

Recuperación de regiones

El proceso de conmutación por recuperación difiere significativamente entre los escenarios de conmutación por error administrados por Microsoft y los administrados por el cliente.

- **Conmutación por error administrada por el cliente (no planeada):** después de una conmutación por error no planeada, la cuenta de almacenamiento se configura con almacenamiento con redundancia local (LRS). Para realizar una conmutación por error, es necesario restablecer la relación de almacenamiento con redundancia geográfica (GRS) y esperar a que se repliquen los datos.
- **Conmutación por error administrada por el cliente (planificada):** después de una conmutación por error planificada, la cuenta de almacenamiento seguirá estando replicada geográficamente. Inicie otra conmutación por error administrada por el cliente para volver a la región principal original. Se aplican las mismas consideraciones de conmutación por error.
- **Conmutación por error administrada por Microsoft:** si Microsoft iniciase una conmutación por error, es probable que se haya producido un desastre importante en la región principal y que esta no se pueda recuperar. Cualquier calendario o plan de recuperación dependerá del alcance del desastre regional y de los esfuerzos de recuperación. Debe supervisar las comunicaciones de Azure Service Health para más información.

Prueba de fallos de región

En el caso de las cuentas de GRS, puede realizar operaciones de conmutación por error planeadas durante las ventanas de mantenimiento para probar el proceso completo de conmutación por error y conmutación por recuperación. La conmutación por error planeada no requiere pérdida de datos, pero requiere tiempo de inactividad durante la conmutación por error y la conmutación por recuperación.

Soluciones personalizadas de varias regiones para la resistencia

Las funcionalidades de conmutación por error entre regiones de Azure Storage podrían no ser adecuadas por los siguientes motivos:

- La cuenta de almacenamiento está en una región no emparejada.
- Los objetivos de tiempo de actividad empresarial no se satisfacen por el tiempo de recuperación o la pérdida de datos que proporciona las opciones de conmutación por error integradas.
- Debe conmutar por error a una región que no esté emparejada con la región principal.
- Necesita una configuración activa o activa entre regiones.
- Se usan tipos de recursos compartidos de archivos que no admiten redundancia geográfica.

En esta sección se proporciona información general de alto nivel de algunos enfoques que se deben tener en cuenta. Una introducción completa de las topologías de implementación de varias regiones para Azure Storage está fuera del ámbito de este artículo.

Tenga en cuenta los siguientes enfoques comunes de alto nivel:

- **Varias cuentas de almacenamiento:** Azure Files se puede implementar en varias regiones mediante cuentas de almacenamiento independientes en cada región. Este enfoque ofrece flexibilidad en la selección de regiones, la posibilidad de utilizar regiones no emparejadas y un control más granular sobre el momento de la replicación y la coherencia de los datos. Al implementar varias cuentas de almacenamiento entre regiones, es necesario configurar la replicación de datos entre regiones, implementar directivas de equilibrio de carga y conmutación por error y garantizar la coherencia de los datos entre regiones.
- **Replicación de nivel de aplicación:** Implemente la lógica de replicación personalizada mediante Azure Data Factory o AzCopy para sincronizar datos entre recursos compartidos de archivos en distintas regiones. Este enfoque requiere mecanismos personalizados de desarrollo y resolución de conflictos.
- **Use Azure File Sync para replicar archivos en un recurso compartido de archivos en otra región de Azure.** Puede usar Azure File Sync para sincronizar entre un recurso compartido de archivos de Azure SMB (punto de *conexión en la nube*), un servidor de archivos de Windows local y un recurso compartido de archivos montado que se ejecuta en una máquina virtual (VM) en otra región de Azure (un punto de *conexión de servidor de recuperación* ante desastres).

  Este enfoque requiere que implemente varios recursos compartidos de archivos y una máquina virtual para coordinar el proceso de sincronización.

Si usa este enfoque para la replicación de archivos de varias regiones:

- Deshabilite la nube por niveles para asegurarse de que todos los datos están presentes localmente en el servidor de archivos.
- Aprovisione suficiente almacenamiento en la máquina virtual de Azure para almacenar todo el conjunto de datos.
- Acceda a los archivos y modifique los archivos en el punto de conexión del servidor, y no en Azure, para asegurarse de que los cambios se replican rápidamente en la región secundaria.

Copias de seguridad y restauración

La copia de seguridad de Azure Files es una integración nativa entre Azure Files y Azure Backup diseñada para proteger los datos frente a ataques accidentales de eliminación, daños y ransomware.

La copia de seguridad de Azure Files crea instantáneas a nivel de compartición almacenadas en la misma cuenta de almacenamiento. Esta funcionalidad permite la recuperación rápida de archivos individuales y de comparticiones de archivos completas. También puede usar *directivas de copia de seguridad* para proporcionar períodos de retención largos con frecuencia de copia de seguridad personalizable.

Puede crear las instantáneas y almacenarlas de dos maneras diferentes:

- **Almacenamiento de nivel de recurso compartido:** En escenarios de recuperación operativos y a corto plazo, puede crear instantáneas de nivel de recurso compartido y almacenarlas en la misma cuenta de almacenamiento. Las instantáneas a nivel de recurso compartido permiten la recuperación rápida de archivos individuales o de recursos compartidos completos en su ubicación original o en una ubicación alternativa.
- **Almacenamiento de copia de seguridad archivado:** Mediante el almacenamiento de respaldo archivado, puede copiar las instantáneas diarias en un almacén de Azure Recovery Services. Para mejorar la seguridad, este almacén está aislado y desconectado físicamente de la cuenta de almacenamiento principal.

  Cuando utilizas una región de Azure emparejada y configuras la bóveda para usar GRS, la bóveda replica los datos en la región emparejada. Esta replicación soporta flujos de trabajo para recuperación entre regiones y recuperación ante desastres.

Acuerdo de nivel de servicio

El acuerdo de nivel de servicio (SLA) para Azure Storage describe la disponibilidad esperada del servicio y las condiciones que se deben cumplir para lograr esa expectativa de disponibilidad. El contrato de nivel de servicio de disponibilidad al que tiene derecho dependerá del nivel de almacenamiento y del tipo de replicación que utilice. Para obtener más información, consulte los SLAs for Online Services.

Contenido relacionado

- Documentación de Azure Files
- Redundancia de Azure Files
- Planificación de una implementación de Azure Files
- Fiabilidad de Azure

ⓘ **Nota:** El autor ha creado este artículo con ayuda de inteligencia artificial. Más información

---

Last updated on 22/01/2026

Confiabilidad en Azure NetApp Files

Azure NetApp Files es una solución nativa de almacenamiento de archivos de nivel empresarial que se integra sin problemas en Azure y permite el uso compartido de archivos entre clientes a través de protocolos network File System (NFS) y Bloque de mensajes del servidor (SMB). Azure NetApp Files está diseñado para un alto rendimiento y proporciona almacenamiento de archivos escalable y seguro administrado como servicio.

Al usar Azure, relibilidad es una responsabilidad compartida. Microsoft proporciona una variedad de funcionalidades para admitir resistencia y recuperación. Es responsable de comprender cómo funcionan esas funcionalidades dentro de todos los servicios que usa y de seleccionar las funcionalidades que necesita para cumplir los objetivos empresariales y los objetivos de tiempo de actividad.

En este artículo se describe cómo hacer que NetApp Files sea resistente a una variedad de posibles interrupciones y problemas, incluidos errores transitorios, interrupciones de zona de disponibilidad y interrupciones de regiones. También se describe cómo puede usar copias de seguridad para recuperarse de otros tipos de problemas y se resalta cierta información clave sobre el acuerdo de nivel de servicio (SLA) de Azure NetApp Files.

Recomendaciones de implementación de producción

Para obtener información sobre cómo implementar Azure NetApp Files para admitir los requisitos de confiabilidad de la solución y cómo afecta la confiabilidad a otros aspectos de la arquitectura, consulte Architecture best practices for Azure NetApp Files in the Azure Well-Architected Framework.

Introducción a la arquitectura de confiabilidad

Para usar Azure NetApp Files, debe configurar una cuenta de NetApp que contenga *pools de capacidad* que alojan *volúmenes*. Es posible configurar la capacidad y el rendimiento de forma independiente y administrar las opciones de protección de datos para que se ajusten a diversas necesidades. Puede habilitar la replicación entre volúmenes, incluso si están en ubicaciones diferentes.

Resistencia a errores transitorios

Los errores transitorios son errores breves e intermitentes en los componentes. Se producen con frecuencia en un entorno distribuido como la nube y son una parte normal de las operaciones. Los errores transitorios se corrigen después de un breve período de tiempo. Es importante que las aplicaciones puedan controlar errores transitorios, normalmente mediante el reintento de solicitudes afectadas.

Todas las aplicaciones hospedadas en la nube deben seguir las instrucciones de control de errores transitorios Azure cuando se comunican con cualquier API, bases de datos y otros componentes hospedados en la nube. Para más información, vea Recomendaciones para gestionar errores temporales.

Además de los tipos de error transitorios que pueden afectar a cualquier solución basada en la nube, un mantenimiento planeado ocasional, como actualizaciones de plataforma, actualizaciones de servicio y actualizaciones de software, también puede afectar a Azure NetApp Files.

Desde una perspectiva del protocolo de archivos, como NFS y SMB, los errores transitorios no son perjudiciales si la aplicación puede controlar las pausas de entrada/salida (E/S) que podrían producirse durante estos eventos. Las pausas de E/S suelen ser cortas, desde unos segundos hasta 30 segundos. Algunas aplicaciones pueden requerir el ajuste para controlar las pausas de E/S.

El protocolo NFS es sólido y las operaciones de archivos de servidor cliente, por lo general, continúan con normalidad. Algunas aplicaciones podrían requerir ajustes para controlar las pausas de E/S durante un tiempo de 30 a 45 segundos. Asegúrese de que conoce la configuración de resistencia de la aplicación para hacer frente a los eventos de mantenimiento del servicio de almacenamiento.

En el caso de las aplicaciones interactivas humanas que usan el protocolo SMB, la configuración de protocolo estándar suele ser suficiente. Azure NetApp Files también admite la disponibilidad continua de SMB, lo que permite la conmutación por error transparente de SMB. La conmutación por error transparente de SMB elimina las interrupciones que provocan los eventos de mantenimiento del servicio. También mejora la confiabilidad y la experiencia del usuario.

La disponibilidad continua de SMB solo está disponible para aplicaciones específicas.

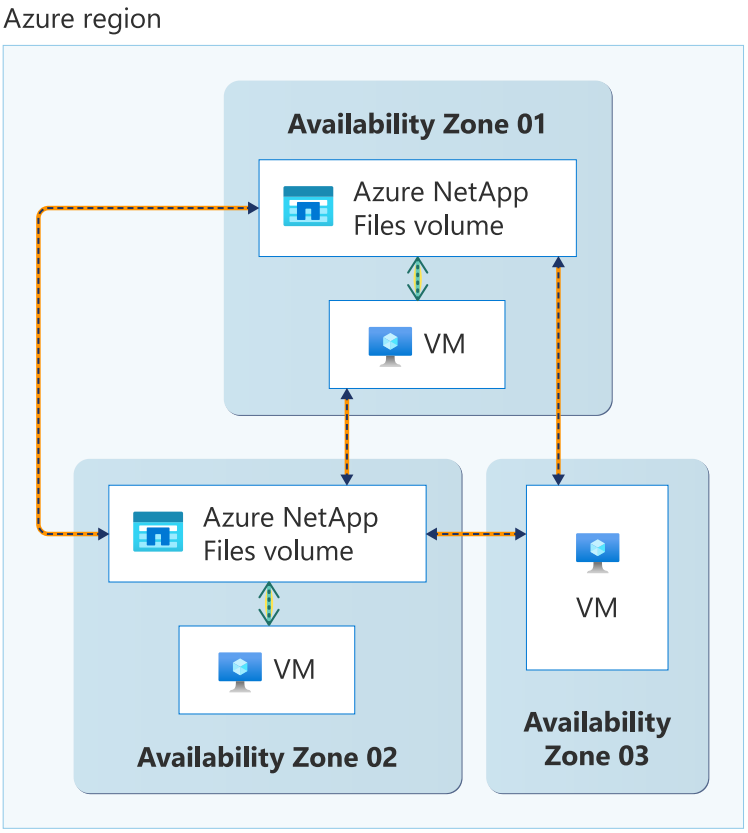
Para obtener más recomendaciones, consulte Preguntas más frecuentes sobre resistencia de la aplicación para Azure NetApp Files.

Resistencia a errores de zona de disponibilidad

Availability zones son grupos de centros de datos físicamente independientes dentro de una región de Azure. Cuando una zona falla, los servicios pueden transferirse a una de las zonas restantes.

Azure NetApp Files admite implementaciones *zonales* de volúmenes. Use la característica availability zone volume placement en Azure NetApp Files para implementar cada volumen en una sola zona de disponibilidad de su elección. Esta característica solo se puede usar si Azure NetApp Files está presente en esa zona de disponibilidad y tiene capacidad suficiente. Si tiene aplicaciones sensibles a la latencia, puede implementar un volumen en la misma zona de disponibilidad que los recursos de proceso de Azure y otros servicios.

En el diagrama siguiente, las flechas naranjas con puntas de flecha sólidas representan cómo todas las máquinas virtuales (VM) dentro de la región de las redes virtuales emparejadas pueden acceder a todos los recursos de Azure NetApp Files. Las flechas verdes representan cómo las máquinas virtuales que acceden a volúmenes de Azure NetApp Files en la misma zona comparten el dominio de fallo de la zona de disponibilidad. No hay replicación entre los distintos volúmenes en el nivel de plataforma.



Una implementación de una sola zona no es suficiente para cumplir los requisitos de alta confiabilidad. Para replicar de forma asincrónica los datos entre volúmenes de diferentes zonas de disponibilidad, puede usar la replicación entre zonas. Debe configurar la replicación entre zonas por separado de la ubicación del volumen de la zona de disponibilidad.

Si se produce un error en una zona de disponibilidad, es responsable de detectar el error y cambiar a un volumen alternativo en una zona diferente.

Requisitos

- **Region support:** La replicación entre zonas está disponible en todas las regiones con zonas de disponibilidad habilitadas que admiten Azure NetApp Files.
- **Inquilino de Microsoft Entra:** La replicación solo se permite entre diferentes suscripciones de Azure si están dentro del mismo inquilino de Microsoft Entra.

Consideraciones

La colocación de volúmenes en una zona de disponibilidad en Azure NetApp Files proporciona una distribución de volúmenes por zona. Verá una latencia baja al conectarse a máquinas virtuales dentro de la misma zona de disponibilidad. Sin embargo, la colocación de volumen de zona de disponibilidad no proporciona colocación de proximidad con máquinas virtuales u otros recursos, y el volumen podría estar en una parte física diferente del centro de datos.

Para obtener más consideraciones sobre las zonas de disponibilidad en Azure NetApp Files, consulte Requisitos y consideraciones sobre el uso de la replicación entre zonas y Administrar la colocación de volúmenes en zonas de disponibilidad.

Cost

No hay ningún cargo adicional para habilitar la colocación de volúmenes en zonas de disponibilidad en Azure NetApp Files. Solo paga por los grupos de capacidad y los recursos que implemente en estas zonas.

Los volúmenes de réplica se hospedan en un grupo de capacidad. El coste de la replicación entre zonas se basa en el tamaño y el nivel del grupo de capacidad aprovisionado. No hay ningún coste adicional para la replicación de datos.

Configurar soporte de zonas de disponibilidad

Debe configurar por separado la ubicación del volumen y la replicación entre zonas.

- **Selección de ubicación del volumen:**
  - **Cree un nuevo volumen o configure un volumen existente con compatibilidad con la zona de disponibilidad.** Para configurar zonas de disponibilidad para volúmenes en

Azure NetApp Files, consulte Administrar ubicación de volumen de zona de disponibilidad para Azure NetApp Files.

Al implementar volúmenes administrados por Terraform con zonas de disponibilidad, se requerirán otras configuraciones. Para más información, consulte Rellenar la zona de disponibilidad para volúmenes administrados por Terraform.

Al usar el control de acceso basado en roles, asegúrese de configurar los permisos correctos.

- **Migrar volúmenes entre zonas de disponibilidad.** Después de configurar un volumen para colocarlo en una zona de disponibilidad, no se podrán modificar la zona de disponibilidad especificada. No se pueden mover volúmenes entre zonas de disponibilidad.
- **Deshabilite la compatibilidad de zona de disponibilidad para un volumen.** Después de configurar un volumen para colocarlo en una zona de disponibilidad, no se podrá deshabilitar la compatibilidad con la zona de disponibilidad.

- **Replicación entre zonas>**
  - **Habilitar la replicación entre zonas.** Para mejorar la resistencia de la solución, configure la replicación entre zonas para otro volumen.
  - **Deshabilitar la replicación entre zonas.** Deshabilite la replicación entre zonas mediante la interrupción del emparejamiento de replicación. Para obtener más información, consulte Administrar la recuperación ante desastres mediante Azure NetApp Files.

Comportamiento cuando todas las zonas están en buen estado

En esta sección se describe qué esperar cuando se implementan varios volúmenes de Azure NetApp Files en zonas de disponibilidad independientes, la replicación entre zonas está habilitada y todas las zonas de disponibilidad están operativas.

- **Enrutamiento de tráfico entre zonas:** las solicitudes entrantes se enrutan al volumen específico, que se encuentra en la zona de disponibilidad que seleccione.
- **Replicación de datos entre zonas:** La replicación entre zonas de Azure NetApp Files significa que todos los cambios en el volumen de origen se replican de forma asincrónica en los volúmenes de destino. Es posible decidir con qué frecuencia se producirá la

replicación. La replicación entre zonas admite tres programaciones de replicación: cada 10 minutos, cada hora y cada día.

> **Importante**
>
> La programación de replicación de 10 minutos no se admite para grandes volúmenes que usen la replicación entre zonas.

Comportamiento durante un fallo de zona

En esta sección se describe qué esperar cuando se implementan varios volúmenes de Azure NetApp Files en zonas de disponibilidad independientes, la replicación entre zonas está habilitada y hay una interrupción de zona de disponibilidad.

- **Detección y respuesta:** Es responsable de detectar la pérdida de una zona de disponibilidad e iniciar una conmutación por error.

  La conmutación por error es un proceso manual. Cuando necesite activar el volumen de destino, como cuando quiera conmutar por error a la zona de disponibilidad de destino, debe interrumpir el emparejamiento de replicación y, a continuación, montar el volumen de destino. Para obtener más información, consulte Conmutación por error al volumen de destino.

- **Notification:** Para supervisar el estado del volumen de Azure NetApp Files, puede usar métricas de Azure Monitor. Azure Monitor detecta anomalías que indican un escenario de zona abajo a través de métricas en tiempo real, como operaciones de entrada y salida por segundo (IOPS), latencia y uso de capacidad. Es posible configurar alertas y notificaciones para enviar a los administradores para que respondan inmediatamente reequilibrando recursos compartidos de archivos o iniciando la conmutación por error u otros protocolos de recuperación ante desastres.

- **Solicitudes activas:** durante un evento de zona inactiva, las solicitudes activas pueden experimentar interrupciones o aumentar las latencias.

- **Pérdida de datos esperada:** la cantidad de pérdida de datos o el objetivo de punto de recuperación (RPO) que se puede esperar durante una conmutación por error de zona dependerá de la programación de replicación entre zonas configurada.

Expandir tabla

Programación de replicación	RPO típico
Cada 10 minutos	20 minutos
Cada hora	Dos horas
Diariamente	Menos de 48 horas

- **Tiempo de inactividad esperado:** la conmutación por error a otra zona requiere que interrumpa la relación de emparejamiento para activar el volumen de destino y proporcionar acceso a datos de lectura y escritura en el segundo sitio. Después de desencadenar el emparejamiento para que se interrumpa, espere que la conmutación por error se complete en un minuto.

  Sin embargo, la cantidad total de tiempo de inactividad o el objetivo de tiempo de recuperación (RTO), que se puede esperar durante una conmutación por error de zona dependerá de varios factores, incluyendo el tiempo que tarden los sistemas o procesos en detectar la pérdida de la zona e iniciar los procesos de conmutación por error. También es importante decidir si se automatizará la respuesta o si se requerirán pasos manuales. Para aquellas configuraciones que estén bien preparadas, el proceso general normalmente requerirá de unos minutos a una hora para completarse.

- **Reenrutamiento del tráfico:** Es responsable de redirigir el tráfico de la aplicación para conectarse al volumen de destino recién activo. Para obtener más información, consulte Conmutación por error al volumen de destino.

Recuperación de zona

La conmutación por recuperación es un proceso manual que requiere que se realice una operación de resincronización, restablecer la replicación y volver a montar el volumen de origen para que el cliente acceda. Para obtener más información, consulte Administrar la recuperación ante desastres mediante Azure NetApp Files.

Prueba de fallos de zona

Puede probar la configuración de replicación entre zonas de forma segura mediante instantáneas del volumen. Para obtener información sobre un enfoque de alto nivel para probar la configuración de replicación entre zonas, consulte Test disaster recovery for Azure NetApp Files .

Resistencia a errores en toda la región

De forma predeterminada, Azure NetApp Files es un servicio de una sola región. Si la región deja de estar disponible, los volúmenes almacenados en esa región tampoco están disponibles. Para mejorar la resistencia si se produce una interrupción regional, Azure NetApp Files admite la replicación entre regiones. Puede replicar datos de forma asincrónica desde un volumen de Azure NetApp Files (el origen) de una región a otro volumen Azure NetApp Files (el destino) en otra región que Microsoft preseleccione. Esta funcionalidad le permite realizar la conmutación por error de la aplicación crítica en caso de un desastre o una interrupción en toda la región.

> ⓘ **Nota:**
>
> También puede replicar un único volumen en otra zona de disponibilidad *y* en otra región. Para más información, consulte **Comprender la replicación de Azure NetApp Files**.

Requisitos

Compatibilidad con regiones: La región secundaria a la que puede replicar los volúmenes depende de la región primaria. Para más información, consulte pares de regiones admitidos.

Consideraciones

La replicación se permite entre distintas suscripciones de Azure solo si están dentro del mismo inquilino de Microsoft Entra.

Para conocer otras consideraciones relacionadas con la replicación entre regiones en Azure NetApp Files, consulte Requirements y consideraciones sobre el uso de la replicación entre regiones.

Cost

Los cargos de replicación entre regiones se basan en la cantidad de datos replicados. Para obtener más información y algunos escenarios de ejemplo, consulte Modelo de costes de la replicación entre regiones.

Configuración de la compatibilidad con varias regiones

- **Habilite la replicación entre regiones:** Para mejorar la resistencia de la solución, configure la replicación entre regiones.

- **Deshabilitar la replicación entre regiones:** Puede deshabilitar la replicación entre regiones mediante la interrupción del emparejamiento de replicación. Para obtener más información, consulte Administrar la recuperación ante desastres mediante Azure NetApp Files.

Comportamiento cuando todas las regiones están en buen estado

En esta sección se describe qué esperar cuando los volúmenes de Azure NetApp Files están configurados para usar la replicación entre regiones y ambas regiones están operativas.

- **Enrutamiento del tráfico entre regiones:** Las solicitudes entrantes se enrutan al volumen específico, que se encuentra en la región primaria.

- **Replicación de datos entre regiones:** Azure NetApp Files replicación entre regiones significa que todos los cambios en el volumen de origen se replican de forma asincrónica en volúmenes de destino. Es posible decidir con qué frecuencia se producirá la replicación. La replicación entre regiones admite tres programaciones de replicación: cada 10 minutos, cada hora y cada día.

  > **Importante**
  >
  > La programación de replicación de 10 minutos no se admite para **grandes volúmenes** que usen la replicación entre regiones.

- **Supervisar el estado de replicación:** Puede supervisar el estado de la relación de emparejamiento y puede configurar alertas para notificarle si el retraso de replicación aumenta más allá del umbral esperado. Para más información, consulte Visualización del estado y supervisión de la relación de replicación.

Comportamiento durante una falla de región

En esta sección se describe qué esperar cuando los volúmenes de Azure NetApp Files están configurados para usar la replicación entre diferentes regiones y hay una interrupción de la región primaria.

- **Detección y respuesta:** Es responsable de detectar la pérdida de una región e iniciar una conmutación por error. La conmutación por error es un proceso manual. Cuando necesite activar el volumen de destino, como cuando quiera conmutar por error a la región de destino, debe interrumpir el emparejamiento de replicación y, a continuación, montar el

volumen de destino. Para obtener más información, consulte Conmutación por error al volumen de destino.

- **Notification:** Para supervisar el estado del volumen de Azure NetApp Files, puede usar métricas de Azure Monitor. Azure Monitor detecta las anomalías que indican un escenario de reducción regional a través de métricas en tiempo real, como IOPS, latencia y uso de capacidad. Es posible configurar alertas y notificaciones para enviar a los administradores para que respondan inmediatamente reequilibrando recursos compartidos de archivos o iniciando la conmutación por error u otros protocolos de recuperación ante desastres.
- **Solicitudes activas:** Durante un evento de región inactiva, las solicitudes activas pueden experimentar interrupciones o mayores latencias.
- **Pérdida de datos esperada:** la cantidad de pérdida de datos o RPO que se puede esperar durante una conmutación por error de regiones dependerá de la programación de replicación entre regiones configurada.

Expandir tabla

Programación de replicación	RPO típico
Cada 10 minutos	Menos de 20 minutos
Cada hora	Menos de dos horas
Diariamente	Menos de 48 horas

- **Tiempo de inactividad esperado:** la conmutación por error a otra región requiere interrumpir la relación de emparejamiento para activar el volumen de destino y proporcionar acceso a datos de lectura y escritura en el segundo sitio. Después de desencadenar el emparejamiento para que se interrumpa, espere que la conmutación por error se complete en un minuto.

  Sin embargo, el tiempo total de inactividad, o tiempo objetivo de recuperación (RTO), que cabe esperar durante una conmutación por error de una región depende de varios factores, entre ellos el tiempo que tardan sus sistemas o procesos en detectar la pérdida de la región y en iniciar los procesos de conmutación por error. También es importante decidir si se automatizará la respuesta o si se requerirán pasos manuales. Para aquellas configuraciones que estén bien preparadas, el proceso general normalmente requerirá de unos minutos a una hora para completarse.

- **Reenrutamiento del tráfico:** Es responsable de redirigir el tráfico de la aplicación para conectarse al volumen de destino recién activo. Para obtener más información, consulte Conmutación por error al volumen de destino.

Recuperación de regiones

Después de que la región primaria se recupere, usted será responsable de la conmutación por restauración. La conmutación por recuperación es un proceso manual que requiere que se realice una operación de resincronización, restablecer la replicación y volver a montar el volumen de origen para que el cliente acceda. Para obtener más información, consulte Administrar la recuperación ante desastres mediante Azure NetApp Files.

Prueba de fallos de región

Puede probar la configuración de replicación entre regiones de forma segura mediante instantáneas del volumen. Para obtener información sobre un enfoque de alto nivel para probar la configuración de replicación entre regiones, consulte Test disaster recovery for Azure NetApp Files.

Copias de seguridad y restauración

Azure NetApp Files copia de seguridad amplía las funcionalidades de protección de datos de Azure NetApp Files proporcionando una solución de copia de seguridad totalmente administrada para la recuperación, el archivo y el cumplimiento a largo plazo. Las copias de seguridad que crea el servicio se almacenan en el almacenamiento de Azure, independientemente de las instantáneas de volumen que están disponibles para la recuperación o la clonación a corto plazo. Las copias de seguridad que realiza el servicio se pueden restaurar en nuevos volúmenes Azure NetApp Files dentro de la región. Azure NetApp Files copia de seguridad admite tanto copias de seguridad basadas en directivas (programadas) como copias de seguridad manuales (a petición).

Para mayor seguridad, los snapshots de Azure NetApp Files agregan estabilidad, escalabilidad y capacidad de recuperación rápida sin afectar al rendimiento. Proporcionan la base para otras soluciones de redundancia, como la copia de seguridad, la replicación entre regiones y la replicación entre zonas.

Para la mayoría de las soluciones, no debe confiar exclusivamente en copias de seguridad. En su lugar, utilice las otras capacidades descritas en esta guía para apoyar los requisitos de resiliencia. Sin embargo, las copias de seguridad protegen contra algunos riesgos que otros enfoques no. Para más información, consulte ¿Qué son la redundancia, la replicación y la copia de seguridad?.

Resistencia al mantenimiento del servicio

Durante el mantenimiento del servicio o la plataforma, pueden producirse interrupciones breves en las operaciones de E/S. Las cargas de trabajo deben ser resistentes a estos tipos de interrupciones. Para obtener más información, consulte Resistencia de aplicaciones: eventos de mantenimiento.

Acuerdo de nivel de servicio

El acuerdo de nivel de servicio (SLA) para Azure servicios describe la disponibilidad esperada de cada servicio y las condiciones que la solución debe cumplir para lograr esa expectativa de disponibilidad. Para obtener más información, vea SLAs for servicios en línea.

Contenido relacionado

- Mejores prácticas de arquitectura para Azure NetApp Files
- Preguntas más frecuentes sobre resistencia de aplicaciones para Azure NetApp Files
- Requisitos y consideraciones para usar la replicación de Azure NetApp Files
- Administrar la colocación de volúmenes en la zona de disponibilidad para Azure NetApp Files
- Crear relaciones de replicación entre zonas para volúmenes de Azure NetApp Files
- Crear relaciones de replicación entre regiones para volúmenes de Azure NetApp Files

ⓘ**Nota:** El autor ha creado este artículo con ayuda de inteligencia artificial. Más información

---

Last updated on 07/07/2026

Confiabilidad en Azure Queue Storage

Azure Queue Storage es un servicio para almacenar y distribuir un gran número de mensajes. Queue Storage se usa normalmente para crear un trabajo pendiente del trabajo que se va a procesar de forma asincrónica. Proporciona una entrega de mensajes confiable para arquitecturas de aplicaciones acopladas de forma flexible. Un mensaje en cola puede tener un tamaño de hasta 64 KB y una cola puede contener millones de mensajes, hasta el límite de capacidad total de una cuenta de almacenamiento.

Cuando se usa Azure, la confiabilidad es una responsabilidad compartida. Microsoft proporciona una variedad de funcionalidades para admitir resistencia y recuperación. Es responsable de comprender cómo funcionan esas funcionalidades dentro de todos los servicios que usa y de seleccionar las funcionalidades que necesita para cumplir los objetivos empresariales y los objetivos de tiempo de actividad.

En este artículo se describe cómo hacer que Queue Storage sea resistente a una variedad de posibles interrupciones y problemas, incluidos errores transitorios, interrupciones de zona de disponibilidad y interrupciones de regiones. También se describe cómo puede usar copias de seguridad para recuperarse de otros tipos de problemas y se resalta cierta información clave sobre el acuerdo de nivel de servicio (SLA) de Queue Storage.

> ⓘ **Nota:**
>
> Queue Storage forma parte de la plataforma Azure Storage. Algunas de las funcionalidades de Queue Storage son comunes en muchos servicios de Azure Storage.

Recomendaciones de implementación de producción para la confiabilidad

Para entornos de producción:

- Habilite el almacenamiento con redundancia de zona (ZRS) para las cuentas de almacenamiento que contienen recursos de Queue Storage. ZRS proporciona una mayor disponibilidad mediante la replicación de los datos de forma sincrónica en varias zonas de disponibilidad de la región primaria. Una mayor disponibilidad ayuda a proteger las cuentas de almacenamiento frente a errores de zona de disponibilidad.

- Si necesitase resistencia a interrupciones de regiones y la región primaria de la cuenta de almacenamiento se encontrase emparejada, considere la posibilidad de habilitar el almacenamiento con redundancia geográfica (GRS). GRS replica los datos de forma asincrónica en la región emparejada. En las regiones admitidas, es posible combinar la

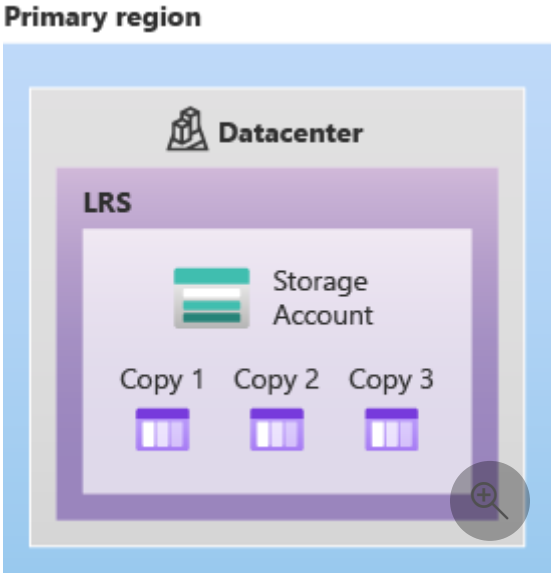
redundancia geográfica con redundancia de zona mediante el almacenamiento con redundancia de zona geográfica (GZRS).

Para los requisitos de mensajería avanzados, considere usar Azure Service Bus. Para obtener información sobre las diferencias entre Queue Storage y Service Bus, consulte Comparación de las colas de Azure Storage y las colas de Service Bus.

Introducción a la arquitectura de confiabilidad

Queue Storage funciona como un servicio de mensajería distribuida dentro de la infraestructura de la plataforma de Azure Storage. El servicio proporciona redundancia a través de varias copias de los datos de la cola y de mensajes. El modelo de redundancia específico depende de la configuración de la cuenta de almacenamiento.

El almacenamiento con redundancia local (LRS) replica los datos de las cuentas de almacenamiento en una o varias zonas de disponibilidad de Azure ubicadas en la región primaria de su elección. Aunque no hay ninguna opción para elegir la zona de disponibilidad preferida, Azure puede mover o expandir cuentas LRS entre zonas para mejorar el equilibrio de carga. No hay ninguna garantía de que los datos se distribuirán entre zonas. Para obtener más información sobre las zonas de disponibilidad, consulte ¿Qué son las zonas de disponibilidad?.



El almacenamiento con redundancia de zona (ZRS), el almacenamiento con redundancia geográfica (GRS) y el almacenamiento con redundancia de zona geográfica (GZRS) proporcionan protecciones adicionales. En este artículo se describen estas opciones al detalle.

Resistencia a errores transitorios

Los errores transitorios son errores breves e intermitentes en los componentes. Se producen con frecuencia en un entorno distribuido como la nube y son una parte normal de las

operaciones. Los errores transitorios se corrigen después de un breve período de tiempo. Es importante que las aplicaciones puedan controlar errores transitorios, normalmente mediante el reintento de solicitudes afectadas.

Todas las aplicaciones hospedadas en la nube deben seguir las instrucciones de control de errores transitorios de Azure cuando se comunican con cualquier API, bases de datos y otros componentes hospedados en la nube. Para obtener más información, consulte Recomendaciones para controlar errores transitorios.

Queue Storage se usa normalmente en aplicaciones para ayudarles a controlar errores transitorios en otros componentes. Mediante el uso de mensajería asincrónica con un servicio como Queue Storage, las aplicaciones podrán recuperarse de errores transitorios gracias al reprocesamiento de mensajes más adelante. Para más información, consulte Introducción a la mensajería asincrónica.

Dentro del propio servicio, Queue Storage controla automáticamente los errores transitorios mediante varios mecanismos que proporcionan la plataforma de Azure Storage y las bibliotecas cliente. El servicio está diseñado para proporcionar funcionalidades resistentes de Message Queue Server incluso durante problemas temporales de infraestructura.

Las bibliotecas cliente de Queue Storage incluyen directivas de reintento integradas que controlan automáticamente errores transitorios comunes, como tiempos de espera de red, falta de disponibilidad temporal del servicio (HTTP 503) y respuestas de limitación (HTTP 429). Cuando la aplicación encuentre estas condiciones transitorias, las bibliotecas cliente reintentarán automáticamente las operaciones mediante estrategias de retroceso exponencial.

Para administrar los errores transitorios de forma eficaz mediante Queue Storage, realice las siguientes acciones:

- **Configure los tiempos de espera adecuados** en el cliente de Queue Storage para equilibrar la capacidad de respuesta con resistencia a ralentizaciones temporales. Los tiempos de espera predeterminados en las bibliotecas cliente de Azure Storage suelen ser adecuados para la mayoría de los escenarios.

- **Implemente patrones de disyuntor** en la aplicación al procesar mensajes de colas. Los patrones de cortacircuitos impiden fallos en cascada cuando los servicios aguas abajo experimentan problemas.

- **Use los tiempos de espera de visibilidad correctamente** cuando la aplicación reciba mensajes. Los tiempos de espera de visibilidad garantizan que los mensajes estén disponibles para reintentos si la aplicación encontrase errores durante el procesamiento.

Para más información sobre la arquitectura de Azure Table Storage y cómo diseñar aplicaciones resistentes y a gran escala, consulte Lista de comprobación de rendimiento y escalabilidad para

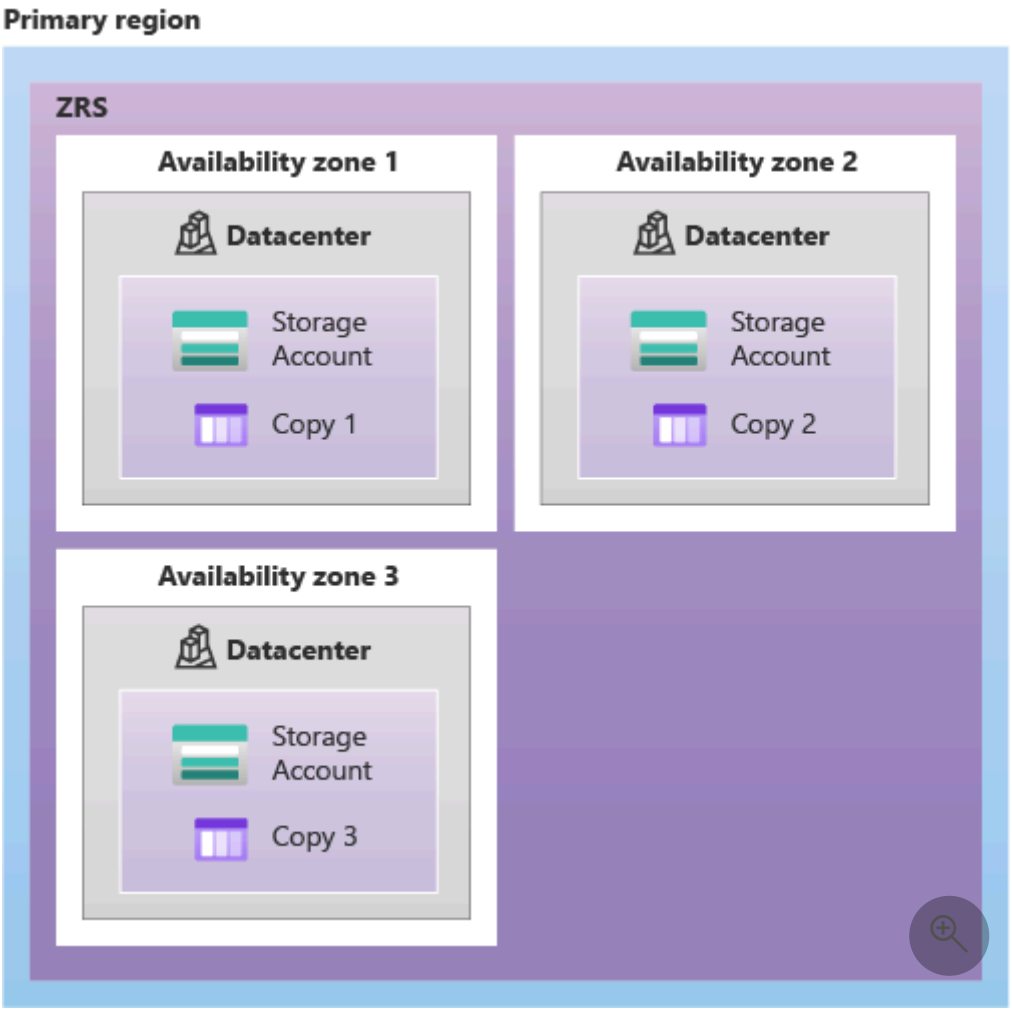
Queue Storage.

Resistencia a errores de zona de disponibilidad

Las zonas de disponibilidad son grupos físicamente independientes de centros de datos dentro de una región de Azure. Cuando una zona falla, los servicios pueden transferirse a una de las zonas restantes.

Azure Queue Storage tiene redundancia de zona cuando se implementa con la configuración de ZRS. A diferencia de LRS, ZRS garantiza que Azure replique de forma sincrónica los datos de la cola en varias zonas de disponibilidad. ZRS garantiza que sus datos sigan estando accesibles incluso si se produce un corte en una zona. ZRS garantiza que las colas sigan siendo accesibles incluso si una zona de disponibilidad completa deja de estar disponible. Todas las operaciones de escritura deben confirmarse en varias zonas antes de que se completen, lo que proporciona garantías de coherencia sólidas.

La redundancia de zona está habilitada en el nivel de cuenta de almacenamiento y se aplica a todos los recursos de Queue Storage de esa cuenta. No se pueden configurar colas individuales para distintos niveles de redundancia. La configuración se aplica a toda la cuenta de almacenamiento. Cuando una zona de disponibilidad experimenta una interrupción, Azure Storage enruta automáticamente las solicitudes a zonas correctas sin necesidad de intervención de la aplicación.



Requisitos

- **Compatibilidad con regiones:** Puede implementar cuentas de Azure Storage con redundancia de zona en cualquier región que admita zonas de disponibilidad.
- **Tipos de cuenta de almacenamiento:** Debe usar una cuenta de almacenamiento estándar de uso general v2 para habilitar ZRS para Queue Storage. Las cuentas de Premium Storage no admiten Queue Storage.

Cost

Cuando se habilita el almacenamiento con redundancia de zona (ZRS), se cobra una tarifa diferente a la del almacenamiento redundante local (LRS) debido a la replicación adicional y los gastos generales de almacenamiento.

Para obtener información detallada sobre los precios, consulte Precios de Queue Storage.

Configurar soporte de zonas de disponibilidad

- **Cree una cuenta de almacenamiento de ZRS y una cola siguiendo los pasos siguientes.**

1. Cree una cuenta de almacenamiento y seleccione ZRS, GZRS o almacenamiento con redundancia de zona geográfica con acceso de lectura (RA-GZRS) como opción de redundancia durante la creación de la cuenta.

2. Cree una cola.

- **Cambiar el tipo de replicación.** Para obtener información sobre cómo cambiar una cuenta de almacenamiento existente a almacenamiento con redundancia de zona (ZRS) y sobre las opciones y requisitos de configuración, consulte Cambiar la forma en que se replica una cuenta de almacenamiento.

- **Desactivar la redundancia de zona.** Convierta las cuentas de ZRS de nuevo a una configuración no zonal, como el almacenamiento con redundancia local (LRS), utilizando el mismo proceso de cambio de configuración de redundancia.

Comportamiento cuando todas las zonas están en buen estado

En esta sección se describe qué esperar cuando una cuenta de Queue Storage está configurada para la redundancia de zona y todas las zonas de disponibilidad están operativas.

- **Enrutamiento del tráfico entre zonas:** Azure Storage con almacenamiento con redundancia de zona (ZRS) distribuye automáticamente las solicitudes entre los clústeres de almacenamiento en varias zonas de disponibilidad. La distribución del tráfico es transparente para las aplicaciones y no requiere ninguna configuración del lado cliente.

- **Replicación de datos entre zonas:** todas las operaciones de escritura en ZRS se replican de forma sincrónica en todas las zonas de disponibilidad dentro de la región. Al cargar o modificar datos, la operación no se considera completa hasta que los datos se hayan replicado correctamente en todas las zonas de disponibilidad. Esta replicación sincrónica garantiza una fuerte coherencia y cero pérdida de datos durante los errores de zona.

Comportamiento durante un fallo de zona

Cuando una zona de disponibilidad deja de estar disponible, Queue Storage controla automáticamente el proceso de conmutación por error al tomar las siguientes acciones.

- **Detección y respuesta:** Microsoft detecta automáticamente errores de zona e inicia procesos de recuperación. No se requiere ninguna acción del cliente para las cuentas de almacenamiento con redundancia de zona (ZRS). Si una zona deja de estar disponible, Azure lleva a cabo actualizaciones de red como el repunto del sistema de nombres de dominio (DNS).

- **Notificación**: Microsoft no le notifica automáticamente cuando una zona está inactiva. Sin embargo, puede usar Azure Resource Health para supervisar el estado de un recurso individual y puede configurar alertas de Resource Health para notificarle problemas. También puede usar Azure Service Health para comprender el estado general del servicio, incluidos los errores de zona, y puede configurar alertas de Service Health para notificarle problemas.

- **Solicitudes activas:** Es posible que las solicitudes en curso se descarten durante el proceso de recuperación y se deban reintentar. Las aplicaciones deben implementar lógica de reintento para controlar estas interrupciones temporales.

- **Pérdida de datos esperada:** No se produce ninguna pérdida de datos durante los errores de zona porque los datos se replican sincrónicamente en varias zonas antes de completar las operaciones de escritura.

- **Tiempo de inactividad previsto:** es posible que se produzca un breve periodo de inactividad, normalmente de unos segundos, durante la recuperación automática, ya que el tráfico se redirige a zonas correctas. Cuando diseñe aplicaciones para ZRS, siga las prácticas de manejo de fallos transitorios, incluyendo la implementación de directivas de reintento con retroceso exponencial.

- **Redireccionar el tráfico.** Si una zona deja de estar disponible, Azure realiza actualizaciones de red, como el redireccionamiento del Sistema de nombres de dominio (DNS), para que las solicitudes se dirijan a las zonas de disponibilidad restantes que funcionan correctamente. El servicio mantiene una funcionalidad completa utilizando las zonas que permanecen operativas sin necesidad de intervención del cliente.

Recuperación de zona

Cuando se recupera la zona de disponibilidad con errores, Azure Storage restaura automáticamente las operaciones normales en todas las zonas de disponibilidad. El servicio garantiza automáticamente la coherencia de los datos mediante la sincronización de las operaciones que se produjeron durante el período de interrupción.

Prueba de fallos de zona

El usar el almacenamiento con redundancia de zona (ZRS), Azure Storage administrará automáticamente la replicación, el enrutamiento del tráfico y las respuestas ante caídas de zonas. Dado que esta característica está totalmente administrada, no es necesario iniciar ni validar los procesos de error de zona de disponibilidad.

Resistencia a errores en toda la región

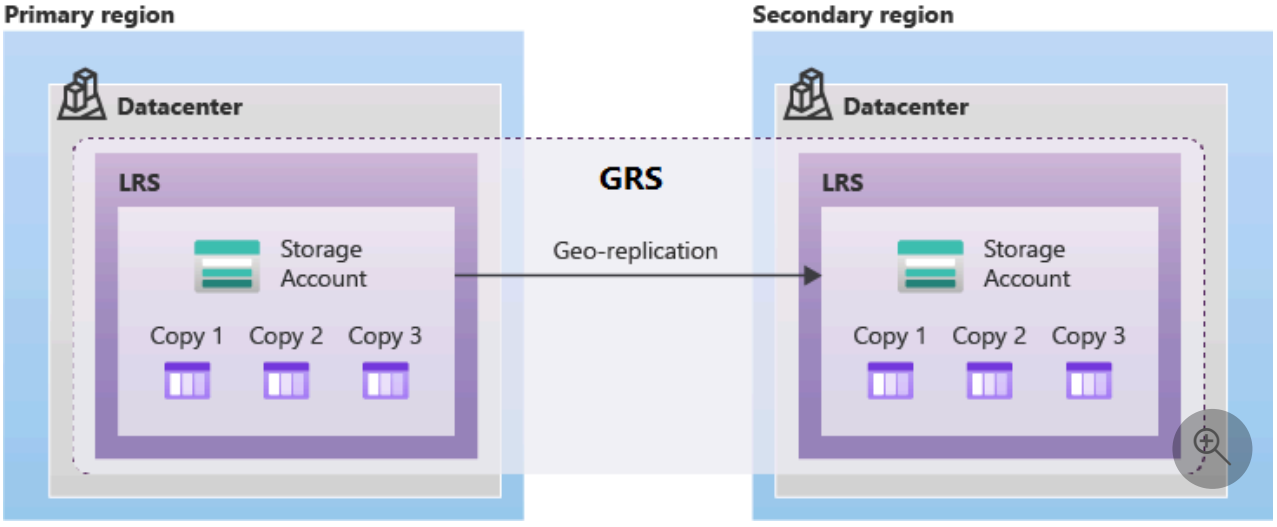
Azure Storage, que incluye Azure Blob Storage, Azure Files, Azure Table Storage y Azure Queue Storage, ofrece una amplia gama de capacidades de redundancia geográfica y conmutación por error para adaptarse a diferentes requisitos.

> ⓘ **Importante**
>
> El GRS solo funciona dentro de regiones emparejadas de Azure. Si la región de la cuenta de almacenamiento no está emparejada, considere la posibilidad de usar las soluciones personalizadas de varias regiones para lograr resistencia.

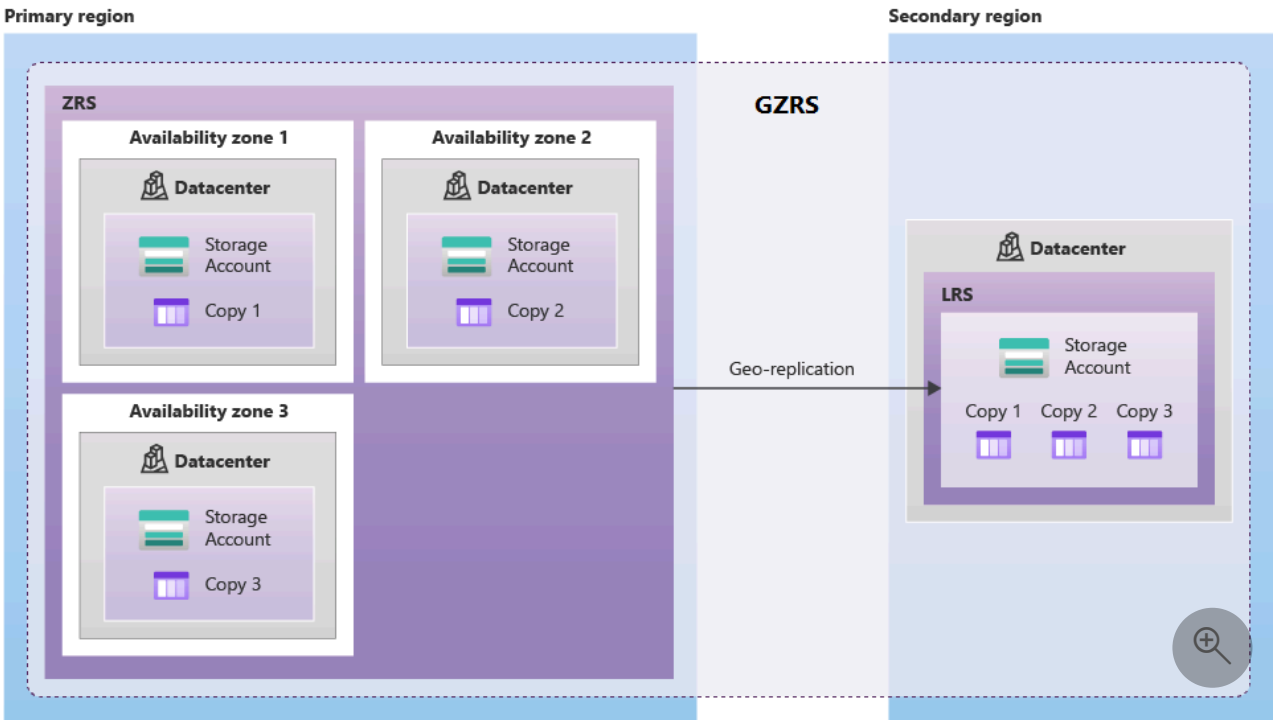
Almacenamiento con redundancia geográfica para regiones emparejadas

Azure Storage ofrece varios tipos de GRS en regiones emparejadas. Independientemente del tipo de GRS que utilice, los datos de la región secundaria siempre se replicarán mediante almacenamiento con redundancia local (LRS). Este enfoque proporciona protección contra fallos de hardware dentro de la región secundaria.

- GRS proporciona soporte para conmutaciones por error planificadas y no planificadas a la región emparejada de Azure cuando se produce una interrupción en la región principal. GRS replica de forma asincrónica los datos de la región primaria a la región emparejada.



- El almacenamiento con redundancia de zona geográfica (GZRS) replica los datos en varias zonas de disponibilidad de la región principal y en la región emparejada.



- El almacenamiento con redundancia geográfica con acceso de lectura (RA-GRS) y el almacenamiento con redundancia de zona geográfica con acceso de lectura (RA-GZRS) amplían el almacenamiento con redundancia geográfica (GRS) y el almacenamiento con redundancia de zona geográfica (GZRS), con la ventaja adicional del acceso de lectura al punto de conexión secundario. Estas opciones son idóneas y están diseñadas para obtener alta disponibilidad en aplicaciones críticas para la empresa. En el improbable caso de que el punto de conexión principal experimente una interrupción, las aplicaciones configuradas para el acceso de lectura a la región secundaria pueden seguir funcionando.

Tipos de conmutación por error

Azure Storage admite tres tipos de conmutación por error para distintos escenarios.

- **Conmutación por error no planeada administrada por el cliente:** es responsable de iniciar la recuperación si hubiera un error de almacenamiento en toda la región principal.

- **Conmutación por error planeada administrada por el cliente:** Usted es responsable de iniciar la recuperación si otra parte de su solución sufre un fallo en la región primaria y necesita trasladar toda su solución a la región secundaria. Use una conmutación por error planeada cuando el almacenamiento permanezca operativo en la región primaria, pero debe realizar una conmutación por error de toda la solución a una región secundaria, como para los simulacros de recuperación ante desastres diseñados para garantizar el cumplimiento de los requisitos de auditoría.

- **Conmutación por error administrada por Microsoft:** en circunstancias excepcionales, Microsoft podría iniciar la conmutación por error para todas las cuentas de almacenamiento con redundancia geográfica (GRS) de una región. Sin embargo, la

conmutación por error administrada por Microsoft es un último recurso y se espera que solo se realice después de un período prolongado de interrupción. No se debe confiar en el failover administrado por Microsoft.

Las cuentas de GRS pueden utilizar cualquiera de estos tipos de conmutación por error. No es necesario preconfigurar una cuenta de almacenamiento para usar ninguno de los tipos de conmutación por error con antelación.

Requisitos

- **Compatibilidad con regiones:** Las configuraciones con redundancia geográfica de Azure Storage usan regiones emparejadas de Azure para la replicación de regiones secundarias. La región secundaria se determina automáticamente en función de la selección de la región principal y no se puede personalizar. Para obtener una lista completa de las regiones emparejadas de Azure, consulte Lista de regiones de Azure.

  Si la región de la cuenta de almacenamiento no está emparejada, considere la posibilidad de usar las soluciones personalizadas de varias regiones para lograr resistencia.

- **Tipos de cuenta de almacenamiento:** El almacenamiento con redundancia geográfica (GRS) y la conmutación por error y la conmutación por recuperación iniciadas por el cliente están disponibles en todas las regiones emparejadas de Azure que admiten cuentas de Azure Storage de uso general v2.

Consideraciones

Al implementar Queue Storage en varias regiones, tenga en cuenta los siguientes factores importantes.

- **Latencia de replicación asincrónica:** la replicación de datos en la región secundaria es asincrónica, lo que significa que hay un retraso entre cuando los datos se escriben en la región primaria y cuando estarán disponibles en la región secundaria. Este retraso podría provocar una posible pérdida de datos si se produjese un fallo en la región principal antes de que se repliquen los datos recientes. La pérdida de datos se mide mediante el objetivo de punto de recuperación (RPO). Espere que el retraso de replicación sea inferior a 15 minutos, pero este tiempo es una estimación y no se garantiza.

  Compruebe la propiedad Última hora de sincronización para saber cuántos datos se podrían perder si su cuenta de almacenamiento sufriese una conmutación por error no planificada.

- **Acceso a la región secundaria:** con las configuraciones de almacenamiento con redundancia geográfica (GRS) y almacenamiento con redundancia de zona geográfica

(GZRS), la región secundaria no será accesible a las lecturas hasta que se produzca una conmutación por error.

Las configuraciones de almacenamiento con redundancia geográfica con acceso de lectura (RA-GRS) y almacenamiento con redundancia de zona geográfica con acceso de lectura (RA-GZRS) proporcionan acceso de lectura a la región secundaria durante las operaciones normales, pero debido a la latencia de replicación asincrónica, podrían devolver datos ligeramente obsoletos.

- **Limitaciones de las características:** algunas características de Azure Storage no son compatibles o tienen limitaciones cuando se utiliza almacenamiento con redundancia geográfica (GRS) o conmutación por error administrada por el cliente. Revise la compatibilidad de las característica antes de implementar la redundancia geográfica.

Cost

Las configuraciones de cuentas de Azure Storage multirregionales conllevan costes adicionales por la replicación entre regiones y el almacenamiento en la región secundaria. La transferencia de datos entre regiones de Azure se cobra en función de las tarifas estándar de ancho de banda entre regiones.

Para obtener información detallada sobre los precios, consulte Precios de Queue Storage.

Configuración de la compatibilidad con varias regiones

- **Cree una nueva cuenta de almacenamiento con redundancia geográfica (GRS).** Para crear una cuenta de GRS, consulte Creación de una cuenta de almacenamiento y seleccione GRS, almacenamiento con redundancia geográfica con acceso de lectura (RA-GRS), almacenamiento con redundancia de zona geográfica (GZRS) o almacenamiento con redundancia de zona geográfica con acceso de lectura (RA-GZRS) durante la creación de la cuenta.

- **Habilite la redundancia geográfica en una cuenta de almacenamiento existente.** Para convertir una cuenta de almacenamiento existente en almacenamiento con redundancia geográfica (GRS), consulte Cambio de cómo se replica una cuenta de almacenamiento.

> ⚠ **Advertencia**
>
> Una vez que su cuenta se haya reconfigurado para la redundancia geográfica, es posible que transcurra un tiempo considerable antes de que los datos existentes en la nueva región principal se copien por completo a la nueva región secundaria.

Para evitar una pérdida importante de datos, compruebe el valor de la **propiedad Última hora de sincronización** antes de iniciar una conmutación por error no planificada. Para evaluar la posible pérdida de datos, compare la última hora de sincronización con la última hora en que se escribieron los datos en la nueva región principal.

- **Deshabilite la redundancia geográfica.** Convierta las cuentas de GRS en configuraciones de una sola región, como el almacenamiento con redundancia local (LRS) o el almacenamiento con redundancia local (ZRS) con el mismo proceso de cambio de configuración de redundancia.

Comportamiento cuando todas las regiones están en buen estado

En esta sección se describe qué esperar cuando una cuenta de almacenamiento está configurada para redundancia geográfica y todas las regiones están operativas.

- **Enrutamiento del tráfico entre regiones:** Azure Storage utiliza un enfoque activo-pasivo en el que todas las operaciones de escritura y la mayoría de las operaciones de lectura se dirigen a la región principal.

  En las configuraciones de almacenamiento con redundancia geográfica con acceso de lectura (RA-GRS) y almacenamiento con redundancia de zona geográfica con acceso de lectura (RA-GZRS), las aplicaciones pueden leer opcionalmente desde la región secundaria accediendo al punto de conexión secundario. Este enfoque requiere una configuración explícita de la aplicación y no es automática. Además, debido al retraso de replicación asincrónica, los datos de la región secundaria podrían estar ligeramente obsoletos.

- **Replicación de datos entre regiones:** las operaciones de escritura se confirman primero en la región principal utilizando los siguientes tipos de redundancia configurados:
  - Almacenamiento con redundancia local (LRS) para almacenamiento con redundancia geográfica (GRS) y RA-GRS
  - Almacenamiento con redundancia de zona (ZRS) para el almacenamiento con redundancia de zona geográfica (GZRS) y RA-GZRS

  Tras completarse con éxito en la región principal, los datos se replican de forma asíncrona en la región secundaria, donde se almacenan mediante LRS.

  La naturaleza asíncrona de la replicación entre regiones significa que suele haber un retraso entre el momento en que los datos se escriben en la región principal y el momento en que están disponibles en la región secundaria. Supervise el tiempo de replicación utilizando la propiedad Última hora de sincronización.

Comportamiento durante una falla de región

En esta sección se describe qué esperar cuando una cuenta de almacenamiento está configurada para la redundancia geográfica y hay una interrupción en la región primaria.

- **Conmutación por error administrada por el cliente (no planificada):** utilice una conmutación por error no planificada cuando el almacenamiento en la región principal no esté disponible.
  - **Detección y respuesta:** En el caso poco probable de que su cuenta de almacenamiento no esté disponible en la región primaria, puede considerar iniciar un failover no planeado administrado por el cliente. Para tomar esta decisión, tenga en cuenta los siguientes factores:
    - Si Azure Resource Health muestra problemas al acceder a la cuenta de almacenamiento en su región principal
    - Si Microsoft le recomienda realizar una conmutación por error a otra región

    > ⚠ **Advertencia**
    >
    > Una conmutación por error no planeada puede **provocar la pérdida de datos**. Antes de iniciar una conmutación por error administrada por el cliente, decida si la restauración del servicio justifica el riesgo de pérdida de datos.

  - **Notificación**: Microsoft no le notifica automáticamente cuando una región está inactiva. Sin embargo:
    - Puede usar Azure Resource Health para supervisar el estado de un recurso individual y puede configurar alertas de Resource Health para notificarle problemas.
    - Puede usar Azure Service Health para comprender el estado general del servicio, incluidos los errores de región, y puede configurar alertas de Service Health para notificarle problemas.
  - **Solicitudes activas:** Durante el proceso de conmutación por error, los puntos de conexión de la cuenta de almacenamiento principal y secundaria dejan de estar disponibles temporalmente para lecturas y escrituras. Las solicitudes activas podrían fallar, y las aplicaciones cliente deben intentar nuevamente después de que se complete la conmutación por error.
  - **Pérdida de datos prevista:** la pérdida de datos es habitual durante una conmutación por error no planificada debido al retraso en la replicación asíncrona, lo que significa que es posible que las escrituras recientes no se repliquen. Consulte la propiedad

Última hora de sincronización para saber cuántos datos se podrían perder durante una conmutación por error no planificada. La pérdida de datos esperada se conoce a menudo como objetivo de punto de recuperación (RPO). Normalmente, puede esperar que el RPO sea inferior a 15 minutos, pero ese tiempo no está garantizado.

  - **Tiempo de inactividad esperado:** La cantidad de tiempo de inactividad esperado se conoce a menudo como objetivo de tiempo de recuperación (RTO). La conmutación por error administrada por el cliente se completa normalmente en un plazo de hasta 60 minutos, dependiendo del tamaño y la complejidad de la cuenta.

  - **Reenrutamiento del tráfico:** A medida que se completa la conmutación por error, Azure actualiza automáticamente los puntos de conexión de la cuenta de almacenamiento para que no sea necesario volver a configurar las aplicaciones. Si su aplicación mantiene entradas del Sistema de nombres de dominio (DNS) almacenadas en caché, puede que sea necesario borrar la caché para asegurarse de que la aplicación envíe el tráfico a la nueva región principal.

  - **Configuración posterior a la conmutación por error:** una vez completada una conmutación por error no planificada, su cuenta de almacenamiento en la región de destino utilizará el nivel de almacenamiento con redundancia local (LRS). Si necesitase volver a replicar geográficamente los datos, deberá volver a habilitar el almacenamiento con redundancia geográfica (GRS) y esperar a que los datos se repliquen en la nueva región secundaria.

  Para obtener más información sobre cómo iniciar una conmutación por error administrada por el cliente, consulte Cómo funciona la conmutación por error administrada por el cliente (no planificada) e Iniciar una conmutación por error de una cuenta de almacenamiento.

- **Conmutación por error administrada por el cliente (planificada):** utilice una conmutación por error planificada cuando el almacenamiento siga operativo en la región principal, pero necesite conmutar por error toda su solución a una región secundaria por otro motivo. Por ejemplo, otro servicio de Azure podría estar experimentando un problema y debe cambiar al uso de una región secundaria para toda la solución. O bien, puede usar una conmutación por error planeada para realizar un simulacro de recuperación ante desastres (DR) con fines de cumplimiento y auditoría.

  - **Detección y respuesta:** Usted es responsable de decidir conmutar por error. Normalmente, debe tomar esta decisión si necesita realizar una conmutación entre regiones, aunque su cuenta de almacenamiento esté en buen estado. Por ejemplo, puede desencadenar una conmutación por error cuando se produce una interrupción importante en otro componente de la aplicación que no se puede recuperar en la región principal.

- **Notificación**: Microsoft no le notifica automáticamente cuando una región está inactiva. Sin embargo:
  - Puede usar Azure Resource Health para supervisar el estado de un recurso individual y puede configurar alertas de Resource Health para notificarle problemas.
  - Puede usar Azure Service Health para comprender el estado general del servicio, incluidos los errores de región, y puede configurar alertas de Service Health para notificarle problemas.
- **Solicitudes activas:** Durante el proceso de conmutación por error, los puntos de conexión de la cuenta de almacenamiento principal y secundaria dejan de estar disponibles temporalmente para lecturas y escrituras. Las solicitudes activas podrían fallar, y las aplicaciones cliente deben intentar nuevamente después de que se complete la conmutación por error.
- **Pérdida de datos esperada:** No se espera ninguna pérdida de datos porque el proceso de conmutación por error se completa solo después de sincronizar todos los datos, lo que da como resultado un RPO de cero.
- **Tiempo de inactividad esperado:** La conmutación por error normalmente se completa en un plazo de 60 minutos, lo que significa que el tiempo de recuperación objetivo (RTO) esperado es de 60 minutos, en función del tamaño y la complejidad de la cuenta. Durante el proceso de conmutación por error, los puntos de conexión de la cuenta de almacenamiento principal y secundaria dejan de estar disponibles temporalmente para lecturas y escrituras.
- **Reenrutamiento del tráfico:** A medida que se completa la conmutación por error, Azure actualiza automáticamente los puntos de conexión de la cuenta de almacenamiento para que no sea necesario volver a configurar las aplicaciones. Si su aplicación mantiene entradas DNS almacenadas en caché, puede que sea necesario borrar la caché para asegurarse de que la aplicación envíe tráfico a la nueva región principal.
- **Configuración posterior a la conmutación por error:** Una vez completada una conmutación por error planeada, la cuenta de almacenamiento de la región de destino sigue siendo replicada geográficamente y permanece en el nivel GRS.

Para obtener más información sobre cómo iniciar una conmutación por error administrada por el cliente, consulte Cómo funciona la conmutación por error administrada por el cliente (planificada) e Iniciar una conmutación por error de una cuenta de almacenamiento.

- **Conmutación por error administrada por Microsoft:** En el caso excepcional de un desastre importante en el que Microsoft determina que la región primaria es irrecuperable permanentemente, es posible que se inicie una conmutación automática por error a la región secundaria. Microsoft controla todo el proceso y no se requiere ninguna acción del cliente. El tiempo que transcurre antes de que ocurra un failover depende de la gravedad del desastre y del tiempo necesario para evaluar la situación.
  - **Notificación**: Microsoft no le notifica automáticamente cuando una región está inactiva. Sin embargo:
    - Puede usar Azure Resource Health para supervisar el estado de un recurso individual y puede configurar alertas de Resource Health para notificarle problemas.
    - Puede usar Azure Service Health para comprender el estado general del servicio, incluidos los errores de región, y puede configurar alertas de Service Health para notificarle problemas.

> **Importante**
>
> Utilice las opciones de conmutación por error administradas por el cliente para desarrollar, probar e implementar tus planes de recuperación ante desastres. **No confíe en la conmutación por error administrada por Microsoft**, que solo se puede utilizar en circunstancias extremas. Es probable que se haya iniciado una conmutación por error administrada por Microsoft para toda una región. No se puede iniciar para cuentas de almacenamiento individuales, suscripciones o clientes. La conmutación por error puede producirse en momentos diferentes para distintos servicios de Azure. Recomendamos que utilice la conmutación por error administrada por el cliente.

Recuperación de regiones

El proceso de conmutación por recuperación difiere significativamente entre los escenarios de conmutación por error administrados por Microsoft y los administrados por el cliente.

- **Conmutación por error administrada por el cliente (no planeada):** después de una conmutación por error no planeada, la cuenta de almacenamiento se configura con almacenamiento con redundancia local (LRS). Para realizar una conmutación por error, es necesario restablecer la relación de almacenamiento con redundancia geográfica (GRS) y esperar a que se repliquen los datos.
- **Conmutación por error administrada por el cliente (planificada):** después de una conmutación por error planificada, la cuenta de almacenamiento seguirá estando

replicada geográficamente. Inicie otra conmutación por error administrada por el cliente para volver a la región principal original. Se aplican las mismas consideraciones de conmutación por error.

- **Conmutación por error administrada por Microsoft:** si Microsoft iniciase una conmutación por error, es probable que se haya producido un desastre importante en la región principal y que esta no se pueda recuperar. Cualquier calendario o plan de recuperación dependerá del alcance del desastre regional y de los esfuerzos de recuperación. Debe supervisar las comunicaciones de Azure Service Health para más información.

Prueba de fallos de región

Simule fallos regionales para probar sus procedimientos de recuperación ante desastres.

- **Pruebas de conmutación por error planificadas:** en el caso de las cuentas de almacenamiento con redundancia geográfica (GRS), es posible realizar operaciones de conmutación por error planificadas durante las ventanas de mantenimiento para probar el proceso completo de conmutación por error y conmutación por recuperación. La conmutación por error planificada no implica pérdida de datos, pero sí conlleva tiempo de inactividad tanto durante la conmutación por error como durante la conmutación por recuperación.

- **Pruebas de puntos de conexión secundarios:** para el almacenamiento con redundancia geográfica con acceso de lectura (RA-GRS) y las configuraciones de almacenamiento con redundancia de zona geográfica con acceso de lectura (RA-GZRS), pruebe periódicamente las operaciones de lectura en el punto de conexión secundario para asegurarse de que la aplicación pueda leer correctamente los datos de la región secundaria.

Soluciones personalizadas de varias regiones para la resistencia

Las funcionalidades de conmutación por error entre regiones de Azure Storage podrían no ser adecuadas por los siguientes motivos:

- La cuenta de almacenamiento está en una región no emparejada.

- Los objetivos de tiempo de actividad empresarial no se satisfacen por el tiempo de recuperación o la pérdida de datos que proporciona las opciones de conmutación por error integradas.

- Debe conmutar por error a una región que no esté emparejada con la región principal.
- Necesita una configuración activa o activa entre regiones.

En esta sección se proporciona información general de alto nivel de algunos enfoques que se deben tener en cuenta. Una introducción completa de las topologías de implementación de varias regiones para Azure Storage está fuera del ámbito de este artículo.

> **Nota:**
>
> En el caso de los requisitos avanzados de varias regiones, considere la posibilidad de usar Service Bus en su lugar, lo que incluye compatibilidad con regiones no emparejadas.

Implemente Azure Storage en varias regiones utilizando cuentas de almacenamiento independientes en cada región. Este enfoque ofrece flexibilidad en la selección de regiones, la posibilidad de utilizar regiones no emparejadas y un control más granular sobre el momento de la replicación y la coherencia de los datos. Al implementar varias cuentas de almacenamiento entre regiones, es necesario configurar la replicación de datos entre regiones, implementar directivas de equilibrio de carga y conmutación por error y garantizar la coherencia de los datos entre regiones.

Este enfoque requiere que administre la distribución de mensajes, controle la sincronización de datos entre las colas de las distintas cuentas de almacenamiento e implemente la lógica de conmutación por error personalizada.

Copias de seguridad y restauración

Queue Storage no proporciona funcionalidades de copia de seguridad tradicionales, como la restauración a un momento dado (PITR). Esto se debe a que las colas están diseñadas para el almacenamiento transitorio de mensajes en lugar de la persistencia de datos a largo plazo. Normalmente, los mensajes se procesan y quitan de las colas durante las operaciones normales de la aplicación.

En escenarios que requieran durabilidad de mensajes más allá de las opciones de redundancia integradas, considere la posibilidad de implementar su propio registro de mensajes de nivel de aplicación o persistencia en un almacén de datos permanente, como Blob Storage o Azure SQL Database. Este enfoque le permite mantener el historial de mensajes mientras usa Queue Storage para su propósito previsto de almacenamiento temporal de búfer de mensajes y coordinación de procesamiento.

Acuerdo de nivel de servicio

El acuerdo de nivel de servicio (SLA) para Azure Storage describe la disponibilidad esperada del servicio y las condiciones que se deben cumplir para lograr esa expectativa de disponibilidad. El contrato de nivel de servicio de disponibilidad al que tiene derecho dependerá del nivel de almacenamiento y del tipo de replicación que utilice. Para obtener más información, consulte los SLAs for Online Services.

Contenido relacionado

- ¿Qué es Queue Storage?
- Redundancia de Azure Storage
- Planeamiento y conmutación por error de recuperación ante desastres de Azure Storage
- ¿Qué son las zonas de disponibilidad?
- Fiabilidad de Azure
- Recomendaciones para controlar errores transitorios

ⓘ **Nota:** El autor ha creado este artículo con ayuda de inteligencia artificial. Más información

---

Last updated on 22/01/2026

Confiabilidad en Acciones de almacenamiento de Azure

En este artículo se describe la compatibilidad con la confiabilidad en Acciones de almacenamiento de Azure y se trata tanto la resistencia dentro de la región con zonas de disponibilidad y recuperación ante desastres entre regiones y continuidad empresarial. Para obtener información general más detallada sobre los principios de confiabilidad de Azure, consulte Confiabilidad de Azure.

Acciones de almacenamiento de Azure es un marco sin servidor que puede usar para realizar operaciones de datos comunes en millones de objetos en varias cuentas de almacenamiento. El propio servicio es regional y no tiene SKU ni compatibilidad con zonas de disponibilidad. Sin embargo, el plano de control del servicio admite automáticamente la redundancia de zona. El plano de datos también puede admitir redundancia en función de si la cuenta de almacenamiento se ejecuta o no en una configuración con redundancia de zona.

Soporte para zonas de disponibilidad

Las zonas de disponibilidad son grupos físicamente independientes de centros de datos dentro de una región de Azure. Cuando una zona falla, los servicios pueden transferirse a una de las zonas restantes.

Aunque el servicio Acciones de almacenamiento de Azure es regional y no ofrece SKU ni zonas de disponibilidad, la redundancia de zona está disponible desde el plano de control y condicionalmente desde el plano de datos:

- **El plano de control del servicio es con redundancia de zona.** Cuando una zona está inactiva en una región, el plano de control sigue estando disponible. Durante un escenario de zona inactiva, puede seguir administrando la definición y la asignación de tareas.
- **El plano de datos (ejecución de la asignación de tareas) hereda las propiedades zonales de la cuenta de almacenamiento primaria.** Si la cuenta de almacenamiento se implementa en una zona con errores, la cuenta deja de estar disponible y, desde la perspectiva del cliente, el plan de datos no está disponible. Si la cuenta de almacenamiento es con redundancia de zona, la cuenta sigue estando disponible y el servicio continúa realizando la operación en la cuenta.

Experiencia a nivel de zona

En un escenario de zona terminada, el servicio Acción de almacenamiento sigue estando disponible. El progreso de las tareas depende de la compatibilidad de la zona de disponibilidad de las cuentas de almacenamiento con las que se ejecutan. Si la cuenta no se ve afectada por la zona inactiva, las tareas siguen progresando. De lo contrario, se produce un error en las tareas.

Preparación y recuperación de interrupciones de zona

El servicio Acción de almacenamiento no es zonal, pero la cuenta de almacenamiento sí. Si la cuenta de almacenamiento se ve afectada por una interrupción de zona, se producirá un error en las tareas de almacenamiento asignadas a la cuenta. Una vez que la zona y la cuenta de almacenamiento estén disponibles, las tareas programadas continúan ejecutándose según la programación. Si la tarea está configurada para ejecutarse una vez, es posible que deba programarla para que se vuelva a ejecutar.

Recuperación ante desastres entre regiones y continuidad empresarial

La recuperación ante desastres (DR) hace referencia a las prácticas que las organizaciones usan para recuperarse de eventos de alto impacto, como desastres naturales o implementaciones con errores que producen tiempo de inactividad y pérdida de datos. Independientemente de la causa, el mejor remedio para un desastre es un plan de recuperación ante desastres bien definido y probado y un diseño de aplicaciones que apoye activamente la recuperación ante desastres. Antes de empezar a crear el plan de recuperación ante desastres, consulte Recomendaciones para diseñar una estrategia de recuperación ante desastres.

Para la recuperación ante desastres, Microsoft usa el modelo de responsabilidad compartida. En este modelo, Microsoft garantiza que la infraestructura de línea base y los servicios de plataforma estén disponibles. Sin embargo, muchos servicios de Azure no replican automáticamente datos ni se revierten de una región con errores para realizar la replicación cruzada en otra región habilitada. En esos servicios, es responsable de configurar un plan de recuperación ante desastres válido para su carga de trabajo. La mayoría de los servicios que se ejecutan en ofertas de plataforma como servicio (PaaS) de Azure proporcionan características e instrucciones para admitir la recuperación ante desastres. Puede usar características específicas del servicio para admitir la recuperación rápida con el fin de contribuir al desarrollo del plan de recuperación ante desastres.

Las cuentas de almacenamiento con GRS y GZRS replican datos en una región secundaria en caso de conmutaciones de error de la cuenta de almacenamiento. La continuidad empresarial de las acciones de almacenamiento depende significativamente de la configuración de redundancia de la cuenta de almacenamiento de destino. Las cuentas de almacenamiento

configuradas con redundancia geográfica se benefician de un proceso de conmutación por error automatizado. Esta gestión automática garantiza que las futuras asignaciones de tareas, ya sean únicas o recurrentes, se ejecuten en la región secundaria sin problemas. Sin embargo, las tareas de almacenamiento que estaban en curso en el momento de la conmutación por error podrían encontrarse con fallos. Las nuevas tareas de almacenamiento y las asignaciones de tareas de almacenamiento siguen funcionando según lo previsto.

La supervisión coherente de la cuenta de almacenamiento es fundamental. Con una conmutación por error, debe revisar exhaustivamente los informes de tareas y la supervisión para comprobar la finalización correcta de todas las operaciones de blobs e identificar las discrepancias que necesitan atención.

Detección, notificación y administración de interrupciones

Las tareas de almacenamiento no envían ninguna notificación cuando hay una interrupción en el propio servicio. Es importante comprobar el estado de la tarea de almacenamiento y volver a intentar las tareas después de que se recupere el servicio o la región.

Pasos siguientes

- Tutorial: Creación de una aplicación de varias regiones de alta disponibilidad en Azure App Service
- Confiabilidad en Azure

Last updated on 22/01/2026

Confiabilidad en la detección de Azure Storage

Este artículo explica cómo Azure Storage Discovery admite la confiabilidad, incluida la resiliencia intrarregional con zonas de disponibilidad y recuperación ante desastres cruzada de regiones y continuidad empresarial. Para conocer los principios generales de confiabilidad en Azure, consulte Azure confiabilidad.

Azure Storage Discovery es un servicio totalmente administrado que proporciona visibilidad de toda la empresa sobre la Azure Blob Storage y Azure Data lake Storage. Desde un solo panel de vidrio, puede analizar tendencias, optimizar los costos, mejorar la seguridad y mejorar la eficiencia operativa.

Comportamiento del servicio durante interrupciones

Expandir tabla

	Interrupción de la región	Interrupción de zona
Gestión del espacio de trabajo de almacenamiento	Si la región en la que se crea el espacio de trabajo de descubrimiento de almacenamiento experimenta una interrupción, los informes del espacio de trabajo no estarán disponibles durante el periodo.	Si una zona está inactiva, el área de trabajo Detección de almacenamiento permanece disponible. Puede seguir administrando el área de trabajo y acceder a los informes.
Informes de detección de almacenamiento	Si una región deja de funcionar, los datos de las cuentas de almacenamiento de esa región se pueden retrasar en los informes de detección hasta que se restaure la región. En el caso poco frecuente de pérdida completa de datos durante el periodo de interrupción, el área de trabajo de detección no puede recuperar los datos perdidos durante ese periodo.	Si la cuenta de almacenamiento se implementa en una zona con errores, la cuenta deja de estar disponible y la información del almacenamiento afectado no está disponible. Si la cuenta de almacenamiento es redundante por zona, el Discovery workspace continúa proporcionando información sobre la cuenta.

Soporte para zonas de disponibilidad

Availability zones son grupos de centros de datos físicamente independientes dentro de una región de Azure. Cuando una zona falla, los servicios pueden transferirse a una de las zonas

restantes.

- La administración del área de trabajo de Detección de almacenamiento tiene redundancia de zona. Cuando una zona está inactiva, la área de trabajo de Detección de almacenamiento permanece disponible. Puede seguir administrando el área de trabajo y acceder a los informes.
- Los informes de Detección de almacenamiento tienen redundancia de zona. Sin embargo, la interrupción del recurso de almacenamiento puede afectar a la información sobre ese recurso. Si un recurso de almacenamiento tiene redundancia de zona, el flujo de información permanece ininterrumpido.

Recuperación ante desastres entre regiones y continuidad empresarial

La recuperación ante desastres (DR) hace referencia a las prácticas que las organizaciones usan para recuperarse de eventos de alto impacto, como desastres naturales o implementaciones con errores que producen tiempo de inactividad y pérdida de datos. Independientemente de la causa, el mejor remedio para un desastre es un plan de recuperación ante desastres bien definido y probado y un diseño de aplicaciones que apoye activamente la recuperación ante desastres. Antes de empezar a crear el plan de recuperación ante desastres, consulte Recomendaciones para diseñar una estrategia de recuperación ante desastres.

Para la recuperación ante desastres, Microsoft usa el modelo de responsabilidad compartida. En este modelo, Microsoft garantiza que la infraestructura de línea base y los servicios de plataforma estén disponibles. Sin embargo, muchos servicios de Azure no replican automáticamente datos ni se revierten de una región con errores para realizar la replicación cruzada en otra región habilitada. En esos servicios, es responsable de configurar un plan de recuperación ante desastres válido para su carga de trabajo. La mayoría de los servicios que se ejecutan en ofertas de plataforma como servicio (PaaS) de Azure proporcionan características e instrucciones para admitir la DR. Puede usar características específicas del servicio para admitir la recuperación rápida con el fin de contribuir al desarrollo del plan de recuperación ante desastres.

Detección y notificación de interrupciones

La detección de almacenamiento no proporciona notificaciones de interrupción para cada incidente. El servicio De detección de almacenamiento utiliza la misma infraestructura de notificación que se usa en Azure para informar a los clientes sobre interrupciones prolongadas o graves.

> ⓘ **Nota:**
>
> Si necesita acceso ininterrumpido a los informes de detección, puede considerar la posibilidad de crear una segunda área de trabajo en otra región. Esta segunda área de trabajo debe configurarse para analizar los mismos recursos de almacenamiento. Este enfoque le protege de una interrupción regional que afecta al área de trabajo principal. Sin embargo, no protege frente a problemas de disponibilidad en los propios recursos de almacenamiento entre regiones. Obtenga más información sobre los precios de Storage Discovery para evaluar los costos adicionales.

Disponibilidad de datos y cobertura de actualización

Factores clave que influyen en la disponibilidad y actualización de los datos en los informes de detección de almacenamiento:

- Después de crear un área de trabajo de detección de almacenamiento, la mayoría de las conclusiones aparecen en los informes en un plazo de 24 horas.
- La mayoría de las actualizaciones, como agregar o quitar una cuenta de almacenamiento en una suscripción o un grupo de recursos vinculados a un área de trabajo de detección de almacenamiento, se reflejarán en los informes de detección de almacenamiento en un plazo de 24 horas.
- Los cambios relacionados con la adición de un nuevo ámbito o la edición de un ámbito existente en el área de trabajo de detección se reflejan normalmente en los informes en un plazo de 24 horas.
- El efecto de cambiar el plan de precios en un área de trabajo de Detección de almacenamiento existente es inmediato.
- Cuando se elimina un ámbito, se pierde inmediatamente la información recopilada para este ámbito.
- Cuando se elimina un área de trabajo de Discovery, se pierden inmediatamente los análisis recopilados para todos los ámbitos contenidos. Tenga cuidado al eliminar un área de trabajo o un ámbito.

Pasos siguientes

- Planeamiento de la implementación de la Detección de almacenamiento
- Preguntas más frecuentes

ⓘ**Nota:** El autor ha creado este artículo con ayuda de inteligencia artificial. Más información

Last updated on 15/04/2026

Confiabilidad en Azure Storage Mover

Azure Storage Mover es un servicio totalmente administrado que migra archivos y carpetas a Azure Storage y mantiene los archivos sincronizados entre cuentas de almacenamiento. Use Storage Mover cuando mueva datos a Azure o cuando necesite mantener los datos sincronizados entre diferentes ubicaciones dentro de Azure.

Cuando se usa Azure, la confiabilidad es una responsabilidad compartida. Microsoft proporciona una variedad de capacidades para apoyar la resiliencia y la recuperación. Es responsable de comprender cómo funcionan esas funcionalidades dentro de todos los servicios que usa y de seleccionar las funcionalidades que necesita para cumplir los objetivos empresariales y los objetivos de tiempo de actividad.

En este artículo se describe cómo Azure Storage Mover responde a una variedad de posibles interrupciones y problemas, incluidos errores transitorios, errores de zona de disponibilidad y errores en toda la región. También se describe cómo proteger la configuración de Storage Mover.

> **ⓘ Importante**
>
> En este artículo se describe la confiabilidad del servicio Azure Storage Mover y sus recursos únicamente. La confiabilidad de una migración de un extremo a otro depende de todos los componentes: el servicio Storage Mover, los agentes de Storage Mover que implemente, el entorno de origen y la conectividad de red y la cuenta de almacenamiento de destino. Es responsable de la confiabilidad de los agentes, los sistemas de origen y el almacenamiento de destino. Para obtener más información sobre la confiabilidad de Azure Storage, consulte Confiabilidad en Azure Blob Storage y Confiabilidad en Azure Files.

Introducción a la arquitectura de confiabilidad

En esta sección se describen algunos de los aspectos importantes de cómo funciona el servicio que es más relevante desde una perspectiva de confiabilidad. En la sección se presenta la arquitectura lógica, que incluye algunos de los recursos y características que se implementan y usan. También se describe la arquitectura física, que proporciona detalles sobre cómo funciona el servicio en segundo plano.

Arquitectura lógica

Azure Storage Mover está diseñado para migrar y sincronizar datos entre ubicaciones de almacenamiento, no para atender solicitudes en la ruta de acceso en tiempo de ejecución de una carga de trabajo de producción. Tiene una jerarquía de recursos que define los componentes que implementa y administra. El recurso de nivel superior se denomina *mover de almacenamiento*. Dentro de un mover de almacenamiento, se definen *proyectos* que contienen *definiciones de trabajo* que describen qué migrar y dónde. *Los puntos de conexión* definen las ubicaciones de origen y destino para un trabajo de migración o sincronización.

En algunos escenarios, como las migraciones desde entornos locales, también se implementan uno o varios *agentes de Storage Mover*. Un agente es software que se ejecuta en una máquina que controla, como una máquina virtual o una máquina física. Algunos escenarios no requieren un agente.

El servicio almacena metadatos de configuración, incluidos proyectos, puntos de conexión, registros de agente, definiciones de trabajo y historial de ejecución de trabajos. Estos metadatos no incluyen los datos que migras.

Arquitectura física

El servicio Azure Storage Mover se ejecuta en una infraestructura administrada por Microsoft. Los agentes se ejecutan en el hardware que administras. Usted es responsable de la confiabilidad de los agentes, que está fuera del ámbito de este artículo.

Resistencia a errores transitorios

Los errores transitorios son errores breves e intermitentes en los componentes. Se producen con frecuencia en un entorno distribuido como la nube y son una parte normal de las operaciones. Los errores transitorios se corrigen después de un breve período de tiempo. Es importante que las aplicaciones puedan controlar errores transitorios, normalmente mediante el reintento de solicitudes afectadas.

Todas las aplicaciones hospedadas en la nube deben seguir las instrucciones de control de errores transitorios de Azure cuando se comunican con cualquier API, bases de datos y otros componentes hospedados en la nube. Para más información, vea Recomendaciones para gestionar errores temporales.

Si un error transitorio afecta a la comunicación entre un agente y el servicio Storage Mover, o al conectarse a un origen o destino, el agente vuelve a intentarlo automáticamente. Para los

trabajos de Azure a Azure, el servicio también es resistente a muchos errores transitorios. Cuando se restaura la conectividad, los trabajos de migración en curso se reanudan.

En algunos casos, los errores transitorios aparecen como errores en el historial de ejecución de trabajos. Para obtener una descripción de los códigos de error, incluidos los errores transitorios, consulte Azure Storage códigos de estado y tipos de error de Mover. Para obtener instrucciones sobre cómo resolver problemas de conectividad de red persistentes, consulte Solución de problemas de conectividad de red de Azure Storage Mover.

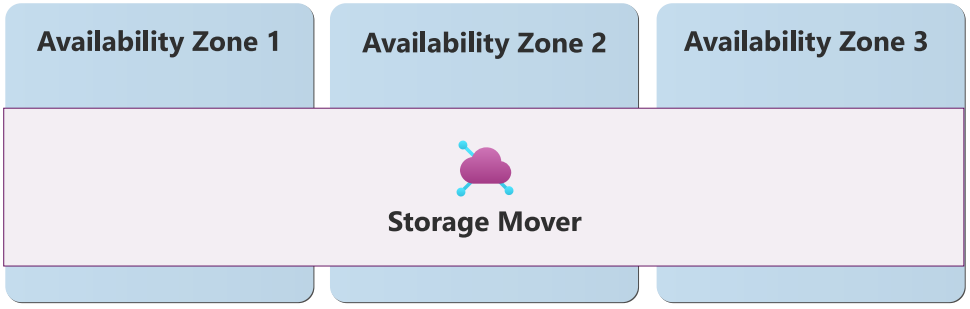
Resistencia a errores de zona de disponibilidad

Las zonas de disponibilidad son grupos físicamente independientes de centros de datos dentro de una región de Azure. Cuando una zona falla, los servicios pueden transferirse a una de las zonas restantes.

En las regiones que admiten zonas de disponibilidad, la plataforma distribuye los metadatos de configuración de un storage mover entre zonas en la medida de lo posible, pero este comportamiento no está garantizado. Si sus migraciones de almacenamiento deben soportar la pérdida de una zona, diseñe el proceso de migración para que tolere la pérdida de un migrador de almacenamiento y revise Resiliencia ante fallos en toda una región.

Tenga en cuenta el efecto de un error de zona en el contexto de cómo se usa Storage Mover. El servicio orquesta la migración y la sincronización de datos, y normalmente no forma parte de la ruta de ejecución de su carga de trabajo de producción. Si un mecanismo de traslado de almacenamiento no está disponible durante un fallo de zona, un trabajo de migración o sincronización suele retrasarse en lugar de provocar una interrupción en producción, y se puede reanudar o reintentar después de que el servicio se recupere. Storage Mover tampoco ofrece un acuerdo de nivel de servicio (SLA) de disponibilidad, por lo que el diseño no debe suponer que el servicio está disponible continuamente. Si la carga de trabajo depende de la sincronización en curso, evalúe si este tipo de retraso es aceptable para su escenario.

En el diagrama siguiente se muestra un mover de almacenamiento con metadatos de infraestructura y configuración distribuidos entre tres zonas:



> **Nota:**
>
> La confiabilidad de cualquier migración de datos también depende de las cuentas de almacenamiento y los agentes que use. Por ejemplo, si la cuenta de almacenamiento de destino usa almacenamiento con redundancia local (LRS), no es resistente a un error de zona. Para que una migración sea resistente a un error de zona, use una cuenta de almacenamiento de destino con redundancia de zona.

Requirements

Compatibilidad regional: La distribución según el criterio de máximo esfuerzo de los metadatos de configuración entre zonas solo puede darse en una región que sea compatible tanto con Storage Mover como con las zonas de disponibilidad. Compruebe la disponibilidad de la región de Storage Mover y compárela con la lista de regiones que admiten zonas de disponibilidad. Incluso en estas regiones, la resiliencia zonal no está garantizada.

Costo

Storage Mover no ofrece compatibilidad con zonas de disponibilidad configurables, por lo que no incurre en ningún costo adicional relacionado con las zonas de disponibilidad. Para obtener más información sobre cómo se factura Storage Mover, consulte Descripción de la facturación de Azure Storage Mover.

Configurar soporte de zonas de disponibilidad

Storage Mover no ofrece compatibilidad con zonas de disponibilidad configurables, por lo que no hay nada para habilitar o participar. Para obtener más información sobre cómo crear un recurso de Storage Mover, consulte Planeamiento de implementación para Azure Storage Mover.

Comportamiento cuando todas las zonas están en buen estado

En esta sección se describe qué esperar cuando un mover de almacenamiento está en una región que admite zonas de disponibilidad y todas las zonas están operativas.

- **Operación entre zonas:** La infraestructura de cualquiera de las zonas de disponibilidad de la región podría servir a las operaciones de administración y al acceso a metadatos. Las conexiones de los agentes pueden acceder al servicio a través de cualquier zona.
- **Distribución de datos entre zonas:** El servicio tiene como objetivo replicar de forma sincrónica los metadatos de configuración en las zonas de disponibilidad de la región.

Comportamiento durante un fallo de zona

En esta sección se describe qué esperar cuando un mover de almacenamiento está en una región que admite zonas de disponibilidad y hay una interrupción en una de las zonas.

- **Detección y respuesta:** La plataforma está diseñada para detectar la pérdida de una zona de disponibilidad y redirigir el tráfico a zonas correctas, pero esta respuesta es el mejor esfuerzo y no está garantizada.
- **Notificación**: Microsoft no le notifica automáticamente cuando una zona está inactiva. Sin embargo, puede usar Azure Resource Health para supervisar el estado de un recurso individual y puede configurar alertas de Resource Health para notificarle problemas. También puede usar Azure Service Health para comprender el estado general del servicio, incluidos los errores de zona, y puede configurar alertas de Service Health para notificarle problemas.
- **Solicitudes activas:** Las operaciones de administración en curso que dependen de la infraestructura de la zona afectada pueden producir errores y debe reintentarlas. Los trabajos de migración de datos activos que se ejecutan en agentes podrían seguir ejecutándose, pero las operaciones de administración que dependen del servicio de metadatos podrían no estar disponibles.
- **Pérdida de datos esperada:** Los datos que está migrando un mover de almacenamiento no se pierden durante un error de zona.

  Dado que Storage Mover no garantiza que los metadatos de configuración se distribuyan entre zonas, un error de zona podría hacer que algunos de los metadatos de configuración del mover de almacenamiento no estén disponibles temporalmente hasta que se recupere la zona.

- **Tiempo de inactividad esperado:** La plataforma intenta restaurar las operaciones mediante otra zona, pero en algunas situaciones es posible que las operaciones no estén disponibles hasta que se recupere la zona afectada. Prepare la carga de trabajo siguiendo las instrucciones de control de errores transitorios.

- **Redistribución:** Si la plataforma redirige el tráfico a zonas de disponibilidad en buen estado, lo hace en la medida de lo posible.

Recuperación de zona

Cuando se recupera una zona de disponibilidad, la plataforma pretende restaurar la capacidad en la zona recuperada y reequilibrar el tráfico entre zonas. Este comportamiento se realiza en la medida de lo posible y no está garantizado. No es necesario realizar ninguna acción para iniciar la recuperación de zona.

Prueba de fallos de zona

No se puede iniciar ni simular un error de zona de disponibilidad para un Storage Mover. Dado que Storage Mover no garantiza la resiliencia de la zona, no se debe dar por sentado que un Storage Mover sobrevivirá a un error de zona. Si el proceso de migración necesita soportar la pérdida de una zona, valide por su cuenta la resiliencia integral de ese proceso y revise Resilience to region-wide failures para conocer enfoques que le permitan controlar la conmutación por error.

Resistencia a errores en toda la región

Azure Storage Mover es un servicio de una sola región. Al implementar un recurso de Azure Storage Mover, seleccione una región para almacenar los metadatos de configuración del recurso. Si la región del mover de almacenamiento experimenta una interrupción, es posible que las operaciones de administración que realice el agente y que dependen de Azure no se completen. Además, es posible que se produzca un error en las migraciones de datos activas a cuentas de almacenamiento ubicadas en la región afectada.

Si su Storage Mover se encuentra en una región de Azure con un par, sus metadatos de configuración se replican en la región de Azure emparejada con fines de recuperación ante desastres, y Microsoft puede activar la conmutación por error a la región emparejada durante un desastre que afecte a su región principal.

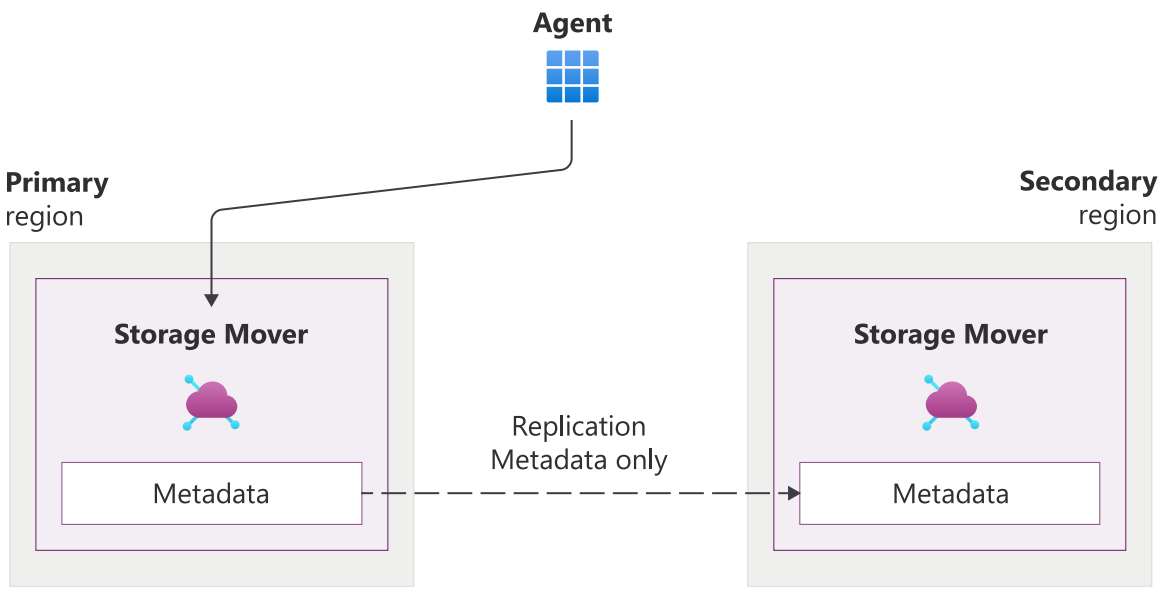
Si su Storage Mover se encuentra en una *región no emparejada*, Microsoft no replica los metadatos de configuración y no existe una conmutación por error integrada a otra región. Sin

embargo, puede implementar recursos independientes en varias regiones. En este escenario, es responsabilidad suya administrar la replicación, la distribución del tráfico y la conmutación por error. Si utiliza una región no emparejada o si la replicación de metadatos integrada no satisface sus necesidades, puede crear una estrategia de conmutación por error multirregional personalizada.

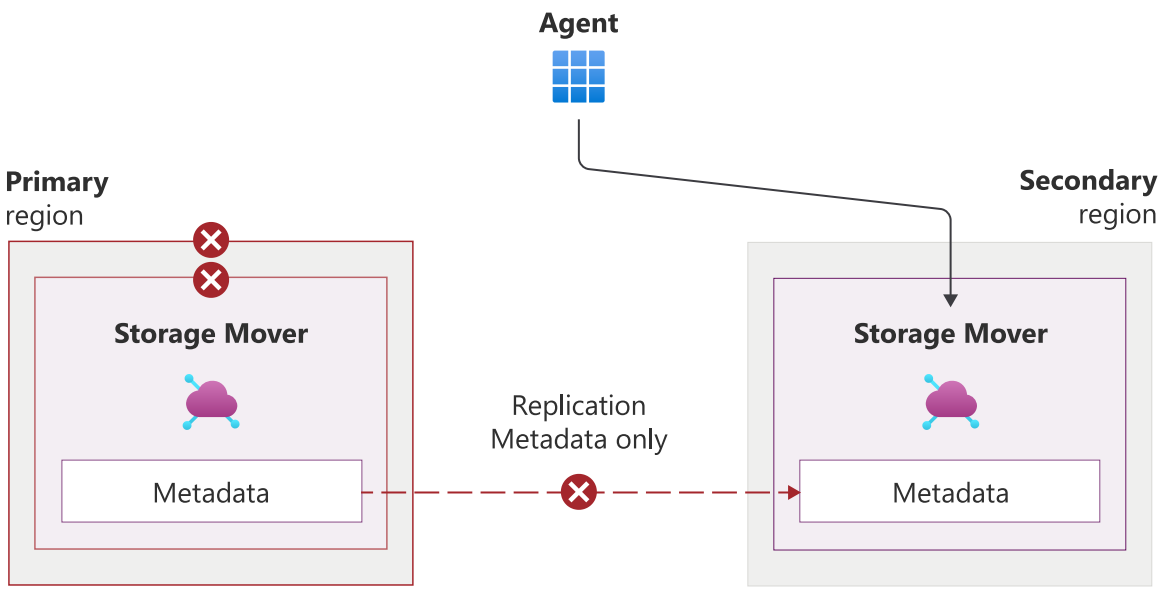
> **Nota:**
>
> Usted es responsable de la recuperación ante desastres de sus orígenes de datos (incluidos los de Azure y los locales), los destinos y los agentes.

Conmutación por error administrada por Microsoft en una región emparejada

Si el recurso de Storage Mover está en una región que se empareja con otra región, Microsoft replica los metadatos de configuración del mover de almacenamiento en la región emparejada.



En caso de interrupción del servicio en una región, Microsoft podría realizar una conmutación por error a la región emparejada utilizando los metadatos de configuración replicados. Este proceso es una opción predeterminada y no requiere intervención de usted.



La conmutación por error de los recursos de Storage Mover puede producirse en un momento diferente a cualquier conmutación por error de otros servicios de Azure.

> **Importante**
>
> Es poco probable que Microsoft inicie la conmutación por error salvo tras un retraso significativo, y esta se lleva a cabo según el principio de mejor esfuerzo. Si necesita cumplir plazos específicos para la recuperación de Storage Mover, o si el comportamiento predeterminado de replicación y conmutación por error no se ajusta a sus necesidades, utilice soluciones multirregionales personalizadas para la resiliencia con el fin de planificar e iniciar su propia conmutación por error.

La replicación entre regiones solo se aplica a los metadatos de configuración. No se aplica a los datos de origen ni a la cuenta de almacenamiento de destino, que tiene sus propias opciones de confiabilidad y replicación. Para obtener más información, consulte Confiabilidad en Azure Blob Storage y Confiabilidad en Azure Files.

Requirements

Compatibilidad con regiones: la replicación entre regiones administrada por Microsoft solo está disponible para los recursos de Storage Mover que implemente en una región que tenga una región emparejada. Para los recursos ubicados en regiones no emparejadas, no se proporciona replicación ni conmutación por error entre regiones. Para lograr resistencia entre regiones en regiones no emparejadas, use una solución personalizada de varias regiones.

Costo

Storage Mover no cobra por la replicación entre regiones administrada por Microsoft de la configuración del mover de almacenamiento. Sin embargo, puede haber un pequeño coste por la replicación entre regiones. Para obtener más información, consulte Precios del ancho de banda.

Configuración de la compatibilidad con varias regiones

La replicación entre regiones administrada por Microsoft se habilita automáticamente para los recursos de Storage Mover en regiones emparejadas. No configuras ni optas por este comportamiento.

Comportamiento cuando todas las regiones están en buen estado

En esta sección se describe qué cabe esperar cuando un migrador de almacenamiento está configurado para la replicación entre regiones y la conmutación por error, mientras la región primaria está operativa.

- **Operación entre regiones:** El recurso de Storage Mover en la región primaria atiende todas las solicitudes. La región emparejada solo se usa en caso de una conmutación por error iniciada por Microsoft.
- **Replicación de datos entre regiones:** La región primaria replica la configuración de forma asincrónica en la región emparejada. Dado que la replicación es asincrónica, es posible que los cambios recientes en la configuración no se reflejen en la región emparejada en el momento de un error.

Comportamiento durante una falla de región

En esta sección se describe qué cabe esperar cuando un servicio de transferencia de almacenamiento está configurado para la replicación entre regiones y la conmutación por error, y se produce una interrupción en la región primaria.

- **Detección y respuesta:** Microsoft detecta los errores en las regiones y decide si iniciar una conmutación por error. Es poco probable que Microsoft inicie la conmutación por error salvo tras un retraso significativo, y dicha conmutación se lleva a cabo según el principio de mejor esfuerzo.
- **Notificación**: Microsoft no le notifica automáticamente cuando una región está inactiva. Sin embargo:

- Puede usar Azure Resource Health para supervisar el estado de un recurso individual y puede configurar alertas de Resource Health para notificarle problemas.
- Puede usar Azure Service Health para comprender el estado general del servicio, incluidos los errores de región, y puede configurar alertas de Service Health para notificarle los problemas.

- **Solicitudes activas:** Las solicitudes de administración activas se descartan y deben volver a intentarse una vez finalizada la conmutación por error. Las tareas de migración de datos activas que se estén ejecutando en los agentes pueden generar un error si dependen de la región que está sufriendo la interrupción del servicio.
- **Pérdida de datos esperada:** Dado que la replicación entre regiones es asincrónica, es posible que se pierdan los cambios de metadatos de configuración que no se replican en la región emparejada en el momento de la interrupción.
- **Tiempo de inactividad esperado:** Una conmutación por error de región puede tardar hasta 24 horas en completarse. Durante este tiempo, el Storage Mover no estará disponible.
- **Redistribución:** Una vez completada la conmutación por error, Storage Mover comienza a ejecutar tareas desde la región emparejada.

  Sin embargo, es necesario volver a registrar los agentes en el Storage Mover de la región emparejada.

Recuperación de regiones

Cuando la región principal original se recupere, Microsoft coordinará la conmutación de retorno. Debes volver a registrar los agentes en el Storage Mover de la región principal.

Prueba de fallos de región

La plataforma Azure Storage Mover administra la replicación entre regiones, la conmutación por error y la recuperación de regiones. Dado que Microsoft administra íntegramente esta característica, no puedes iniciar ni probar una conmutación por error de región.

Soluciones personalizadas de varias regiones para la resistencia

Si necesita controlar cuándo se produce la conmutación por error, o si se encuentra en una región no emparejada pero aun así necesita que su recurso de Storage Mover sea resistente a

caídas regionales, implemente recursos de Storage Mover independientes en varias regiones de Azure. Usted es responsable de todos los aspectos de este enfoque, entre los que se incluyen:

- Crear y mantener proyectos, puntos de conexión, agentes y definiciones de trabajo equivalentes en cada región.
- Detección de errores en las regiones y decisión sobre cuándo realizar la conmutación por error.
- Redirigir agentes y trabajos de migración a la región secundaria.
- Sincronización del estado de los trabajos y del historial de ejecuciones entre regiones.

Una solución multirregional personalizada funciona tanto para regiones emparejadas como para regiones no emparejadas, y ofrece un control total sobre el proceso de conmutación por error.

Para obtener más información, consulte Recuperación ante desastres iniciada por el cliente para Azure Storage Mover.

Copias de seguridad y restauración

Azure Storage Mover es un servicio de orquestación de movimiento de datos y migración. No almacena los datos migrados. El servicio solo almacena metadatos de configuración, como proyectos, puntos de conexión, definiciones de trabajo y historial de ejecución de trabajos. No hay datos de migración para realizar copias de seguridad.

Para proteger tu configuración de Storage Mover, define tus recursos mediante infraestructura como código, como los archivos Bicep, y almacena esas definiciones en el control de código fuente. Si necesita volver a crear un recurso, puede volver a implementarlo desde la configuración almacenada.

Para la mayoría de las soluciones, no debe confiar exclusivamente en copias de seguridad. En su lugar, utilice las otras capacidades descritas en esta guía para apoyar los requisitos de resiliencia. Sin embargo, las copias de seguridad protegen contra algunos riesgos que otros enfoques no. Para más información, consulte ¿Qué son la redundancia, la replicación y la copia de seguridad?.

Resistencia al mantenimiento del servicio

Microsoft aplica periódicamente actualizaciones de servicio y realiza otro mantenimiento. La plataforma Azure controla estas actividades automáticamente, lo que garantiza que el mantenimiento sea transparente y sin problemas. No se prevé ningún tiempo de inactividad durante las operaciones de mantenimiento, a menos que se te haya informado a través de el mantenimiento planificado de Azure Service Health.

Los agentes de Storage Mover se actualizan automáticamente.

Acuerdo de nivel de servicio

Storage Mover es un servicio de migración y no ofrece un contrato de nivel de servicio (SLA) de disponibilidad. Sin embargo, la documentación de Storage Mover describe los objetivos de escala y rendimiento esperados. Estos objetivos se basan en migraciones simuladas y no constituyen una garantía ni un compromiso.

Contenido relacionado

- Confiabilidad en Azure
- planeación de la implementación de Azure Storage Mover
- Redundancia de Azure Storage

ⓘ**Nota:** El autor ha creado este artículo con ayuda de inteligencia artificial. Más información

---

Last updated on 28/07/2026

Confiabilidad en Azure Table Storage

Azure Table Storage es un servicio que almacena datos NoSQL estructurados en la nube. Proporciona un almacén sin esquema donde se accede a cada entidad a través de una clave y contiene un conjunto de atributos. Una sola tabla puede contener entidades que tienen diferentes conjuntos de propiedades y las propiedades pueden constar de varios tipos de datos.

Cuando se usa Azure, la confiabilidad es una responsabilidad compartida. Microsoft proporciona una variedad de funcionalidades para admitir resistencia y recuperación. Es responsable de comprender cómo funcionan esas funcionalidades dentro de todos los servicios que usa y de seleccionar las funcionalidades que necesita para cumplir los objetivos empresariales y los objetivos de tiempo de actividad.

En este artículo se describe cómo hacer que Table Storage sea resistente a una variedad de posibles interrupciones y problemas, incluidos errores transitorios, interrupciones de zona de disponibilidad y interrupciones de regiones. También se describe cómo puede usar copias de seguridad para recuperarse de otros tipos de problemas y se resalta cierta información clave sobre el acuerdo de nivel de servicio (SLA) de Table Storage.

> **Nota:**
>
> Table Storage forma parte de la plataforma Azure Storage. Algunas de las funcionalidades de Table Storage son comunes en muchos servicios de Azure Storage. En este artículo, usamos *Azure Storage* o *Storage* para hacer referencia a estas funcionalidades comunes.

Recomendaciones de implementación de producción para la confiabilidad

Para entornos de producción, realice las siguientes acciones:

- Habilite el almacenamiento con redundancia de zona (ZRS) para las cuentas de almacenamiento que contienen recursos de Table Storage. ZRS proporciona una mayor disponibilidad mediante la replicación de los datos de forma sincrónica en varias zonas de disponibilidad de la región primaria. Esta replicación protege frente a errores de zona de disponibilidad.

- Si necesita resistencia a interrupciones de regiones y la región primaria de la cuenta de almacenamiento está emparejada, considere la posibilidad de habilitar el almacenamiento con redundancia geográfica (GRS) para replicar datos de forma asincrónica en la región emparejada. En las regiones admitidas, es posible combinar la redundancia geográfica

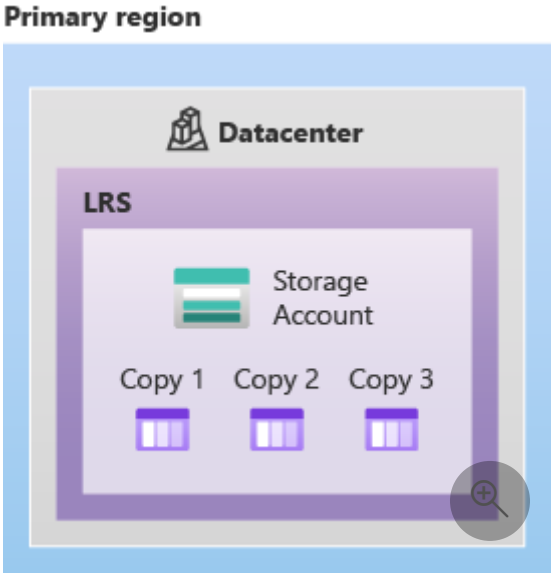
con redundancia de zona mediante el almacenamiento con redundancia de zona geográfica (GZRS).

- Para cargas de trabajo de producción a gran escala o si tiene requisitos de resistencia elevados, considere la posibilidad de usar Azure Cosmos DB para Table. Azure Cosmos DB para Table es compatible con las aplicaciones escritas para Table Storage. Admite operaciones de lectura y escritura de baja latencia a gran escala y proporciona una distribución global sólida en varias regiones con modelos de coherencia flexibles. También proporciona copias de seguridad integradas y otras funcionalidades que mejoran la resistencia y el rendimiento de la aplicación.

Introducción a la arquitectura de confiabilidad

Table Storage funciona como una base de datos NoSQL distribuida dentro de la infraestructura de la plataforma de Azure Storage. El servicio proporciona redundancia a través de varias copias de los datos de la tabla y el modelo de redundancia específico depende de la configuración de la cuenta de almacenamiento.

El almacenamiento con redundancia local (LRS) replica los datos de las cuentas de almacenamiento en una o varias zonas de disponibilidad de Azure ubicadas en la región primaria de su elección. Aunque no hay ninguna opción para elegir la zona de disponibilidad preferida, Azure puede mover o expandir cuentas LRS entre zonas para mejorar el equilibrio de carga. No hay ninguna garantía de que los datos se distribuirán entre zonas. Para obtener más información sobre las zonas de disponibilidad, consulte ¿Qué son las zonas de disponibilidad?.



El almacenamiento con redundancia de zona (ZRS), el almacenamiento con redundancia geográfica (GRS) y el almacenamiento con redundancia de zona geográfica (GZRS) proporcionan protecciones adicionales. En este artículo se describen estas opciones al detalle.

Resistencia a errores transitorios

Los errores transitorios son errores breves e intermitentes en los componentes. Se producen con frecuencia en un entorno distribuido como la nube y son una parte normal de las operaciones. Los errores transitorios se corrigen después de un breve período de tiempo. Es importante que las aplicaciones puedan controlar errores transitorios, normalmente mediante el reintento de solicitudes afectadas.

Todas las aplicaciones hospedadas en la nube deben seguir las instrucciones de control de errores transitorios de Azure cuando se comunican con cualquier API, bases de datos y otros componentes hospedados en la nube. Para obtener más información, consulte Recomendaciones para controlar errores transitorios.

Las bibliotecas y los SDK de cliente de Table Storage incluyen directivas de reintento integradas que controlan automáticamente errores transitorios comunes, como tiempos de espera de red, falta de disponibilidad temporal del servicio (HTTP 503), respuestas de limitación (HTTP 429) y condiciones de sobrecarga del servidor de particiones. Cuando la aplicación experimenta estas condiciones transitorias, las bibliotecas cliente reintentarán automáticamente las operaciones mediante estrategias de retroceso exponencial.

Para administrar los errores transitorios de forma eficaz al usar Table Storage, realice las siguientes acciones:

- **Configure los tiempos de espera adecuados** en el cliente de Table Storage para equilibrar la capacidad de respuesta con resistencia a ralentizaciones temporales. Los tiempos de espera predeterminados en las bibliotecas cliente de Azure Storage suelen ser adecuados para la mayoría de los escenarios.

- **Implemente el retroceso exponencial** para las directivas de reintento, especialmente cuando la aplicación encuentre errores de tiempo de espera de operación HTTP 503 ocupado o HTTP 500. Table Storage podría limitar las solicitudes cuando las particiones individuales se vuelven activas o cuando se aproximan los límites de la cuenta de almacenamiento.

- **Diseñe la lógica de reintento compatible con particiones** en aplicaciones a gran escala. La lógica de reintento compatible con particiones es un enfoque más avanzado que considera la arquitectura con particiones en Table Storage y distribuye las operaciones entre varias particiones para reducir la probabilidad de encontrar la limitación en servidores de particiones individuales.

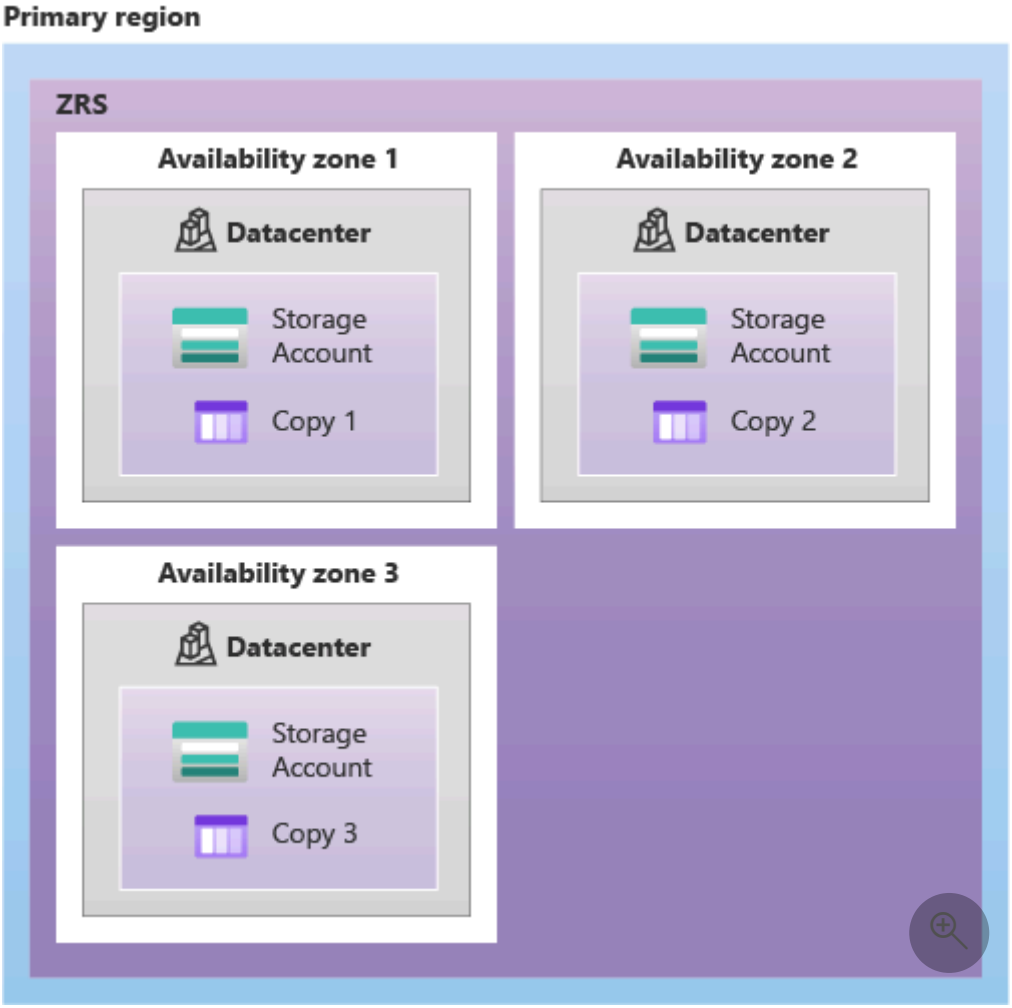
Para más información sobre la arquitectura de Table Storage y cómo diseñar aplicaciones resistentes y a gran escala, consulte Lista de comprobación de rendimiento y escalabilidad para Table Storage.

Resistencia a errores de zona de disponibilidad

Las zonas de disponibilidad son grupos físicamente independientes de centros de datos dentro de una región de Azure. Cuando una zona falla, los servicios pueden transferirse a una de las zonas restantes.

Table Storage tiene redundancia de zona al implementarlo con la configuración de ZRS. A diferencia del almacenamiento con redundancia local (LRS), ZRS garantiza que Azure replique de forma sincrónica los datos de tablas en varias zonas de disponibilidad. Esta configuración garantiza que las tablas sigan siendo accesibles incluso si una zona de disponibilidad completa deja de estar disponible. Todas las operaciones de escritura deben confirmarse en varias zonas antes de que el servicio complete la escritura, lo que proporciona garantías de coherencia sólidas.

La redundancia de zona está habilitada en el nivel de cuenta de almacenamiento y se aplica a todos los recursos de Table Storage de esa cuenta. Dado que la configuración se aplica a toda la cuenta de almacenamiento, no puede configurar entidades individuales para distintos niveles de redundancia. Cuando una zona de disponibilidad sufre una interrupción, Azure Storage redirige automáticamente las solicitudes a zonas que funcionan correctamente sin necesidad de intervención por su parte ni por parte de su aplicación.



Requisitos

- **Compatibilidad con regiones:** Puede implementar cuentas de Azure Storage con redundancia de zona en cualquier región que admita zonas de disponibilidad.
- **Tipos de cuenta de almacenamiento:** Debe usar una cuenta de almacenamiento estándar de uso general v2 para habilitar ZRS para Table Storage. Las cuentas de Premium Storage no admiten Table Storage.

Cost

Cuando se habilita el almacenamiento con redundancia de zona (ZRS), se cobra una tarifa diferente a la del almacenamiento redundante local (LRS) debido a la replicación adicional y los gastos generales de almacenamiento.

Para información detallada sobre los precios, consulte Precios de Table Storage.

Configurar soporte de zonas de disponibilidad

- **Cree una cuenta de almacenamiento con redundancia de zona y una tabla:**
    1. Crear una cuenta de almacenamiento. Asegúrese de seleccionar ZRS, GZRS o almacenamiento con redundancia geográfica con acceso de lectura (RA-GZRS) como opción de redundancia.
    2. Cree una tabla.
- **Cambiar el tipo de replicación.** Para obtener información sobre cómo cambiar una cuenta de almacenamiento existente a almacenamiento con redundancia de zona (ZRS) y sobre las opciones y requisitos de configuración, consulte Cambiar la forma en que se replica una cuenta de almacenamiento.
- **Desactivar la redundancia de zona.** Convierta las cuentas de ZRS de nuevo a una configuración no zonal, como el almacenamiento con redundancia local (LRS), utilizando el mismo proceso de cambio de configuración de redundancia.

Comportamiento cuando todas las zonas están en buen estado

En esta sección, se describe qué esperar cuando se configura una cuenta de Table Storage para la redundancia de zona y todas las zonas de disponibilidad están operativas.

- **Enrutamiento del tráfico entre zonas:** Azure Storage con almacenamiento con redundancia de zona (ZRS) distribuye automáticamente las solicitudes entre los clústeres de almacenamiento en varias zonas de disponibilidad. La distribución del tráfico es transparente para las aplicaciones y no requiere ninguna configuración del lado cliente.

- **Replicación de datos entre zonas:** todas las operaciones de escritura en ZRS se replican de forma sincrónica en todas las zonas de disponibilidad dentro de la región. Al cargar o modificar datos, la operación no se considera completa hasta que los datos se hayan replicado correctamente en todas las zonas de disponibilidad. Esta replicación sincrónica garantiza una fuerte coherencia y cero pérdida de datos durante los errores de zona.

Comportamiento durante un fallo de zona

Cuando una zona de disponibilidad deja de estar disponible, Table Storage controla automáticamente el proceso de conmutación por error respondiendo con los comportamientos siguientes:

- **Detección y respuesta:** Microsoft detecta automáticamente errores de zona e inicia procesos de recuperación. No se requiere ninguna acción del cliente para las cuentas de almacenamiento con redundancia de zona (ZRS). Si una zona deja de estar disponible, Azure lleva a cabo actualizaciones de red como el repunto del sistema de nombres de dominio (DNS).

- **Notificación**: Microsoft no le notifica automáticamente cuando una zona está inactiva. Sin embargo, puede usar Azure Resource Health para supervisar el estado de un recurso individual y puede configurar alertas de Resource Health para notificarle problemas. También puede usar Azure Service Health para comprender el estado general del servicio, incluidos los errores de zona, y puede configurar alertas de Service Health para notificarle problemas.

- **Solicitudes activas:** Es posible que las solicitudes en curso se descarten durante el proceso de recuperación y se deban reintentar. Las aplicaciones deben implementar lógica de reintento para controlar estas interrupciones temporales.

- **Pérdida de datos esperada:** No se produce ninguna pérdida de datos durante los errores de zona porque los datos se replican sincrónicamente en varias zonas antes de completar las operaciones de escritura.

- **Tiempo de inactividad previsto:** es posible que se produzca un breve periodo de inactividad, normalmente de unos segundos, durante la recuperación automática, ya que el tráfico se redirige a zonas correctas. Cuando diseñe aplicaciones para ZRS, siga las prácticas de manejo de fallos transitorios, incluyendo la implementación de directivas de reintento con retroceso exponencial.

- **Redireccionamiento de tráfico:** Si una zona deja de estar disponible, Azure realiza actualizaciones de red, como el redireccionamiento del Sistema de nombres de dominio (DNS) para que las solicitudes se dirijan a las zonas de disponibilidad restantes que funcionan correctamente. El servicio mantiene toda la funcionalidad mediante las zonas en buen estado y no requiere intervención del cliente.

Recuperación de zona

Cuando se recupera la zona de disponibilidad con errores, Azure Storage restaura automáticamente las operaciones normales en todas las zonas de disponibilidad. El servicio garantiza automáticamente la coherencia de los datos mediante la sincronización de las operaciones que se produjeron durante el período de interrupción.

Prueba de fallos de zona

El usar el almacenamiento con redundancia de zona (ZRS), Azure Storage administrará automáticamente la replicación, el enrutamiento del tráfico y las respuestas ante caídas de zonas. Dado que esta característica está totalmente administrada, no es necesario iniciar ni validar los procesos de error de zona de disponibilidad.

Resistencia a errores en toda la región

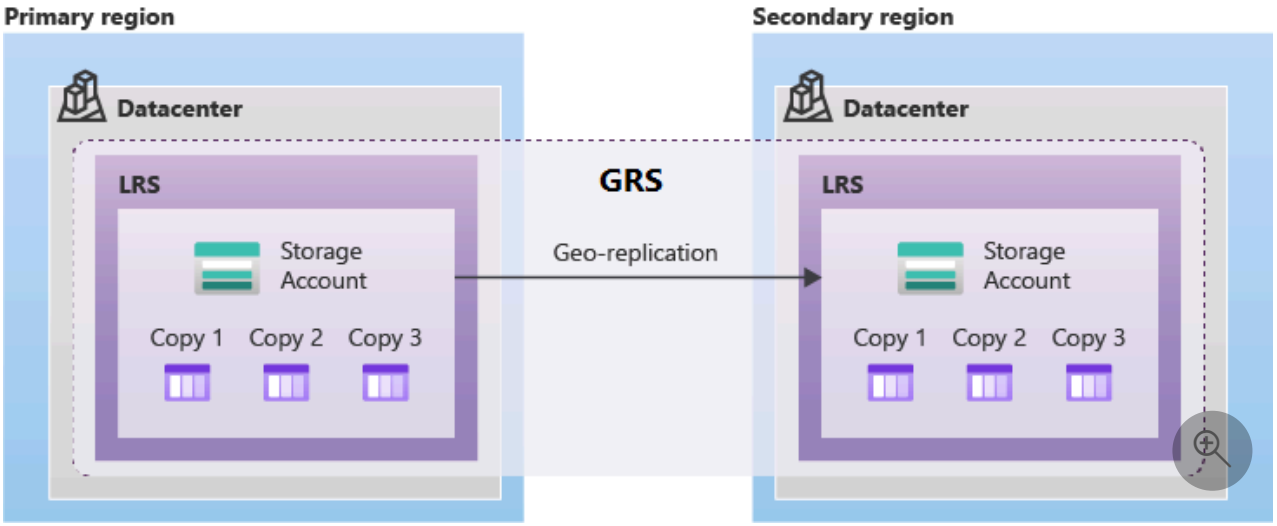
Azure Storage, que incluye Azure Blob Storage, Azure Files, Azure Table Storage y Azure Queue Storage, ofrece una amplia gama de capacidades de redundancia geográfica y conmutación por error para adaptarse a diferentes requisitos.

> **Importante**
>
> El GRS solo funciona dentro de regiones emparejadas de Azure. Si la región de la cuenta de almacenamiento no está emparejada, considere la posibilidad de usar las soluciones personalizadas de varias regiones para lograr resistencia.

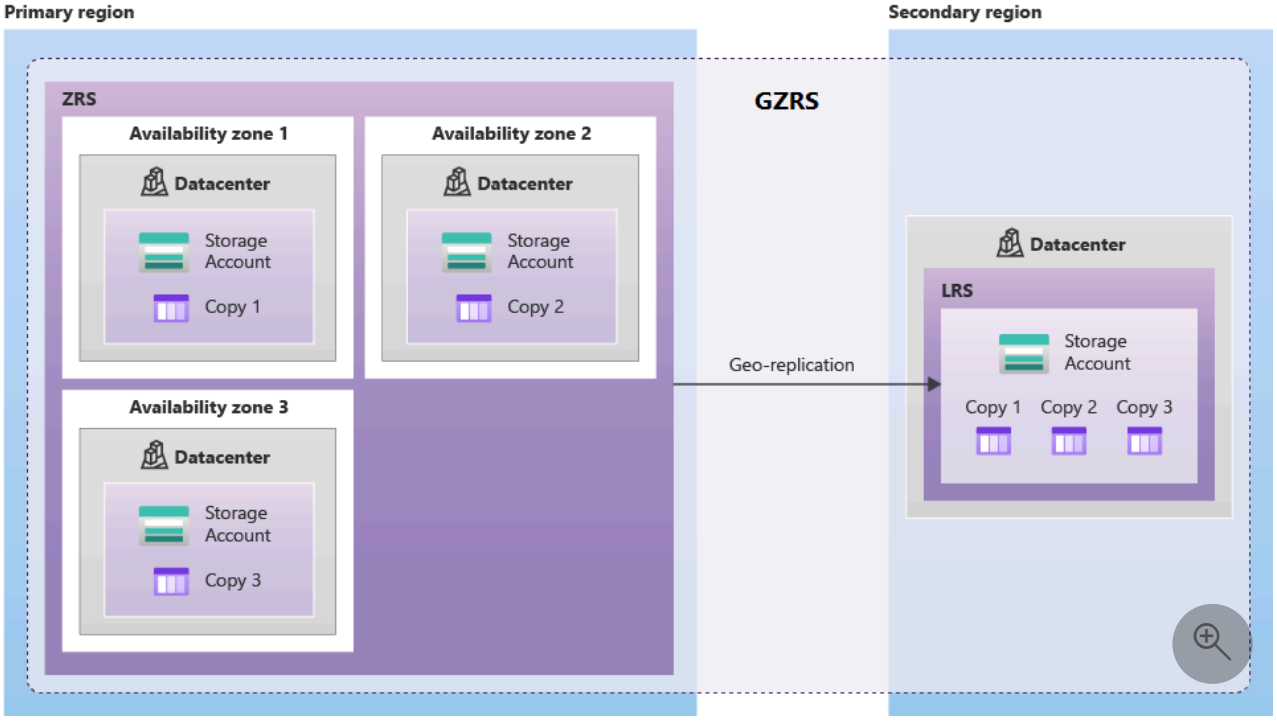
Almacenamiento con redundancia geográfica para regiones emparejadas

Azure Storage ofrece varios tipos de GRS en regiones emparejadas. Independientemente del tipo de GRS que utilice, los datos de la región secundaria siempre se replicarán mediante almacenamiento con redundancia local (LRS). Este enfoque proporciona protección contra fallos de hardware dentro de la región secundaria.

- GRS proporciona soporte para conmutaciones por error planificadas y no planificadas a la región emparejada de Azure cuando se produce una interrupción en la región principal. GRS replica de forma asincrónica los datos de la región primaria a la región emparejada.



- El almacenamiento con redundancia de zona geográfica (GZRS) replica los datos en varias zonas de disponibilidad de la región principal y en la región emparejada.



- El almacenamiento con redundancia geográfica con acceso de lectura (RA-GRS) y el almacenamiento con redundancia de zona geográfica con acceso de lectura (RA-GZRS) amplían el almacenamiento con redundancia geográfica (GRS) y el almacenamiento con redundancia de zona geográfica (GZRS), con la ventaja adicional del acceso de lectura al punto de conexión secundario. Estas opciones son idóneas y están diseñadas para obtener alta disponibilidad en aplicaciones críticas para la empresa. En el improbable caso de que el punto de conexión principal experimente una interrupción, las aplicaciones configuradas para el acceso de lectura a la región secundaria pueden seguir funcionando.

Tipos de conmutación por error

Azure Storage admite tres tipos de conmutación por error para distintos escenarios.

- **Conmutación por error no planeada administrada por el cliente:** es responsable de iniciar la recuperación si hubiera un error de almacenamiento en toda la región principal.
- **Conmutación por error planeada administrada por el cliente:** Usted es responsable de iniciar la recuperación si otra parte de su solución sufre un fallo en la región primaria y necesita trasladar toda su solución a la región secundaria. Use una conmutación por error planeada cuando el almacenamiento permanezca operativo en la región primaria, pero debe realizar una conmutación por error de toda la solución a una región secundaria, como para los simulacros de recuperación ante desastres diseñados para garantizar el cumplimiento de los requisitos de auditoría.
- **Conmutación por error administrada por Microsoft:** en circunstancias excepcionales, Microsoft podría iniciar la conmutación por error para todas las cuentas de almacenamiento con redundancia geográfica (GRS) de una región. Sin embargo, la conmutación por error administrada por Microsoft es un último recurso y se espera que solo se realice después de un período prolongado de interrupción. No se debe confiar en el failover administrado por Microsoft.

Las cuentas de GRS pueden utilizar cualquiera de estos tipos de conmutación por error. No es necesario preconfigurar una cuenta de almacenamiento para usar ninguno de los tipos de conmutación por error con antelación.

Requisitos

- **Compatibilidad con regiones:** Las configuraciones con redundancia geográfica de Azure Storage usan regiones emparejadas de Azure para la replicación de regiones secundarias. La región secundaria se determina automáticamente en función de la selección de la región principal y no se puede personalizar. Para obtener una lista completa de las regiones emparejadas de Azure, consulte Lista de regiones de Azure.

  Si la región de la cuenta de almacenamiento no está emparejada, considere la posibilidad de usar las soluciones personalizadas de varias regiones para lograr resistencia.

- **Tipos de cuenta de almacenamiento:** El almacenamiento con redundancia geográfica (GRS) y la conmutación por error y la conmutación por recuperación iniciadas por el cliente están disponibles en todas las regiones emparejadas de Azure que admiten cuentas de Azure Storage de uso general v2.

Consideraciones

Al implementar Table Storage en varias regiones, tenga en cuenta los siguientes factores importantes:

- **Latencia de replicación asincrónica:** la replicación de datos en la región secundaria es asincrónica, lo que significa que hay un retraso entre cuando los datos se escriben en la región primaria y cuando estarán disponibles en la región secundaria. Este retraso podría provocar una posible pérdida de datos si se produjese un fallo en la región principal antes de que se repliquen los datos recientes. La pérdida de datos se mide mediante el objetivo de punto de recuperación (RPO). Espere que el retraso de replicación sea inferior a 15 minutos, pero este tiempo es una estimación y no se garantiza.

  Compruebe la propiedad Última hora de sincronización para saber cuántos datos se podrían perder si su cuenta de almacenamiento sufriese una conmutación por error no planificada.

- **Acceso a la región secundaria:** con las configuraciones de almacenamiento con redundancia geográfica (GRS) y almacenamiento con redundancia de zona geográfica (GZRS), la región secundaria no será accesible a las lecturas hasta que se produzca una conmutación por error.

  Las configuraciones de almacenamiento con redundancia geográfica con acceso de lectura (RA-GRS) y almacenamiento con redundancia de zona geográfica con acceso de lectura (RA-GZRS) proporcionan acceso de lectura a la región secundaria durante las operaciones normales, pero debido a la latencia de replicación asincrónica, podrían devolver datos ligeramente obsoletos.

- **Limitaciones de las características:** algunas características de Azure Storage no son compatibles o tienen limitaciones cuando se utiliza almacenamiento con redundancia geográfica (GRS) o conmutación por error administrada por el cliente. Revise la compatibilidad de las característica antes de implementar la redundancia geográfica.

Cost

Las configuraciones de cuentas de Azure Storage multirregionales conllevan costes adicionales por la replicación entre regiones y el almacenamiento en la región secundaria. La transferencia de datos entre regiones de Azure se cobra en función de las tarifas estándar de ancho de banda entre regiones.

Para información detallada sobre los precios, consulte Precios de Table Storage.

Configuración de la compatibilidad con varias regiones

- **Cree una nueva cuenta de almacenamiento con redundancia geográfica (GRS).** Para crear una cuenta de GRS, consulte Creación de una cuenta de almacenamiento y seleccione GRS, almacenamiento con redundancia geográfica con acceso de lectura (RA-GRS), almacenamiento con redundancia de zona geográfica (GZRS) o almacenamiento con redundancia de zona geográfica con acceso de lectura (RA-GZRS) durante la creación de la cuenta.

- **Habilite la redundancia geográfica en una cuenta de almacenamiento existente.** Para convertir una cuenta de almacenamiento existente en almacenamiento con redundancia geográfica (GRS), consulte Cambio de cómo se replica una cuenta de almacenamiento.

  > ⚠ **Advertencia**
  >
  > Una vez que su cuenta se haya reconfigurado para la redundancia geográfica, es posible que transcurra un tiempo considerable antes de que los datos existentes en la nueva región principal se copien por completo a la nueva región secundaria.
  >
  > **Para evitar una pérdida importante de datos**, compruebe el valor de la **propiedad Última hora de sincronización** antes de iniciar una conmutación por error no planificada. Para evaluar la posible pérdida de datos, compare la última hora de sincronización con la última hora en que se escribieron los datos en la nueva región principal.

- **Deshabilite la redundancia geográfica.** Convierta las cuentas de GRS en configuraciones de una sola región, como el almacenamiento con redundancia local (LRS) o el almacenamiento con redundancia local (ZRS) con el mismo proceso de cambio de configuración de redundancia.

Comportamiento cuando todas las regiones están en buen estado

En esta sección se describe qué esperar cuando una cuenta de almacenamiento está configurada para redundancia geográfica y todas las regiones están operativas.

- **Enrutamiento del tráfico entre regiones:** Azure Storage utiliza un enfoque activo-pasivo en el que todas las operaciones de escritura y la mayoría de las operaciones de lectura se dirigen a la región principal.

  En las configuraciones de almacenamiento con redundancia geográfica con acceso de lectura (RA-GRS) y almacenamiento con redundancia de zona geográfica con acceso de lectura (RA-GZRS), las aplicaciones pueden leer opcionalmente desde la región secundaria accediendo al punto de conexión secundario. Este enfoque requiere una configuración explícita de la aplicación y no es automática. Además, debido al retraso de

replicación asincrónica, los datos de la región secundaria podrían estar ligeramente obsoletos.

- **Replicación de datos entre regiones:** las operaciones de escritura se confirman primero en la región principal utilizando los siguientes tipos de redundancia configurados:
  - Almacenamiento con redundancia local (LRS) para almacenamiento con redundancia geográfica (GRS) y RA-GRS
  - Almacenamiento con redundancia de zona (ZRS) para el almacenamiento con redundancia de zona geográfica (GZRS) y RA-GZRS

  Tras completarse con éxito en la región principal, los datos se replican de forma asíncrona en la región secundaria, donde se almacenan mediante LRS.

  La naturaleza asíncrona de la replicación entre regiones significa que suele haber un retraso entre el momento en que los datos se escriben en la región principal y el momento en que están disponibles en la región secundaria. Supervise el tiempo de replicación utilizando la propiedad Última hora de sincronización.

Comportamiento durante una falla de región

En esta sección se describe qué esperar cuando una cuenta de almacenamiento está configurada para la redundancia geográfica y hay una interrupción en la región primaria.

- **Conmutación por error administrada por el cliente (no planificada):** utilice una conmutación por error no planificada cuando el almacenamiento en la región principal no esté disponible.

  - **Detección y respuesta:** En el caso poco probable de que su cuenta de almacenamiento no esté disponible en la región primaria, puede considerar iniciar un failover no planeado administrado por el cliente. Para tomar esta decisión, tenga en cuenta los siguientes factores:

    - Si Azure Resource Health muestra problemas al acceder a la cuenta de almacenamiento en su región principal

    - Si Microsoft le recomienda realizar una conmutación por error a otra región

    > ⚠ **Advertencia**
    >
    > Una conmutación por error no planeada puede **provocar la pérdida de datos**. Antes de iniciar una conmutación por error administrada por el cliente, decida si la restauración del servicio justifica el riesgo de pérdida de datos.

- **Notificación**: Microsoft no le notifica automáticamente cuando una región está inactiva. Sin embargo:
  - Puede usar Azure Resource Health para supervisar el estado de un recurso individual y puede configurar alertas de Resource Health para notificarle problemas.
  - Puede usar Azure Service Health para comprender el estado general del servicio, incluidos los errores de región, y puede configurar alertas de Service Health para notificarle problemas.
- **Solicitudes activas:** Durante el proceso de conmutación por error, los puntos de conexión de la cuenta de almacenamiento principal y secundaria dejan de estar disponibles temporalmente para lecturas y escrituras. Las solicitudes activas podrían fallar, y las aplicaciones cliente deben intentar nuevamente después de que se complete la conmutación por error.
- **Pérdida de datos prevista:** la pérdida de datos es habitual durante una conmutación por error no planificada debido al retraso en la replicación asíncrona, lo que significa que es posible que las escrituras recientes no se repliquen. Consulte la propiedad Última hora de sincronización para saber cuántos datos se podrían perder durante una conmutación por error no planificada. La pérdida de datos esperada se conoce a menudo como objetivo de punto de recuperación (RPO). Normalmente, puede esperar que el RPO sea inferior a 15 minutos, pero ese tiempo no está garantizado.
- **Tiempo de inactividad esperado:** La cantidad de tiempo de inactividad esperado se conoce a menudo como objetivo de tiempo de recuperación (RTO). La conmutación por error administrada por el cliente se completa normalmente en un plazo de hasta 60 minutos, dependiendo del tamaño y la complejidad de la cuenta.
- **Reenrutamiento del tráfico:** A medida que se completa la conmutación por error, Azure actualiza automáticamente los puntos de conexión de la cuenta de almacenamiento para que no sea necesario volver a configurar las aplicaciones. Si su aplicación mantiene entradas del Sistema de nombres de dominio (DNS) almacenadas en caché, puede que sea necesario borrar la caché para asegurarse de que la aplicación envíe el tráfico a la nueva región principal.
- **Configuración posterior a la conmutación por error:** una vez completada una conmutación por error no planificada, su cuenta de almacenamiento en la región de destino utilizará el nivel de almacenamiento con redundancia local (LRS). Si necesitase volver a replicar geográficamente los datos, deberá volver a habilitar el almacenamiento con redundancia geográfica (GRS) y esperar a que los datos se repliquen en la nueva región secundaria.

Para obtener más información sobre cómo iniciar una conmutación por error administrada por el cliente, consulte Cómo funciona la conmutación por error administrada por el cliente (no planificada) e Iniciar una conmutación por error de una cuenta de almacenamiento.

- **Conmutación por error administrada por el cliente (planificada):** utilice una conmutación por error planificada cuando el almacenamiento siga operativo en la región principal, pero necesite conmutar por error toda su solución a una región secundaria por otro motivo. Por ejemplo, otro servicio de Azure podría estar experimentando un problema y debe cambiar al uso de una región secundaria para toda la solución. O bien, puede usar una conmutación por error planeada para realizar un simulacro de recuperación ante desastres (DR) con fines de cumplimiento y auditoría.
  - **Detección y respuesta:** Usted es responsable de decidir conmutar por error. Normalmente, debe tomar esta decisión si necesita realizar una conmutación entre regiones, aunque su cuenta de almacenamiento esté en buen estado. Por ejemplo, puede desencadenar una conmutación por error cuando se produce una interrupción importante en otro componente de la aplicación que no se puede recuperar en la región principal.
    - **Notificación**: Microsoft no le notifica automáticamente cuando una región está inactiva. Sin embargo:
      - Puede usar Azure Resource Health para supervisar el estado de un recurso individual y puede configurar alertas de Resource Health para notificarle problemas.
      - Puede usar Azure Service Health para comprender el estado general del servicio, incluidos los errores de región, y puede configurar alertas de Service Health para notificarle problemas.
  - **Solicitudes activas:** Durante el proceso de conmutación por error, los puntos de conexión de la cuenta de almacenamiento principal y secundaria dejan de estar disponibles temporalmente para lecturas y escrituras. Las solicitudes activas podrían fallar, y las aplicaciones cliente deben intentar nuevamente después de que se complete la conmutación por error.
  - **Pérdida de datos esperada:** No se espera ninguna pérdida de datos porque el proceso de conmutación por error se completa solo después de sincronizar todos los datos, lo que da como resultado un RPO de cero.
  - **Tiempo de inactividad esperado:** La conmutación por error normalmente se completa en un plazo de 60 minutos, lo que significa que el tiempo de recuperación objetivo (RTO) esperado es de 60 minutos, en función del tamaño y la complejidad de la cuenta.

Durante el proceso de conmutación por error, los puntos de conexión de la cuenta de almacenamiento principal y secundaria dejan de estar disponibles temporalmente para lecturas y escrituras.

- **Reenrutamiento del tráfico:** A medida que se completa la conmutación por error, Azure actualiza automáticamente los puntos de conexión de la cuenta de almacenamiento para que no sea necesario volver a configurar las aplicaciones. Si su aplicación mantiene entradas DNS almacenadas en caché, puede que sea necesario borrar la caché para asegurarse de que la aplicación envíe tráfico a la nueva región principal.
- **Configuración posterior a la conmutación por error:** Una vez completada una conmutación por error planeada, la cuenta de almacenamiento de la región de destino sigue siendo replicada geográficamente y permanece en el nivel GRS.

Para obtener más información sobre cómo iniciar una conmutación por error administrada por el cliente, consulte Cómo funciona la conmutación por error administrada por el cliente (planificada) e Iniciar una conmutación por error de una cuenta de almacenamiento.

- **Conmutación por error administrada por Microsoft:** En el caso excepcional de un desastre importante en el que Microsoft determina que la región primaria es irrecuperable permanentemente, es posible que se inicie una conmutación automática por error a la región secundaria. Microsoft controla todo el proceso y no se requiere ninguna acción del cliente. El tiempo que transcurre antes de que ocurra un failover depende de la gravedad del desastre y del tiempo necesario para evaluar la situación.
  - **Notificación**: Microsoft no le notifica automáticamente cuando una región está inactiva. Sin embargo:
    - Puede usar Azure Resource Health para supervisar el estado de un recurso individual y puede configurar alertas de Resource Health para notificarle problemas.
    - Puede usar Azure Service Health para comprender el estado general del servicio, incluidos los errores de región, y puede configurar alertas de Service Health para notificarle problemas.

> **ⓘ Importante**
>
> Utilice las opciones de conmutación por error administradas por el cliente para desarrollar, probar e implementar tus planes de recuperación ante desastres. **No confíe en la conmutación por error administrada por Microsoft**, que solo se puede utilizar en circunstancias extremas. Es probable que se haya iniciado una

conmutación por error administrada por Microsoft para toda una región. No se puede iniciar para cuentas de almacenamiento individuales, suscripciones o clientes. La conmutación por error puede producirse en momentos diferentes para distintos servicios de Azure. Recomendamos que utilice la conmutación por error administrada por el cliente.

Recuperación de regiones

El proceso de conmutación por recuperación difiere significativamente entre los escenarios de conmutación por error administrados por Microsoft y los administrados por el cliente.

- **Conmutación por error administrada por el cliente (no planeada):** después de una conmutación por error no planeada, la cuenta de almacenamiento se configura con almacenamiento con redundancia local (LRS). Para realizar una conmutación por error, es necesario restablecer la relación de almacenamiento con redundancia geográfica (GRS) y esperar a que se repliquen los datos.

- **Conmutación por error administrada por el cliente (planificada):** después de una conmutación por error planificada, la cuenta de almacenamiento seguirá estando replicada geográficamente. Inicie otra conmutación por error administrada por el cliente para volver a la región principal original. Se aplican las mismas consideraciones de conmutación por error.

- **Conmutación por error administrada por Microsoft:** si Microsoft iniciase una conmutación por error, es probable que se haya producido un desastre importante en la región principal y que esta no se pueda recuperar. Cualquier calendario o plan de recuperación dependerá del alcance del desastre regional y de los esfuerzos de recuperación. Debe supervisar las comunicaciones de Azure Service Health para más información.

Prueba de fallos de región

Simule fallos regionales para probar sus procedimientos de recuperación ante desastres.

- **Pruebas de conmutación por error planificadas:** en el caso de las cuentas de almacenamiento con redundancia geográfica (GRS), es posible realizar operaciones de conmutación por error planificadas durante las ventanas de mantenimiento para probar el proceso completo de conmutación por error y conmutación por recuperación. La conmutación por error planificada no implica pérdida de datos, pero sí conlleva tiempo de inactividad tanto durante la conmutación por error como durante la conmutación por recuperación.

- **Pruebas de puntos de conexión secundarios:** para el almacenamiento con redundancia geográfica con acceso de lectura (RA-GRS) y las configuraciones de almacenamiento con redundancia de zona geográfica con acceso de lectura (RA-GZRS), pruebe periódicamente las operaciones de lectura en el punto de conexión secundario para asegurarse de que la aplicación pueda leer correctamente los datos de la región secundaria.

Soluciones personalizadas de varias regiones para la resistencia

Las funcionalidades de conmutación por error entre regiones de Azure Storage podrían no ser adecuadas por los siguientes motivos:

- La cuenta de almacenamiento está en una región no emparejada.
- Los objetivos de tiempo de actividad empresarial no se satisfacen por el tiempo de recuperación o la pérdida de datos que proporciona las opciones de conmutación por error integradas.
- Debe conmutar por error a una región que no esté emparejada con la región principal.
- Necesita una configuración activa o activa entre regiones.

En esta sección se proporciona información general de alto nivel de algunos enfoques que se deben tener en cuenta. Una introducción completa de las topologías de implementación de varias regiones para Azure Storage está fuera del ámbito de este artículo.

> ⓘ **Nota:**
>
> Para que las aplicaciones creadas usen Table Storage, considere la posibilidad de usar Azure Cosmos DB for Table. Azure Cosmos DB for Table admite requisitos avanzados de varias regiones, incluida la compatibilidad con regiones no emparejadas. También está diseñado para la compatibilidad con las aplicaciones creadas para Table Storage.

Implemente Azure Storage en varias regiones utilizando cuentas de almacenamiento independientes en cada región. Este enfoque ofrece flexibilidad en la selección de regiones, la posibilidad de utilizar regiones no emparejadas y un control más granular sobre el momento de la replicación y la coherencia de los datos. Al implementar varias cuentas de almacenamiento entre regiones, es necesario configurar la replicación de datos entre regiones, implementar directivas de equilibrio de carga y conmutación por error y garantizar la coherencia de los datos entre regiones.

Para Table Storage, un enfoque de varias cuentas requiere que administre la distribución de datos, controle la sincronización entre tablas entre regiones, incluida la resolución de conflictos, e implemente lógica de conmutación por error personalizada.

Copias de seguridad y restauración

Table Storage no proporciona funcionalidades de copia de seguridad tradicionales, como la restauración a un momento dado (PITR). Sin embargo, puede implementar estrategias de copia de seguridad personalizadas para los datos de tabla.

Si necesita funcionalidades de copia de seguridad integradas, considere la posibilidad de pasar a Azure Cosmos DB for Table, lo que proporciona compatibilidad con copias de seguridad periódicas y continuas. Para más información, consulte Copia de seguridad en línea y restauración de datos a petición en Azure Cosmos DB.

Para escenarios que requieren copia de seguridad de datos de Table Storage, tenga en cuenta los siguientes enfoques:

- **Exportación mediante Azure Data Factory.** Use el conector de Azure Data Factory para Table Storage para exportar las entidades a otra ubicación. Por ejemplo, puede realizar una copia de seguridad de cada entidad en un archivo JSON almacenado en Azure Blob Storage.

- **Realice una copia de seguridad de nivel de aplicación.** Implemente lógica de copia de seguridad personalizada dentro de las aplicaciones para exportar entidades de tabla críticas a otros servicios de almacenamiento, como Azure SQL Database o Azure Cosmos DB para funcionalidades de copia de seguridad y restauración más sólidas.

Al diseñar estrategias de copia de seguridad para Table Storage, tenga en cuenta la naturaleza particionada de los datos y asegúrese de que los procesos de copia de seguridad pueden controlar tablas de gran tamaño de forma eficaz mediante el procesamiento de varias particiones en paralelo.

Para la mayoría de las soluciones, no debe confiar exclusivamente en copias de seguridad. En su lugar, utilice las otras capacidades descritas en esta guía para apoyar los requisitos de resiliencia. Sin embargo, las copias de seguridad protegen contra algunos riesgos que otros enfoques no. Para más información, consulte ¿Qué son la redundancia, la replicación y la copia de seguridad?.

Acuerdo de nivel de servicio

El acuerdo de nivel de servicio (SLA) para Azure Storage describe la disponibilidad esperada del servicio y las condiciones que se deben cumplir para lograr esa expectativa de disponibilidad. El contrato de nivel de servicio de disponibilidad al que tiene derecho dependerá del nivel de almacenamiento y del tipo de replicación que utilice. Para obtener más información, consulte los SLAs for Online Services.

Contenido relacionado

- ¿Qué es Table Storage?
- Diseño de tablas escalables y eficaces
- Redundancia de Azure Storage
- Planeamiento y conmutación por error de recuperación ante desastres de Azure Storage
- Lista de comprobación de rendimiento y de escalabilidad para Table Storage
- Fiabilidad de Azure
- Recomendaciones para controlar errores transitorios

ⓘ **Nota:** El autor ha creado este artículo con ayuda de inteligencia artificial. Más información

---

Last updated on 22/01/2026

Seguridad en Azure AI Search

Azure AI Search es una infraestructura de búsqueda escalable que indexa contenido heterogéneo y permite la recuperación a través de API, aplicaciones y agentes de IA. Es adecuado para escenarios de búsqueda empresarial y experiencias de clientes con tecnología de inteligencia artificial que requieren generación dinámica de contenido a través de modelos de finalización de chat. Como servicio de Azure, AI Search proporciona una variedad de funcionalidades para satisfacer sus requisitos de confiabilidad.

Cuando se usa Azure, la confiabilidad es una responsabilidad compartida. Microsoft proporciona una variedad de funcionalidades para admitir resistencia y recuperación. Es responsable de comprender cómo funcionan esas funcionalidades dentro de todos los servicios que usa y de seleccionar las funcionalidades que necesita para cumplir los objetivos empresariales y los objetivos de tiempo de actividad.

En este artículo se describe cómo hacer que Azure AI Search sea resistente a una variedad de posibles interrupciones y problemas, incluidos errores transitorios, interrupciones de zona de disponibilidad, interrupciones de regiones y mantenimiento del servicio. También se describe cómo puede usar copias de seguridad para recuperarse de otros tipos de problemas y se resalta cierta información clave sobre el acuerdo de nivel de servicio (SLA) de Azure AI Search.

Recomendaciones de implementación de producción para la confiabilidad

En el caso de las cargas de trabajo de producción, se recomienda que usted:

- ✓ Use un nivel facturable que tenga al menos dos réplicas. Esta configuración hace que su servicio de búsqueda sea más resistente a errores transitorios y operaciones de mantenimiento. También cumple el acuerdo de nivel de servicio (SLA) para la búsqueda de IA. El Acuerdo de Nivel de Servicio requiere dos réplicas para cargas de trabajo de solo lectura y tres o más réplicas para cargas de trabajo de lectura y escritura.
- ✓ No use el nivel Gratis para su uso en producción. Ai Search no proporciona un Acuerdo de Nivel de Servicio para el nivel Gratis, que está limitado a una réplica.

Introducción a la arquitectura de confiabilidad

Cuando se usa AI Search, se crea un *servicio de búsqueda*. Cada servicio de búsqueda admite muchos *índices de búsqueda* que almacenan el contenido que se puede buscar.

La búsqueda de IA no está diseñada como un almacén de datos principal. En su lugar, se usan *indexadores* para conectar el servicio de búsqueda a orígenes de datos externos. Un indexador

rastrea los datos de origen, invoca *aptitudes* que realizan el procesamiento y el enriquecimiento, y rellena el índice con las salidas de aptitud.

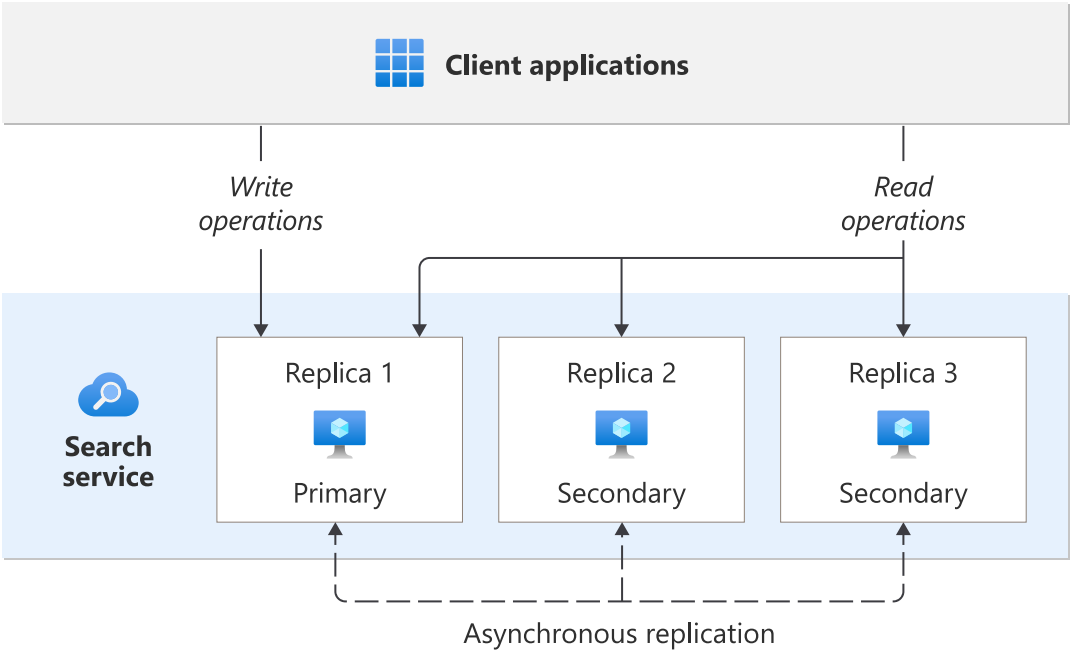
También configurará el número de *réplicas* para el servicio. En AI Search, una réplica es una copia del motor de búsqueda de tu servicio. Puede pensar en una réplica como si representara una sola máquina virtual (VM). Cada servicio Search puede tener entre 1 y 12 réplicas.

La adición de varias réplicas permite que AI Search pueda:

- Aumente la disponibilidad del servicio Search.
- Realice el mantenimiento en una réplica mientras las consultas continúan ejecutándose en otras réplicas.
- Controle cargas de trabajo de consulta e indexación más altas.
- Mejore la resistencia intentando aprovisionar réplicas en diferentes zonas de disponibilidad, si la región las admite.

Ai Search asigna automáticamente una réplica para que sea la *réplica principal*. Todas las operaciones de escritura se realizan en esa réplica. Las otras réplicas se usan para las operaciones de lectura.

En el diagrama siguiente se muestra cómo un servicio Search con tres réplicas se puede distribuir entre tres zonas de disponibilidad:



También puede configurar el número de *particiones*, que representan el almacenamiento que usan los índices de búsqueda.

Es importante comprender el impacto de agregar réplicas y particiones, ya que cada uno afecta al rendimiento de lectura y escritura de maneras diferentes. Para obtener más información sobre las réplicas y las particiones, consulte Estimación y administración de la capacidad de un servicio Search.

Resistencia a errores transitorios

Los errores transitorios son errores breves e intermitentes en los componentes. Se producen con frecuencia en un entorno distribuido como la nube y son una parte normal de las operaciones. Los errores transitorios se corrigen después de un breve período de tiempo. Es importante que las aplicaciones puedan controlar errores transitorios, normalmente mediante el reintento de solicitudes afectadas.

Todas las aplicaciones hospedadas en la nube deben seguir las instrucciones de control de errores transitorios de Azure cuando se comunican con cualquier API, bases de datos y otros componentes hospedados en la nube. Para obtener más información, consulte Recomendaciones para controlar errores transitorios.

Los indexadores de AI Search tienen un control de errores transitorios integrado. Si un origen de datos no está disponible brevemente, el indexador está diseñado para recuperarse y volver a intentarlo. Usa el seguimiento de cambios para reanudar la indexación del último documento indizado correctamente.

Los servicios de búsqueda pueden experimentar errores transitorios durante las operaciones de mantenimiento estándar y no programadas. Búsqueda de Azure AI no proporciona notificaciones anticipadas ni permite la programación de mantenimiento en momentos específicos. Aunque se hace todo lo posible por minimizar el tiempo de inactividad, incluso en los servicios de réplica única, pueden producirse breves interrupciones. Para mejorar la resistencia frente a estos errores transitorios, se recomienda usar dos o más réplicas.

Si construyes aplicaciones que interactúan con AI Search, estas deben manejar errores transitorios. Use una estrategia de reintento con retrocesos exponenciales para las operaciones de lectura y escritura.

Resistencia a errores de zona de disponibilidad

Las zonas de disponibilidad son grupos físicamente independientes de centros de datos dentro de una región de Azure. Cuando una zona falla, los servicios pueden transferirse a una de las zonas restantes.

AI Search tiene redundancia de zona, lo que significa que las réplicas se distribuyen a través de varias zonas de disponibilidad dentro de la región del servicio de búsqueda.

Al agregar dos o más réplicas al servicio, AI Search intenta colocar cada réplica en una zona de disponibilidad diferente. En el caso de los servicios que tienen más réplicas que las zonas disponibles, las réplicas se distribuyen entre zonas lo más uniformemente posible.

En el diagrama siguiente se muestra cómo se puede implementar un servicio de búsqueda de ejemplo con cuatro réplicas en tres zonas de disponibilidad:

> ⓘ **Importante**
>
> La búsqueda de IA no garantiza la ubicación exacta de las réplicas. La selección de ubicación está sujeta a restricciones de capacidad, operaciones de escalado y otros factores.

Requisitos

La redundancia de zona se habilita automáticamente cuando el servicio Search cumple todos los criterios siguientes:

- **Compatibilidad con regiones:** La compatibilidad con zonas de disponibilidad depende de la infraestructura y el almacenamiento. Para obtener una lista de las regiones admitidas, consulte Elección de una región para la búsqueda de IA.
- **Nivel:** El servicio debe estar en el nivel Básico o superior.
- **Número de réplicas:** El servicio debe tener al menos dos réplicas.

  > ⓘ **Nota:**
  >
  > Ai Search intenta distribuir réplicas entre varias zonas cuando tiene dos o más réplicas. Sin embargo, para las cargas de trabajo de lectura y escritura, debería usar tres o más réplicas para lograr el SLA de disponibilidad más alto posible.

Distribución de instancias a través de zonas

AI Search intenta colocar réplicas en diferentes zonas de disponibilidad. Sin embargo, en ocasiones hay situaciones en las que todas las réplicas de un servicio de búsqueda pueden colocarse en la misma zona de disponibilidad. Esta situación puede ocurrir cuando las réplicas se quitan del servicio, como cuando se *reduce horizontalmente* configurando el servicio para que use menos réplicas. La eliminación de réplicas no hace que las réplicas restantes se redistribuyan automáticamente entre las zonas de disponibilidad.

Para reducir la probabilidad de que todas las réplicas se coloquen en una sola zona de disponibilidad, puede desencadenar manualmente una operación de escalabilidad horizontal inmediatamente después de una operación de reducción horizontal. Por ejemplo, supongamos que el servicio Search tiene 10 réplicas y desea reducir horizontalmente a 7 réplicas. En lugar de realizar una sola operación de escalado, puede escalar temporalmente a 6 instancias y, a continuación, escalar inmediatamente a 7 instancias para desencadenar el reequilibrio de zona.

Cost

Cada servicio de búsqueda comienza con una réplica. La redundancia de zona requiere dos o más réplicas, lo que aumenta el costo de ejecutar el servicio. Para comprender las implicaciones de facturación de las réplicas, utilice la calculadora de precios.

Configurar soporte de zonas de disponibilidad

Si su servicio de búsqueda cumple los requisitos de redundancia de zona, no es necesaria ninguna configuración adicional. Siempre que sea posible, AI Search procura colocar las réplicas en diferentes zonas de disponibilidad.

Planeamiento y administración de capacidad

Para prepararse para un fallo en la zona de disponibilidad, considere sobreaprovisionar el número de réplicas. El aprovisionamiento excesivo permite al servicio de búsqueda tolerar cierta pérdida de capacidad y seguir funcionando sin un rendimiento degradado. Agregar réplicas durante una interrupción es difícil, por lo que el aprovisionamiento excesivo ayuda a garantizar que el servicio de búsqueda pueda controlar los volúmenes de solicitudes normales, incluso con una capacidad reducida. Para más información, consulte Administración de la capacidad mediante el sobreaprovisionamiento.

Comportamiento cuando todas las zonas están en buen estado

En esta sección se describe qué esperar cuando los servicios Search están configurados para la redundancia de zona y todas las zonas de disponibilidad están operativas.

- **Enrutamiento de tráfico entre zonas:** AI Search ejecuta el equilibrio de carga automático de todas las consultas y escrituras a través de todas las réplicas disponibles. AI Search puede enviar operaciones de lectura a cualquier réplica de cualquier zona de disponibilidad. Envía operaciones de escritura a una sola réplica principal que el servicio AI Search selecciona.

- **Replicación de datos entre zonas:** Los cambios en los datos se replican automáticamente entre réplicas entre zonas de disponibilidad. La replicación se produce de forma asincrónica, lo que significa que las escrituras se confirman en una réplica principal antes de que se repliquen en otras réplicas.

Comportamiento durante un fallo de zona

En esta sección se describe qué esperar cuando los servicios de búsqueda están configurados para la redundancia de zona y se produce una interrupción de zona de disponibilidad.

- **Detección y respuesta:** Ai Search es responsable de detectar un error en una zona de disponibilidad. No es necesario hacer nada para iniciar una conmutación por error de la zona.

- **Notificación**: Microsoft no le notifica automáticamente cuando una zona está inactiva. Sin embargo, puede usar Azure Resource Health para supervisar el estado de un recurso individual y puede configurar alertas de Resource Health para notificarle problemas. También puede usar Azure Service Health para comprender el estado general del servicio, incluidos los errores de zona, y puede configurar alertas de Service Health para notificarle problemas.

- **Solicitudes activas:** Se finalizan las solicitudes que las réplicas procesan en la zona con errores. Los clientes deben reintentar las solicitudes siguiendo las instrucciones para controlar errores transitorios.

- **Pérdida de datos esperada:** Si la zona de disponibilidad afectada solo contiene réplicas de lectura, no se espera ninguna pérdida de datos.

  Si se pierde la réplica principal porque estaba en la zona afectada, es posible que se pierdan las operaciones de escritura que aún no se hayan replicado.

- **Tiempo de inactividad esperado:** En la mayoría de las situaciones, no se espera que un error de zona provoque tiempo de inactividad en el servicio de búsqueda para las

operaciones de lectura porque las réplicas de lectura de otras zonas de disponibilidad siguen atendiendo solicitudes.

Si se pierde la réplica principal porque estaba en la zona afectada, AI Search promueve automáticamente otra réplica para convertirse en la nueva principal para que las operaciones de escritura se puedan reanudar. Normalmente, la promoción de la réplica tarda unos segundos en producirse. Durante este tiempo, es posible que las operaciones de escritura no se realicen correctamente. Asegúrese de que las aplicaciones están preparadas siguiendo las instrucciones de control de errores transitorios.

Sin embargo, hay algunas situaciones poco probables en las que todas las réplicas de tu servicio de búsqueda podrían encontrarse en una única zona de disponibilidad. En este escenario, puede experimentar una interrupción del servicio hasta que se recupere la zona. Para obtener más información y comprender una solución alternativa, consulte Distribución de instancias.

- **Reenrutamiento del tráfico:** Cuando se produce un error en una zona, AI Search detecta el error y enruta las solicitudes a réplicas activas en las zonas supervivientes. Si se pierde la réplica principal, se promueve otra réplica para que sea la nueva principal.

Recuperación de zona

Cuando se recupera la zona de disponibilidad, AI Search restaura automáticamente las operaciones normales y comienza a enrutar el tráfico a las réplicas disponibles en todas las zonas, incluida la zona recuperada.

Prueba de fallos de zona

AI Search administra el enrutamiento del tráfico para los servicios con redundancia de zona. No es necesario iniciar ni validar ningún proceso de error de zona.

Resistencia a errores en toda la región

AI Search es un servicio de una sola región. Si la región deja de estar disponible, su servicio de búsqueda también dejará de estar disponible.

Soluciones personalizadas de varias regiones para la resistencia

Opcionalmente, puede implementar varios servicios de búsqueda de IA en diferentes regiones. Es responsable de implementar y configurar servicios independientes en cada región. Si crea

una implementación idéntica en una región secundaria de Azure que usa una arquitectura de varias regiones, la aplicación será menos susceptible a un desastre de una sola región.

Cuando se sigue este enfoque, es necesario sincronizar los índices entre regiones para recuperar el último estado de la aplicación. También debe configurar las directivas de equilibrio de carga y conmutación por error.

Para optimizar el rendimiento de la solución general, busque oportunidades para realizar la indexación en réplicas de solo lectura de los orígenes de datos. Por ejemplo, algunos indexadores admiten la lectura desde las réplicas de lectura de un origen de datos distribuido geográficamente.

Para más información, consulte Implementaciones de varias regiones en Búsqueda de Azure AI.

Copias de seguridad y restauración

Dado que AI Search no es una solución de almacenamiento de datos principal, no proporciona opciones de copia de seguridad y restauración de autoservicio. Sin embargo, puede usar la muestra `index-backup-restore` para .NET o Python para hacer una copia de seguridad de la definición del índice y sus documentos en una serie de archivos JSON, que luego se usan para restaurar el índice.

Sin embargo, si elimina accidentalmente el índice y no tiene una copia de seguridad, puede volver a generar el índice. La regeneración implica volver a crear el índice en su servicio de búsqueda y luego volver a cargarlo recuperando los datos de su almacén de datos principal.

Acuerdo de nivel de servicio

El contrato de nivel de servicio (SLA) para los servicios de Azure describe la disponibilidad esperada de cada servicio y las condiciones que la solución deberá cumplir para lograr esa expectativa de disponibilidad. Para obtener más información, consulte Acuerdos de Nivel de Servicio para servicios en línea.

En AI Search, el Acuerdo de Nivel de Servicio de disponibilidad se aplica a los servicios de búsqueda que:

- Están configurados para usar un nivel facturable.
- Tener al menos dos réplicas para cargas de trabajo de solo lectura (consultas).
- Tener al menos tres réplicas para cargas de trabajo de lectura y escritura (consultas e indexación).

Contenido relacionado

- Confiabilidad en Azure
- Límites de servicio en AI Search
- Elección de una región para la búsqueda de IA
- Elección de un plan de tarifa para AI Search
- Planeamiento o adición de capacidad en la búsqueda de IA

ⓘ **Nota:** El autor ha creado este artículo con ayuda de inteligencia artificial. Más información

---

Last updated on 22/01/2026

Confiabilidad en Azure API Management

Azure API Management es un servicio totalmente administrado que ayuda a las organizaciones a publicar, proteger, transformar, mantener y supervisar las API. Como servicio de Azure, API Management proporciona una variedad de funcionalidades para satisfacer los requisitos de confiabilidad.

Cuando se usa Azure, la confiabilidad es una responsabilidad compartida. Microsoft proporciona una variedad de funcionalidades para admitir resistencia y recuperación. Es responsable de comprender cómo funcionan esas funcionalidades dentro de todos los servicios que usa y de seleccionar las funcionalidades que necesita para cumplir los objetivos empresariales y los objetivos de tiempo de actividad.

En este artículo se describe cómo hacer que API Management sea resistente a una variedad de posibles interrupciones y problemas, incluidos errores transitorios, interrupciones de zona de disponibilidad, interrupciones de regiones y mantenimiento del servicio. También se describe cómo puede usar copias de seguridad para recuperarse de otros tipos de problemas y se resalta cierta información clave sobre el acuerdo de nivel de servicio (SLA) de API Management.

Introducción a la arquitectura de confiabilidad

API Management usa una arquitectura basada en unidades de escalado para proporcionar redundancia y escalabilidad integradas. Al implementar una instancia de API Management, se configuran una o varias *unidades de escalado* o *unidades*. Cada unidad es una representación lógica de la capacidad que contiene los recursos de proceso necesarios para controlar las solicitudes de API.

Unidades

Cada unidad consta de dos recursos de proceso (máquinas virtuales o servidores similares, en función del nivel de servicio) que controlan las solicitudes de API juntas. Usted no ve estas máquinas virtuales ni otros servidores. La plataforma administra automáticamente su creación y monitorización del estado. Si se produce un error en un recurso de proceso, la unidad continúa funcionando pero con una capacidad reducida, lo que proporciona cierta protección de confiabilidad integrada.

Al configurar una instancia con dos o más unidades, las unidades disponibles funcionan juntas para procesar solicitudes y proporcionar equilibrio de carga automático. Si una de las unidades deja de estar disponible, las unidades restantes siguen controlando el tráfico, pero con capacidad potencialmente reducida.

Para obtener mayores niveles de confiabilidad, API Management admite la distribución de unidades entre zonas de disponibilidad dentro de una región y entre varias regiones.

> ⓘ **Nota:**
>
> API Management usa unidades para los componentes del gateway. Las unidades no son aplicables al portal para desarrolladores ni al plano de administración.

Niveles de servicio

Los niveles de servicio de API Management proporcionan distintos niveles de confiabilidad:

- **Nivel Premium (clásico):** Admite varias unidades que se pueden distribuir entre zonas de disponibilidad y regiones para lograr una máxima resistencia.
- **Nivel Premium v2**: admite varias unidades que se pueden distribuir entre zonas de disponibilidad. Actualmente no admite implementaciones de varias regiones.
- **Niveles Básico v2, Estándar y Estándar v2:** Todas admiten varias unidades dentro de un único centro de datos. No admiten zonas de disponibilidad ni implementaciones de varias regiones.
- **Nivel de desarrollador:** Solo admite una sola unidad y no proporciona compatibilidad con zonas de disponibilidad ni con varias regiones. Este nivel está diseñado para escenarios de desarrollo y pruebas. No es adecuado para cargas de trabajo de producción.
- **Nivel de consumo:** Tiene funcionalidades de resistencia integradas y es resistente a una serie de errores dentro de un único centro de datos de Azure. Sin embargo, el nivel Consumo no proporciona compatibilidad con zonas de disponibilidad ni implementaciones de varias regiones. Para comprender el tiempo de actividad esperado de una instancia de API Management de nivel de consumo, revise el acuerdo de nivel de servicio (SLA).

> ⓘ **Nota:**
>
> Los niveles Desarrollador y Premium de API Management admiten puertas de enlace autohospedada, que puede ejecutar en su propia infraestructura. Al usar puertas de enlace autohospedadas, es responsable de configurarlas para cumplir los requisitos de confiabilidad. Las puertas de enlace autohospedados están fuera del ámbito de este artículo.

Recomendaciones de implementación de producción

Azure Well-Architected Framework proporciona recomendaciones sobre confiabilidad, rendimiento, seguridad, costo y operaciones. Para comprender cómo estas áreas influyen entre sí y contribuyen a una solución confiable de API Management, consulte Procedimientos recomendados de arquitectura para API Management.

Resistencia a errores transitorios

Los errores transitorios son errores breves e intermitentes en los componentes. Se producen con frecuencia en un entorno distribuido como la nube y son una parte normal de las operaciones. Los errores transitorios se corrigen después de un breve período de tiempo. Es importante que las aplicaciones puedan controlar errores transitorios, normalmente mediante el reintento de solicitudes afectadas.

Todas las aplicaciones hospedadas en la nube deben seguir las instrucciones de control de errores transitorios de Azure cuando se comunican con cualquier API, bases de datos y otros componentes hospedados en la nube. Para obtener más información, consulte Recomendaciones para controlar errores transitorios.

Al usar API Management delante de una API, es posible que tenga que reintentar solicitudes que produzcan un error debido a errores transitorios. Para proteger la API de back-end de ser sobrecargada por demasiadas solicitudes, API Management proporciona directivas de reintento, límite de velocidad y cuota. También puede configurar las funcionalidades de equilibrio de carga y disyuntor mediante el uso de recursos de backend.

Resistencia a errores de zona de disponibilidad

Las zonas de disponibilidad son grupos físicamente independientes de centros de datos dentro de una región de Azure. Cuando una zona falla, los servicios pueden transferirse a una de las zonas restantes.

API Management proporciona dos tipos de compatibilidad con zonas de disponibilidad al implementar una instancia de API Management Premium (clásica) en una región admitida:

- **Automático** (recomendado): API Management proporciona compatibilidad con zonas de disponibilidad automáticas cuando no se especifican las zonas de disponibilidad que se van a usar.

- **Manual:** API Management proporciona compatibilidad con la zona de disponibilidad manual cuando se especifican explícitamente las zonas de disponibilidad que se van a usar.

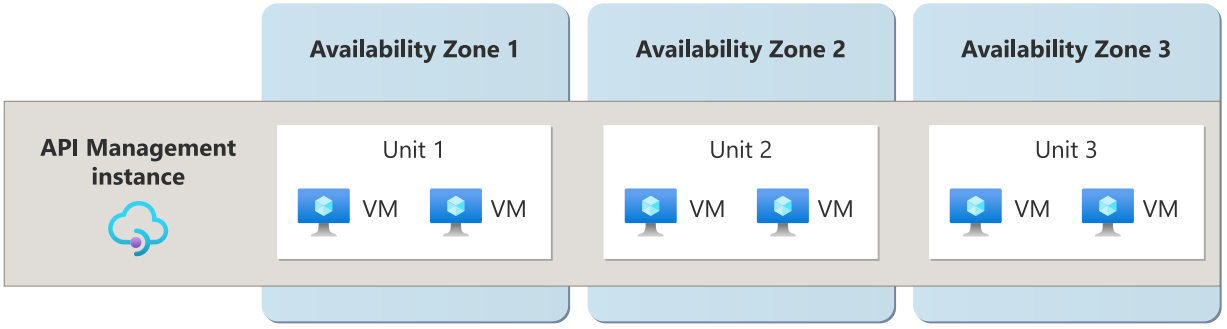
Con la compatibilidad con zonas de disponibilidad, API Management replica componentes de servicio entre zonas para lograr alta disponibilidad. En la región primaria, estos componentes incluyen la puerta de enlace (unidades de escalado), el plano de administración y el portal para desarrolladores. En regiones secundarias, solo se replican las unidades de puerta de enlace. Para más información sobre las regiones secundarias, consulte Resistencia a errores en toda la región.

Compatibilidad con zonas de disponibilidad automáticas

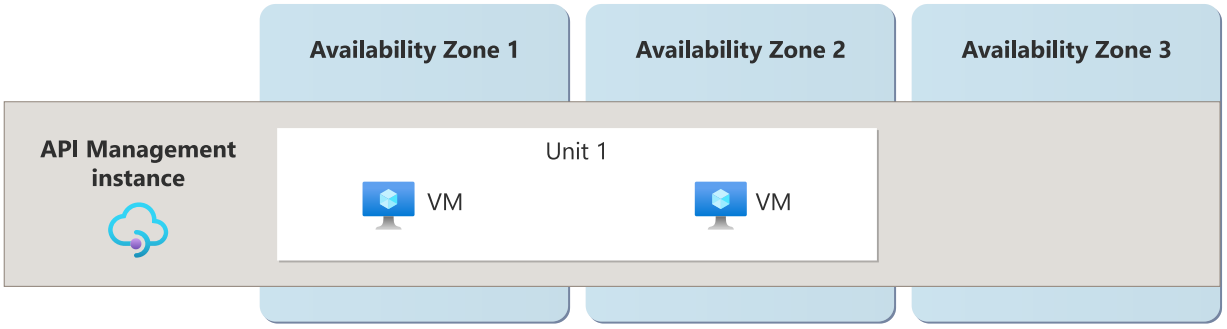
Puede usar la compatibilidad automática con zonas de disponibilidad para elegir una configuración de instancia de una sola unidad o de varias unidades con el fin de lograr la redundancia de zonas:

- **Configuración** de varias unidades (recomendada): si la instancia tiene dos o más unidades, API Management realiza un mejor intento de distribuir las unidades de la instancia entre las zonas de disponibilidad de la región. No puede determinar en qué zonas de disponibilidad se colocan las unidades. Implemente un mínimo de dos unidades, que se pueden distribuir entre dos zonas.

  En el diagrama siguiente se muestra una instancia de API Management con tres unidades configuradas para la compatibilidad automática con zonas de disponibilidad:



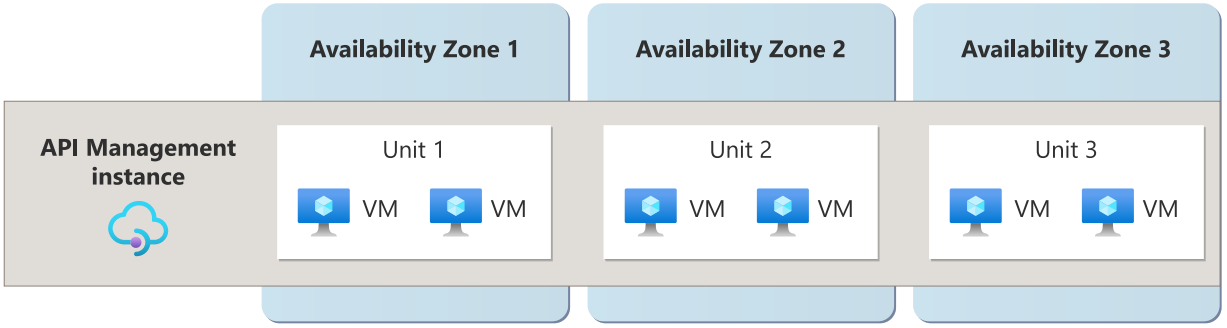
- **Configuración de una sola unidad:** Si la instancia tiene una sola unidad, los recursos de proceso subyacentes de la unidad se distribuyen a dos zonas de disponibilidad. No puede determinar en qué zonas de disponibilidad se colocan los recursos de proceso de la unidad.



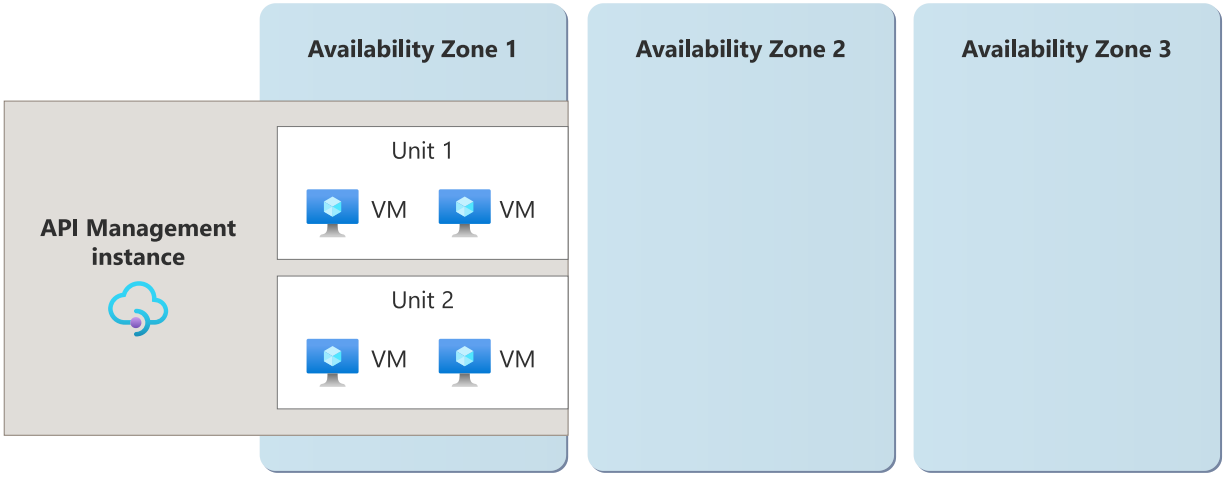
Compatibilidad con zonas de disponibilidad configuradas manualmente

Si desea seleccionar explícitamente las zonas de disponibilidad que se van a usar, puede elegir entre configuraciones con redundancia de zona y zonales:

- **Redundancia de zona:** Configure manualmente la redundancia de zona para una instancia de API Management en una región compatible con el fin de proporcionar redundancia a los componentes del servicio. Al seleccionar dos o más zonas de disponibilidad que se van a usar, Azure replica automáticamente los componentes de servicio en las zonas seleccionadas.



- **Zonal:** Los componentes del servicio API Management se implementan en una sola zona que se selecciona dentro de una región de Azure. Todas las unidades se colocan en la misma zona de disponibilidad.



> **Importante**
>
> Ancle a una sola zona de disponibilidad solo si la latencia entre zonas es demasiado alta para sus necesidades y después de comprobar que la latencia no cumple sus requisitos. Por sí mismo, una instancia zonal no proporciona resistencia a una interrupción de zona de disponibilidad. Para mejorar la resistencia de una implementación de API Management zonal, debe implementar explícitamente instancias independientes en varias zonas de disponibilidad y configurar el enrutamiento y la conmutación por error del tráfico.

Requisitos

- **Compatibilidad con regiones:** API Management admite zonas de disponibilidad para los niveles Premium (clásico) y Premium v2 en regiones en las que el nivel api Management está disponible y la región admite zonas de disponibilidad.
- **Requisito de nivel:** Debe usar el nivel Premium (clásico) o Premium v2 para configurar la compatibilidad con zonas de disponibilidad. API Management no admite actualmente zonas de disponibilidad en los niveles consumo clásico, desarrollador, básico y estándar o en los niveles Básico v2 y Estándar v2. Para ver las opciones de actualización, consulte Actualización y escalado de una instancia de API Management.

Consideraciones

- **Número de unidades para instancias con redundancia de zona:** Si configura manualmente la redundancia de zona para una instancia, también debe configurar una serie de unidades de API Management que se pueden distribuir uniformemente en todas las zonas de disponibilidad seleccionadas. Por ejemplo, si selecciona dos zonas, debe configurar al menos dos unidades. En su lugar, puede configurar cuatro unidades u otras

varias de dos unidades. Si selecciona tres zonas de disponibilidad, debe configurar tres unidades, seis unidades u otro múltiplo de tres unidades.

Si usa la compatibilidad con la zona de disponibilidad automática, no es necesario usar un número específico de unidades. Las unidades que implemente se distribuyen entre las zonas de disponibilidad de una manera más adecuada. Para obtener redundancia máxima de zona, use al menos dos unidades para que una interrupción de zona de disponibilidad no afecte al rendimiento de la puerta de enlace.

Para determinar el número de unidades que proporcionan el rendimiento necesario de su puerta de enlace, use las métricas de capacidad y sus propias pruebas. Para más información sobre el escalado y la actualización de la instancia de servicio, consulte Actualización y escalado de una instancia de API Management.

- **Escalado automático:** Si configura manualmente zonas de disponibilidad en una instancia de API Management configurada con el escalado automático, es posible que tenga que ajustar la configuración de escalado automático después de la configuración. En este caso, el número de unidades de API Management en reglas y límites de escalado automático debe ser un múltiplo del número de zonas. Si usa la compatibilidad con la zona de disponibilidad automática, no es necesario ajustar la configuración de escalado automático.

- **Requisitos de dirección IP:** Cuando habilita la compatibilidad con zonas de disponibilidad en una instancia de API Management implementada en una red virtual externa o interna, debe especificar un recurso de dirección IP pública para que lo utilice la instancia. En una red virtual interna, la dirección IP pública solo se usa para las operaciones de administración, no para las solicitudes de API. Para más información, consulte Direcciones IP en API Management.

Cost

Independientemente de la configuración de la zona de disponibilidad, si agrega más unidades, incurrirá en más costos. Para más información, consulte Precios de API Management.

Configurar soporte de zonas de disponibilidad

En esta sección se explica cómo configurar la compatibilidad con zonas de disponibilidad para la instancia de API Management. Para obtener más información, consulte Habilitación de la compatibilidad con zonas de disponibilidad en instancias de API Management.

- **Cree una instancia de API Management que admita zonas de disponibilidad:** Al crear una instancia de API Management Premium (clásica) en una región que admita zonas de

disponibilidad, la instancia admite zonas de disponibilidad de forma predeterminada. Puede seleccionar la compatibilidad automática con zonas de disponibilidad o configurar manualmente el soporte zonal o el soporte con redundancia de zona.

> ⓘ **Nota:**
>
> Cuando selecciona qué zonas de disponibilidad usar, en realidad está seleccionando la *zona de disponibilidad lógica*. Si implementa otros componentes de la carga de trabajo en una suscripción de Azure diferente, podrían usar un número de zona de disponibilidad lógica distinto para acceder a la misma zona de disponibilidad física. Para obtener más información, consulte zonas de disponibilidad física y lógica.

- **Habilite o vuelva a configurar la compatibilidad con la zona de disponibilidad:** Puede cambiar la configuración de zona de disponibilidad de una instancia de API Management, incluida la adición de zonas de disponibilidad y el traslado de una instancia zonal entre zonas de disponibilidad. Para obtener información sobre cómo configurar la compatibilidad con zonas de disponibilidad en una instancia de API Management, consulte Habilitar la compatibilidad con zonas de disponibilidad en instancias de API Management. Ninguna de las opciones de configuración requiere tiempo de inactividad.

  Al cambiar la configuración de la zona de disponibilidad, los cambios pueden tardar entre 15 y 45 minutos o más en aplicarse. La puerta de enlace de API Management puede seguir controlando las solicitudes de API durante ese tiempo.

  Cambiar la configuración de la zona de disponibilidad desencadena un cambio de dirección IP pública y privada.

Planeamiento y administración de capacidad

En un escenario de reducción de zona, no hay ninguna garantía de que las solicitudes de más capacidad en otra zona de disponibilidad tengan éxito. El reposición de unidades perdidas se produce de forma más adecuada. Si necesita capacidad garantizada cuando se produce un error en una zona de disponibilidad, cree y configure la instancia de API Management para que tenga en cuenta la pérdida de una zona mediante la realización de todas las acciones siguientes:

- Aprovisionamiento excesivo de las unidades de la instancia de API Management.
- Use la configuración de zona de disponibilidad automática o redundante por zonas.

Para más información, consulte Administración de la capacidad con sobreaprovisionamiento.

Use métricas de capacidad y sus propias pruebas para determinar el número de unidades que proporcionan el rendimiento de la puerta de enlace necesaria. Para más información sobre cómo escalar y actualizar la instancia de servicio, consulte Actualización y escalado de una instancia de API Management.

Comportamiento cuando todas las zonas están en buen estado

En esta sección se describe qué esperar cuando las instancias de API Management están configuradas con compatibilidad con zonas de disponibilidad y todas las zonas de disponibilidad están operativas.

- **Enrutamiento de tráfico entre zonas:** Durante las operaciones normales, el tráfico se enruta entre todas las unidades de API Management disponibles en todas las zonas de disponibilidad seleccionadas.
- **Replicación de datos entre zonas:** API Management almacena y replica los siguientes datos.
  - *La configuración de la puerta de enlace*, como las API y las definiciones de directiva, se sincroniza periódicamente entre las zonas de disponibilidad que seleccione para la instancia. La propagación de actualizaciones entre las zonas de disponibilidad normalmente tarda menos de 10 segundos.
  - *Datos de la memoria caché interna*, si usa la memoria caché interna que proporciona API Management. Las entradas de caché se distribuyen entre zonas de disponibilidad. La caché interna es volátil y no se garantiza que los datos se conserven. Considere la posibilidad de usar una caché externa si necesita conservar los datos almacenados en caché.
  - *Contadores de límite de velocidad*, si usa las funcionalidades de limitación de velocidad que proporciona API Management. Los contadores de límite de velocidad se replican de forma asíncrona entre las zonas de disponibilidad que seleccione para la instancia.

Comportamiento durante un fallo de zona

En esta sección se describe qué esperar cuando las instancias de API Management están configuradas con compatibilidad con la zona de disponibilidad y hay una interrupción de zona de disponibilidad.

- **Detección y respuesta:** La responsabilidad de la detección y la respuesta depende de la configuración de zona de disponibilidad que usa la instancia.

- *Automático y con redundancia de zona:* En el caso de las instancias configuradas para utilizar la compatibilidad automática con zonas de disponibilidad o configuradas manualmente para utilizar la redundancia de zona, la plataforma API Management se encarga de detectar los fallos en una zona de disponibilidad y responder a ellos. No es necesario hacer nada para iniciar una conmutación por error de la zona.
- *Zonal:* para que las instancias configuradas sean zonales, debe detectar la pérdida de una zona de disponibilidad e iniciar una conmutación por error a una instancia secundaria que cree en otra zona de disponibilidad.

- **Solicitudes activas:** Cuando una zona de disponibilidad no está disponible, las solicitudes en curso que están conectadas a una unidad de API Management en la zona de disponibilidad defectuosa se finalizan y deben reintentarse.
- **Notificación**: Microsoft no le notifica automáticamente cuando una zona está inactiva. Sin embargo, puede usar Azure Resource Health para supervisar el estado de un recurso individual y puede configurar alertas de Resource Health para notificarle problemas. También puede usar Azure Service Health para comprender el estado general del servicio, incluidos los errores de zona, y puede configurar alertas de Service Health para notificarle problemas.
- **Pérdida de datos esperada:** API Management almacena los siguientes datos.
  - *Cambios en la configuración de la puerta de enlace*, y que se replican en cada zona de disponibilidad seleccionada en aproximadamente 10 segundos. Si se produce una interrupción de una zona de disponibilidad, puede perder los cambios de configuración que no se replican.
  - *Datos de la memoria caché interna*, si usa la característica de caché interna. La caché interna es volátil y no se garantiza que los datos se conserven. Durante una interrupción del servicio en una zona de disponibilidad, es posible que se pierda parte o la totalidad de los datos almacenados en caché. Considere la posibilidad de usar una caché externa si necesita conservar los datos almacenados en caché.
  - *Contadores de límite de velocidad*, si usa la característica de límite de velocidad. Durante una interrupción de zona de disponibilidad, es posible que los contadores de límite de frecuencia no estén actualizados en las zonas supervivientes.
- **Tiempo de inactividad esperado:** El tiempo de inactividad esperado depende de la configuración de zona de disponibilidad que usa la instancia.
  - *Automático:* Puede esperar que las instancias que usen compatibilidad con zonas de disponibilidad automáticas no tengan tiempo de inactividad durante una interrupción

de la zona de disponibilidad. Las unidades de la zona o zonas no afectadas siguen funcionando.

También puede esperar instancias que usen compatibilidad automática con zonas de disponibilidad, pero que tengan una sola unidad, sin tiempo de inactividad. En este caso, API Management distribuye los recursos de proceso subyacentes de la unidad a dos zonas. El recurso de la zona no afectada sigue funcionando.

- *Con redundancia de zona:* Puede esperar que las instancias con redundancia de zona no sufran interrupciones durante una interrupción de la zona de disponibilidad.
- *Zonal:* En el caso de las instancias zonales, cuando una zona no está disponible, la instancia no está disponible hasta que se recupere la zona de disponibilidad.

- **Reenrutamiento del tráfico:** El comportamiento de reenrutamiento del tráfico depende de la configuración de zona de disponibilidad que usa la instancia.
  - *Automático y con redundancia de zona:* En el caso de las instancias que están configuradas para utilizar la compatibilidad automática con zonas de disponibilidad o que están configuradas manualmente para utilizar la redundancia de zona, cuando una zona no está disponible, ninguna de las unidades de la zona afectada estará disponible. Puede optar por escalar la instancia para agregar más unidades.
  - *Zonal:* En el caso de las instancias zonales, cuando una zona no está disponible, la instancia no está disponible. Si tiene una instancia secundaria en otra zona de disponibilidad, es responsable de redirigir el tráfico a esa instancia secundaria.

Recuperación de zona

- **Automático y con redundancia de zona:** En el caso de las instancias que están configuradas para utilizar la compatibilidad automática con zonas de disponibilidad o que están configuradas manualmente para utilizar la redundancia de zona, cuando la zona de disponibilidad se recupera, API Management restaura automáticamente las unidades de la zona de disponibilidad y redirige el tráfico entre sus unidades con normalidad.
- **Zonal:** En el caso de las instancias zonales, usted es responsable de redirigir el tráfico a la instancia en la zona de disponibilidad original una vez que la zona de disponibilidad se haya recuperado.

Prueba de fallos de zona

- **Automático y con redundancia de zona:** En el caso de las instancias que están configuradas para utilizar la compatibilidad automática con zonas de disponibilidad o que están configuradas manualmente para utilizar la redundancia de zona, la plataforma API Management administra el enrutamiento del tráfico, la conmutación por error y la conmutación por recuperación. Esta característica está totalmente administrada, por lo que no es necesario iniciar ni validar los procesos de error de zona de disponibilidad.

- **Zonal:** En el caso de las instancias zonales, no se puede simular una interrupción de la zona de disponibilidad que contiene la instancia de API Management. Sin embargo, puede configurar manualmente puertas de enlace ascendentes o equilibradores de carga para redirigir el tráfico a una instancia diferente en una zona de disponibilidad diferente.

Resistencia a errores en toda la región

Mediante una implementación de varias regiones, puede agregar puertas de enlace de API regionales a una instancia de API Management existente en una o varias regiones de Azure admitidas. La implementación en varias regiones ayuda a reducir la latencia de las solicitudes que perciben los consumidores de API distribuidos geográficamente. Una implementación de varias regiones también mejora la disponibilidad del servicio si una región se queda sin conexión.

> ⓘ **Importante**
>
> Las implementaciones de varias regiones solo se admiten en el nivel Premium (clásico) de API Management.

Implementación multirregión administrada por Microsoft

Al agregar una región, configure:

- Número de unidades que hospeda esa región.

- Resistencia a errores de zona de disponibilidad, si esa región proporciona zonas de disponibilidad.

- La configuración de red virtual en la región agregada, si las redes están configuradas en la región o regiones existentes.

Requisitos

- **Compatibilidad con regiones:** Puede crear implementaciones de varias regiones en el nivel Premium (clásico) con cualquier región de Azure que admita API Management. Para ver qué regiones admiten implementaciones de varias regiones, consulte Disponibilidad del producto por región.

- **Requisito de nivel:** Debe usar el nivel Premium (clásico) para configurar la compatibilidad con varias regiones. Para actualizar la instancia al nivel Premium (clásico), consulte Actualización al nivel Premium.

Consideraciones

- **Solo puerta de enlace:** Únicamente el componente de puerta de enlace de su instancia de gestión de API se replica en múltiples regiones. El plano de administración y el portal para desarrolladores de la instancia permanecen hospedados solo en la región primaria donde implementó originalmente el servicio.

- **Requisitos de red:** Si desea configurar una ubicación secundaria para su instancia de API Management cuando se implementa (inyecta) en una red virtual, la red virtual y la región de la subred deben coincidir con la ubicación secundaria que configure. Si agrega, quita o habilita la zona de disponibilidad en la región primaria, o si cambia la subred de la región primaria, cambia la dirección IP virtual (VIP) de la instancia de API Management. Para obtener más información, consulte Cambios en las direcciones IP. Sin embargo, si se agrega una región secundaria, el VIP de la región principal no cambia, ya que cada región tiene su propio VIP privado.

- **Nombres del sistema de nombres de dominio (DNS):** La puerta de enlace de cada región, incluida la región primaria, tiene un nombre DNS regional que sigue el patrón de dirección URL de `https://<service-name>-<region>-01.regional.azure-api.net`, por ejemplo `https://contoso-westus2-01.regional.azure-api.net`.

Cost

Agregar regiones implica costos. Para más información, consulte Precios de API Management.

Configuración de la compatibilidad con varias regiones

Para configurar la compatibilidad con varias regiones en una instancia de API Management, consulte Implementar una instancia de API Management en varias regiones de Azure.

Para quitar una región de una instancia de API Management, consulte Eliminación de una región del servicio API Management.

Planeamiento y administración de capacidad

En un escenario de caída regional, no hay garantía de que las solicitudes de más capacidad en otra región se realicen con éxito. Si necesita capacidad garantizada cuando se produce un error en una región, debe crear y configurar la instancia de API Management para que tenga en cuenta la pérdida de una región. Puede hacerlo mediante el aprovisionamiento excesivo de la capacidad de la instancia de API Management. Para más información sobre el principio de sobreaprovisionamiento, consulte Gestionar la capacidad con sobreaprovisionamiento.

En las implementaciones de varias regiones, el escalado automático solo se aplica a la región primaria. Las regiones secundarias requieren ajustes de escalado manual o herramientas personalizadas que controle.

Comportamiento cuando todas las regiones están en buen estado

En esta sección se describe qué se puede esperar cuando las instancias de API Management se configuran con compatibilidad con varias regiones y todas las regiones están operativas.

- **Enrutamiento del tráfico entre regiones:** API Management enruta automáticamente las solicitudes recibidas a una puerta de enlace regional. Una solicitud se enruta a la puerta de enlace regional con la latencia más baja del cliente. Si necesita usar un enfoque de enrutamiento diferente, puede configurar su propio Administrador de tráfico con reglas de enrutamiento personalizadas. Para obtener más información, consulte Uso del enrutamiento personalizado en las puertas de enlace regionales de API Management.

  Cuando una solicitud llega a una puerta de enlace regional de API Management, se enruta a la API backend, a menos que una directiva devuelva una respuesta directamente desde la puerta de enlace, como una respuesta almacenada en caché o un código de error. En una solución de varias regiones, debe tener cuidado de enrutar a una instancia de la API de back-end que cumpla los requisitos de rendimiento. Para obtener más información, consulte Enrutamiento de llamadas API a servicios back-end regionales.

- **Replicación de datos entre regiones:** La configuración de puerta de enlace, como las API y las definiciones de directiva, se sincronizan periódicamente entre las regiones primarias y secundarias que agregue. La propagación de actualizaciones a las puertas de enlace regionales normalmente tarda menos de 10 segundos.

  Los contadores de límite de velocidad y los datos de la caché interna son específicos de la región, por lo que no se replican entre regiones.

Comportamiento durante una falla de región

En esta sección se describe qué puede ocurrir cuando las instancias de API Management se configuran con compatibilidad con varias regiones y se produce una interrupción del servicio en una de las regiones que utiliza.

- **Detección y respuesta:** La API Management se encarga de detectar un error en una región y de realizar automáticamente una conmutación por error a una puerta de enlace en una de las otras regiones que usted configure.

- **Solicitudes activas:** Las solicitudes activas que se están procesando en la región defectuosa se pueden quitar y el cliente debe reintentarlas.

- **Pérdida de datos esperada:** API Management no almacena datos, excepto los contadores de configuración, caché y límite de velocidad.

  Los cambios de configuración se replican en cada región en aproximadamente 10 segundos. Si se produce una interrupción del servicio en su región principal, es posible que pierda los cambios de configuración que no se hayan replicado.

  Los contadores de límite de velocidad y los datos de la caché interna son específicos de la región, por lo que no se replican entre regiones.

- **Tiempo de inactividad esperado:** No se espera ningún tiempo de inactividad de puerta de enlace.

  Si la región primaria se queda sin conexión, el plano de administración de API Management y el portal para desarrolladores no estarán disponibles. Sin embargo, las regiones secundarias siguen atendiendo solicitudes de API mediante la configuración de puerta de enlace más reciente.

- **Reenrutamiento del tráfico:** Si una región se queda sin conexión, las solicitudes de API se enrutan automáticamente evitando la región fallida hacia la siguiente puerta de enlace más cercana.

Recuperación de regiones

Cuando se recupera la región primaria, API Management restaura automáticamente las unidades de la región y vuelve a enrutar el tráfico entre las unidades.

Prueba de fallos de región

Para estar preparado para interrupciones inesperadas de regiones, realice pruebas periódicas de sus respuestas a los fallos de la región. Puede simular algunos aspectos de un error de región deshabilitando el enrutamiento a una puerta de enlace regional.

Copias de seguridad y restauración

API Management no almacena la mayoría de los datos en tiempo de ejecución. Sin embargo, puede realizar una copia de seguridad de la configuración del servicio API Management. También puede usar operaciones de copia de seguridad y restauración para replicar las configuraciones del servicio API Management entre entornos operativos, como el desarrollo y el almacenamiento provisional.

> **Importante**
>
> En un procedimiento de copia de seguridad, se incluyen datos en tiempo de ejecución, como usuarios y suscripciones, que podrían no ser siempre deseables.

La copia de seguridad se admite en los niveles Desarrollador, Básico, Estándar y Premium.

Para obtener más información, consulte Cómo implementar la recuperación ante desastres mediante la copia de seguridad y restauración del servicio en API Management.

Para realizar copias de seguridad o restaurar algunos componentes o recursos del servicio, también puede considerar opciones administradas por el cliente, como las herramientas APIOps y las soluciones de infraestructura como código (IaC).

Resistencia al mantenimiento del servicio

API Management realiza actualizaciones de servicio periódicas y otras formas de mantenimiento.

En los niveles Básico, Estándar y Premium (clásico), puede personalizar cuándo en el proceso de actualización la instancia recibe una actualización. Si necesita validar el efecto de las actualizaciones en la carga de trabajo, considere la posibilidad de configurar una instancia de prueba para recibir actualizaciones al principio de un ciclo de actualización y establecer la instancia de producción para recibir actualizaciones en un momento posterior al ciclo. También puede especificar una ventana de mantenimiento, que es la hora del día en la que desea que la instancia aplique las actualizaciones del servicio.

Para obtener más información, consulte Configurar los ajustes de actualización del servicio para sus instancias de API Management.

Acuerdo de nivel de servicio

El contrato de nivel de servicio (SLA) para los servicios de Azure describe la disponibilidad esperada de cada servicio y las condiciones que la solución deberá cumplir para lograr esa expectativa de disponibilidad. Para obtener más información, consulte Acuerdos de Nivel de Servicio para servicios en línea.

Al implementar una instancia de API Management en varias zonas de disponibilidad o regiones, aumenta el porcentaje de tiempo de actividad definido en el Acuerdo de Nivel de Servicio.

El servicio proporciona su propio Acuerdo de Nivel de Servicio, pero también debe tener en cuenta la confiabilidad prevista de otros componentes de carga de trabajo, como los back-end de API.

Contenido relacionado

- Recuperación ante desastres y continuidad empresarial para API Management
- Uso de zonas de disponibilidad en API Management
- Implementación en varias regiones de API Management
- Supervisar API Management
- Planeamiento de la capacidad de API Management
- Procedimientos recomendados de arquitectura para API Management

ⓘ **Nota:** El autor ha creado este artículo con ayuda de inteligencia artificial. Más información

---

Last updated on 22/01/2026

Confiabilidad en Azure App Configuration

Azure App Configuration almacena y administra centralmente las opciones de configuración de la aplicación y las marcas de características, que reemplazan los archivos de configuración incrustados directamente dentro de las aplicaciones. Con este enfoque, puede actualizar dinámicamente los valores de configuración, realizar un seguimiento del historial de versiones y mantener un registro de los cambios de configuración a lo largo del tiempo. La disponibilidad y confiabilidad de App Configuration son consideraciones importantes porque el comportamiento de la aplicación puede depender directamente del acceso a los datos de configuración en tiempo de ejecución.

Al usar Azure, relibilidad es una responsabilidad compartida. Microsoft proporciona una variedad de capacidades para admitir resiliencia y recuperación. Es responsable de comprender cómo funcionan esas funcionalidades dentro de todos los servicios que usa y de seleccionar las funcionalidades que necesita para cumplir los objetivos empresariales y los objetivos de tiempo de actividad.

En este artículo se describe la arquitectura de confiabilidad de App Configuration y cómo el servicio está diseñado para permanecer disponible durante errores transitorios, errores de zona de disponibilidad y interrupciones de región.

Recomendaciones de implementación de producción para la confiabilidad

Para la mayoría de las implementaciones de producción de App Configuration, tenga en cuenta las siguientes recomendaciones:

- ✓ **SKU:** Use la SKU Estándar o Premium.
- ✓ **Eliminación temporal y protección de purga:** Active la eliminación temporal y la protección de purga para ayudar a evitar la eliminación de datos.
- ✓ **Para escenarios críticos:** Use la SKU Premium y configure la réplica incluida para replicar en varias regiones. Este enfoque mejora la alta disponibilidad y la resistencia a las interrupciones de la región.

Para obtener una lista de los procedimientos recomendados y la configuración de las cargas de trabajo de producción, consulte Compilación de aplicaciones que tienen alta resistencia.

Introducción a la arquitectura de confiabilidad

Al implementar App Configuration, se implementa una *tienda*. Tu tienda contiene varios tipos de configuración que la aplicación puede usar, incluidas las claves y los valores, y las banderas de características. El servicio también incluye funcionalidades integradas para organizar, proteger, control de versiones y implementar de forma segura los cambios de configuración en todos los entornos. Para obtener más información, consulte ¿Qué es App Configuration?

App Configuration es un servicio totalmente administrado. Microsoft es responsable de realizar el mantenimiento en el servicio y almacenar y administrar la configuración.

Al crear aplicaciones cliente que se conectan a App Configuration, opcionalmente puede usar App Configuration con Azure Front Door (versión preliminar) para habilitar el caché y la aceleración global del tráfico. Esta configuración presenta otras consideraciones para la replicación geográfica, que se resaltan en este artículo cuando corresponda.

Resistencia a errores transitorios

Los errores transitorios son errores breves e intermitentes en los componentes. Se producen con frecuencia en un entorno distribuido como la nube y son una parte normal de las operaciones. Los errores transitorios se corrigen después de un breve período de tiempo. Es importante que las aplicaciones puedan controlar errores transitorios, normalmente mediante el reintento de solicitudes afectadas.

Todas las aplicaciones hospedadas en la nube deben seguir las instrucciones de control de errores transitorios Azure cuando se comunican con cualquier API, bases de datos y otros componentes hospedados en la nube. Para obtener más información, consulte Recomendaciones para controlar errores transitorios.

Al usar App Configuration, tenga en cuenta los siguientes procedimientos recomendados para minimizar el efecto de los errores transitorios en el acceso a la configuración, especialmente dentro de las rutas de acceso de código críticas.

- **Proveedores de configuración:** Use proveedores de App Configuration, que tienen integradas funcionalidades de reintento, almacenamiento en caché y otras capacidades de resiliencia.

- **SDK de Azure:** Utilice los SDK de App Configuration si su aplicación necesita enviar solicitudes de escritura. Los SDKs reintentan automáticamente en las respuestas de código de estado HTTP 429 y otros errores transitorios.

- **Lógica de reintento:** Incluya lógica de reintento en clientes personalizados si no puede utilizar proveedores o SDK de App Configuration. El `retry-after-ms` encabezado de la respuesta proporciona un tiempo de espera sugerido en milisegundos antes de que el cliente vuelva a intentar la solicitud.

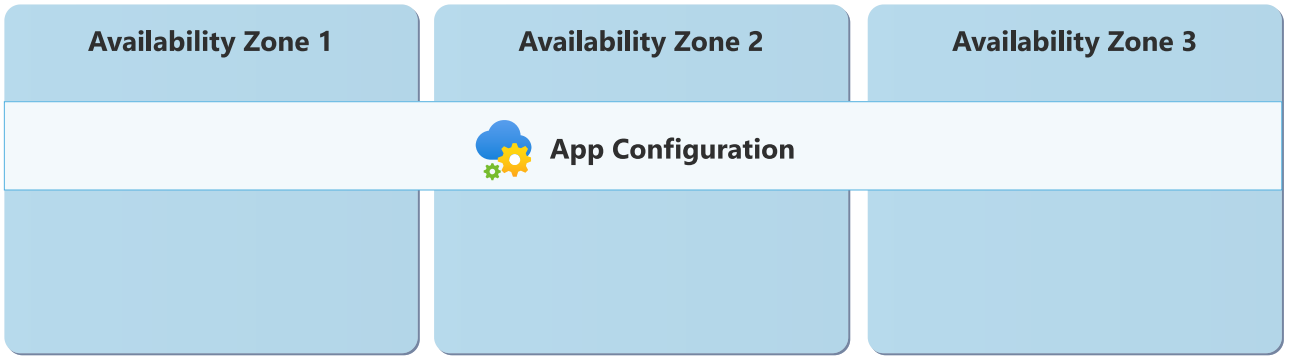
- **Almacenamiento en caché:** Configure los ajustes de caché en la memoria cuando sea posible para reducir las solicitudes directas a su tienda.

Para obtener otras instrucciones de configuración de aplicaciones, consulte Preguntas más frecuentes sobre App Configuration.

Resistencia a errores de zona de disponibilidad

Availability zones son grupos de centros de datos físicamente independientes dentro de una región de Azure. Cuando una zona falla, los servicios pueden transferirse a una de las zonas restantes.

App Configuration proporciona automáticamente redundancia de zona en regiones que admiten zonas de disponibilidad. Esta redundancia proporciona alta disponibilidad dentro de una región sin necesidad de ninguna configuración específica.



Cuando una zona de disponibilidad deja de estar disponible, App Configuration redirige automáticamente las solicitudes a otras zonas de disponibilidad correctas para garantizar una alta disponibilidad.

Requisitos

Compatibilidad con regiones: Los almacenes implementados en las siguientes regiones son automáticamente redundantes de zona.

Expandir tabla

Americas	Europa	Oriente Medio	África	Asia Pacífico
Sur de Brasil	Centro de Francia	Israel Central		Australia East
Canada Central	Centro-oeste de Alemania	Qatar Central		Centro de la India
Central US	Norte de Italia	Norte de Emiratos Árabes Unidos		Norte de China 3
East US	Norte de Europa			Este de Asia
Este de EE. UU. 2	Norway East			Japón Oriental
Centro de México	Centro de Polonia			Centro de Corea del Sur
Centro-sur de EE. UU.	Centro de España			Sudeste de Asia
Gobierno de EE. UU. - Virginia	Centro de Suecia			
Oeste de EE. UU. 2	Norte de Suiza			
Oeste de EE. UU. 3	Sur de Reino Unido			
	Oeste de Europa			

Cost

No hay ningún costo adicional para la redundancia de zona para App Configuration.

Configurar soporte de zonas de disponibilidad

Microsoft configura automáticamente la redundancia de zona para un almacén cuando se encuentra en a región que admite zonas de disponibilidad.

Si App Configuration agrega compatibilidad con zonas de disponibilidad a una región existente, no es necesario realizar ninguna acción para beneficiarse de la compatibilidad con la zona de disponibilidad. La tienda se beneficia de la compatibilidad con la zona de disponibilidad que está disponible para los almacenes de App Configuration en la región.

Comportamiento cuando todas las zonas están en buen estado

En esta sección describimos lo que puede esperar cuando tiene un almacén de App Configuration con redundancia de zona y todas las zonas están operativas.

- **Operación entre zonas:** App Configuration administra automáticamente el enrutamiento de tráfico entre zonas de disponibilidad. Durante las operaciones normales, distribuye de forma transparente las solicitudes entre zonas.
- **Replicación de datos entre zonas:** En las regiones que admiten zonas, App Configuration replica de forma sincrónica los datos entre zonas de disponibilidad. Esta replicación garantiza que la configuración siga siendo coherente y disponible incluso si una zona deja de estar disponible.

Comportamiento durante un fallo de zona

En esta sección se describe qué cabe esperar al usar un almacén de App Configuration con redundancia de zona y se produce una interrupción en una de las zonas.

- **Detección y respuesta:** El servicio App Configuration detecta errores de zona y responde automáticamente a ellos. No es necesario realizar ninguna acción ante un fallo de zona.
- **Notification:** Microsoft no le notifica automáticamente cuando una zona está inactiva. Sin embargo, puede usar Azure Service Health para comprender el estado general del servicio, incluidos los errores de zona, y puede configurar alertas de Service Health para notificarle problemas.
- **Solicitudes activas:** durante un error de zona, es posible que la zona afectada no controle las solicitudes en curso, lo que requiere que las aplicaciones cliente vuelvan a intentarlas. Las aplicaciones cliente deben seguir los procedimientos transitorios de control de errores para asegurarse de que pueden reintentar solicitudes si se produce un error de zona.
- **Pérdida de datos esperada:** No se espera ninguna pérdida de datos durante un error de zona debido a la replicación sincrónica entre zonas.
- **Tiempo de inactividad esperado:** No se espera ningún tiempo de inactividad.
- **Redistribución:** App Configuration vuelve a enrutar automáticamente el tráfico de la zona afectada a zonas correctas sin necesidad de intervención del cliente.

Recuperación de zona

Cuando se recupera una zona no disponible anteriormente, App Configuration restaura automáticamente las operaciones normales en todas las zonas de disponibilidad. No es necesario realizar ninguna acción para recuperar de una falla de zona.

Prueba de fallos de zona

La plataforma App Configuration gestiona el enrutamiento del tráfico, la conmutación ante fallos y la recuperación de zona para almacenes con redundancia de zona. Microsoft administra completamente este proceso, por lo que no es necesario validar los procesos de error de zona de disponibilidad.

Resistencia a errores en toda la región

App Configuration proporciona capacidades nativas de geo-replicación para admitir la resiliencia durante interrupciones regionales. La replicación geográfica permite que los datos de configuración se repliquen entre regiones como una característica de servicio administrado.

Replicación geográfica

Con la replicación geográfica, puede replicar un almacén en varias regiones de Azure. Cada almacén puede tener varias *réplicas* en regiones diferentes. El almacén original también es una réplica. Esta funcionalidad ayuda a proteger las aplicaciones frente a interrupciones en toda la región.

Requisitos

- **Region support:** Puede crear réplicas en cualquier región de Azure compatible con App Configuration, incluso si las regiones no son regiones emparejadas de Azure.
- **Nivel:** El almacén de configuración debe usar un nivel admitido para habilitar la replicación geográfica. Para obtener más información, consulte Habilitación de la replicación geográfica.

Consideraciones

Al habilitar la replicación geográfica, tenga en cuenta los siguientes factores:

- **Réplicas redundantes por zona:** Cualquier réplica creada en una región donde App Configuration admita zonas de disponibilidad será automáticamente redundante por zona.
- **Azure Front Door:** Para habilitar la entrega de configuración con redundancia geográfica con Azure Front Door, configure réplicas de App Configuration como orígenes dentro de un grupo de origen. Los orígenes configurados correctamente son necesarios para que Azure Front Door realice el enrutamiento basado en la salud, el equilibrio de carga y la

conmutación automática por error entre regiones. Para obtener más información, consulte Métodos de enrutamiento del tráfico al origen.

Cost

Cada región con replicación geográfica se factura por separado según los precios del plan y la región correspondientes. No se aplican cargos de salida de datos para la replicación entre regiones. Para obtener más información sobre los precios, consulte Precios de App Configuration.

Configuración de la compatibilidad con varias regiones

Para configurar la replicación para un almacén de configuración recién creado, consulte Habilitación de la replicación geográfica.

Comportamiento cuando todas las regiones están en buen estado

En esta sección se describe qué esperar al configurar un almacén de App Configuration para la replicación geográfica y todas las regiones están operativas.

- **Operación entre regiones:** Cada réplica es direccionable individualmente y tiene su propio nombre del sistema de nombres de dominio (DNS). Todas las réplicas pueden aceptar operaciones de lectura y escritura.

  App Configuration no enruta automáticamente el tráfico entre regiones. Cuando se usan proveedores de configuración de App Configuration, la aplicación puede usar opcionalmente la detección automática de réplicas. Alternativamente, puede especificar una lista priorizada de réplicas, y App Configuration selecciona la primera réplica saludable. Este enfoque permite a la aplicación controlar qué réplica usa.

  > **Nota:**
  >
  > Si usa Azure Front Door, el comportamiento del enrutamiento del tráfico es diferente. Para obtener más información, consulte Conmutación por error y equilibrio de carga.

- **Replicación de datos entre regiones:** Los datos se replican de forma asincrónica y finalmente son coherentes. Puede usar la métrica de latencia de replicación en Azure Monitor para supervisar la latencia de replicación actual entre réplicas.

Comportamiento durante una falla de región

En esta sección se describe qué se puede esperar al configurar una tienda de configuración de aplicaciones para la replicación geográfica, y cuando se produce una interrupción en una de las regiones de réplica.

- **Detección y respuesta:** Microsoft es responsable de detectar errores de región o réplica e iniciar procesos de recuperación.

  Cuando se usan proveedores de configuración de App Configuration con detección automática de réplicas o una lista de varias réplicas, la aplicación detecta automáticamente errores y realiza una conmutación en caso de error a una réplica sana.

  Si no usa proveedores de App Configuration, es responsable de cambiar su aplicación a una réplica saludable.

- **Notification:** Microsoft no le notifica automáticamente cuando una región está inactiva. Sin embargo, puede usar Azure Service Health para comprender el estado general del servicio, incluidos los errores de región, y puede configurar alertas de Service Health para notificarle problemas.

- **Solicitudes activas:** Las solicitudes activas contra una réplica en la región pueden fallar. Las aplicaciones cliente deben reintentar las solicitudes en una réplica diferente.

- **Pérdida de datos esperada:** Si se produce un error en una réplica, es posible que los cambios recientes realizados en esa réplica aún no se repliquen en otras réplicas. Esos cambios pueden permanecer no disponibles hasta que se recupere la réplica. Para calcular la posible pérdida de datos, supervise la métrica de latencia de replicación en Azure Monitor.

- **Tiempo de inactividad esperado:** Cuando una réplica deja de estar disponible, permanece sin conexión hasta que se recupera su región. Otras réplicas siguen gestionando las solicitudes. Las aplicaciones pueden experimentar un breve tiempo de inactividad mientras detectan el error y cambian a una réplica correcta. La duración depende de la rapidez con la que cada aplicación realiza esta detección y conmutación por error.

- **Redistribución:** Las aplicaciones deben enrutar el tráfico a una réplica saludable cuando se produce un error.

  Si utiliza proveedores de configuración de App Configuration, estos gestionan automáticamente la selección de réplica y la conmutación por error.

Si coloca Azure Front Door delante del almacenamiento de datos y configura el grupo de origen con varias réplicas como orígenes para la conmutación por error, Azure Front Door vuelve a enrutar automáticamente las solicitudes a una réplica en buen estado.

Recuperación de regiones

Una vez que la región se recupera, App Configuration sincroniza la réplica con las demás réplicas sin intervención del usuario.

Usted es responsable de volver a configurar su aplicación para enrutar el tráfico de regreso a la instancia de la región recuperada. Las aplicaciones que utilizan App Configuration reanudan automáticamente el uso de la réplica.

Prueba de fallos de región

No se puede simular una conmutación por error de réplica directamente en App Configuration. Sin embargo, dado que las aplicaciones controlan la selección de réplica, puede probar el comportamiento de la conmutación por error forzando la aplicación a un estado en el que debe cambiar las réplicas.

Para validar el comportamiento de failover de réplica de su aplicación, puede introducir un error de conectividad controlado en un entorno de no producción y observar la respuesta de la aplicación.

Un enfoque consiste en usar la máquina local u otro entorno en el que tenga acceso administrativo. Siga estos pasos:

1. Active el registro exhaustivo para el SDK de Azure. En .NET, use la clase `AzureEventSourceListener` para configurar un registrador. Para obtener más información, vea el documento sobre supervisión y registro.

2. Configure manualmente el `hosts` archivo para que las solicitudes al almacén de App Configuration se enruten a una dirección IP que no pueda recibirlos, como `127.0.0.1` (localhost).

   > ⚠ **Advertencia**
   >
   > Este paso bloquea eficazmente el acceso desde tu equipo a tu almacén de Configuración de la Aplicación. Siga estos pasos solo en un entorno que no sea de producción.

3. Supervise los registros de un mensaje similar al ejemplo siguiente:

   Consola

   ```
   [Warning] Microsoft-Extensions-Configuration-AzureAppConfiguration-Refresh:
   Failed to get configuration settings from endpoint
   'https://myappconfigstore.azconfig.io'. Failing over to endpoint
   https://myappconfigstore-eus.azconfig.io'.
   ```

   Este mensaje indica que la aplicación conmutó con éxito para usar otra réplica de su almacén.

4. Después de completar la prueba, deshaga los cambios en el archivo `hosts`.

Copias de seguridad y restauración

Puede usar App Configuration para exportar datos de configuración desde un almacén y usarlos como parte de una estrategia de copia de seguridad más amplia.

Para la mayoría de las soluciones, no debe confiar exclusivamente en copias de seguridad. En su lugar, utilice las otras capacidades descritas en esta guía para apoyar los requisitos de resiliencia. Sin embargo, las copias de seguridad protegen contra algunos riesgos que otros enfoques no. Para más información, consulte ¿Qué son la redundancia, la replicación y la copia de seguridad?.

Resistencia a la eliminación accidental

App Configuration proporciona dos características clave de recuperación para evitar la eliminación accidental o malintencionada:

- **Eliminación temporal:** Al activar la eliminación temporal, puede recuperar almacenes y configuraciones eliminados durante un período de retención configurable. Las funciones de borrado suave actúan como una papelera de reciclaje para los recursos de App Configuration.

- **Protección de purga:** Al activar la protección de purga, el servicio evita la eliminación permanente del almacén y su configuración hasta que transcurre el período de retención. Esta protección evita que los actores malintencionados destruyan permanentemente la configuración.

Use ambas características para entornos de producción. Para más información, consulte Eliminación temporal y protección contra purgas.

Resistencia al mantenimiento del servicio

Microsoft realiza periódicamente actualizaciones de servicio y otro mantenimiento. El servicio controla estas actividades automáticamente, lo que hace que el mantenimiento sea transparente y sin problemas. No se espera ningún tiempo de inactividad durante los eventos de mantenimiento a menos que Azure Service Health proporcione un aviso de mantenimiento planeado.

Resistencia a problemas de configuración

Los cambios de configuración incorrectos o accidentales pueden provocar tiempo de inactividad de la aplicación. Utiliza instantáneas de configuración para desplegar los cambios en la configuración de forma segura. Supervise el estado de la aplicación después de cualquier cambio de configuración y vuelva a la última instantánea de configuración correcta conocida si los cambios presentan un problema.

Acuerdo de nivel de servicio

El acuerdo de nivel de servicio (SLA) para Azure servicios describe la disponibilidad esperada de cada servicio y las condiciones que la solución debe cumplir para lograr esa expectativa de disponibilidad. Para obtener más información, vea SLAs for servicios en línea.

Contenido relacionado

- Confiabilidad en Azure
- Documentación de Configuración de la aplicación

ⓘ**Nota:** El autor ha creado este artículo con ayuda de inteligencia artificial. Más información

---

Last updated on 15/04/2026

Confiabilidad en Azure App Service

Azure App Service es un servicio basado en HTTP para hospedar aplicaciones web, API REST y back-end móvil. App Service se integra con Microsoft Azure para proporcionar seguridad, equilibrio de carga, escalado automático y administración automatizada para aplicaciones. Como servicio de Azure, App Service proporciona una variedad de funcionalidades para satisfacer los requisitos de confiabilidad.

Cuando se usa Azure, la confiabilidad es una responsabilidad compartida. Microsoft proporciona una variedad de funcionalidades para admitir resistencia y recuperación. Es responsable de comprender cómo funcionan esas funcionalidades dentro de todos los servicios que usa y de seleccionar las funcionalidades que necesita para cumplir los objetivos empresariales y los objetivos de tiempo de actividad.

En este artículo se describe cómo hacer que App Service sea resistente a una variedad de posibles interrupciones y problemas, incluidos errores transitorios, interrupciones de zona de disponibilidad, interrupciones de regiones y mantenimiento del servicio. También se describe cómo puede usar copias de seguridad para recuperarse de otros tipos de problemas y se resalta cierta información clave sobre el acuerdo de nivel de servicio (SLA) de App Service.

> **Nota:**
>
> Si busca información sobre el soporte de confiabilidad en App Service Environment, consulte Confiabilidad en App Service Environment.

Recomendaciones de implementación de producción

Azure Well-Architected Framework proporciona recomendaciones sobre confiabilidad, rendimiento, seguridad, costo y operaciones. Para comprender cómo estas áreas influyen entre sí y contribuir a una solución confiable de App Service, consulte Procedimientos recomendados de arquitectura para App Service (Web Apps) en Azure Well-Architected Framework.

Introducción a la arquitectura de confiabilidad

Al crear una aplicación web de App Service, se especifica el plan de App Service que ejecuta la aplicación.

Un plan de App Service define un conjunto de recursos de proceso que ejecutan las aplicaciones web. Todas las Web Apps deben ejecutarse dentro de un plan. Puede escalar un

plan para ejecutarse en varias *instancias* de máquina virtual, también denominadas *trabajadores*. Estas instancias proporcionan los recursos de proceso que ejecutan el código de su aplicación. Un único plan de App Service puede hospedar varias aplicaciones. Todas las aplicaciones se ejecutan en el mismo conjunto compartido de instancias de máquina virtual.

App Service ofrece las siguientes características de redundancia:

- **Distribución entre dominios de error:** En el nivel de plataforma, Azure distribuye automáticamente las instancias de máquina virtual del plan de App Service entre dominios de error dentro de la región de Azure. Esta distribución minimiza el riesgo de errores de hardware localizados mediante la agrupación de máquinas virtuales que comparten una fuente de alimentación común y un conmutador de red.

- **Distribución entre zonas de disponibilidad:** Si habilita la redundancia de zona en un plan de App Service compatible, Azure distribuye las instancias entre zonas de disponibilidad dentro de la región. Esta configuración proporciona una mayor resistencia si se produce una interrupción de zona. Para obtener más información sobre la redundancia de zonas, consulte Compatibilidad con zonas de disponibilidad.

- **Escalado de aplicaciones:** Al configurar el plan de App Service para ejecutar varias instancias de máquina virtual, todas las aplicaciones del plan se ejecutan en todas las instancias de forma predeterminada. Si configura su plan para el escalado automático, todas las aplicaciones se escalarán juntas según la configuración de escalado automático. Sin embargo, puede personalizar el número de instancias de plan que ejecutan una aplicación específica mediante el escalado por aplicación.

- **Unidades de escalado:** Internamente, App Service se ejecuta en una infraestructura de plataforma denominada *unidades de escalado*, también conocida como *stamps* o *espacios web*. Una unidad de escalado incluye todos los componentes necesarios para hospedar y ejecutar App Service, incluido el proceso, el almacenamiento, las redes y el equilibrio de carga. Azure administra las unidades de escalado para garantizar una distribución equilibrada de cargas de trabajo, realizar un mantenimiento rutinario y mantener la confiabilidad general de la plataforma.

  Es posible que algunas funcionalidades solo se apliquen a unidades de escalado específicas. Por ejemplo, algunas unidades de escalado de App Service pueden admitir la redundancia de zona, mientras que otras unidades de escala de la misma región no.

Resistencia a errores transitorios

Los errores transitorios son errores breves e intermitentes en los componentes. Se producen con frecuencia en un entorno distribuido como la nube y son una parte normal de las

operaciones. Los errores transitorios se corrigen después de un breve período de tiempo. Es importante que las aplicaciones puedan controlar errores transitorios, normalmente mediante el reintento de solicitudes afectadas.

Todas las aplicaciones hospedadas en la nube deben seguir las instrucciones de control de errores transitorios de Azure cuando se comunican con cualquier API, bases de datos y otros componentes hospedados en la nube. Para obtener más información, consulte Recomendaciones para controlar errores transitorios.

Los SDK proporcionados por Microsoft suelen controlar errores transitorios. Dado que hospeda sus propias aplicaciones en App Service, siga los pasos necesarios para reducir la posibilidad de errores transitorios:

- **Implemente varias instancias en el plan.** App Service realiza actualizaciones automatizadas y otras formas de mantenimiento en instancias del plan. Si una instancia se vuelve incorrecta, el servicio puede reemplazar automáticamente esa instancia por una nueva instancia correcta. Durante el proceso de reemplazo, puede haber un breve período cuando la instancia anterior no está disponible y una nueva instancia no está lista para atender el tráfico. Para mitigar estos efectos, implemente varias instancias del plan de App Service.

- **Uso de ranuras de implementación.** Las ranuras de implementación de App Service permiten implementaciones sin tiempo de inactividad de las aplicaciones. Use ranuras de implementación para minimizar el efecto de las implementaciones y los cambios de configuración para los usuarios. Las ranuras de implementación también reducen la probabilidad de que se reinicie la aplicación. Reiniciar la aplicación produce un error transitorio.

- **Evite escalar o reducir verticalmente.** Estas operaciones modifican la CPU, la memoria y otros recursos asignados a cada instancia, y pueden provocar el reinicio de la aplicación empresarial. En su lugar, seleccione un nivel y un tamaño de instancia que cumplan los requisitos de rendimiento bajo una carga típica. Para escalar y reducir horizontalmente, agregue y elimine instancias de forma dinámica para administrar los cambios en el volumen de tráfico.

Resistencia a errores de zona de disponibilidad

Las zonas de disponibilidad son grupos físicamente independientes de centros de datos dentro de una región de Azure. Cuando una zona falla, los servicios pueden transferirse a una de las zonas restantes.

En el caso de los **niveles Premium v2 a v4**, puede configurar App Service como *redundante de zona*, lo que significa que los recursos se distribuyen entre varias zonas de disponibilidad. La distribución entre varias zonas ayuda a las cargas de trabajo de producción a lograr resistencia y confiabilidad. Al configurar la redundancia de zona en los planes de App Service, todas las aplicaciones que utilizan el plan pasan a ser redundantes en la zona.

Requisitos

Para habilitar la redundancia de zona, debe cumplir los siguientes requisitos:

- **Compatibilidad con regiones:** En el caso de los planes de App Service **Premium v2 y v3** , la redundancia de zona se admite en cualquier región que admita zonas de disponibilidad.
- **Tipo de plan:** Use tipos de plan Premium v2 a v4.

  > ⓘ **Importante**
  >
  > Para habilitar la redundancia de zona para los planes **de App Service Premium v4** , debe **confirmar que la región deseada admite planes v4** y que **admite zonas de disponibilidad**.

- **Número mínimo de instancias:** Implemente un mínimo de dos instancias en el plan.
- **Unidad de escalado:** La aplicación debe implementarse en una unidad de escalado que admita zonas de disponibilidad. No controla directamente la unidad de escalado que utiliza el plan. En su lugar, al crear un plan de App Service, el plan se asigna a una unidad de escalado en función del grupo de recursos del propio plan. Para determinar si la unidad de escalado del plan de App Service admite redundancia de zona, consulte Comprobar la compatibilidad con redundancia de zona para un plan de App Service.

  Si tu plan de App Service está en una unidad de escalado que no admite redundancia de zona, no puedes habilitar la redundancia de zona en tu plan. En su lugar, debe reimplementar tus aplicaciones en un nuevo plan en una unidad de escala diferente.

Distribución de instancias a través de zonas

Al crear un plan de App Service con redundancia de zona, Azure distribuye las instancias del plan entre zonas de disponibilidad de la región. Esta distribución garantiza que sus aplicaciones sigan estando disponibles incluso si una zona sufre una interrupción del servicio.

La distribución de instancias en una implementación con redundancia de zona sigue reglas específicas. Estas reglas también se aplican a medida que la aplicación se escala y se reduce horizontalmente:

- **Instancias mínimas:** El plan de App Service debe tener un mínimo de dos instancias para la redundancia de zona.
- **Zonas de disponibilidad máxima compatibles con el plan:** Azure determina el número de zonas de disponibilidad que puede usar el plan, lo que se conoce como *maximumNumberOfZones*. Para ver el número de zonas de disponibilidad que puede usar su plan específico, consulte Comprobar la compatibilidad con la redundancia de zonas para un plan de App Service.
- **Distribución de instancias:** Cuando se habilita la redundancia de zona, Azure distribuye instancias de plan entre varias zonas de disponibilidad automáticamente. La distribución se basa en las reglas siguientes:
  - Si el número de instancias supera el *maximumNumberOfZones* y divide uniformemente, Azure distribuye las instancias uniformemente entre zonas.
  - Si el número de instancias no se divide de manera uniforme, Azure distribuye las instancias restantes entre las zonas restantes.
  - Cuando la plataforma de App Service asigna instancias para un plan de App Service con redundancia de zona, usa el equilibrio de zona de mejor esfuerzo que proporcionan los conjuntos de escalado de máquinas virtuales de Azure subyacentes. Un plan se equilibra si cada zona tiene el mismo número de máquinas virtuales o difiere en una instancia de todas las demás zonas. Para obtener más información, consulte Equilibrio de zona.
- **Ubicación de zona física:** Puede ver la zona de disponibilidad física que se usa para cada una de las instancias del plan de App Service. Para más información, consulte Visualización de zonas físicas para un plan de App Service.

Consideraciones

En el caso de los planes **Premium v2 a v4**, una interrupción de zona de disponibilidad podría afectar a algunos aspectos de Azure App Service, aunque la aplicación siga sirviendo tráfico. Estos comportamientos incluyen el escalado del plan de App Service, la creación de aplicaciones, la configuración de la aplicación y la publicación de aplicaciones.

Al habilitar la redundancia de zona en su plan App Service **Premium v2 a v4**, también mejora la resistencia durante las actualizaciones de la plataforma. Para obtener más información,

consulte Resistencia al mantenimiento del servicio.

En los planes de App Service que no están configurados como redundantes por zona, las instancias de máquina virtual (VM) subyacentes no son resistentes a los errores de la zona de disponibilidad. Pueden experimentar tiempo de inactividad durante una interrupción en cualquier zona de esa región.

Cost

Cuando use los **planes Premium v2 a v4 de App Service**, habilitar las zonas de disponibilidad no supone ningún coste adicional si tiene dos o más instancias. Los cargos se basan en la SKU de su plan de App Service, la capacidad que especifique y las instancias a las que escale según sus criterios de escalado automático.

Si habilita zonas de disponibilidad pero especifica una capacidad de menos de dos, la plataforma aplica un recuento mínimo de instancias de dos. La plataforma le cobra por esas dos instancias.

Configurar soporte de zonas de disponibilidad

- **Cree un nuevo plan de App Service con redundancia de zona.** Para más información, consulte Crear un nuevo plan de App Service que incluya redundancia de zona.
- **Habilite o deshabilite la redundancia de zona en un plan de App Service existente.** Para obtener más información, consulte Configurar la redundancia de zona para un plan de App Service existente.

Planeamiento y administración de capacidad

Para prepararse para errores de zona de disponibilidad, considere la posibilidad de *sobreaprovisionar* la capacidad del plan de App Service. Este enfoque permite que la solución tolere cierta pérdida de capacidad y siga funcionando sin que se vea afectado su rendimiento. Para obtener más información, consulte Administración de la capacidad mediante el aprovisionamiento excesivo.

Comportamiento cuando todas las zonas están en buen estado

En la siguiente lista se describe qué esperar cuando los planes de App Service están configurados para la redundancia de zona y todas las zonas de disponibilidad están operativas:

- **Enrutamiento del tráfico entre zonas:** Durante el funcionamiento normal, el tráfico se enruta entre todas las instancias disponibles del plan de App Service en todas las zonas de disponibilidad.

- **Replicación de datos entre zonas:** Durante las operaciones normales, cualquier estado almacenado en el sistema de archivos de la aplicación se almacena en el almacenamiento con redundancia de zona y se replica sincrónicamente entre zonas de disponibilidad.

Comportamiento durante un fallo de zona

Una interrupción de la zona de disponibilidad puede afectar a algunos aspectos de App Service, aunque la aplicación siga sirviendo al tráfico. Estos comportamientos incluyen el escalado del plan de App Service, la creación de aplicaciones, la configuración de la aplicación y la publicación de aplicaciones.

La siguiente lista describe qué puede ocurrir cuando los planes de App Service están configurados para la redundancia de zona y una o varias zonas de disponibilidad no están disponibles:

- **Detección y respuesta:** La plataforma de App Service detecta automáticamente errores en una zona de disponibilidad e inicia una respuesta. No se requiere ninguna intervención manual para iniciar una conmutación por error de zona.

- **Notificación**: Microsoft no le notifica automáticamente cuando una zona está inactiva. Sin embargo, puede usar Azure Resource Health para supervisar el estado de un recurso individual y puede configurar alertas de Resource Health para notificarle problemas. También puede usar Azure Service Health para comprender el estado general del servicio, incluidos los errores de zona, y puede configurar alertas de Service Health para notificarle problemas.

- **Solicitudes activas:** Las solicitudes en curso que se conectan a una instancia del plan de App Service en la zona de disponibilidad defectuosa se finalizan. Reintentar esas solicitudes.

- **Enrutamiento del tráfico:**App Service detecta las instancias perdidas de esa zona e intenta encontrar nuevas instancias de reemplazo Después de que App Service encuentre reemplazos, distribuye el tráfico entre las nuevas instancias según sea necesario.

  Si se configura el ajuste automático y se determina que se necesitan más instancias, se solicitan instancias a App Service. El comportamiento de escalabilidad automática funciona independientemente del comportamiento de la plataforma de App Service. Por lo tanto, la especificación del número de instancias no tiene por qué ser un múltiplo de

dos. Para más información, consulte Escalado vertical de una aplicación en App Service y Información general sobre el escalado automático.

> **Importante**
>
> Azure no garantiza que las solicitudes de más instancias se realicen correctamente en un escenario de zona descendente. La plataforma intenta reponer instancias perdidas de forma óptima. Si necesita capacidad garantizada durante un error de zona de disponibilidad, cree y configure sus planes de App Service para tener en cuenta la pérdida de zona mediante el **sobreaprovisionamiento de la capacidad**.

- **Comportamientos fuera del tiempo de ejecución:** Las aplicaciones en un plan de App Service con redundancia de zona continúan ejecutándose y prestando servicio al tráfico incluso si una zona de disponibilidad sufre una interrupción. Sin embargo, los comportamientos que no son de ejecución pueden verse afectados durante una interrupción de la zona de disponibilidad. Estos comportamientos incluyen el escalado del plan de App Service, la creación de aplicaciones, la configuración de la aplicación y la publicación de aplicaciones.

Recuperación de zona

Cuando se recupera la zona de disponibilidad, App Service crea automáticamente instancias en la zona de disponibilidad recuperada, quita las instancias temporales creadas en las otras zonas de disponibilidad y enruta el tráfico entre las instancias como de costumbre.

Prueba de fallos de zona

La plataforma de App Service administra el enrutamiento del tráfico, la conmutación por error y la conmutación por recuperación para planes de App Service con redundancia de zona. Esta característica está totalmente administrada, por lo que no es necesario iniciar ni validar los procesos de error de zona de disponibilidad.

Resistencia a errores en toda la región

App Service es un servicio de una sola región. Si la región deja de estar disponible, la aplicación tampoco está disponible.

Soluciones personalizadas de varias regiones para la resistencia

Para reducir el riesgo de un error de una sola región que afecte a la aplicación, puede implementar planes en varias regiones. Los pasos siguientes ayudan a reforzar la resistencia:

- Despliegue su aplicación en los planes de cada región.
- Configure directivas de equilibrio de carga y conmutación por error.
- Replique los datos entre regiones para que pueda recuperar el último estado de la aplicación.

Considere los siguientes recursos relacionados:

- Arquitectura de referencia: aplicación web de alta disponibilidad en múltiples regiones
- Enfoques que se deben tener en cuenta
- Tutorial: Crear una aplicación de alta disponibilidad en varias regiones en App Service

Copias de seguridad y restauración

Cuando use el nivel Básico o superior, puede hacer una copia de seguridad de sus aplicaciones de App Service en un archivo usando las funciones de copia de seguridad y restauración de App Service.

Estas funcionalidades resultan útiles cuando resulta complicado volver a implementar el código o cuando se almacena el estado en el disco. La mayoría de las soluciones no deberían depender exclusivamente de las copias de seguridad. En su lugar, use las otras capacidades de esta guía para satisfacer sus requisitos de resistencia. Sin embargo, las copias de seguridad protegen contra algunos riesgos que otros enfoques no.

> ⓘ **Importante**
>
> A partir del **31 de marzo de 2028**, las copias de seguridad personalizadas de Azure App Service **ya no admiten la copia de seguridad de bases de datos vinculadas**. Consulte Desuso de copias de seguridad de bases de datos vinculadas para obtener más información.
>
> En su lugar, use las herramientas nativas de copia de seguridad y restauración de la base de datos vinculada. Para más información, consulte Copia de seguridad y restauración de la aplicación en App Service.

Resistencia al mantenimiento del servicio

App Service realiza actualizaciones de servicio periódicas y otras tareas de mantenimiento. Para mantener la capacidad esperada durante una actualización, la plataforma agrega

automáticamente instancias adicionales del plan de App Service durante el proceso de actualización.

Habilitación de la redundancia de zona. Al habilitar la redundancia de zona en su plan de App Service, también mejora la resiliencia durante las actualizaciones de la plataforma. *Los dominios de actualización* constan de colecciones de máquinas virtuales que se desconectan durante una actualización y se asignan a zonas de disponibilidad. La implementación de varias instancias en su plan de App Service y la habilitación de la redundancia de zona para su plan agregan una capa adicional de resistencia en caso de que una instancia o zona sea incorrecta durante una actualización.

Para obtener más información, consulte Mantenimiento planeado rutinario para App Service y Mantenimiento rutinario para App Service, reinicios y tiempo de inactividad.

Acuerdo de nivel de servicio

El contrato de nivel de servicio (SLA) para los servicios de Azure describe la disponibilidad esperada de cada servicio y las condiciones que la solución deberá cumplir para lograr esa expectativa de disponibilidad. Para obtener más información, consulte Acuerdos de Nivel de Servicio para servicios en línea.

Al implementar un plan de App Service con redundancia de zona, aumenta el porcentaje de tiempo de actividad definido en el Acuerdo de Nivel de Servicio.

Contenido relacionado

- Confiabilidad en Azure
- Confiabilidad en App Service Environment

Last updated on 22/01/2026

Confiabilidad en Azure App Service Environment

App Service Environment es una característica de Azure App Service que proporciona un entorno totalmente aislado y dedicado para ejecutar aplicaciones de App Service de forma segura a gran escala. Como servicio de Azure, App Service Environment proporciona una variedad de funcionalidades para satisfacer sus requisitos de confiabilidad.

Cuando se usa Azure, la confiabilidad es una responsabilidad compartida. Microsoft proporciona una variedad de funcionalidades para admitir resistencia y recuperación. Es responsable de comprender cómo funcionan esas funcionalidades dentro de todos los servicios que usa y de seleccionar las funcionalidades que necesita para cumplir los objetivos empresariales y los objetivos de tiempo de actividad.

En este artículo se describe el soporte de confiabilidad en App Service Environment, incluida la resistencia a errores transitorios, fallos en zonas de disponibilidad e interrupciones a nivel regional. También se describen las estrategias de copia de seguridad y la resistencia al mantenimiento del servicio y se resalta cierta información clave sobre el acuerdo de nivel de servicio (SLA) de App Service Environment.

> ⓘ **Nota:**
>
> A diferencia de la oferta multiinquilino pública de App Service que comparte la infraestructura de soporte, una instancia de App Service Environment proporciona recursos de proceso dedicados y aislamiento de red mejorado para un único cliente.
>
> Un entorno ofrece las siguientes ventajas clave en materia de confiabilidad:
>
> - Recursos de proceso dedicados que no se comparten con otros clientes
> - Aislamiento de red mejorado para mejorar la seguridad y la estabilidad
> - La capacidad de implementar en su propia red virtual para un mayor control sobre el enrutamiento del tráfico y las directivas de seguridad.
>
> Para más información sobre la compatibilidad con la confiabilidad en App Service, consulte Confiabilidad en App Service.

Recomendaciones de implementación de producción para la confiabilidad

Se recomienda habilitar la redundancia de zona en el entorno y los planes de App Service, lo que requiere que los planes usen un mínimo de dos instancias.

Introducción a la arquitectura de confiabilidad

Cuando implementa un App Service Environment, implemente el entorno como contenedor para sus planes de App Service y aplicaciones web. Durante la configuración, configure los ajustes básicos de red y el aislamiento opcional del hardware. Elija si desea admitir la redundancia de zona en el entorno si la región admite zonas de disponibilidad.

Después de crear su entorno, puede crear uno o varios planes de App Service.

Un plan de App Service define un conjunto de recursos de proceso que ejecutan las aplicaciones web. Todas las Web Apps deben ejecutarse dentro de un plan. Puede escalar un plan para ejecutarse en varias *instancias* de máquina virtual, también denominadas *trabajadores*. Estas instancias proporcionan los recursos de proceso que ejecutan el código de su aplicación. Un único plan de App Service puede hospedar varias aplicaciones. Todas las aplicaciones se ejecutan en el mismo conjunto compartido de instancias de máquina virtual.

Para usar App Service Environment, sus planes deben usar el nivel de precios Aislado v2. Este nivel admite redundancia de zona y aplicaciones críticas a gran escala.

App Service ofrece las siguientes características de redundancia:

- **Distribución entre dominios de error:** En el nivel de plataforma, Azure distribuye automáticamente las instancias de máquina virtual del plan de App Service entre dominios de error dentro de la región de Azure. Esta distribución minimiza el riesgo de errores de hardware localizados mediante la agrupación de máquinas virtuales que comparten una fuente de alimentación común y un conmutador de red.

- **Distribución entre zonas de disponibilidad:** Si habilita la redundancia de zona en un plan de App Service compatible, Azure distribuye las instancias entre zonas de disponibilidad dentro de la región. Esta configuración proporciona una mayor resistencia si se produce una interrupción de zona. Para obtener más información sobre la redundancia de zonas, consulte Compatibilidad con zonas de disponibilidad.

- **Escalado de aplicaciones:** Al configurar el plan de App Service para ejecutar varias instancias de máquina virtual, todas las aplicaciones del plan se ejecutan en todas las instancias de forma predeterminada. Si configura su plan para el escalado automático, todas las aplicaciones se escalarán juntas según la configuración de escalado automático. Sin embargo, puede personalizar el número de instancias de plan que ejecutan una aplicación específica mediante el escalado por aplicación.

- **Unidades de escalado:** Internamente, App Service se ejecuta en una infraestructura de plataforma denominada *unidades de escalado*, también conocida como *stamps* o *espacios web*. Una unidad de escalado incluye todos los componentes necesarios para hospedar y ejecutar App Service, incluido el proceso, el almacenamiento, las redes y el equilibrio de carga. Azure administra las unidades de escalado para garantizar una distribución equilibrada de cargas de trabajo, realizar un mantenimiento rutinario y mantener la confiabilidad general de la plataforma.

  Es posible que algunas funcionalidades solo se apliquen a unidades de escalado específicas. Por ejemplo, algunas unidades de escalado de App Service pueden admitir la redundancia de zona, mientras que otras unidades de escala de la misma región no.

Resistencia a errores transitorios

Los errores transitorios son errores breves e intermitentes en los componentes. Se producen con frecuencia en un entorno distribuido como la nube y son una parte normal de las operaciones. Los errores transitorios se corrigen después de un breve período de tiempo. Es importante que las aplicaciones puedan controlar errores transitorios, normalmente mediante el reintento de solicitudes afectadas.

Todas las aplicaciones hospedadas en la nube deben seguir las instrucciones de control de errores transitorios de Azure cuando se comunican con cualquier API, bases de datos y otros componentes hospedados en la nube. Para obtener más información, consulte Recomendaciones para controlar errores transitorios.

Los SDK proporcionados por Microsoft suelen controlar errores transitorios. Dado que hospeda sus propias aplicaciones en App Service, siga los pasos necesarios para reducir la posibilidad de errores transitorios:

- **Implemente varias instancias en el plan.** App Service realiza actualizaciones automatizadas y otras formas de mantenimiento en instancias del plan. Si una instancia se vuelve incorrecta, el servicio puede reemplazar automáticamente esa instancia por una nueva instancia correcta. Durante el proceso de reemplazo, puede haber un breve período cuando la instancia anterior no está disponible y una nueva instancia no está lista para atender el tráfico. Para mitigar estos efectos, implemente varias instancias del plan de App Service.

- **Uso de ranuras de implementación.** Las ranuras de implementación de App Service permiten implementaciones sin tiempo de inactividad de las aplicaciones. Use ranuras de implementación para minimizar el efecto de las implementaciones y los cambios de configuración para los usuarios. Las ranuras de implementación también reducen la

probabilidad de que se reinicie la aplicación. Reiniciar la aplicación produce un error transitorio.

- **Evite escalar o reducir verticalmente.** Estas operaciones modifican la CPU, la memoria y otros recursos asignados a cada instancia, y pueden provocar el reinicio de la aplicación empresarial. En su lugar, seleccione un nivel y un tamaño de instancia que cumplan los requisitos de rendimiento bajo una carga típica. Para escalar y reducir horizontalmente, agregue y elimine instancias de forma dinámica para administrar los cambios en el volumen de tráfico.

Resistencia a errores de zona de disponibilidad

Las zonas de disponibilidad son grupos físicamente independientes de centros de datos dentro de una región de Azure. Cuando una zona falla, los servicios pueden transferirse a una de las zonas restantes.

Puede configurar su App Service Environment como *redundante por zona*. También puede configurar sus planes de App Service para que sean redundantes por zona, lo que los distribuye entre varias zonas de disponibilidad.

Sin embargo, puede habilitar o deshabilitar la redundancia de zona en cada plan. Esto significa que puede tener algunos planes en el entorno con redundancia de zona y otros que no.

Al crear un plan de App Service con redundancia de zona en el entorno, las instancias del plan de App Service se distribuyen entre las zonas de disponibilidad de la región. Para obtener más información, consulte Distribución de instancias entre zonas.

Requisitos

Para habilitar la redundancia de zona para su App Service Environment, debe cumplir los siguientes requisitos:

- **Compatibilidad con regiones:** Para ver qué regiones admiten zonas de disponibilidad para App Service Environment v3, consulte Regiones.
- **Tipo de plan:** Utilice tipos de plan Aislado versión 2.
- **Número mínimo de instancias:** Implemente un mínimo de dos instancias en el plan.
- **Unidad de escalado:** El entorno debe implementarse en una unidad de escalado que admita zonas de disponibilidad. No controla directamente la unidad de escalado que usa el entorno. En su lugar, al crear un entorno de App Service, el entorno se asigna a una unidad de escalado en función del grupo de recursos del entorno. Para determinar si la

unidad de escalado de App Service Environment admite redundancia de zona, compruebe si hay compatibilidad con la redundancia de zona para un entorno de App Service.

Si su entorno está en una unidad de escalado que no admite zonas de disponibilidad, no puede habilitar la redundancia de zona en su entorno ni en sus planes. En su lugar, debe crear un nuevo entorno en un nuevo grupo de recursos y volver a implementar las aplicaciones en nuevos planes dentro de ese entorno.

- **Requisitos de configuración:** Configure tanto su entorno del servicio de aplicaciones como sus planes para admitir la redundancia de zona. Puede habilitar la redundancia de zona durante la creación del entorno o actualizando un entorno existente.

Distribución de instancias a través de zonas

Al crear un plan de App Service con redundancia de zona, Azure distribuye las instancias del plan entre zonas de disponibilidad de la región. Esta distribución garantiza que sus aplicaciones sigan estando disponibles incluso si una zona sufre una interrupción del servicio.

La distribución de instancias en una implementación con redundancia de zona sigue reglas específicas. Estas reglas también se aplican a medida que la aplicación se escala y se reduce horizontalmente:

- **Instancias mínimas:** El plan de App Service debe tener un mínimo de dos instancias para la redundancia de zona.

- **Zonas de disponibilidad máxima compatibles con el plan:** Azure determina el número de zonas de disponibilidad que puede usar el plan, lo que se conoce como *maximumNumberOfZones*. Para ver el número de zonas de disponibilidad que puede usar su plan específico, consulte Comprobar la compatibilidad con la redundancia de zonas para un plan de App Service.

- **Distribución de instancias:** Cuando se habilita la redundancia de zona, Azure distribuye instancias de plan entre varias zonas de disponibilidad automáticamente. La distribución se basa en las reglas siguientes:

  - Si el número de instancias supera el *maximumNumberOfZones* y divide uniformemente, Azure distribuye las instancias uniformemente entre zonas.

  - Si el número de instancias no se divide de manera uniforme, Azure distribuye las instancias restantes entre las zonas restantes.

  - Cuando la plataforma de App Service asigna instancias para un plan de App Service con redundancia de zona, usa el equilibrio de zona de mejor esfuerzo que

proporcionan los conjuntos de escalado de máquinas virtuales de Azure subyacentes. Un plan se equilibra si cada zona tiene el mismo número de máquinas virtuales o difiere en una instancia de todas las demás zonas. Para obtener más información, consulte Equilibrio de zona.

- **Ubicación de zona física:** Puede ver la zona de disponibilidad física que se usa para cada una de las instancias del plan de App Service. Para más información, consulte Visualización de zonas físicas para un plan de App Service.

Consideraciones

Una interrupción de la zona de disponibilidad puede afectar a algunos aspectos de App Service, aunque la aplicación siga sirviendo al tráfico. Estos comportamientos incluyen el escalado del plan de App Service, la creación de aplicaciones, la configuración de la aplicación y la publicación de aplicaciones.

Al habilitar la redundancia de zona en su plan de App Service, también mejora la resiliencia durante las actualizaciones de la plataforma. Para obtener más información, consulte Resistencia al mantenimiento del servicio.

En los planes de App Service que no son redundantes por zona, las instancias de máquina virtual subyacentes no son resistentes a los errores de la zona de disponibilidad. Pueden experimentar tiempo de inactividad durante una interrupción en cualquier zona de esa región.

Cost

Puede habilitar la redundancia de zona en un App Service Environment o sus planes sin coste adicional. Sin embargo, la redundancia de zona de un plan requiere que tenga dos o más instancias. Se le cobrará en función de la SKU del plan de App Service, la capacidad que especifique y las instancias a las que escale según los criterios de escalado automático.

Si habilita las zonas de disponibilidad pero especifica una capacidad inferior a dos instancias, la plataforma impone un mínimo de dos instancias. La plataforma le cobra por esas dos instancias.

> **ⓘ Importante**
>
> Al habilitar zonas de disponibilidad para un App Service Environment, todos los planes de App Service con menos de 3 instancias se escalan a 3 instancias. Cualquier plan con 3 o más instancias permanece sin cambios. Una vez completada la operación para habilitar las

zonas de disponibilidad, puede ajustar los planes de App Service según sea necesario, incluyendo menos de 3 instancias.

Configurar soporte de zonas de disponibilidad

Para obtener información sobre cómo crear, habilitar o deshabilitar un nuevo App Service Environment redundante en zonas y nuevos planes de App Service redundantes en zonas, consulte Configurar App Service Environments y planes de App Service aislados v2 para la redundancia en zonas.

> **Nota:**
>
> Un cambio en el estado de redundancia de zona de un App Service Environment tarda entre 12 y 24 horas en completarse. Durante el proceso de actualización, no se produce inactividad ni problemas de rendimiento.

Planeamiento y administración de capacidad

Para prepararse para errores de zona de disponibilidad, considere la posibilidad de *sobreaprovisionar* la capacidad del plan de App Service. Este enfoque permite que la solución tolere cierta pérdida de capacidad y siga funcionando sin que se vea afectado su rendimiento. Para obtener más información, consulte Administración de la capacidad mediante el aprovisionamiento excesivo.

Comportamiento cuando todas las zonas están en buen estado

En la siguiente lista se describe qué esperar cuando los planes de App Service están configurados para la redundancia de zona y todas las zonas de disponibilidad están operativas:

- **Enrutamiento del tráfico entre zonas:** Durante el funcionamiento normal, el tráfico se enruta entre todas las instancias disponibles del plan de App Service en todas las zonas de disponibilidad.
- **Replicación de datos entre zonas:** Durante las operaciones normales, cualquier estado almacenado en el sistema de archivos de la aplicación se almacena en el almacenamiento con redundancia de zona y se replica sincrónicamente entre zonas de disponibilidad.

Comportamiento durante un fallo de zona

Una interrupción de la zona de disponibilidad puede afectar a algunos aspectos de App Service, aunque la aplicación siga sirviendo al tráfico. Estos comportamientos incluyen el escalado del plan de App Service, la creación de aplicaciones, la configuración de la aplicación y la publicación de aplicaciones.

La siguiente lista describe qué puede ocurrir cuando los planes de App Service están configurados para la redundancia de zona y una o varias zonas de disponibilidad no están disponibles:

- **Detección y respuesta:** La plataforma de App Service detecta automáticamente errores en una zona de disponibilidad e inicia una respuesta. No se requiere ninguna intervención manual para iniciar una conmutación por error de zona.

- **Notificación**: Microsoft no le notifica automáticamente cuando una zona está inactiva. Sin embargo, puede usar Azure Resource Health para supervisar el estado de un recurso individual y puede configurar alertas de Resource Health para notificarle problemas. También puede usar Azure Service Health para comprender el estado general del servicio, incluidos los errores de zona, y puede configurar alertas de Service Health para notificarle problemas.

- **Solicitudes activas:** Las solicitudes en curso que se conectan a una instancia del plan de App Service en la zona de disponibilidad defectuosa se finalizan. Reintentar esas solicitudes.

- **Enrutamiento del tráfico:**App Service detecta las instancias perdidas de esa zona e intenta encontrar nuevas instancias de reemplazo Después de que App Service encuentre reemplazos, distribuye el tráfico entre las nuevas instancias según sea necesario.

  Si se configura el ajuste automático y se determina que se necesitan más instancias, se solicitan instancias a App Service. El comportamiento de escalabilidad automática funciona independientemente del comportamiento de la plataforma de App Service. Por lo tanto, la especificación del número de instancias no tiene por qué ser un múltiplo de dos. Para más información, consulte Escalado vertical de una aplicación en App Service y Información general sobre el escalado automático.

  > ⓘ **Importante**
  >
  > Azure no garantiza que las solicitudes de más instancias se realicen correctamente en un escenario de zona descendente. La plataforma intenta reponer instancias perdidas de forma óptima. Si necesita capacidad garantizada durante un error de zona de disponibilidad, cree y configure sus planes de App Service para tener en cuenta la pérdida de zona mediante el **sobreaprovisionamiento de la capacidad**.

- **Comportamientos fuera del tiempo de ejecución:** Las aplicaciones en un plan de App Service con redundancia de zona continúan ejecutándose y prestando servicio al tráfico incluso si una zona de disponibilidad sufre una interrupción. Sin embargo, los comportamientos que no son de ejecución pueden verse afectados durante una interrupción de la zona de disponibilidad. Estos comportamientos incluyen el escalado del plan de App Service, la creación de aplicaciones, la configuración de la aplicación y la publicación de aplicaciones.

Recuperación de zona

Cuando se recupera la zona de disponibilidad, App Service crea automáticamente instancias en la zona de disponibilidad recuperada, quita las instancias temporales creadas en las otras zonas de disponibilidad y enruta el tráfico entre las instancias como de costumbre.

Prueba de fallos de zona

La plataforma de App Service administra el enrutamiento del tráfico, la conmutación por error y la conmutación por recuperación para planes de App Service con redundancia de zona. Esta característica está totalmente administrada, por lo que no es necesario iniciar ni validar los procesos de error de zona de disponibilidad.

Resistencia a errores en toda la región

App Service es un servicio de una sola región. Si la región deja de estar disponible, su entorno y sus planes y aplicaciones también dejarán de estar disponibles.

Soluciones personalizadas de varias regiones para la resistencia

Para reducir el riesgo de un error de una sola región que afecte a la aplicación, implemente varias instancias de App Service Environment en varias regiones. Los pasos siguientes ayudan a reforzar la resistencia:

- Implemente la aplicación en los entornos de App Service en cada región.
- Configure directivas de equilibrio de carga y conmutación por error.
- Replique los datos entre regiones para que pueda recuperar el último estado de la aplicación.

Para obtener un enfoque de ejemplo que ilustra esta arquitectura, consulte Implementación empresarial de alta disponibilidad mediante App Service Environment.

Copias de seguridad y restauración

Para realizar una copia de seguridad de las aplicaciones de App Service en un archivo, use las funcionalidades de copia de seguridad y restauración de App Service.

Estas funcionalidades resultan útiles cuando resulta complicado volver a implementar el código o cuando se almacena el estado en el disco. La mayoría de las soluciones no deberían depender exclusivamente de las copias de seguridad. En su lugar, use las otras capacidades de esta guía para satisfacer sus requisitos de resistencia. Sin embargo, las copias de seguridad protegen contra algunos riesgos que otros enfoques no.

> ⓘ **Importante**
>
> A partir del **31 de marzo de 2028**, las copias de seguridad personalizadas de Azure App Service **ya no admiten la copia de seguridad de bases de datos vinculadas**. Consulte Desuso de copias de seguridad de bases de datos vinculadas para obtener más información.
>
> En su lugar, use las herramientas nativas de copia de seguridad y restauración de la base de datos vinculada. Para más información, consulte Copia de seguridad y restauración de la aplicación en App Service.

Resistencia al mantenimiento del servicio

App Service realiza actualizaciones de servicio periódicas y otras tareas de mantenimiento. Para mantener la capacidad esperada durante una actualización, la plataforma agrega automáticamente instancias adicionales del plan de App Service durante el proceso de actualización.

Habilitación de la redundancia de zona. Al habilitar la redundancia de zona en su plan de App Service, también mejora la resiliencia durante las actualizaciones de la plataforma. *Los dominios de actualización* constan de colecciones de máquinas virtuales que se desconectan durante una actualización y se asignan a zonas de disponibilidad. La implementación de varias instancias en su plan de App Service y la habilitación de la redundancia de zona para su plan agregan una capa adicional de resistencia en caso de que una instancia o zona sea incorrecta durante una actualización.

Personalice el ciclo de actualización. Puede personalizar el ciclo de actualización de un entorno de App Service. Si necesita validar el efecto de las actualizaciones en su carga de trabajo, habilite las actualizaciones manuales. Este enfoque le permite realizar validaciones y

pruebas en una instancia que no es de producción antes de aplicarlas a su instancia de producción.

Para obtener más información sobre las preferencias de mantenimiento, consulte Preferencias de actualización para el mantenimiento planificado de App Service Environment.

Acuerdo de nivel de servicio

El contrato de nivel de servicio (SLA) para los servicios de Azure describe la disponibilidad esperada de cada servicio y las condiciones que la solución deberá cumplir para lograr esa expectativa de disponibilidad. Para obtener más información, consulte Acuerdos de Nivel de Servicio para servicios en línea.

Al implementar un plan de App Service con redundancia de zona, aumenta el porcentaje de tiempo de actividad definido en el Acuerdo de Nivel de Servicio.

Contenido relacionado

- Confiabilidad en Azure
- Confiabilidad en App Service

Last updated on 22/01/2026

Confiabilidad en Azure Container Apps

Azure Container Apps es un servicio de hospedaje de contenedores sin servidor totalmente administrado para implementar microservicios y aplicaciones en contenedor.

Cuando se usa Azure, la confiabilidad es una responsabilidad compartida. Microsoft proporciona una variedad de funcionalidades para admitir resistencia y recuperación. Es responsable de comprender cómo funcionan esas funcionalidades dentro de todos los servicios que usa y de seleccionar las funcionalidades que necesita para cumplir los objetivos empresariales y los objetivos de tiempo de actividad.

En este artículo se describe cómo hacer que Las aplicaciones de contenedor sean resistentes a varias posibles interrupciones y problemas, incluidos errores transitorios, interrupciones de zona de disponibilidad, interrupciones de regiones y mantenimiento del servicio. También se describe cómo usar copias de seguridad para recuperarse de otros tipos de problemas y se resalta información clave sobre el acuerdo de nivel de servicio (SLA) de Container Apps.

Recomendaciones de implementación de producción

Para obtener información sobre cómo implementar Container Apps para admitir los requisitos de confiabilidad de la solución y cómo afecta la confiabilidad a otros aspectos de la arquitectura, consulte Procedimientos recomendados de arquitectura para Aplicaciones de contenedor en Azure Well-Architected Framework.

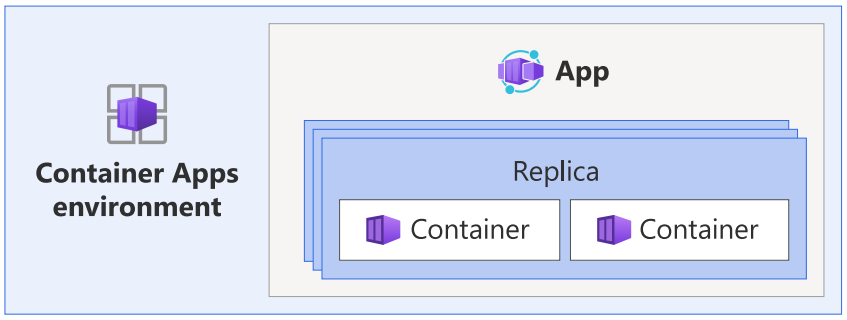
Introducción a la arquitectura de confiabilidad

Al usar Container Apps, implementará un *entorno* que actúa como unidad de implementación básica y define un límite seguro alrededor de un grupo de aplicaciones contenedoras. El entorno es donde se configuran los valores principales, incluida la compatibilidad de la zona de disponibilidad y la configuración de red. Los dos tipos de entornos son entornos de perfiles de carga de trabajo y entornos de solo consumo. Para obtener más información, consulte Estructuras de proceso y facturación en Container Apps.

Puede implementar varias *aplicaciones* en un único entorno. Cada aplicación ejecuta uno o varios *contenedores*. Un entorno también puede ejecutar uno o varios *trabajos*, que representan tareas no interactivas. Para obtener más información, consulte Contenedores en Aplicaciones de contenedor y trabajos en Aplicaciones de contenedor.

Cada aplicación tiene una o varias *réplicas*, que representan las instancias en ejecución de la aplicación. Puedes controlar cómo se escala la aplicación, incluido el número mínimo y máximo

de réplicas y cómo la aplicación agrega y quita réplicas dinámicamente. El planificador de la plataforma garantiza una distribución óptima entre hosts físicos al mismo tiempo que cumple con los requisitos mínimos de cantidad de réplicas. Para obtener más información, consulte Establecimiento de reglas de escalado en Container Apps.



Container Apps apoya la fiabilidad de las aplicaciones mediante diferentes capacidades.

- **Supervisión automática del estado:** El controlador de ingreso integrado equilibra automáticamente la carga del tráfico entre réplicas saludables. Si se produce un error en las comprobaciones de estado de una réplica o su infraestructura subyacente deja de estar disponible durante un tiempo prolongado, el servicio reinicia automáticamente los contenedores con errores o crea réplicas de reemplazo. También redistribuye el tráfico fuera de las réplicas no saludables y administra los reintentos de red en el clúster. Este proceso de recuperación automática no requiere ninguna intervención del cliente y mantiene el recuento de réplicas especificado. Para más información, consulte Sondeos de estado.

- **Resistencia de la aplicación a través de Dapr:** Container Apps proporciona una estrecha integración con Dapr, que es un marco que admite microservicios de nivel de producción y aplicaciones en contenedor. Dapr incluye características que ayudan a mejorar la resistencia, incluido el control de errores en otros servicios. Para obtener más información, consulte Microservicios con Aplicaciones de Contenedores.

- **Resistencia de la infraestructura para los componentes del sistema:** Esta resistencia incluye el plano de control, los controladores de entrada y el entorno de ejecución del contenedor. En las regiones que tienen zonas de disponibilidad, Container Apps proporciona redundancia de zona. Para obtener más información, consulte Resilience to availability zone failures (Resistencia a errores de zona de disponibilidad).

Resistencia a errores transitorios

Los errores transitorios son errores breves e intermitentes en los componentes. Se producen con frecuencia en un entorno distribuido como la nube y son una parte normal de las operaciones. Los errores transitorios se corrigen después de un breve período de tiempo. Es

importante que las aplicaciones puedan controlar errores transitorios, normalmente mediante el reintento de solicitudes afectadas.

Todas las aplicaciones hospedadas en la nube deben seguir las instrucciones de control de errores transitorios de Azure cuando se comunican con cualquier API, bases de datos y otros componentes hospedados en la nube. Para obtener más información, consulte Recomendaciones para controlar errores transitorios.

Container Apps gestiona automáticamente muchos errores transitorios a través de sus mecanismos de reintento a nivel de plataforma y supervisión de salud. Para asegurarse de que las aplicaciones son resistentes a errores transitorios, realice las siguientes acciones:

- **Configure sondeos de estado** que permitan a la plataforma detectar y responder a condiciones de error específicas de la aplicación. Establezca los umbrales de error y los valores de tiempo de espera adecuados en función de las características de inicio de la aplicación. Por ejemplo, para evitar reinicios prematuros del contenedor durante problemas temporales, use un umbral de fallos de 3 con un período de 10 segundos para los probes de vivacidad. Para más información, consulte Sondeos de estado.

- **Utilice directivas de resiliencia para el descubrimiento de servicios (versión preliminar)** para prevenir, detectar y recuperar proactivamente de los fallos en las solicitudes de servicio. Por ejemplo, cuando se usa una directiva de resistencia, cada solicitud entrante a la aplicación se puede reintentar automáticamente si hay un error transitorio que impide que la aplicación responda. Para más información, consulte Resistencia de detección de servicios (versión preliminar).

- **Implemente la lógica de reintento** en las aplicaciones para llamadas de servicio externas, conexiones de base de datos y solicitudes de API.

  Si la aplicación usa Dapr para integrarse con los servicios en la nube, use la resistencia de componentes de Dapr (versión preliminar) para configurar reintentos, tiempos de espera y disyuntores.

  Para otras dependencias, la aplicación debe controlar los errores transitorios. Use estrategias de retroceso exponencial y patrones de cortacircuitos al llamar a servicios externos para evitar fallos en cascada durante interrupciones en servicios posteriores. Las características integradas de detección y equilibrio de carga de Container Apps enrutan automáticamente el tráfico de las instancias con errores, pero las políticas de reintento a nivel de aplicación garantizan una gestión eficaz de manera elegante de los errores transitorios antes de que las comprobaciones de estado a nivel de plataforma desencadenen el reinicio del contenedor.

- **Diseñe trabajos para ser resistentes a errores transitorios**, incluidos los errores durante la ejecución del trabajo o en sus dependencias. Diseñe los trabajos para reanudarlos si se

reinician, o diseñe para la idempotencia de modo que se puedan ejecutar nuevamente de forma segura.

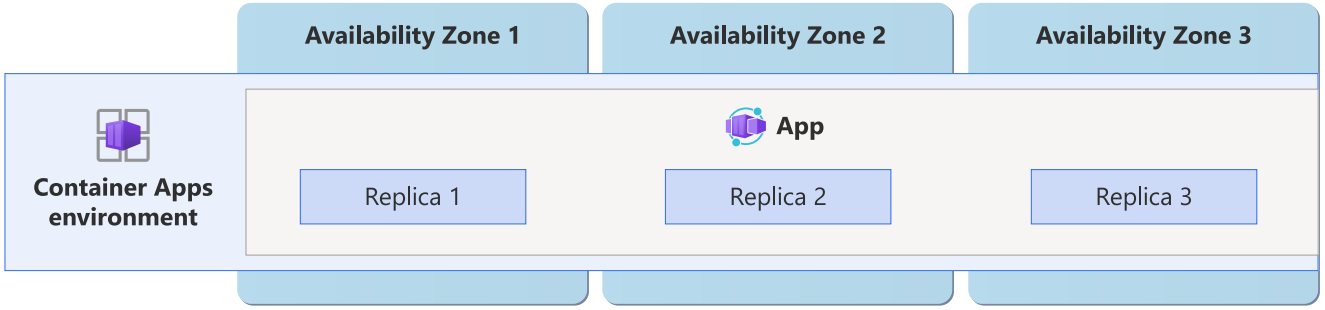
Resistencia a errores de zona de disponibilidad

Las zonas de disponibilidad son grupos físicamente independientes de centros de datos dentro de una región de Azure. Cuando una zona falla, los servicios pueden transferirse a una de las zonas restantes.

Al crear un entorno de Container Apps, puede habilitar la *redundancia de zona* para distribuir la infraestructura subyacente en varias zonas de disponibilidad en la región de Azure elegida. Container Apps programa automáticamente las réplicas de las aplicaciones entre zonas. Esta distribución se produce de forma transparente, lo que significa que no es necesario especificar la ubicación de zona para las réplicas individuales.

La redundancia de zona mejora la resistencia de la aplicación a errores de nivel de zona al garantizar que las réplicas de la aplicación contenedor se reparten entre varias zonas.

En el diagrama siguiente se muestra una aplicación contenedora con redundancia de zona con tres réplicas. Cada réplica se ejecuta en una zona de disponibilidad independiente.



Requisitos

- **Compruebe la compatibilidad con regiones.** La redundancia de zona está disponible en todas las regiones que admiten aplicaciones de contenedores y zonas de disponibilidad.

  Para ver qué regiones admiten zonas de disponibilidad, consulte Regiones de Azure con compatibilidad con zonas de disponibilidad.

  Para ver qué regiones admiten Container Apps, consulte Disponibilidad del producto por región.

- **Use perfiles de carga de trabajo.** La redundancia de zona está disponible para todos los planes de Aplicaciones de Contenedor, incluidos los perfiles de carga de trabajo de consumo y dedicado.

- **Habilite la redundancia de zona durante la creación del entorno.** Esta configuración no se puede cambiar después de crear el entorno.
- **Implemente un entorno de Container Apps en una red virtual.** La red virtual debe estar en una región que admita zonas de disponibilidad. Asegúrese de que la red virtual tiene una subred de tamaño adecuado. Los entornos de solo consumo necesitan una subred con un `/23` intervalo de enrutamiento de Inter-Domain sin clases (CIDR) o mayor, mientras que los entornos de perfiles de carga de trabajo requieren un `/27` intervalo CIDR o superior.
- **Establezca el número mínimo de réplicas en al menos dos para garantizar la distribución entre varias zonas de disponibilidad.** Considere la posibilidad de establecer un número mínimo de réplicas mayor si se aplica alguna de las condiciones siguientes:
  - La carga máxima esperada necesita más de dos réplicas.
  - Debe ser resistente a varias interrupciones de zona simultáneas.
  - Quieres minimizar el tiempo que esperas a que se creen nuevas réplicas en otras zonas durante una interrupción de zona.

Cost

No incurre en cargos adicionales más allá de los precios estándar de Container Apps al habilitar la redundancia de zona. Se pagan las mismas tarifas por los recursos de computación, las solicitudes y los segundos de vCore, independientemente de si la redundancia de zona está habilitada o no. Para más información, consulte Precios de Container Apps y facturación de Container Apps.

Configurar soporte de zonas de disponibilidad

- **Cree un entorno de Container Apps con zonas redundantes.** Para obtener instrucciones de implementación que abarcan Azure Portal, la CLI de Azure y Azure PowerShell, consulte Creación de una aplicación de contenedor con redundancia de zona.
- **Migre a una implementación redundante por zonas.** No se puede habilitar la redundancia de zona en un entorno de Container Apps existente. Para actualizar los entornos existentes que no tienen redundancia de zona, cree un nuevo entorno con redundancia de zona habilitada en una región admitida. A continuación, vuelva a implementar las aplicaciones de contenedor.
- **Desactivar la redundancia de zona.** La redundancia de zona no se puede deshabilitar después de habilitarla durante la creación del entorno. Si necesita una implementación

sin redundancia de zona, debe crear un nuevo entorno sin habilitar la opción de redundancia de zona ni implementarla en una región que no admita zonas de disponibilidad.

- **Compruebe la redundancia de zona.** Puede usar Azure Portal, la CLI de Azure y Azure PowerShell para comprobar el estado de redundancia de zona del entorno.

Planeamiento y administración de capacidad

Si una zona de disponibilidad deja de estar disponible, la plataforma Container Apps usa las reglas de escalado para decidir cuándo reemplazar las réplicas perdidas de esa zona. Es importante configurar correctamente las reglas de escalado para que el programador pueda tomar las decisiones de programación adecuadas.

Para configurar correctamente las reglas de escalado, siga estos principios:

- **Establezca un número mínimo de réplicas** que la aplicación pueda tolerar. Puede tardar un breve período de tiempo en reemplazar las réplicas perdidas porque la plataforma debe detectar que las réplicas antiguas han desaparecido. A continuación, las nuevas réplicas deben iniciarse y devolver un estado de sondeo de preparación correcto para poder recibir solicitudes entrantes. Si no puede tolerar ningún período con menos del número mínimo de réplicas que ha especificado, considere la posibilidad de sobreaprovisionar para mantener el rendimiento de la aplicación incluso si una zona deja de estar disponible.

- **Establezca las solicitudes de recursos y los límites** para ayudar al programador de Container Apps a tomar decisiones de ubicación óptimas entre zonas. Los requisitos de recursos infraespecificados pueden provocar errores de distribución o ubicación desiguales durante una carga alta.

Para obtener más información sobre las opciones de configuración, consulte Establecimiento de reglas de escalado.

Comportamiento cuando todas las zonas están en buen estado

En esta sección se describe qué esperar cuando los recursos de Container Apps están configurados para la redundancia de zona y todas las zonas de disponibilidad están operativas.

- **Enrutamiento de tráfico entre zonas:** Con Container Apps con redundancia de zonas, la plataforma funciona en un modelo activo-activo donde varias réplicas sirven tráfico simultáneamente. El controlador de entrada distribuye las solicitudes entrantes entre

todas las réplicas funcionales, independientemente de su zona, y utiliza el equilibrio de carga 'round robin' de forma predeterminada. Cada zona procesa solicitudes de forma independiente y la plataforma no da prioridad a ninguna zona específica para la distribución del tráfico. Los sondeos de salud se originan en todas las zonas para garantizar una evaluación precisa de la salud de cada réplica desde varias perspectivas.

- **Replicación de datos entre zonas:** Container Apps no replica los datos de la aplicación entre zonas porque está diseñado para cargas de trabajo sin estado. Los datos que la aplicación almacena en almacenamiento efímero, incluido el almacenamiento con ámbito de contenedor y el almacenamiento con ámbito de réplica, se eliminan cuando se cierra el contenedor o la réplica.

  Para los requisitos de datos con estado, monte un recurso compartido de archivos de Azure Files configurado para el almacenamiento con redundancia de zona o use otros servicios de Azure, como Azure Cosmos DB o Azure SQL Database, que proporcionan sus propias funcionalidades de replicación entre zonas.

  La plataforma solo replica los metadatos del plano de control, incluidas las configuraciones de la aplicación, las reglas de escalado y los secretos entre zonas para lograr una alta disponibilidad. Las imágenes de contenedor se extraen del registro de contenedor en cada zona según sea necesario cuando se crean réplicas.

Comportamiento durante un fallo de zona

En esta sección se describe qué esperar cuando los recursos de Container Apps están configurados para redundancia de zona y hay una interrupción en una zona de disponibilidad.

- **Detección y respuesta:** Azure detecta automáticamente errores de zona. Container Apps detiene inmediatamente la programación de nuevas réplicas en la zona fallida y comienza a redistribuir el tráfico a réplicas saludables en las zonas restantes. La plataforma gestiona todas las operaciones de conmutación por error automáticamente, sin necesidad de tu intervención.

- **Notificación:** Microsoft no le notifica automáticamente cuando una zona está inactiva. Sin embargo, puede usar Azure Service Health para comprender el estado general del servicio, incluidos los errores de zona, y puede configurar alertas de Service Health para notificarle problemas.

  También puede supervisar el estado de las aplicaciones a través de las métricas de Container Apps en Azure Monitor. Configure alertas en las caídas del recuento de réplicas y las tasas de error de solicitudes para recibir notificaciones inmediatas cuando se produzcan problemas relacionados con la zona.

- **Solicitudes activas:** Es posible que las solicitudes en vuelo a las réplicas de la zona fallida puedan eliminarse o experimenten tiempos de espera o errores de conexión. Las ejecuciones de trabajos en la zona afectada se anulan y se marcan como fallidas.

- **Pérdida de datos esperada:** No se produce ninguna pérdida de datos en el nivel de plataforma container Apps porque el servicio está diseñado para cargas de trabajo sin estado. Los datos almacenados en el almacenamiento efímero dentro de la zona de disponibilidad se pierden cuando finaliza la réplica y el almacenamiento efímero solo se debe usar para los datos temporales.

- **Tiempo de inactividad esperado:** Las aplicaciones experimentan un tiempo de inactividad mínimo o sin tiempo de inactividad durante los errores de zona. El impacto real depende de la configuración del sondeo de estado de la aplicación y del número de réplicas en zonas correctas. Asegúrese de que los clientes sigan las instrucciones de control de errores transitorios para minimizar los efectos.

  Los trabajos que se ejecutan en la zona afectada son abortados y se marcan como fallidos. Si necesita que un trabajo sea resistente a un error de zona, configure reintentos o configure paralelismo para que el trabajo ejecute varias copias de la misma ejecución. Para obtener más información, consulte Configuración avanzada del trabajo.

- **Reenrutamiento del tráfico:** Las sondas de estado del controlador de ingreso detectan rápidamente las réplicas inaccesibles y las eliminan del pool de equilibrio de carga. En función de la configuración del sondeo de estado de la aplicación, este proceso de conmutación por error suele producirse en unos 30 segundos. El tráfico entrante subsiguiente se distribuye entre las réplicas saludables restantes. Esta redirección de tráfico se produce de forma transparente a los clientes, que siguen usando la misma dirección URL de la aplicación.

  Si la afinidad de sesión está habilitada y una zona deja de funcionar, los clientes que se enrutaron previamente a réplicas de esa zona se enrutan a nuevas réplicas porque las réplicas anteriores ya no están disponibles. Se pierde cualquier estado asociado a las réplicas anteriores.

  No se iniciarán nuevas instancias de trabajos en la zona defectuosa.

- **Administración de instancias:** Es posible que se creen nuevas instancias de réplica en zonas saludables si las reglas de escalado automático se activan debido a un aumento en la carga.

Recuperación de zona

Cuando una zona de disponibilidad se recupera de un error, Container Apps vuelve a reinsertar automáticamente la zona en un servicio activo sin necesidad de intervención. Los sondeos de estado de la plataforma detectan cuándo la infraestructura de la zona recuperada está disponible y Container Apps comienza a programar nuevas réplicas en esa zona en función de la configuración de escalado. Las réplicas existentes en zonas saludables siguen atendiendo el tráfico durante el proceso de reintegración, lo que ayuda a evitar la interrupción del servicio.

Container Apps equilibra gradualmente la distribución de réplicas entre todas las zonas disponibles como parte de las operaciones de escalado estandarizados. Este reajuste automático ocurre cuando se crean réplicas en respuesta a eventos de escalado o cuando se reemplazan las réplicas defectuosas. La plataforma no fuerza la redistribución inmediata de réplicas correctas existentes, lo que evita reinicios innecesarios del contenedor y mantiene la estabilidad de la aplicación durante la recuperación.

Prueba de fallos de zona

La plataforma Container Apps administra el enrutamiento del tráfico, la conmutación por error y la conmutación por recuperación para aplicaciones de contenedor con redundancia de zona. Esta característica está totalmente administrada, por lo que no es necesario iniciar ni validar los procesos de error de zona de disponibilidad.

Para validar la resistencia de la aplicación a errores de zona, simule interrupciones de nivel de zona en el nivel de aplicación mediante enfoques de prueba controlados. Detenga o quite réplicas de zonas específicas mediante el escalado descendente de su aplicación y supervise cómo las réplicas restantes manejan la carga aumentada. Supervise las métricas clave durante las pruebas de resistencia, incluido el recuento de réplicas, las tasas de éxito de las solicitudes, los tiempos de respuesta y el comportamiento de escalado automático. Asegúrese de que el número mínimo de réplicas mantiene la disponibilidad del servicio cuando se eliminan las réplicas y compruebe que las reglas de escalado pueden manejar el aumento de carga en las réplicas restantes. Pruebe las configuraciones de sondeo de estado de salud fallando deliberadamente los puntos de conexión de estado para confirmar que la plataforma quita las instancias no saludables del ciclo de rotación durante los períodos de tiempo esperados.

Resistencia a errores en toda la región

Container Apps es un servicio de una sola región. Si la región deja de estar disponible, el entorno y las aplicaciones tampoco están disponibles.

Soluciones personalizadas de varias regiones para la resistencia

Para reducir el riesgo de un error de una sola región que afecte a la aplicación, puede implementar entornos en varias regiones. Los pasos siguientes ayudan a reforzar la resistencia:

- Implemente las aplicaciones en entornos de cada región. Cada entorno requiere su propia configuración de red virtual y los requisitos de subred se aplican de forma independiente a cada implementación regional. Las imágenes de contenedor deben estar disponibles en todas las regiones, lo que puede lograr mediante Azure Container Registry con la replicación geográfica habilitada.
- Configure las directivas de equilibrio de carga y conmutación por error usando servicios como Azure Front Door o Azure Traffic Manager.
- Replique los datos entre regiones para que pueda recuperar el último estado de la aplicación.

Copias de seguridad y restauración

Container Apps no proporciona funcionalidades de copia de seguridad integradas para las aplicaciones o los datos. Como plataforma de hospedaje de contenedores sin estado, Container Apps espera que las aplicaciones administren sus propias estrategias de persistencia y recuperación de datos a través de servicios externos. Los contenedores de aplicaciones y sus sistemas de archivos locales son efímeros y los datos almacenados localmente se pierden cuando las réplicas se reinician o se mueven.

Resistencia durante las actualizaciones de la aplicación

Use la administración de revisiones para implementar actualizaciones en la aplicación sin tiempo de inactividad. Puede crear nuevas revisiones con imágenes de contenedor actualizadas y realizar una transición mediante una estrategia de implementación azul-verde o cambiar gradualmente el tráfico mediante reglas de división de tráfico. Durante las actualizaciones de aplicaciones, la plataforma mantiene los recuentos mínimos de réplicas mediante la creación de nuevos contenedores antes de retirar los antiguos, lo que ayuda a evitar interrupciones del servicio.

Para obtener más información, consulte Actualización e implementación de cambios en Container Apps.

Resistencia al mantenimiento del servicio

Container Apps realiza el mantenimiento automático de la plataforma para aplicar actualizaciones de seguridad, implementar nuevas características y mejorar la confiabilidad del servicio. La plataforma usa actualizaciones graduales en dominios de error y zonas de disponibilidad para reducir la interrupción de las aplicaciones en ejecución. Durante las ventanas de mantenimiento, los contenedores siguen ejecutándose sin interrupción porque las actualizaciones se aplican a la infraestructura subyacente en fases.

Puede especificar sus propias ventanas de mantenimiento, que son períodos de tiempo que desea realizar el mantenimiento en las aplicaciones. Tenga en cuenta que las actualizaciones críticas pueden producirse fuera de las ventanas de mantenimiento. Para obtener más información, consulte Mantenimiento planeado de Container Apps.

Acuerdo de nivel de servicio

El contrato de nivel de servicio (SLA) para los servicios de Azure describe la disponibilidad esperada de cada servicio y las condiciones que la solución deberá cumplir para lograr esa expectativa de disponibilidad. Para obtener más información, consulte Acuerdos de Nivel de Servicio para servicios en línea.

El Acuerdo de Nivel de Servicio de disponibilidad para las Aplicaciones de Contenedor se basa en las reglas de escalado que establezca en las aplicaciones.

Contenido relacionado

- Introducción a Container Apps
- Sondas de estado en Container Apps
- Escalado automático en aplicaciones de contenedores
- Observabilidad en Aplicaciones de Contenedores
- Documentación sobre confiabilidad de Azure
- Guía de confiabilidad deWell-Architected Framework

ⓘ **Nota:** El autor ha creado este artículo con ayuda de inteligencia artificial. Más información

---

Last updated on 22/01/2026

Confiabilidad de Azure Functions

Azure Functions es un servicio de proceso controlado por eventos que ejecuta pequeños bloques de código o *funciones* sin aprovisionamiento o administración de infraestructuras explícitas. Las funciones pueden responder a eventos como solicitudes HTTP, temporizadores, mensajes de cola y cambios en otros servicios de Azure. Esta funcionalidad hace que Functions sea adecuado para el procesamiento de datos, la integración del sistema y las tareas en segundo plano.

Cuando se usa Azure, la confiabilidad es una responsabilidad compartida. Microsoft proporciona una variedad de funcionalidades para admitir resistencia y recuperación. Es responsable de comprender cómo funcionan esas funcionalidades dentro de todos los servicios que usa y de seleccionar las funcionalidades que necesita para cumplir los objetivos empresariales y los objetivos de tiempo de actividad.

En este artículo se describe cómo hacer que Functions sea resistente a varias posibles interrupciones y problemas, incluidos errores transitorios, errores de zona de disponibilidad y errores en toda la región. También resalta información clave sobre el acuerdo de nivel de servicio (SLA) de Functions.

Recomendaciones de implementación de producción

Azure Well-Architected Framework proporciona recomendaciones sobre confiabilidad, seguridad, costo, operaciones y rendimiento. Para obtener información sobre cómo estas áreas influyen entre sí y contribuyen a una solución confiable de Functions, consulte Procedimientos recomendados de arquitectura para Functions.

Introducción a la arquitectura de confiabilidad

Al implementar Functions, es importante familiarizarse con estos conceptos:

- **Planes de hospedaje:** Los planes representan el entorno de hospedaje de las aplicaciones de funciones. El plan determina los recursos de proceso disponibles, el modelo de precios y el comportamiento de escalado.
- **Cuentas de almacenamiento:** Al crear una aplicación de funciones, debe especificar una cuenta de almacenamiento de host. La cuenta de almacenamiento gestiona aspectos de las operaciones internas de la aplicación de funciones, incluidos el almacenamiento de código

de funciones, el registro y la gestión de concurrencia (como en el caso de las concesiones de blobs para tipos específicos de desencadenadores).

También puede usar una cuenta de almacenamiento para la implementación. Esta cuenta de almacenamiento puede ser la misma que la cuenta de almacenamiento de host o una cuenta de almacenamiento diferente.

> ⓘ **Importante**
>
> Las cuentas de almacenamiento son una parte fundamental de la arquitectura de confiabilidad de Functions. Configúrelos para satisfacer los requisitos de resistencia de la aplicación de funciones.

- **Desencadenadores y enlaces**: Los desencadenadores y enlaces permiten que la función responda a eventos, reciba datos de otros servicios y escriba datos en otros servicios.
- **Durable Functions**: Durable Functions es una característica de Functions. Proporciona funciones con estado, como orquestaciones de larga duración y entidades con estado.

  Cuando se usan funciones duraderas, se configura un proveedor de almacenamiento que almacena el estado. Evalúe las características de confiabilidad del almacén de estado que elija y configúrela para satisfacer los requisitos de resistencia.

Resistencia a errores transitorios

Los errores transitorios son errores breves e intermitentes en los componentes. Se producen con frecuencia en un entorno distribuido como la nube y son una parte normal de las operaciones. Los errores transitorios se corrigen después de un breve período de tiempo. Es importante que las aplicaciones puedan controlar errores transitorios, normalmente mediante el reintento de solicitudes afectadas.

Todas las aplicaciones hospedadas en la nube deben seguir las instrucciones de control de errores transitorios de Azure cuando se comunican con cualquier API, bases de datos y otros componentes hospedados en la nube. Para obtener más información, consulte Recomendaciones para controlar errores transitorios.

Tenga en cuenta las siguientes recomendaciones para controlar errores transitorios en las aplicaciones de funciones:

- **Desencadenadores y enlaces:** La plataforma Functions incluye el control de errores transitorios integrado para desencadenadores y enlaces. Cuando se produce un error transitorio mientras se desencadena un desencadenador compatible o se leen o escriben datos de enlace admitidos, la plataforma puede reintentar automáticamente la operación. Este comportamiento de reintento integrado garantiza que los problemas de conectividad temporal o las interrupciones del servicio no impidan que la función se ejecute. Para obtener más información, consulte Control de errores y reintentos de Functions.

  Esta protección solo cubre errores transitorios. La plataforma no reintenta los errores persistentes, como un cadena de conexión mal configurado o un recurso eliminado.

  La plataforma trata los errores persistentes y los errores transitorios repetidos como errores. Configure el registro para capturar información sobre los errores de ejecución de funciones. Para obtener más información, consulte Configuración de la supervisión de Funciones.

- **Tu código de función:** En tu código de función, eres responsable de controlar errores transitorios cuando tu función llame a servicios externos. Implemente la lógica de reintento, los tiempos de espera y los patrones de cortacircuito según corresponda para las llamadas de servicio externas realizadas en su código de función. Diseñe las funciones para que sean idempotentes siempre que sea posible, de modo que los reintentos no creen efectos secundarios repetidos.

- **Clientes:** Las aplicaciones cliente que se conectan a funciones de forma sincrónica, como a través de una conexión HTTP, deben ser resistentes a errores transitorios.

Resistencia a errores de zona de disponibilidad

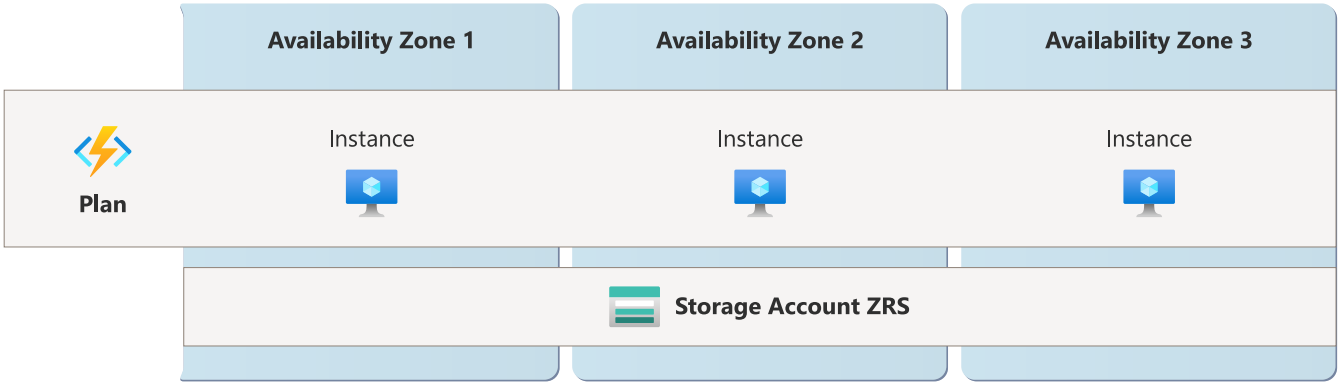
Las zonas de disponibilidad son grupos físicamente independientes de centros de datos dentro de una región de Azure. Cuando una zona falla, los servicios pueden transferirse a una de las zonas restantes.

Los planes Flex Consumption admiten implementaciones con redundancia de zona.

Cuando se habilita la redundancia de zona, la plataforma distribuye automáticamente las instancias del plan en todas las zonas de disponibilidad de la región seleccionada. Si alguna zona de disponibilidad de la región tiene un problema, las funciones continúan ejecutándose en instancias de zonas saludables.

Debe habilitar el almacenamiento con redundancia de zona (ZRS) en la cuenta de almacenamiento del host, para garantizar que también sea resistente a las interrupciones de

zona.



El plan es *no zonal* o *regional* cuando no se habilita la redundancia de zona. La plataforma puede ubicar instancias de plan en cualquier zona de disponibilidad de la región o en la misma zona. Las instancias del plan no son tolerantes a fallos en las zonas de disponibilidad. Es posible que su plan experimente tiempo de inactividad durante una interrupción en cualquier zona de la región.

Requisitos

- **Compatibilidad con regiones:** Puede implementar planes de consumo flexible con redundancia de zona en un conjunto específico de regiones. Puede recuperar la lista actual de regiones admitidas mediante la CLI de Azure. Para más información, consulte Visualización de regiones que admiten zonas de disponibilidad.

- **Cuenta de almacenamiento de host:** Debe configurar la cuenta de almacenamiento de host predeterminada de la aplicación de funciones para usar ZRS. Si usa una cuenta de almacenamiento de host que no está configurada para ZRS, es posible que la aplicación se comporte inesperadamente durante una interrupción de zona.

- **Cuenta de almacenamiento de contenedor de implementación:** Si usa una cuenta de almacenamiento independiente para el contenedor de implementación de la aplicación, actualícela también para que sea redundante de zona.

Consideraciones

Comportamientos no en tiempo de ejecución: La redundancia de zona solo garantiza un tiempo de actividad continuado para las aplicaciones implementadas. Una interrupción de zona de disponibilidad puede afectar a aspectos de Functions, aunque la aplicación siga sirviendo tráfico. Estos comportamientos incluyen el escalado del plan, la creación de aplicaciones, la configuración de la aplicación y la publicación de aplicaciones.

Distribución de instancias a través de zonas

Al configurar las aplicaciones del plan de consumo Flex como redundantes por zona, la plataforma distribuye automáticamente las instancias del plan en varias zonas de la región seleccionada, aplicando reglas diferentes para instancias siempre activas en comparación con las instancias bajo demanda.

- **Las instancias siempre disponibles** se distribuyen entre al menos dos zonas mediante la distribución circular.

  Para garantizar la resistencia de zona, la plataforma mantiene automáticamente al menos dos instancias listas para siempre para cada función o grupo de escalado por función, independientemente de la configuración siempre lista para la aplicación. Las instancias que crea la plataforma son administradas por la plataforma, facturadas como instancias listas para siempre y no cambian las opciones de configuración siempre preparadas.

- **Las instancias a petición** se generan a partir de volúmenes fuente de eventos cuando la aplicación se escala más allá del número de instancias siempre listas. Las instancias a petición se distribuyen entre zonas de disponibilidad de la mejor manera posible. La plataforma prioriza un escalado horizontal más rápido en lugar de una distribución uniforme entre zonas. La plataforma intenta equilibrar la distribución a lo largo del tiempo.

Costo

No se incurre en ningún costo adicional al habilitar la redundancia de zona. Los precios de un plan con redundancia de zona son los mismos que los de un plan de zona única. Sin embargo, habilitar la redundancia de zona afecta al número mínimo de instancias del plan.

Cuando se habilitan zonas de disponibilidad en una aplicación con una configuración de instancia del tipo siempre lista, y hay menos de dos instancias para cada función o grupo de escalado por función, la plataforma crea automáticamente dos instancias del tipo siempre lista para cada función o grupo de escalado por función. Estas nuevas instancias también se facturan como instancias siempre listas.

Para más información sobre los precios completos, consulte Precios de Functions.

Configurar soporte de zonas de disponibilidad

- **Cree un nuevo plan de Funciones con redundancia de zona.** Puede habilitar la redundancia de zona al crear un nuevo plan. Para obtener más información, consulte Creación de una

aplicación de funciones con redundancia de zona.

- **Habilite la redundancia de zona en un plan existente.** Puede activar o desactivar las zonas de disponibilidad para los planes de Elastic Premium existentes. Los planes elastic Premium tienen un comportamiento de capacidad específico que difiere de los planes dedicados (App Service) y requiere pasos de configuración adicionales. Para ver los pasos detallados, consulte Habilitación de la redundancia de zona en un plan existente.

Planeamiento y administración de capacidad

Las aplicaciones de funciones con redundancia de zona siguen ejecutándose incluso cuando las zonas de la región experimentan una interrupción.

Durante una interrupción en la zona, Functions detecta instancias perdidas e intenta automáticamente buscar o crear instancias de reemplazo en las zonas saludables. La plataforma realiza este proceso en base al mejor esfuerzo y no garantiza el éxito. Si la carga de trabajo requiere un número específico de instancias para mantener el nivel de servicio esperado, considere *la posibilidad de sobreaprovisionar* el número de instancias siempre listas. El sobreaprovisionamiento permite a la solución tolerar la pérdida de capacidad y seguir funcionando sin reducir el rendimiento. Para más información, consulte Administración de la capacidad mediante el sobreaprovisionamiento.

Comportamiento cuando todas las zonas están en buen estado

En esta sección se describe qué esperar cuando un plan es con redundancia de zona, la cuenta de almacenamiento de host usa ZRS y todas las zonas de disponibilidad están operativas.

- **Operación entre zonas:** Al configurar la redundancia de zona en Functions, las solicitudes se distribuyen automáticamente entre las instancias de cada zona de disponibilidad. Una solicitud puede ir a cualquier instancia de cualquier zona de disponibilidad.

- **Replicación de datos entre zonas:** Functions es un servicio de proceso sin estado, por lo que no hay datos para replicar entre zonas. La plataforma replica la configuración entre zonas automáticamente.

  Si la cuenta de almacenamiento de host usa ZRS, Azure Storage replica sincrónicamente sus datos en varias zonas de disponibilidad.

  Para funciones duraderas, revise el proveedor de almacenamiento para obtener información sobre cómo replica datos entre zonas.

Comportamiento durante un fallo de zona

En esta sección se describe qué esperar cuando un plan tiene redundancia de zona, la cuenta de almacenamiento del host usa ZRS y hay una interrupción en una zona de disponibilidad.

- **Detección y respuesta:** La plataforma Functions es responsable de detectar un error en una zona de disponibilidad. No es necesario iniciar una conmutación por error de zona.
- **Notificación**: Microsoft no le notifica automáticamente cuando una zona está inactiva. Sin embargo, puede usar Azure Resource Health para supervisar el estado de un recurso individual y puede configurar alertas de Resource Health para notificarle problemas. También puede usar Azure Service Health para comprender el estado general del servicio, incluidos los errores de zona, y puede configurar alertas de Service Health para notificarle problemas.
- **Solicitudes activas:** Cuando una zona de disponibilidad no está disponible, las solicitudes en curso que se conectan a una instancia de la zona de disponibilidad errónea finalizan y se deben reintentar. Asegúrese de que las aplicaciones están preparadas siguiendo las instrucciones de control de errores transitorios.
- **Pérdida de datos esperada:** No se espera que los errores de zona causen pérdida de datos porque Functions es un servicio sin estado.

  Si la cuenta de almacenamiento de host utiliza ZRS, el almacenamiento garantiza que no se pierdan datos debido a un fallo de zona.

  Para las funciones duraderas, compruebe su proveedor de almacenamiento para saber si la pérdida de datos sea posible durante un fallo de zona.
- **Tiempo de inactividad esperado:** Durante las interrupciones de zona, las conexiones pueden experimentar interrupciones breves que suelen durar unos segundos a medida que se redistribuye el tráfico. Asegúrese de que las aplicaciones están preparadas siguiendo las instrucciones de control de errores transitorios.
- **Reenrutamiento del tráfico:** Functions detecta las instancias perdidas de esa zona e intenta buscar nuevas instancias de reemplazo. Una vez que Functions encuentra reemplazos, distribuye el tráfico entre las nuevas instancias según sea necesario.

  > ⓘ **Importante**

> Azure no garantiza que las solicitudes de más instancias se realicen correctamente en un escenario de zona descendente. La plataforma intenta reponer instancias perdidas de forma óptima. Si necesita capacidad garantizada durante un error de zona de disponibilidad, cree y configure los planes para tener en cuenta la pérdida de zona mediante el sobreaprovisionamiento de la capacidad.

- **Comportamientos no en tiempo de ejecución:** Las aplicaciones de un plan de aplicación de funciones con redundancia de zona siguen ejecutándose y atendiendo el tráfico incluso si una zona de disponibilidad experimenta una interrupción. Sin embargo, los comportamientos que no son de ejecución pueden verse afectados durante una interrupción de la zona de disponibilidad. Estos comportamientos incluyen el escalado de aplicaciones de funciones, la creación de aplicaciones, la configuración de la aplicación y la publicación de aplicaciones.

Recuperación de zona

Cuando se recupera la zona de disponibilidad, Functions restaura automáticamente las instancias de la zona de disponibilidad, quita las instancias temporales creadas en las otras zonas de disponibilidad y vuelve a enrutar el tráfico entre las instancias como normal.

Prueba de fallos de zona

La plataforma Functions administra el enrutamiento del tráfico, la conmutación por error y la recuperación de zona para los recursos con redundancia de zona. No es necesario iniciar nada. Azure administra completamente esta característica, por lo que no es necesario validar los procesos de error de zona de disponibilidad.

Resistencia a errores en toda la región

Functions es un servicio de una sola región. Si la región deja de estar disponible, el recurso de Functions tampoco está disponible.

Soluciones personalizadas de varias regiones para la resistencia

Para evitar interrupciones en su servicio durante caídas regionales, puede implementar de forma redundante las mismas funciones en aplicaciones de funciones en varias regiones.

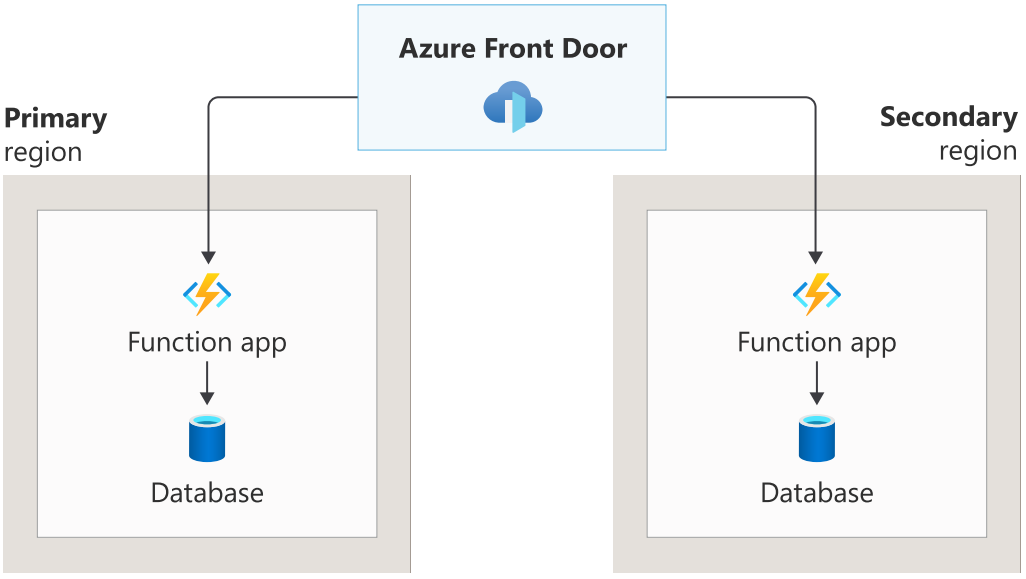
Usted es responsable de:

- Implementación de aplicaciones de funciones en varias regiones.
- Administración de la distribución del tráfico entre regiones.
- Implementación de mecanismos de conmutación por error.
- Garantizar la coherencia de los datos entre regiones (si procede).
- Supervisión y administración de implementaciones entre regiones.

Al ejecutar el mismo código de función en varias regiones, tenga en cuenta los patrones *activo-activo* y *activo-pasivo* . En las secciones siguientes se presentan estos patrones, pero no se proporcionan instrucciones detalladas ni pasos de configuración.

Patrón activo-activo para funciones de desencadenador HTTP

En un patrón activo-activo, las funciones en las dos regiones ejecutan activamente y procesan eventos, ya sea de forma duplicada o de manera alternada. Utilice un patrón activo-activo en combinación con Azure Front Door para sus funciones críticas desencadenadas por HTTP, que pueden enrutar y distribuir solicitudes HTTP entre funciones que se están ejecutando en varias regiones. Azure Front Door también puede comprobar periódicamente el estado de cada punto de conexión. Si un servicio de una región deja de responder a las comprobaciones de estado, Azure Front Door lo quita de la rotación y solo reenvía el tráfico a los servicios saludables restantes.



Patrón activo-pasivo para funciones de desencadenador que no son HTTP

En el caso de funciones dirigidas por eventos que no se activan mediante HTTP (como activadores de Azure Service Bus y Azure Event Hubs), use un patrón activo-pasivo. En un patrón

activo-pasivo, las instancias de función se ejecutan en la región que recibe eventos, mientras que las instancias de la región secundaria permanecen inactivas. Este patrón garantiza que solo una función procese cada mensaje, lo que ayuda a mantener la coherencia de los datos. Permite conmutar por error a la región secundaria durante un desastre, como una interrupción de la región.

Considere la conmutación por error de la app de funciones junto con los comportamientos de conmutación por error de otros servicios que use, como:

- Service Bus, replicación geográfica y recuperación ante desastres
- Replicación geográfica de Event Hubs y recuperación ante desastres geográfica

Considere una topología de ejemplo que usa un activador de Event Hubs, donde el espacio de nombres de Event Hubs está configurado para la recuperación ante desastres geográficos. En este caso, el patrón activo-pasivo requiere los siguientes componentes:

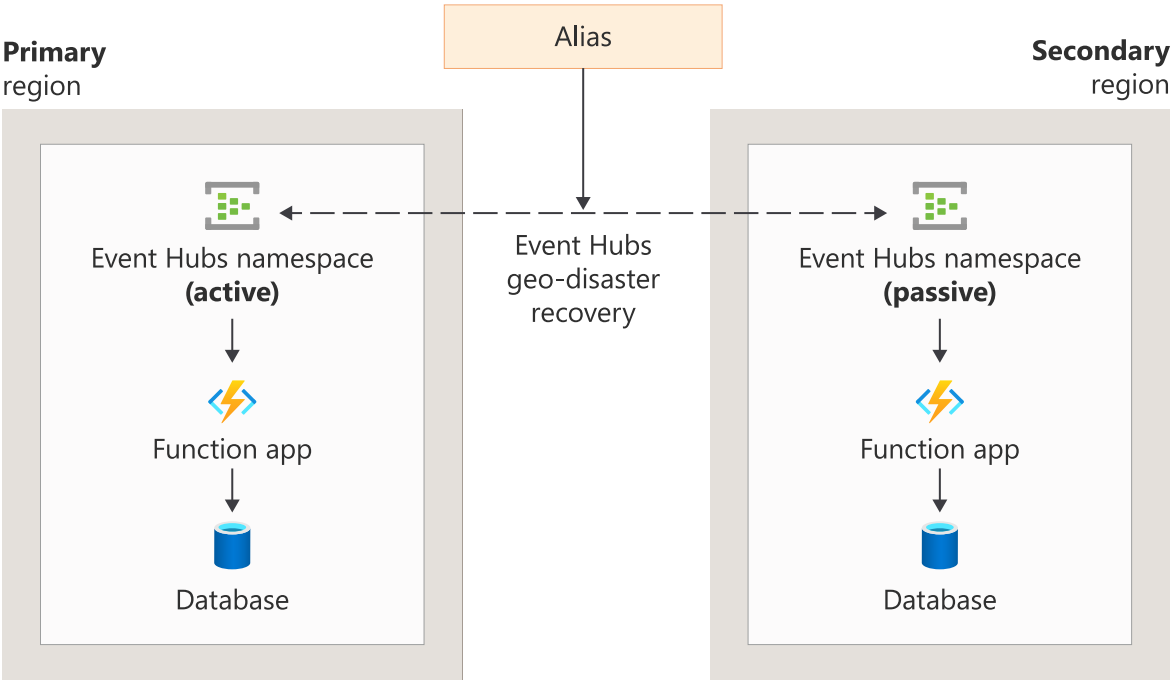
- Espacios de nombres de Event Hubs implementados tanto en una región primaria como en una secundaria.

- Se habilitó la recuperación ante desastres geográfica para vincular los centros de eventos principales y secundarios. Esta configuración también crea un *alias* que puede usar para conectarse al espacio de nombres de Event Hubs y cambiar el alias de la base de datos principal a la secundaria sin cambios en la información de conexión.

- Aplicaciones de funciones desplegadas tanto en la región primaria como en la secundaria. La aplicación de la región secundaria permanece inactiva porque no recibe mensajes.

- Los desencadenadores de cada aplicación de funciones utilizan la cadena de conexión *direct* (noalias) para su espacio de nombres correspondiente de Event Hubs.

- Los publicadores del espacio de nombres de Event Hubs publican al alias de la cadena de conexión.



Antes de la conmutación por error, los publicadores que envían eventos al alias compartido dirigen el tráfico hacia el centro de eventos principal. La aplicación de funciones principal escucha exclusivamente al hub de eventos principal. La aplicación de funciones secundaria permanece pasiva e inactiva.

Cuando se inicia el failover, los publicadores que envían eventos al alias compartido son dirigidos al centro de eventos secundario. La aplicación de funciones secundaria se activa y se desencadena automáticamente. El centro de eventos puede controlar todo el proceso de conmutación por error y cada aplicación de funciones solo se ejecuta cuando su centro de eventos correspondiente está activo.

Durable Functions

Para la recuperación ante desastres en varias regiones para funciones duraderas, consulte Disaster recovery and geo-distribution in Azure durable functions.

Resistencia al mantenimiento del servicio

Functions realiza actualizaciones de servicio periódicas y otras tareas de mantenimiento.

- **Resistencia de errores transitorios:** Durante el mantenimiento del servicio, las instancias que ejecutan la aplicación de funciones pueden reiniciarse o experimentar interrupciones temporales. Asegúrese de que las aplicaciones cliente que interactúan con la aplicación de funciones son resistentes a errores transitorios.

- **Habilite la redundancia de zona:** Al habilitar la redundancia de zona en el plan, también se mejora la resistencia durante las actualizaciones de la plataforma. Implementar varias instancias en tu plan y habilitar la redundancia de zonas proporciona una capa adicional de resiliencia si una instancia o una zona se vuelven defectuosas durante un proceso de actualización.

Resistencia a las implementaciones de aplicaciones

Las implementaciones de aplicaciones presentan el riesgo de problemas en un entorno de producción. Prepárese para revertir una actualización si provoca problemas. Controlar cómo se implementan las actualizaciones para reducir la interrupción de los reinicios de la aplicación.

Los planes de consumo flexible admiten estrategias de actualización del sitio, que proporcionan varias maneras de implementar las actualizaciones de la aplicación. Estas estrategias incluyen actualizaciones graduales para implementaciones sin tiempo de inactividad.

Acuerdo de nivel de servicio

El acuerdo de nivel de servicio (SLA) para Azure servicios describe la disponibilidad esperada de cada servicio y las condiciones que la solución debe cumplir para lograr esa expectativa de disponibilidad. Para obtener más información, consulte Acuerdos de Nivel de Servicio para servicios en línea.

Functions proporciona acuerdos de nivel de servicio de disponibilidad distintos para el plan de consumo y para otros tipos de plan.

Contenido relacionado

- Recuperación ante desastres y distribución geográfica en funciones duraderas de Azure
- Creación de una instancia de Azure Front Door
- Consideraciones sobre la conmutación por error de Event Hubs
- Estrategias de arquitectura para usar regiones y zonas de disponibilidad
- Confiabilidad en Azure

Last updated on 28/04/2026

Confiabilidad en Azure Notification Hubs

Azure Notification Hubs le ayuda a administrar las notificaciones push en varios sistemas de notificaciones de plataforma (PNS), como apple Push Notification Service (APN), Firebase Cloud Messaging (FCM) y Windows Push Notification Service (WNS).

Cuando se usa Azure, la confiabilidad es una responsabilidad compartida. Microsoft proporciona una variedad de capacidades para apoyar la resiliencia y la recuperación. Es responsable de comprender cómo funcionan esas funcionalidades dentro de todos los servicios que usa y de seleccionar las funcionalidades que necesita para cumplir los objetivos empresariales y los objetivos de tiempo de actividad.

En este artículo se describe cómo hacer que Notification Hubs sea resistente a varias posibles interrupciones y problemas, incluidos errores transitorios, errores de zona de disponibilidad, errores de toda la región y mantenimiento del servicio. También se describen las opciones de copia de seguridad y restauración y la información clave sobre el acuerdo de nivel de servicio (SLA) de Notification Hubs.

Recomendaciones de implementación de producción

Para cargas de trabajo de producción, siga estas recomendaciones:

- ✓ Use el nivel Básico o Estándar para que el espacio de nombres sea apto para el Acuerdo de Nivel de Servicio.
- ✓ Cuando sea posible, use instalaciones en lugar de registros en aplicaciones de dispositivo.
- ✓ Use los SDK proporcionados por Microsoft para interactuar con Notification Hubs.
- ✓ Habilitación de la redundancia de zona.
- ✓ Para prepararse para interrupciones en toda la región, habilite la recuperación ante desastres de metadatos en otra región de Azure. Planee cómo realizar copias de seguridad y restaurar registros e instalaciones de dispositivos.

Introducción a la arquitectura de confiabilidad

Azure Notification Hubs se organiza en torno a *espacios de nombres* y *centros de notificaciones*. Un espacio de nombres es un límite administrativo que contiene uno o varios hubs. Los

concentradores representan puntos de conexión para una aplicación. Los dispositivos se registran con esos puntos de conexión mediante *registros* o *instalaciones*, lo que permite al servicio enviar notificaciones push a los dispositivos. Para obtener más información, consulte Administración de registros.

Notification Hubs envía notificaciones push a sistemas de notificaciones de plataforma (PNS), como Apple Push Notification Service (APN) y Firebase Cloud Messaging (FCM). La entrega de notificaciones de un extremo a otro depende de la disponibilidad de Notification Hubs y del comportamiento de los proveedores de PNS de nivel inferior.

Para planear la confiabilidad, es importante distinguir entre los siguientes tipos de datos que Administra Notification Hubs:

- **Metadatos:** Espacio de nombres y configuración del hub, incluida la información de conexión y la configuración de recuperación ante desastres.
- **Datos de registro:** Registros de dispositivos e instalaciones que asignan usuarios y dispositivos a etiquetas y plantillas.

Resistencia a errores transitorios

Los errores transitorios son errores breves e intermitentes en los componentes. Se producen con frecuencia en un entorno distribuido como la nube y son una parte normal de las operaciones. Los errores transitorios se corrigen después de un breve período de tiempo. Es importante que las aplicaciones puedan controlar errores transitorios, normalmente mediante el reintento de solicitudes afectadas.

Todas las aplicaciones hospedadas en la nube deben seguir las instrucciones de control de errores transitorios de Azure cuando se comunican con cualquier API, bases de datos y otros componentes hospedados en la nube. Para más información, vea Recomendaciones para gestionar errores temporales.

Notification Hubs controla automáticamente los errores transitorios que se producen al conectarse a un PNS. Sin embargo, usted es responsable de gestionar los fallos transitorios cuando sus servicios o los dispositivos de sus usuarios interactúan con Notification Hubs. Los errores transitorios pueden producirse durante las operaciones de registro, las operaciones de envío de notificaciones y las operaciones de administración. Siga estas instrucciones:

- **Registros e instalaciones:** Las aplicaciones de los dispositivos deben reintentar las operaciones de registro e instalación que producen errores debido a errores transitorios. los SDK proporcionados por Microsoft controlan los reintentos automáticamente. Si no puedes

utilizar los SDK proporcionados, implementa una lógica de reintento con retroceso exponencial y fluctuación, y haz que las operaciones de registro sean idempotentes siempre que sea posible.

La creación o actualización de una instalación es idempotente, por lo que puedes reintentar la operación con total seguridad. Cuando sea posible, use instalaciones en lugar de registros.

- **Operaciones de envío y administración de notificaciones:** Use un SDK proporcionado por Microsoft para enviar notificaciones push y realizar operaciones de administración. Estos SDK reintentan automáticamente cuando se producen errores transitorios.

  Si no puedes utilizar los SDK proporcionados, implementa una lógica de reintento con retroceso exponencial y fluctuación, y haz que las operaciones de envío de notificaciones sean idempotentes siempre que sea posible.

Resistencia a errores de zona de disponibilidad

Las zonas de disponibilidad son grupos físicamente independientes de centros de datos dentro de una región de Azure. Cuando una zona falla, los servicios pueden transferirse a una de las zonas restantes.

En regiones que admiten zonas de disponibilidad, los espacios de nombres de Notification Hubs admiten una configuración *con redundancia de zona* . Notification Hubs habilita automáticamente la redundancia de zona para todos los espacios de nombres de algunas regiones. Cuando se habilita la redundancia de zona, Microsoft replica tanto los metadatos como los datos de registro en todas las zonas de disponibilidad de la región.

Requirements

- **Compatibilidad con regiones:**

Notification Hubs habilita automáticamente la redundancia de zona para todos los espacios de nombres de las siguientes regiones. No se puede deshabilitar la redundancia de zona en estas regiones:

Expandir tabla

Europe	Oriente Medio	Africa	Asia Pacífico
Centro de Francia	Qatar Central	Norte de Sudáfrica	Norte de China 3
Norte de Italia			Centro de Corea del Sur
Este de Noruega			
Centro de Polonia			
Centro de Suecia			
Norte de Suiza			

En otras regiones que admiten Notification Hubs y tienen zonas de disponibilidad, la redundancia de zona es opcional. Solo puedes habilitarlo al crear un espacio de nombres.

- **Compatibilidad con niveles:** Puede usar zonas de disponibilidad con todos los niveles de Notification Hubs.

Cost

La redundancia de zonas conlleva un cargo adicional al precio del nivel. Para más información, consulte Precios de Notification Hubs.

Configurar soporte de zonas de disponibilidad

- **Cree un nuevo espacio de nombres con redundancia de zona:** El proceso para crear un nuevo espacio de nombres con redundancia de zona depende de la región que use:
  - En las regiones en las que Notification Hubs habilita automáticamente la redundancia de zona, no es necesario configurarla.

> **Importante**
>
> En estas regiones, Notification Hubs siempre crea espacios de nombres con redundancia de zona habilitada, incluso si una implementación basada en código,

como un archivo Bicep o una plantilla de Azure Resource Manager, indica que la redundancia de zona está deshabilitada.

Si no desea un espacio de nombres con redundancia de zona, créelo en una región que admita redundancia de zona opcional.

  - En las regiones en las que la redundancia de zona es opcional, solo puede habilitarla cuando se crea un espacio de nombres. Para obtener información sobre cómo configurar un nuevo espacio de nombres con redundancia de zona, consulte Creación de un centro de notificaciones de Azure en el portal de Azure.

- **Hacer redundante una zona de espacio de nombres existente:** Notification Hubs no admite la migración in situ de un espacio de nombres existente a la compatibilidad con zonas de disponibilidad. Debe implementar un nuevo espacio de nombres y mover los registros a ese espacio de nombres. Siga las instrucciones de Mover recursos entre regiones de Azure, que también se aplican si implementa el nuevo espacio de nombres en la misma región.

Comportamiento cuando todas las zonas están en buen estado

En esta sección se describe qué esperar al configurar un espacio de nombres de Notification Hubs con redundancia de zona, cuando todas las zonas están operativas.

- **Operación entre zonas:** Notification Hubs distribuye y atiende automáticamente las solicitudes mediante el uso de la infraestructura en cualquier zona de la región.

- **Replicación de datos entre zonas:** Tanto los datos de registro como los metadatos se replican sincrónicamente en todas las zonas de la región especificada.

Comportamiento durante un fallo de zona

En esta sección se describe qué cabe esperar al configurar un espacio de nombres de Notification Hubs con redundancia de zona, si se produce una interrupción en una de las zonas.

- **Detección y respuesta:** Microsoft detecta fallos de zona y gestiona la conmutación por error dentro de la región. No es necesario iniciar la conmutación por error.

- **Notificación:** Microsoft no le notifica automáticamente cuando una zona está inactiva. Sin embargo, puede usar Azure Service Health para comprender el estado general del servicio,

incluidos los errores de zona, y puede configurar alertas de Service Health para notificarle problemas.

- **Solicitudes activas:** Las operaciones de administración en curso, los registros de dispositivos y las nuevas solicitudes para enviar notificaciones pueden fallar durante la conmutación por error. Las aplicaciones deben reintentar las operaciones con errores siguiendo las instrucciones de control de errores transitorios.

- **Pérdida de datos esperada:** No se espera la pérdida de datos durante una interrupción de una sola zona porque Notification Hubs replica de forma sincrónica el espacio de nombres y la configuración del centro y los datos de registro en todas las zonas de disponibilidad.

  Esta replicación no es una copia de seguridad. En el modelo de responsabilidad compartida, es responsable de realizar copias de seguridad de los datos de registro e instalación. Para obtener más información, consulte Copia de seguridad y restauración.

- **Tiempo de inactividad esperado:** Una breve interrupción del servicio es posible mientras Microsoft redirecciona el tráfico. Siga las instrucciones de control de errores transitorios para preparar las aplicaciones para estas interrupciones.

- **Redistribución:** El servicio redirige automáticamente las solicitudes a zonas correctas.

Recuperación de zona

Cuando se recupera la zona afectada, no es necesario realizar ninguna acción. Microsoft restaura y vuelve a equilibrar la infraestructura de Notification Hubs para usar la zona recuperada.

Prueba de fallos de zona

No se puede activar directamente una conmutación por error de zona de Notification Hubs. Para probar el comportamiento de la carga de trabajo, ejecute pruebas de resiliencia para reintentos, idempotencia y errores de dependencia en entornos no productivos. También puede usar Azure Chaos Studio para probar los componentes de la aplicación circundantes.

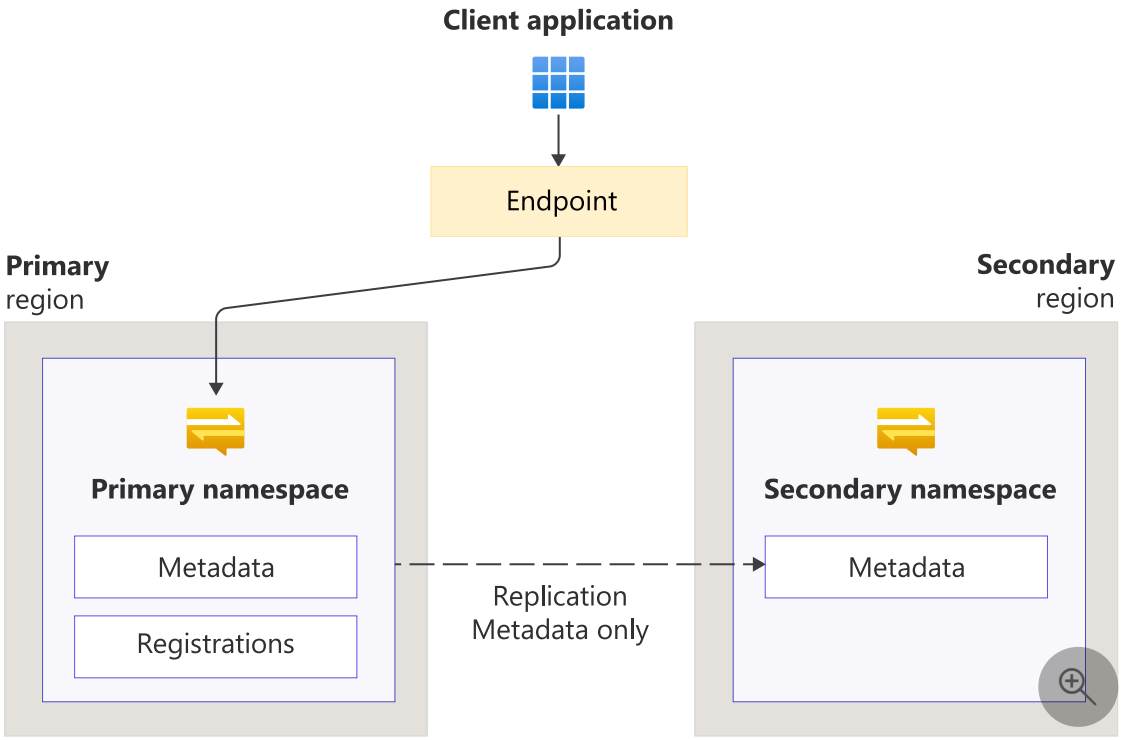
Resistencia a errores en toda la región

Notification Hubs proporciona recuperación ante desastres de metadatos mediante la replicación de los metadatos del espacio de nombres entre regiones, pero no replica los datos de registro de dispositivos. Esta funcionalidad requiere intervención manual durante una interrupción de la región e implica cierto tiempo de inactividad para el centro de notificaciones.

Si necesita reducir el tiempo de inactividad y la intervención manual durante la conmutación por error, considere la posibilidad de usar una solución de varias regiones personalizada.

recuperación geográfica ante desastres de metadatos administrados por Microsoft

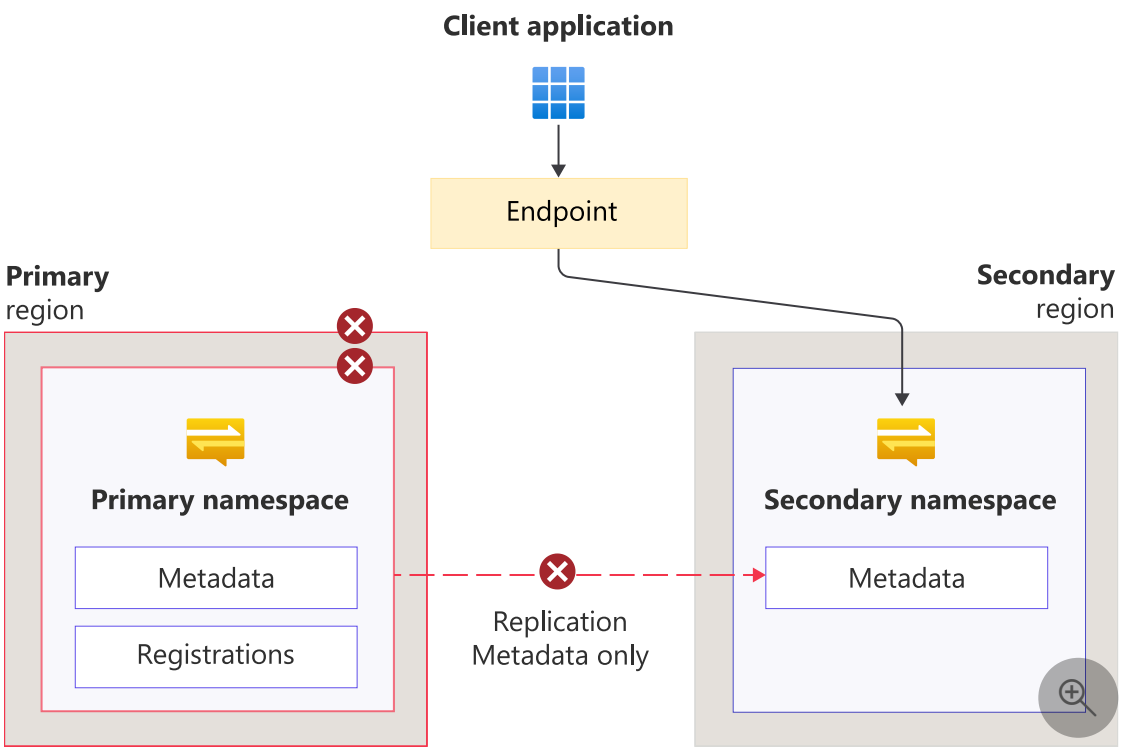
Notification Hubs admite la recuperación ante desastres de metadatos administrados por Microsoft en una región de Azure secundaria. Si la región primaria tiene una región emparejada, puede seleccionar esa región emparejada. Independientemente del estado de emparejamiento de la región primaria, también puede elegir una región secundaria en una lista de *regiones de recuperación flexibles*. A continuación, Notification Hubs replica los metadatos del espacio de nombres, como el nombre del espacio de nombres, las cadenas de conexión y otra información crítica.



> **Importante**
>
> La recuperación ante desastres geográfica de metadatos no replica los datos de registro. Si se desencadena un escenario de recuperación ante desastres, se pueden perder los datos de registro e instalación. Usted es responsable de implementar una solución para restablecer los datos de registro en su hub después de la recuperación.

Microsoft es responsable de declarar un desastre e iniciar la conmutación por error. Cuando esto sucede, Microsoft crea un nuevo espacio de nombres en la región secundaria. Dado que usa los

metadatos de la región primaria, las aplicaciones pueden conectarse a ese espacio de nombres usando el nombre del espacio de nombres existente, la cadena de conexión y los nombres del centro de eventos.



Requirements

- **Compatibilidad con regiones:** En regiones de Azure emparejadas, el espacio de nombres puede usar la región emparejada Azure como región secundaria.

  Si el espacio de nombres está en una región no emparejada o si desea replicar datos en otra región, puede seleccionar una de las siguientes regiones de recuperación flexible como región secundaria:

  Expandir tabla

Americas	Europe	Africa	Asia Pacífico
Sur de Brasil	Norte de Europa	Norte de Sudáfrica	Australia East
Oeste de EE. UU. 2			Sudeste asiático

- **Compatibilidad con niveles:** Las opciones de recuperación ante desastres de metadatos están disponibles en todos los niveles de Notification Hubs.

Cost

Notification Hubs no cobra ningún cargo adicional para configurar o usar la recuperación ante desastres geográfica de metadatos. Sin embargo, paga por el ancho de banda entre regiones que se usa para replicar metadatos. Para más información sobre los precios, consulte Precios de ancho de banda y Precios de Notification Hubs.

Configuración de la compatibilidad con varias regiones

- **Habilite la recuperación ante desastres geográfica de metadatos para un nuevo espacio de nombres:** Siga el procedimiento descrito en Creación de un centro de notificaciones de Azure en el portal de Azure. Seleccione la configuración de recuperación ante desastres al crear el espacio de nombres.
- **Habilite o deshabilite la recuperación ante desastres geográfica de metadatos para un espacio de nombres existente:** Siga el procedimiento descrito en Habilitar la recuperación ante desastres para un espacio de nombres Azure Notification Hubs existente.
- **Realice una copia de seguridad de los datos de registro del dispositivo:** Consulte Exportación e importación de registros de Azure Notification Hubs de forma masiva.

Comportamiento cuando todas las regiones están en buen estado

En esta sección se describe qué cabe esperar al configurar un espacio de nombres de Notification Hubs para la recuperación geográfica ante desastres de metadatos, cuando tanto la región primaria como la secundaria están operativas.

- **Operación entre regiones:** La región primaria atiende todas las solicitudes. La región secundaria no atiende solicitudes a menos que se produzca una conmutación por error.
- **Replicación de datos entre regiones:** Los metadatos, como el nombre del espacio de nombres, la configuración del centro de conectividad, las cadenas de conexión y otra información crítica, se replican de forma asincrónica entre regiones. Los datos de registro no se replican. Usted es responsable de exportarlo periódicamente para mantener una copia de seguridad.

Comportamiento durante una falla de región

En esta sección se describe qué cabe esperar al configurar un espacio de nombres de Notification Hubs para la recuperación ante desastres geográficos de metadatos, y cuando se produce una interrupción del servicio en la región principal.

- **Detección y respuesta:** Microsoft es responsable de detectar el error de la región y decidir si se debe desencadenar la conmutación por error en la región secundaria configurada.
- **Notificación:** Microsoft no le notifica automáticamente cuando una región está inactiva. Sin embargo, puede usar Azure Service Health para comprender el estado general del servicio, incluidos los errores de región, y puede configurar alertas de Service Health para notificarle problemas.
- **Solicitudes activas:** Las solicitudes en curso al espacio de nombres de la región primaria pueden producir un error cuando la región se queda sin conexión. Los clientes deben reintentar las operaciones una vez completada la conmutación por error.
- **Pérdida de datos esperada:** Se conservan los metadatos. No se realiza una copia de seguridad automática de los datos de registro, pero puede realizar una copia de seguridad. Para obtener más información, consulte Exportación e importación de registros de Azure Notification Hubs de forma masiva. Si no lo hace, los datos de registro no estarán disponibles hasta que se recupere la región primaria.
- **Tiempo de inactividad previsto:** Microsoft tarda un tiempo en activar la conmutación por error de los metadatos y, posteriormente, en completar dicha conmutación. Aunque el tiempo puede variar, suele tardar varias horas.

  Una vez completada la conmutación por error, es responsabilidad del usuario restaurar cualquier copia de seguridad de los datos de registro.
- **Redistribución:** Tras la conmutación por error, las solicitudes se redirigen a un espacio de nombres de la región secundaria que utiliza los datos replicados desde la región primaria. Una vez completada la conmutación por error, los clientes se conectan automáticamente al espacio de nombres de la región secundaria.

Recuperación de regiones

Si la región primaria se recupera, podría ser posible volver al espacio de nombres primario de la región primaria. El espacio de nombres principal conservaría los datos de registro antes de la interrupción. Este sería un proceso manual y Microsoft se comunicaría con usted para explicar cómo funciona esto.

Una vez que la región primaria se haya recuperado, debes:

- Comprueba el estado de tu espacio de nombres y sus datos.

- Determine si se van a sincronizar los cambios de datos de registro recientes desde la región secundaria a la región primaria.

Prueba de fallos de región

No se puede iniciar una conmutación por error geográfica. Sin embargo, debe probar sus propios procedimientos de recuperación ante desastres. Compruebe que se realiza una copia de seguridad de los registros y que puede restaurarlos en un nuevo espacio de nombres.

Soluciones de varias regiones personalizadas para la resistencia

La recuperación geográfica ante desastres de metadatos administrados por Microsoft solo replica metadatos. La característica puede recuperar esos metadatos en un espacio de nombres secundario, pero es responsable de importar registros de dispositivos en ese espacio de nombres para que la aplicación pueda seguir funcionando. Este enfoque requiere intervención manual durante un desastre e implica tiempo de inactividad.

Si tus objetivos de recuperación requieren menos tiempo de inactividad o intervención manual, puedes implementar una solución multirregional activa-activa personalizada. Implemente un segundo espacio de nombres de Notification Hubs en otra región de Azure con antelación.

> **Nota:**
>
> En esta sección se proporcionan instrucciones básicas para diseñar este tipo de solución. Usted es responsable de diseñar, implementar, probar, implementar, realizar la conmutación por error y administrar la solución.

- **Conmutación por error:** Dado que el segundo espacio de nombres es un recurso operativo, puede implementar lógica para detectar un fallo en una región y cambiar a ese espacio de nombres.

- **Sincronización:** Para mantener un segundo centro de notificaciones sincronizado con el centro de notificaciones principal, use una de las siguientes opciones:

  - **Para las instalaciones:** Use un back-end de aplicación que cree y actualice simultáneamente instalaciones en ambos centros de notificaciones. Las instalaciones permiten especificar su propio identificador de dispositivo único, que admite este escenario de replicación. Para obtener más información, consulte el ejemplo redundantehub .

- **Para registros:** Use un back-end de aplicación que exporte regularmente los registros desde el centro de notificaciones principal como copia de seguridad y los importe en bloque en el centro de notificaciones secundario. Para obtener más información, consulte Exportación e importación de registros de Azure Notification Hubs de forma masiva.

  Como alternativa, si no tiene un back-end, configure la aplicación para crear instalaciones en ambos centros cuando la aplicación se inicie en los dispositivos de destino. Los dispositivos crean nuevos registros en ambos centros de notificaciones. Finalmente, el centro de notificaciones secundario tiene registrados todos los dispositivos activos.

- **Registros e instalaciones expirados:** Es posible que el centro de notificaciones secundario haya expirado los registros y las instalaciones. Cuando se envía una notificación push a un identificador expirado, Notification Hubs elimina automáticamente el registro o registro de instalación asociado en el centro de notificaciones, en función de la respuesta recibida del servidor PNS. Puede limpiar los registros expirados de la solución de copia de seguridad que prefiera agregando lógica personalizada que procesa los comentarios de cada envío y quita los registros y las instalaciones expirados.

- **Aplicaciones sin abrir:** Hay un período de tiempo durante el que los dispositivos con aplicaciones sin abrir no reciben notificaciones.

- **Costo:** Si usa su propio centro secundario para proteger los datos de registro, ese centro incurre en cargos de servicio normales. Del mismo modo, si implementa cualquier otro recurso de Azure en la región secundaria para respaldar la recuperación, pagará por ellos según las tarifas normales del servicio.

Copias de seguridad y restauración

Notification Hubs no proporciona una única característica integrada de copia de seguridad y restauración para todos los datos almacenados en su espacio de nombres. Usted es responsable de combinar los enfoques siguientes:

- Utilice infraestructura como código (IaC), como Bicep, para definir la configuración de su espacio de nombres, hub y directivas. Almacene esas definiciones en el control de código fuente para que pueda volver a implementar los recursos cuando sea necesario.
- Realice una copia de seguridad de los datos de registro del dispositivo exportando los registros de Azure Notification Hubs de manera masiva.

Resistencia al mantenimiento del servicio

Microsoft aplica periódicamente actualizaciones de servicio y realiza otro mantenimiento. La plataforma Azure controla estas actividades automáticamente, lo que garantiza que el mantenimiento sea transparente y sin problemas. No se prevé ningún tiempo de inactividad durante las operaciones de mantenimiento, a menos que se te haya informado a través de el mantenimiento planificado de Azure Service Health.

Acuerdo de nivel de servicio

El acuerdo de nivel de servicio (SLA) para Azure servicios describe la disponibilidad esperada de cada servicio y las condiciones que la solución debe cumplir para lograr esa expectativa de disponibilidad. Para obtener más información, consulte los SLAs de los servicios en línea.

En el caso de Notification Hubs, el acuerdo de nivel de servicio (SLA) de disponibilidad se aplica a los espacios de nombres que utilizan los niveles Básico y Estándar.

Contenido relacionado

- Confiabilidad en Azure

ⓘ**Nota:** El autor ha creado este artículo con ayuda de inteligencia artificial. Más información

---

Last updated on 27/08/2026

Confiabilidad en Azure SignalR Service

Azure SignalR Service es un servicio totalmente administrado que permite la comunicación bidireccional en tiempo real en aplicaciones web y móviles. Abstrae el mecanismo de transporte subyacente. Cuando WebSockets está disponible, el servicio los usa. Cuando no lo son, recurre a Server-Sent Events o sondeos largos. Esta abstracción significa que el código de cliente y servidor se puede comunicar en tiempo real sin estar vinculado a un protocolo de transporte específico.

Cuando se usa Azure, la confiabilidad es una responsabilidad compartida. Microsoft proporciona una variedad de capacidades para apoyar la resiliencia y la recuperación. Es responsable de comprender cómo funcionan esas funcionalidades dentro de todos los servicios que usa y de seleccionar las funcionalidades que necesita para cumplir los objetivos empresariales y los objetivos de tiempo de actividad.

En este artículo se describe cómo hacer que Azure SignalR Service sean resistentes a una variedad de posibles interrupciones y problemas, incluidos errores transitorios, interrupciones de zona de disponibilidad, interrupciones de región y mantenimiento del servicio. También destaca información clave sobre el acuerdo de nivel de servicio (SLA) de Azure SignalR Service.

Recomendaciones de implementación de producción para la confiabilidad

En el caso de las cargas de trabajo de producción, se recomienda que usted:

- ✓ Use el nivel Premium. El nivel Premium es resistente a los errores de zona de disponibilidad en regiones admitidas y permite configurar la replicación geográfica.
- ✓ Diseñe aplicaciones cliente y servidores de aplicaciones para volver a conectarse de forma segura automáticamente cuando se quiten las conexiones. Las conmutaciones por error de zona, las conmutaciones por error de región y los errores transitorios interrumpen las conexiones activas.
- ✓ Habilite la replicación geográfica para protegerse frente a errores en toda la región. Dimensione cada réplica con suficientes unidades para manejar toda la carga de tráfico esperada durante un evento de failover.

Introducción a la arquitectura de confiabilidad

En esta sección se describen algunos de los aspectos importantes de cómo funciona el servicio que es más relevante desde una perspectiva de confiabilidad. En la sección se presenta la arquitectura lógica, que incluye algunos de los recursos y características que se implementan y usan. También se describe la arquitectura física, que proporciona detalles sobre cómo funciona el servicio en segundo plano.

Arquitectura lógica

El recurso que crea es un recurso *SignalR Service*. Configure un recurso con una serie de *unidades*, que representa la capacidad del recurso, incluido el número máximo de conexiones simultáneas. Para obtener más información, consulte Guía de rendimiento para Azure SignalR Service.

Azure SignalR Service admite dos modos *service*. En el modo predeterminado, los servidores de aplicaciones se conectan al recurso Azure SignalR Service y contienen la lógica del centro de operaciones. En modo sin servidor, el servicio se integra con Azure Functions y Functions actúa como controladores de mensajes controlados por eventos para que no administre directamente los servidores de aplicaciones. Para obtener más información, vea Modo de servicio en Azure SignalR Service.

Un recurso de SignalR Service tiene un punto de conexión único globalmente similar a `contoso.service.signalr.net`. Los clientes establecen conexiones con este punto final. Los servidores de aplicaciones se conectan al mismo punto de conexión para enviar mensajes y recibir eventos de los clientes.

Arquitectura física

Azure SignalR Service administra el estado de conexión y el enrutamiento de mensajes en un conjunto de recursos de proceso. Microsoft administra la infraestructura subyacente. No ve ni interactúa directamente con máquinas virtuales individuales que usa el servicio u otros componentes de infraestructura.

Resistencia a errores transitorios

Los errores transitorios son errores breves e intermitentes en los componentes. Se producen con frecuencia en un entorno distribuido como la nube y son una parte normal de las operaciones. Los errores transitorios se corrigen después de un breve período de tiempo. Es importante que las aplicaciones puedan controlar errores transitorios, normalmente mediante el reintento de solicitudes afectadas.

Todas las aplicaciones hospedadas en la nube deben seguir las instrucciones de control de errores transitorios de Azure cuando se comunican con cualquier API, bases de datos y otros componentes hospedados en la nube. Para obtener más información, consulte Recomendaciones para controlar errores transitorios.

Azure SignalR Service usa conexiones de larga duración entre clientes, servidores de aplicaciones y el servicio. Estas conexiones se pueden quitar debido a errores transitorios, como inestabilidad de red, reconfiguraciones del equilibrador de carga o suspensiones de pestañas del explorador. Diseñe las aplicaciones cliente y los servidores de aplicaciones para controlar las caídas de conexión y volver a conectarse automáticamente.

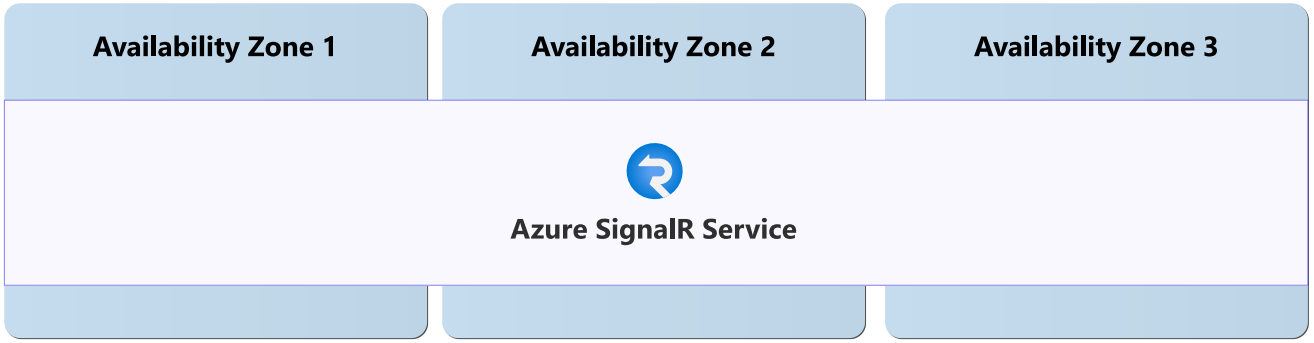
Los SDK de Azure SignalR Service incluyen el control de reconexión integrado para las conexiones de servidor al servicio. La reconexión del lado cliente depende de la biblioteca cliente que use. Los clientes de ASP.NET Core SignalR admiten la reconexión con estado, lo que permite a un cliente reanudar su conexión anterior sin perder el estado si se vuelve a conectar rápidamente usando el mismo identificador de conexión. Si la reconexión da como resultado un nuevo identificador de conexión, por ejemplo, después de una interrupción más larga, el servicio trata al cliente como una nueva conexión. En este caso, el cliente debe volver a unir los grupos en los que se encontraba anteriormente y restaurar cualquier estado de sesión.

Para obtener instrucciones detalladas sobre cómo diseñar la aplicación para controlar las desconexiones de cliente y las reconexión, consulte Conexiones de cliente y reconexión en Azure SignalR Service.

Resistencia a errores de zona de disponibilidad

Las zonas de disponibilidad son grupos físicamente independientes de centros de datos dentro de una región de Azure. Cuando una zona falla, los servicios pueden transferirse a una de las zonas restantes.

Azure SignalR Service admite implementaciones con redundancia de zona en el nivel Premium. Al crear o actualizar a un recurso de nivel Premium en una región que admita zonas de disponibilidad, Azure SignalR Service distribuye automáticamente su capacidad de proceso en todas las zonas de disponibilidad de la región. Si se produce un error en una zona de disponibilidad, Azure SignalR Service enruta las nuevas conexiones a las instancias de las zonas correctas restantes.



Requisitos

- **Compatibilidad con regiones:** La redundancia de zona se admite en la mayoría de las regiones en las que se aplican ambas condiciones:
  - Azure SignalR Service está disponible. Para obtener una lista de las regiones en las que el servicio está disponible, consulte Disponibilidad del producto por región.
  - La región admite zonas de disponibilidad. Para obtener una lista de regiones con zonas de disponibilidad, consulte Azure lista de regiones.

  Sin embargo, El oeste de Japón no admite actualmente redundancia de zona para Azure SignalR Service.

- **Nivel:** La redundancia de zona está disponible en el nivel Premium.

Costo

La redundancia de zona no implica un costo adicional, y pagas la tarifa estándar del nivel Premium. Para obtener más información, consulte precios de Azure SignalR Service.

Configurar soporte de zonas de disponibilidad

La redundancia de zona no requiere ninguna configuración más allá de seleccionar el nivel Premium. Se habilita automáticamente en ambos casos:

- **Cree un nuevo recurso de servicio de SignalR con redundancia de zona.** Seleccione una SKU de nivel Premium al crear el recurso. Para obtener más información, consulte los inicios rápidos, como Quickstart: Creación de un salón de chat mediante SignalR Service.

- **Actualice un recurso existente al nivel Premium.** La redundancia de zona se habilita automáticamente al actualizar un recurso existente a una SKU de nivel Premium. La actualización de Estándar a Premium no provoca tiempo de inactividad del servicio. Para obtener más información, consulte Change your Azure SignalR Service tier.

Comportamiento cuando todas las zonas están en buen estado

En esta sección se describe qué esperar al configurar un recurso de Azure SignalR Service para la redundancia de zona y todas las zonas de disponibilidad están operativas.

- **Operación entre zonas:** Azure SignalR Service administra automáticamente cómo se distribuyen las conexiones y las operaciones a través de las zonas de disponibilidad. La infraestructura de varias zonas procesa el tráfico en un modelo activo-activo. No es necesario configurar nada para aprovechar este comportamiento. El servicio enruta los mensajes entre instancias entre zonas automáticamente, por lo que un mensaje enviado por un cliente de una zona se entrega a los clientes conectados en cualquier otra zona.

- **Replicación de datos entre zonas:** Azure SignalR Service no conserva los datos del cliente, por lo que no hay datos para replicar entre zonas. El estado de conexión es efímero y está asociado a cada conexión activa.

Comportamiento durante un fallo de zona

En esta sección se describe qué esperar al configurar un recurso de Azure SignalR Service para la redundancia de zona y se produce una interrupción en una de las zonas de disponibilidad.

- **Detección y respuesta:** La plataforma Azure SignalR Service es responsable de detectar un error en una zona de disponibilidad. No es necesario realizar ninguna acción para iniciar una conmutación por error de zona.

- **Notificación**: Microsoft no le notifica automáticamente cuando una zona está inactiva. Sin embargo, puede usar Azure Resource Health para supervisar el estado de un recurso individual y puede configurar alertas de Resource Health para notificarle problemas. También puede usar Azure Service Health para comprender el estado general del servicio, incluidos los errores de zona, y puede configurar alertas de Service Health para notificarle problemas.

- **Solicitudes activas:** Durante un fallo de zona, se interrumpen las conexiones activas a los nodos de la zona afectada. Si los clientes controlan los errores transitorios correctamente, por ejemplo, volviendo a conectarse después de un breve período de tiempo, normalmente evitan un impacto significativo.

- **Expected data loss:** Azure SignalR Service no conserva los mensajes, por lo que no se espera que un error de zona cause pérdida de datos en el Azure SignalR service. Sin

embargo, las conexiones activas se interrumpen durante un evento de caída de zona, por lo que es posible que se pierdan los datos que se están transmitiendo activamente.

- **Tiempo de inactividad esperado:** La reconexión de conexiones activas eliminadas suele tardar unos segundos. Los clientes que implementan la lógica de reconexión experimentan una interrupción mínima.
- **Redistribution:** Azure SignalR Service detecta la pérdida de la zona y redistribuye automáticamente el tráfico entre las zonas correctas. No tiene que realizar ninguna acción.

Recuperación de zona

Cuando se recupera una zona de disponibilidad, Azure SignalR Service vuelve a reinsertarla automáticamente en la topología de servicio activa. No es necesario realizar ninguna acción para la recuperación de zona.

Una vez recuperada una zona, es posible que se dirijan nuevas conexiones a la infraestructura de la zona recuperada. Las conexiones existentes no se moverán ni se reequilibrarán en la zona recuperada, pero se volverán a equilibrar gradualmente a medida que las conexiones existentes bajen y se vuelvan a conectar con el tiempo. El desequilibrio de conexiones entre zonas no tiene ningún impacto en la carga de trabajo.

Prueba de fallos de zona

Azure SignalR Service administra automáticamente el enrutamiento del tráfico, la conmutación por error y la recuperación de zona para los recursos de nivel Premium con redundancia de zona. No es necesario iniciar nada. Dado que la redundancia de zona está totalmente administrada, no es necesario validar los procesos de fallo de zona de disponibilidad.

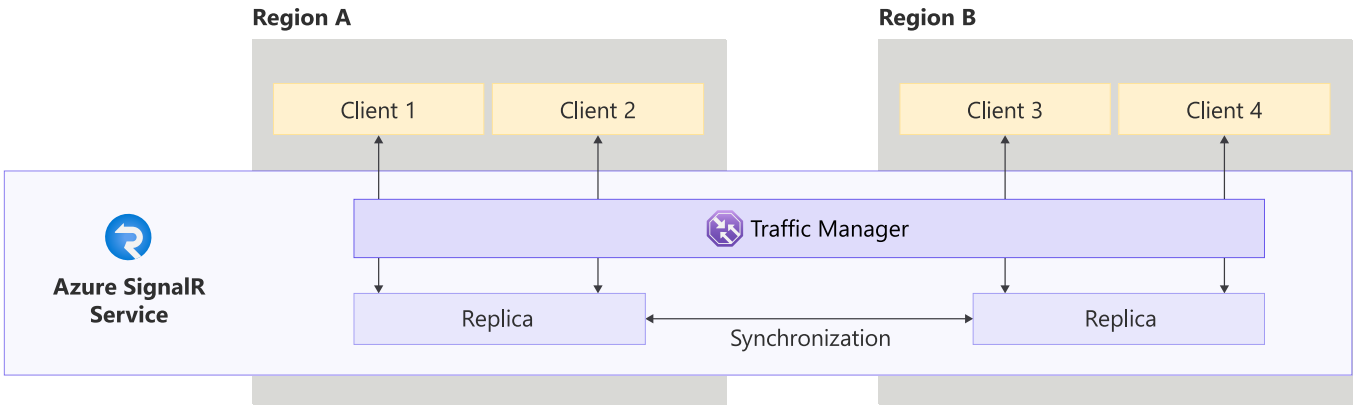
Resistencia a errores en toda la región

Azure SignalR Service es un servicio de una sola región. Si la región deja de estar disponible, el recurso de SignalR Service tampoco está disponible.

Para proteger la aplicación frente a un error en toda la región, puede usar la *replicación geográfica*, que está disponible en el nivel Premium. Como alternativa, puede crear una solución de varias regiones personalizada mediante la implementación de varios recursos de SignalR Service en diferentes regiones.

Geo-replication

La replicación geográfica permite agregar *replicas* del recurso de SignalR Service en otras regiones de Azure. Azure Traffic Manager usa el enrutamiento basado en DNS para dirigir cada cliente a la réplica regional correcta más cercana. Si se produce un error en una región, el Traffic Manager detecta el error a través de comprobaciones de estado y deja de dirigir a los clientes a esa réplica. Las nuevas conexiones de cliente se enrutan automáticamente a la réplica saludable más cercana.



La región en la que creó el recurso de Azure SignalR Service se denomina *región principal* y su réplica es la réplica *principal*. El plano *control* del recurso principal administra la configuración del recurso de Azure SignalR Service.

Requisitos

- **Region support:** Puede agregar réplicas en cualquier región en la que Azure SignalR Service esté disponible.
- **Nivel:** Debe usar el nivel Premium para habilitar la replicación geográfica.
- **Replica limit:** Cada recurso de SignalR Service principal admite hasta ocho réplicas.

Consideraciones

- **Herencia de configuración:** Las réplicas heredan la mayoría de las opciones de configuración del recurso principal. Algunas opciones deben configurarse por separado en cada réplica. Para obtener la lista completa de configuraciones que no se heredan, consulte Geo-replication en Azure SignalR Service.

- **Cambios de configuración:** El plano de control principal, en la región primaria, procesa cualquier cambio de configuración en el recurso SignalR Service. Si el plano de control principal no está disponible, no podrá actualizar la configuración del recurso, aunque las réplicas existentes seguirán procesando el tráfico de datos sin interrupción.

Costo

Cada réplica se factura por separado en función de su propio recuento de unidades y del volumen de mensajes salientes. Las tarifas de salida entre regiones se aplican cuando los mensajes se transfieren entre réplicas y se entregan a clientes o servidores de otra región. Para obtener más información, consulte los precios del Azure SignalR Service.

Configuración de la replicación geográfica

Para agregar o quitar una réplica a un recurso de SignalR Service, consulte Geo-replication en Azure SignalR Service.

Planeamiento y administración de capacidad

Cada réplica controla el tráfico de forma independiente. Durante una conmutación por error regional, los clientes de la región que ha fallado se reconectan a la réplica saludable más cercana. Para asegurarse de que las réplicas supervivientes tienen suficiente capacidad para absorber esta carga adicional, configure cada réplica con unidades que puedan controlar el tráfico total esperado de la carga de trabajo, no solo la parte que normalmente sirve.

Como alternativa, habilite el escalado automático en cada réplica para que las unidades se amplíen automáticamente en respuesta a una mayor carga. El escalado automático sigue funcionando cuando una réplica secundaria no está disponible, pero el escalado automático no funciona si el plano de control principal no está disponible. Para obtener más información sobre el escalado automático, consulte Autoscaling en Azure SignalR Service.

Para obtener instrucciones generales sobre el sobreaprovisionamiento como estrategia, consulte Administración de la capacidad mediante el sobreaprovisionamiento.

Comportamiento cuando todas las regiones están en buen estado

En esta sección se describe qué esperar al configurar Azure SignalR Service para la replicación geográfica y todas las réplicas están operativas.

- **Operación entre regiones:** Azure Traffic Manager enruta cada cliente a la réplica regional saludable más cercana. Los clientes de diferentes áreas geográficas pueden conectarse a diferentes réplicas. SignalR Service sincroniza los mensajes entre réplicas para que los clientes conectados a cualquier réplica puedan comunicarse entre sí.

- **Replicación de datos entre regiones:** Cuando se envía un mensaje a una réplica, el servicio transfiere sincrónicamente ese mensaje a otras réplicas para que los clientes conectados a otro lugar puedan recibirlo. La sobrecarga de sincronización es mínima para los patrones de mensajería más comunes, como la difusión a grupos grandes o la mensajería de una sola

conexión. La mensajería a grupos pequeños (menos de 10 miembros) puede producir una sobrecarga de sincronización ligeramente mayor.

Azure SignalR Service no conserva los mensajes; solo la entrega activa se sincroniza entre réplicas.

Comportamiento durante una falla de región

En esta sección se describe qué esperar al configurar Azure SignalR Service para la replicación geográfica y hay una interrupción en una de las regiones de réplica.

- **Detección y respuesta:** SignalR Service tiene la responsabilidad de detectar una falla en una región y redirigir automáticamente el tráfico entrante a una réplica en una de las demás regiones que configure.
- **Notificación:** Microsoft no le notifica automáticamente cuando una región está inactiva. Sin embargo, puede usar Azure Service Health para comprender el estado general del servicio, incluidos los errores de región, y puede configurar alertas de Service Health para notificarle problemas.
- **Solicitudes activas:** Se quitan las conexiones activas a la réplica en la región con errores. Los clientes deben volver a conectarse después de que la réplica conmute por error.
- **Pérdida de datos esperada:** Azure SignalR Service no conserva los mensajes. Es posible que se pierdan los mensajes que estaban en tránsito a los clientes en la región que falló en el momento del fallo. No se espera ninguna pérdida de datos persistente porque el servicio no almacena los datos del cliente.
- **Tiempo de inactividad esperado:** Azure Traffic Manager realiza verificaciones de salud en cada réplica. Cuando una interrupción en una región provoca que una réplica falle en su verificación de estado, Traffic Manager elimina el endpoint de esa réplica de los resultados de la resolución DNS. Después de quitar el punto de conexión, deben transcurrir 90 segundos del TTL DNS antes de que los clientes vean los registros DNS actualizados. En total, la transición suele tardar unos minutos. Los clientes bien diseñados que implementan la lógica de reconexión pueden reanudar el funcionamiento normal después de volver a conectarse a la réplica sana.

  Si el plano de control principal no está disponible, no puede realizar ningún cambio en la configuración del recurso de SignalR Service o sus réplicas. Sin embargo, las conexiones siguen funcionando en réplicas saludables.

- **Redistribution:** Azure Traffic Manager dirige la solicitud entrante a las réplicas saludables. Sin embargo, si un cliente intenta volver a conectarse antes de que Azure Traffic Manager haya detectado el fallo de conmutación de réplica y se hayan propagado las entradas DNS actualizadas al cliente, el intento de reconexión de un cliente podría continuar apuntando a la región no disponible, lo que provocaría un fallo.

  Una vez que la actualización de DNS se propaga, los clientes que se reconectan se enrutan automáticamente a la réplica saludable más cercana.

Recuperación de regiones

Cuando se recupera la región fallida, la verificación de salud del Traffic Manager detecta la réplica restaurada y vuelve a incluir su punto de conexión en la resolución DNS. Los clientes conectados actualmente a otras réplicas no se ven afectados y permanecen conectados hasta que se desconectan. Las nuevas conexiones se enrutan a la réplica de la región recuperada cuando es la réplica más cercana y en funcionamiento.

Prueba de fallos de región

Para simular un fallo de región y con el fin de probar el comportamiento de reconexión de su aplicación cliente, puede deshabilitar el punto de conexión de una réplica. Esta acción hace que Traffic Manager deje de enrutar el tráfico a esa réplica, lo que le permite observar cómo se comportan los clientes cuando la réplica a la que se conectan deja de estar disponible. Para obtener pasos detallados, consulte Deshabilitar o habilitar el punto de conexión de réplica.

Soluciones personalizadas de varias regiones para la resistencia

Si necesita resistencia entre regiones, pero no usa la replicación geográfica, puede implementar y administrar recursos de SignalR Service independientes en varias regiones e implementar su propia lógica de conmutación por error en el servidor de aplicaciones. Este enfoque es más complejo que la replicación geográfica y no admite la tolerancia a fallos sin tiempo de inactividad para la conexión de cliente a cliente. Para obtener información general detallada sobre la arquitectura, los patrones de conmutación por error y las instrucciones de prueba, consulte Resiliency y recuperación ante desastres en Azure SignalR Service.

Copias de seguridad y restauración

Azure SignalR Service es un servicio de mensajería sin estado. No conserva los mensajes del cliente y no tiene capacidad de copia de seguridad ni restauración.

Para proteger la configuración de recursos, defina los recursos de SignalR Service mediante la infraestructura como código (como Bicep o plantillas de ARM) y almacene esas definiciones en el control de código fuente. Si necesita volver a crear un recurso, vuelva a implementarlo desde la configuración almacenada.

Resistencia al mantenimiento del servicio

Microsoft aplica periódicamente actualizaciones de servicio y realiza otro mantenimiento. La plataforma Azure controla estas actividades automáticamente, lo que garantiza que el mantenimiento sea transparente y sin problemas. No se espera ningún tiempo de inactividad durante los eventos de mantenimiento a menos que se le haya informado a través del mantenimiento planeado de Azure Service Health.

Durante el mantenimiento planeado, Azure SignalR Service usa una estrategia de apagado correcta para reducir el impacto en los clientes conectados. Las conexiones se desconectan gradualmente en un período de tiempo establecido, lo que permite a los clientes volver a conectarse progresivamente en lugar de hacerlo todo a la vez. Para obtener más información, consulte Desconexiones durante el mantenimiento del servicio.

Los eventos de mantenimiento se presentan a los clientes como caídas de conexión. Asegúrese de que las aplicaciones cliente implementen la lógica de reconexión para que puedan recuperarse de desconexiones relacionadas con el mantenimiento sin interrupciones visibles para el usuario.

Acuerdo de nivel de servicio

El acuerdo de nivel de servicio (SLA) para Azure servicios describe la disponibilidad esperada de cada servicio y las condiciones que la solución debe cumplir para lograr esa expectativa de disponibilidad. Para obtener más información, consulte Acuerdos de Nivel de Servicio para servicios en línea.

Contenido relacionado

- Confiabilidad en Azure
- Información general sobre Azure SignalR Service
- Geo-replication en Azure SignalR Service
- Resiliencia y recuperación ante desastres en Azure SignalR Service
- Comprensión de las desconexiones y reconexiones del cliente en Azure SignalR Service

ⓘ **Nota:** El autor ha creado este artículo con ayuda de inteligencia artificial. Más información

---

Last updated on 30/04/2026

Confiabilidad en Azure Web PubSub Service

Azure Web PubSub Service es un servicio de mensajería totalmente administrado en tiempo real que permite la comunicación bidireccional entre servidores y clientes mediante el protocolo WebSocket. Un único recurso de Web PubSub puede escalar a un millón de conexiones WebSocket simultáneas. El servicio admite varios patrones de mensajería, como difusión de servidor a cliente, mensajería a grupos con nombre, pub/sub de cliente a cliente y streaming de tokens de IA.

Cuando se usa Azure, la confiabilidad es una responsabilidad compartida. Microsoft proporciona una variedad de capacidades para apoyar la resiliencia y la recuperación. Es responsable de comprender cómo funcionan esas funcionalidades dentro de todos los servicios que usa y de seleccionar las funcionalidades que necesita para cumplir los objetivos empresariales y los objetivos de tiempo de actividad.

En este artículo se describe cómo hacer que el servicio Azure Web PubSub sea resistente a una variedad de posibles interrupciones y problemas, incluidos errores transitorios, errores de zona de disponibilidad y errores en toda la región. También describe cómo el servicio controla el mantenimiento y resalta la información clave sobre el contrato de nivel de servicio (SLA) de Azure Web PubSub servicio.

Recomendaciones de implementación de producción para la confiabilidad

Para cargas de trabajo de producción, siga estas recomendaciones:

- ✓ Use el nivel Premium. El nivel Premium es resistente a los errores de zona de disponibilidad en regiones admitidas y permite configurar la replicación geográfica.
- ✓ Use el SDK de cliente de Azure Web PubSub al compilar aplicaciones cliente o siga las instrucciones de control de errores transitorios mediante la reconexión segura. Las conmutaciones por error de zona, las conmutaciones por error de región y los errores transitorios interrumpen las conexiones activas.
- ✓ Habilite la replicación geográfica para protegerse frente a errores en toda la región. Dimensione cada réplica con suficientes unidades para manejar toda la carga de tráfico esperada durante un evento de failover.

Introducción a la arquitectura de confiabilidad

En esta sección se describen algunos de los aspectos importantes de cómo funciona el servicio que es más relevante desde una perspectiva de confiabilidad. En la sección se presenta la arquitectura lógica, que incluye algunos de los recursos y características que se implementan y usan. También se describe la arquitectura física, que proporciona detalles sobre cómo funciona el servicio en segundo plano.

Arquitectura lógica

El recurso que crea es un *recurso de Web PubSub*. Configure un recurso con una serie de *unidades*, que representa la capacidad del recurso, incluido el número máximo de conexiones simultáneas. Para obtener más información, consulte Performance guide for Azure Web PubSub Service.

Un recurso de Web PubSub tiene un punto de conexión único globalmente similar a `contoso.webpubsub.azure.com`. Los clientes establecen conexiones de WebSocket a este punto de conexión. Los servidores de aplicaciones se conectan al mismo punto de conexión para enviar mensajes y recibir eventos de los clientes.

Para obtener más información, consulte Azure Web PubSub service internals.

Arquitectura física

Azure Web PubSub Service administra el estado de conexión de WebSocket y el enrutamiento de mensajes en un conjunto de recursos de proceso. Microsoft administra la infraestructura subyacente. No ve ni interactúa directamente con máquinas virtuales individuales que usa el servicio u otros componentes de infraestructura.

Resistencia a errores transitorios

Los errores transitorios son errores breves e intermitentes en los componentes. Se producen con frecuencia en un entorno distribuido como la nube y son una parte normal de las operaciones. Los errores transitorios se corrigen después de un breve período de tiempo. Es importante que las aplicaciones puedan controlar errores transitorios, normalmente mediante el reintento de solicitudes afectadas.

Todas las aplicaciones hospedadas en la nube deben seguir las instrucciones de control de errores transitorios de Azure cuando se comunican con cualquier API, bases de datos y otros componentes hospedados en la nube. Para obtener más información, consulte Recomendaciones para controlar errores transitorios.

WebSocket es un protocolo de conexión de larga duración. Los eventos de red transitorios, los reinicios de la infraestructura de back-end y las operaciones de mantenimiento del servicio pueden quitar una conexión activa. Una reconexión básica restaura la conexión, pero sin lógica adicional, el cliente pierde los mensajes que estaban en curso o en cola durante la interrupción.

Azure Web PubSub Service soluciona este problema a través de *subprotocolos fiables* que operan sobre la conexión WebSocket sin procesar. Los subprotocolos realizan un seguimiento de la secuencia de mensajes y el estado de conexión para que, cuando una conexión cae, el cliente renegocia con el servicio y se reanuda desde donde se dejó.

Normalmente, no hay pérdida de mensajes después de que una conexión se corte y vuelva a establecerse. Sin embargo, hay algunas situaciones en las que puede producirse una pérdida de mensajes. Por ejemplo, si el cliente se desconecta durante más de un minuto y, a continuación, se vuelve a conectar con el mismo identificador de conexión, la operación de reconexión muestra un estado de error para indicar que puede producirse una pérdida de mensajes.

Para aprovechar los subprotocolos confiables, siga estas recomendaciones:

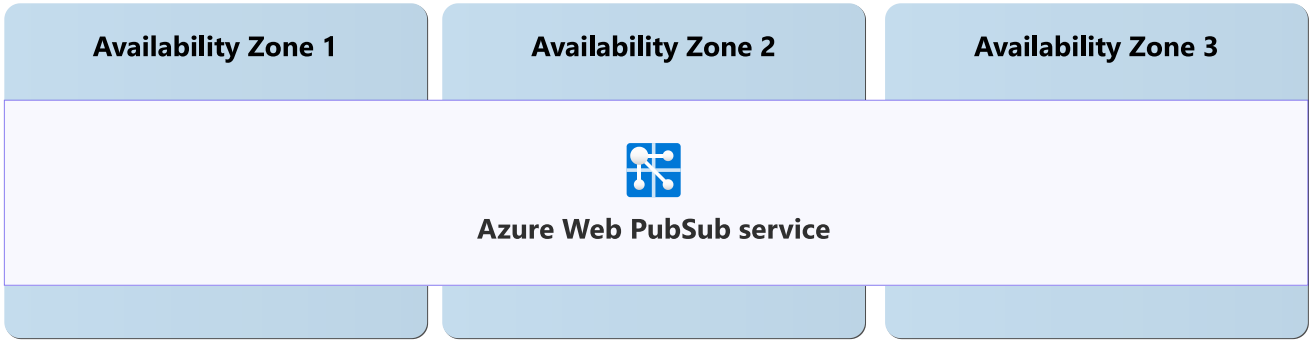
- **Siempre que sea posible, use un SDK de cliente de Azure Web PubSub.** El SDK implementa automáticamente el subprotocolo confiable. No se requiere configuración adicional. Para obtener más información, consulte:
  - SDK del lado cliente de Web PubSub para JavaScript
  - Azure Web PubSub biblioteca cliente para .NET
  - Azure Web PubSub biblioteca cliente para Python
  - Biblioteca cliente de Azure WebPubSub para Java
- **Si no puede usar el SDK, implemente uno de los subprotocolos confiables** directamente en el código de cliente de WebSocket. Para obtener instrucciones completas sobre la especificación y la implementación, consulte Creación de clientes de WebSocket confiables.

Resistencia a errores de zona de disponibilidad

Las zonas de disponibilidad son grupos físicamente independientes de centros de datos dentro de una región de Azure. Cuando una zona falla, los servicios pueden transferirse a una de las zonas restantes.

Azure Web PubSub Service admite implementaciones con redundancia de zona cuando se usa el nivel Premium. Al crear o actualizar un recurso de Web PubSub de nivel Premium en una región que admita zonas de disponibilidad, la redundancia de zona se habilita automáticamente. El

servicio distribuye su infraestructura en varias zonas de disponibilidad de la región. Si se produce un error en una zona, el servicio enruta el tráfico a la infraestructura en la zona en buen estado.



Requisitos

- **Compatibilidad con regiones:** La redundancia de zona se admite en la mayoría de las regiones en las que se aplican ambas condiciones:
  - Servicio de Azure Web PubSub está disponible. Para obtener una lista de las regiones en las que el servicio está disponible, consulte Disponibilidad del producto por región.
  - La región admite zonas de disponibilidad. Para obtener una lista de regiones con zonas de disponibilidad, consulte Azure lista de regiones.

  Sin embargo, El oeste de Japón no admite actualmente redundancia de zona para Azure Web PubSub.

- **Nivel:** La redundancia de zona está disponible en el nivel Premium.

Costo

La redundancia de zona no implica un costo adicional, y pagas la tarifa estándar del nivel Premium. Para obtener más información, consulte precios de servicio Azure Web PubSub.

Configurar soporte de zonas de disponibilidad

La redundancia de zona no requiere ninguna configuración más allá de seleccionar el nivel Premium. Se habilita automáticamente en ambos casos:

- **Cree un nuevo recurso de Web PubSub con redundancia de zona.** Seleccione una SKU de nivel Premium al crear el recurso. Para obtener más información, consulte Crear un recurso de Azure Web PubSub.

- **Actualice un recurso existente al nivel Premium.** La redundancia de zona se habilita automáticamente al actualizar un recurso existente a una SKU de nivel Premium. La

actualización de Estándar a Premium no provoca tiempo de inactividad del servicio. Para obtener más información, consulte Escalar una instancia de servicio de Azure Web PubSub.

Comportamiento cuando todas las zonas están en buen estado

En esta sección se describe qué esperar al configurar un recurso de Azure Web PubSub para la redundancia de zona y todas las zonas de disponibilidad están operativas.

- **Operación entre zonas:** Azure Web PubSub Service gestiona automáticamente cómo se distribuyen las conexiones y las operaciones entre zonas de disponibilidad. La infraestructura de varias zonas procesa el tráfico en un modelo activo-activo. No es necesario configurar nada para aprovechar este comportamiento. El servicio enruta los mensajes entre instancias entre zonas automáticamente, por lo que un mensaje enviado por un cliente de una zona se entrega a los clientes conectados en cualquier otra zona.

- **Replicación de datos entre zonas:** Azure Web PubSub Service no conserva datos de clientes. El servicio mantiene los metadatos de sesión, como el estado de conexión y la información de secuencia de mensajes para las conexiones activas. Estos metadatos se replican sincrónicamente entre zonas de disponibilidad.

Comportamiento durante un fallo de zona

En esta sección se describe qué esperar al configurar un recurso de Azure Web PubSub para la redundancia de zona y hay una interrupción en una de las zonas de disponibilidad.

- **Detección y respuesta:** La plataforma del servicio Azure Web PubSub es responsable de detectar un error en una zona de disponibilidad. No es necesario realizar ninguna acción para iniciar una conmutación por error de zona.

- **Notificación**: Microsoft no le notifica automáticamente cuando una zona está inactiva. Sin embargo, puede usar Azure Resource Health para supervisar el estado de un recurso individual y puede configurar alertas de Resource Health para notificarle problemas. También puede usar Azure Service Health para comprender el estado general del servicio, incluidos los errores de zona, y puede configurar alertas de Service Health para notificarle problemas.

- **Solicitudes activas:** Durante un fallo de zona, las conexiones activas de WebSocket a la infraestructura de la zona afectada se interrumpen. Si los clientes controlan los errores transitorios correctamente, por ejemplo, volviendo a conectarse después de un breve período de tiempo, normalmente evitan un impacto significativo.

- **Expected data loss:** Azure Web PubSub Service no conserva los mensajes, por lo que no se espera que un error de zona cause pérdida de datos en el servicio Azure Web PubSub. Sin embargo, las conexiones activas se interrumpen durante un evento de caída de zona, por lo que es posible que se pierdan los datos que se están transmitiendo activamente.

  Si los publicadores usan un SDK de cliente de Azure Web PubSub o implementan los subprotocolos confiables, el servicio los reconoce después de que el servicio los reciba. Cuando se confirma un mensaje, se replica en todas las zonas de disponibilidad, por lo que el error de zona del publicador no hace que se pierda el mensaje. Sin embargo, si un suscriptor no recibe el mensaje antes de quitarlo, es posible que no reciba el mensaje.

- **Tiempo de inactividad esperado:** La reconexión de conexiones activas eliminadas suele tardar unos segundos. Los clientes que implementan la lógica de reconexión experimentan una interrupción mínima.

- **Redistribution:** Azure Web PubSub Service detecta la pérdida de la zona y redistribuye automáticamente el tráfico entre las zonas correctas. No tiene que realizar ninguna acción.

Recuperación de zona

Cuando se recupera una zona de disponibilidad, Azure Web PubSub Service lo vuelve a reinsertar automáticamente en la topología de servicio activa. No es necesario realizar ninguna acción para la recuperación de zona.

Una vez recuperada una zona, es posible que se dirijan nuevas conexiones a la infraestructura de la zona recuperada. Las conexiones existentes no se moverán ni se reequilibrarán en la zona recuperada, pero se volverán a equilibrar gradualmente a medida que las conexiones existentes bajen y se vuelvan a conectar con el tiempo. El desequilibrio de conexiones entre zonas no tiene ningún impacto en la carga de trabajo.

Prueba de fallos de zona

Azure Web PubSub Service administra automáticamente el enrutamiento del tráfico, la conmutación por error y la recuperación de zona para los recursos de nivel Premium con redundancia de zona. No es necesario iniciar nada. Dado que la redundancia de zona está totalmente administrada, no es necesario validar los procesos de fallo de zona de disponibilidad.

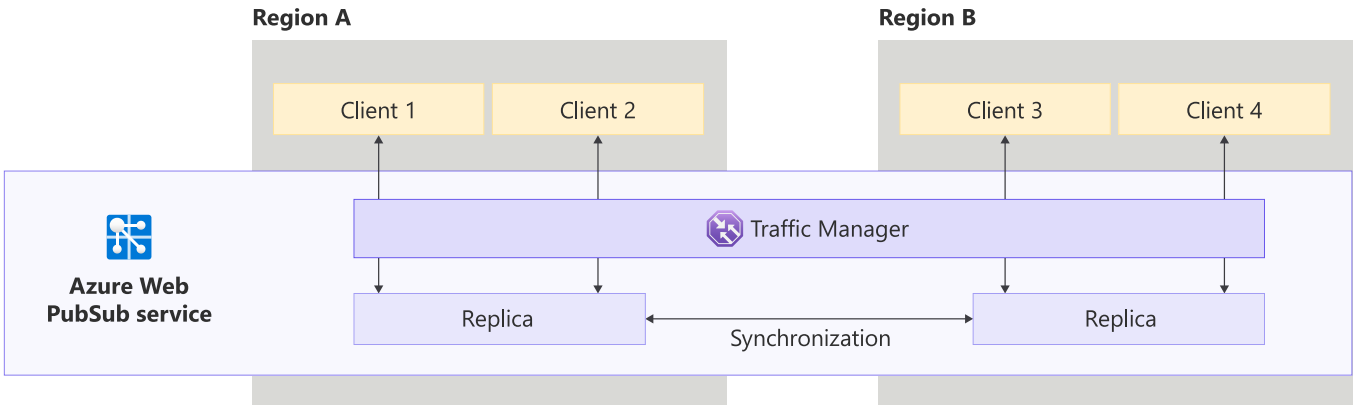
Resistencia a errores en toda la región

Azure Web PubSub Service es un servicio de una sola región. Si la región deja de estar disponible, el recurso de Web PubSub tampoco está disponible.

Para proteger la aplicación frente a un error en toda la región, puede usar la *replicación geográfica*, que está disponible en el nivel Premium. Como alternativa, puede crear una solución de varias regiones personalizada mediante la implementación de varios recursos de Web PubSub en diferentes regiones.

Replicación geográfica

La replicación geográfica permite agregar *replicas* del recurso de Web PubSub en otras regiones de Azure. Todas las réplicas comparten un único punto de conexión (`contoso.webpubsub.azure.com`). Detrás de este punto de conexión, Azure Traffic Manager utiliza el enrutamiento basado en DNS para dirigir a cada cliente a la réplica regional más cercana que esté en buen estado. Si se produce un error en una región, el Traffic Manager detecta el error a través de comprobaciones de estado y deja de dirigir a los clientes a esa réplica. Las nuevas conexiones de cliente se enrutan automáticamente a la réplica saludable más cercana.



La región en la que creó el recurso Web PubSub en se denomina *región primaria* y su réplica es la *réplica principal*. El plano de control del recurso principal administra la configuración del recurso de Web PubSub.

Requisitos

- **Region support:** Puede agregar réplicas en cualquier región en la que Azure Web PubSub Service esté disponible.
- **Nivel:** Debe usar el nivel Premium para habilitar la replicación geográfica.
- **Límite de réplica:** Cada recurso principal de Web PubSub admite hasta ocho réplicas.

Consideraciones

- **Herencia de configuración:** Las réplicas heredan la mayoría de las opciones de configuración del recurso principal. Algunas opciones deben configurarse por separado en cada réplica. Para obtener la lista completa de opciones de configuración que no se heredan, consulte Geo-replication en Azure Web PubSub.
- **Cambios de configuración:** El plano de control principal, en la región primaria, procesa los cambios de configuración en el recurso de Web PubSub. Si el plano de control principal no está disponible, no podrá actualizar la configuración del recurso, aunque las réplicas existentes seguirán procesando el tráfico de datos sin interrupción.

Costo

Cada réplica se factura por separado en función de su propio recuento de unidades y del volumen de mensajes salientes. Si un mensaje se transfiere entre réplicas y, a continuación, se entrega a un cliente o servidor de otra región, se factura como mensaje saliente. Para obtener más información, consulte precios de servicio Azure Web PubSub.

Configuración de la replicación geográfica

Para agregar o quitar una réplica a un recurso de Web PubSub, consulte Geo-replication en Azure Web PubSub.

Planeamiento y administración de capacidad

Cada réplica controla el tráfico de forma independiente. Durante una conmutación por error regional, los clientes de la región que ha fallado se reconectan a la réplica saludable más cercana. Para asegurarse de que las réplicas supervivientes tienen suficiente capacidad para absorber esta carga adicional, configure cada réplica con unidades que puedan controlar el tráfico total esperado de la carga de trabajo, no solo la parte que normalmente sirve.

Como alternativa, habilite el escalado automático en cada réplica para que las unidades se amplíen automáticamente en respuesta a una mayor carga. El escalado automático sigue funcionando cuando una réplica secundaria no está disponible, pero el escalado automático no funciona si el plano de control principal no está disponible. Para obtener más información sobre el escalado automático, consulte Unidades de escalado automático de un servicio de Azure Web PubSub.

Para obtener instrucciones generales sobre el sobreaprovisionamiento como estrategia, consulte Administración de la capacidad mediante el sobreaprovisionamiento.

Comportamiento cuando todas las regiones están en buen estado

En esta sección se describe qué esperar al configurar Azure Web PubSub Service para la replicación geográfica y todas las regiones están operativas.

- **Operación entre regiones:** Azure Traffic Manager enruta cada cliente a la réplica regional saludable más cercana. Los clientes de diferentes áreas geográficas pueden conectarse a diferentes réplicas. El servicio Web PubSub sincroniza los mensajes entre réplicas para que los clientes conectados a cualquier réplica puedan comunicarse entre sí.

- **Replicación de datos entre regiones:** Cuando se envía un mensaje a una réplica, el servicio transfiere sincrónicamente ese mensaje a otras réplicas para que los clientes conectados a otro lugar puedan recibirlo. La sobrecarga de sincronización es mínima para los patrones de mensajería más comunes, como la difusión a grupos grandes o la mensajería de una sola conexión. La mensajería a grupos pequeños (menos de 10 miembros) puede producir una sobrecarga de sincronización ligeramente mayor.

  Azure Web PubSub Servicio no conserva los mensajes; solo la entrega activa se sincroniza entre réplicas.

Comportamiento durante una falla de región

En esta sección se describe qué esperar al configurar Azure Web PubSub Service para la replicación geográfica y se produce una interrupción en una de las regiones de réplica.

- **Detección y respuesta:** El servicio Web PubSub es responsable de detectar un error en una región y redirigir automáticamente el tráfico entrante a una réplica en una de las demás regiones que configure.

- **Notificación**: Microsoft no le notifica automáticamente cuando una región está inactiva. Sin embargo:

  - Puede usar Azure Resource Health para supervisar el estado de un recurso individual y puede configurar alertas de Resource Health para notificarle problemas.

  - Puede usar Azure Service Health para comprender el estado general del servicio, incluidos los errores de región, y puede configurar alertas de Service Health para notificarle los problemas.

- **Solicitudes activas:** Se quitan las conexiones WebSocket activas a la réplica de la región con errores. Los clientes deben volver a conectarse después de que la réplica conmute por error.

- **Pérdida de datos esperada:** Azure Web PubSub Service no almacena los mensajes. Es posible que se pierdan los mensajes que estaban en tránsito a los clientes en la región que falló en el momento del fallo. No se espera ninguna pérdida de datos persistente porque el servicio no almacena los datos del cliente.

- **Tiempo de inactividad esperado:** Azure Traffic Manager realiza verificaciones de salud en cada réplica. Cuando una interrupción en una región provoca que una réplica falle en su verificación de estado, Traffic Manager elimina el endpoint de esa réplica de los resultados de la resolución DNS. Después de quitar el punto de conexión, deben transcurrir 90 segundos del TTL DNS antes de que los clientes vean los registros DNS actualizados. En total, la transición suele tardar unos minutos. Los clientes bien diseñados que implementan la lógica de reconexión pueden reanudar el funcionamiento normal después de volver a conectarse a la réplica sana.

  Si el plano de control principal no está disponible, no puede realizar ningún cambio en la configuración del recurso de Web PubSub ni en sus réplicas. Sin embargo, las conexiones de WebSocket siguen funcionando en réplicas saludables.

- **Redistribution:** Azure Traffic Manager dirige la solicitud entrante a las réplicas saludables. Sin embargo, si un cliente intenta volver a conectarse antes de que Azure Traffic Manager haya detectado el fallo de conmutación de réplica y se hayan propagado las entradas DNS actualizadas al cliente, el intento de reconexión de un cliente podría continuar apuntando a la región no disponible, lo que provocaría un fallo.

  Una vez que la actualización de DNS se propaga, los clientes que se reconectan se enrutan automáticamente a la réplica saludable más cercana.

Recuperación de regiones

Cuando se recupera la región fallida, la verificación de salud del Traffic Manager detecta la réplica restaurada y vuelve a incluir su punto de conexión en la resolución DNS. Los clientes conectados actualmente a otras réplicas no se ven afectados y permanecen conectados hasta que se desconectan. Las nuevas conexiones se enrutan a la réplica de la región recuperada cuando es la réplica más cercana y en funcionamiento.

Prueba de fallos de región

Para simular un fallo de región y con el fin de probar el comportamiento de reconexión de su aplicación cliente, puede deshabilitar el punto de conexión de una réplica. Esta acción hace que Traffic Manager deje de enrutar el tráfico a esa réplica, lo que le permite observar cómo se

comportan los clientes cuando la réplica a la que se conectan deja de estar disponible. Para obtener pasos detallados, consulte Deshabilitar o habilitar el punto de conexión de réplica.

Soluciones de varias regiones personalizadas para la resistencia

Si necesita resistencia entre regiones, pero no usa la replicación geográfica, puede implementar y administrar recursos independientes de Web PubSub en varias regiones e implementar su propia lógica de conmutación por error en el servidor de aplicaciones. Este enfoque es más complejo que la replicación geográfica y no admite la tolerancia a fallos sin tiempo de inactividad para la conexión de cliente a cliente. Para una visión general detallada de la arquitectura, patrones de conmutación por error y guías para la realización de pruebas, consulte Resiliencia y recuperación ante desastres en Azure Web PubSub Service.

Copias de seguridad y restauración

Azure Web PubSub Service es un servicio de mensajería sin estado. No conserva los mensajes del cliente y no tiene capacidad de copia de seguridad ni restauración.

Para proteger la configuración de sus recursos, defina sus recursos de Web PubSub mediante la infraestructura como código (como Bicep o plantillas de ARM) y almacene esas definiciones en el control de versiones. Si necesita volver a crear un recurso, vuelva a implementarlo desde la configuración almacenada.

Resistencia al mantenimiento del servicio

Microsoft aplica periódicamente actualizaciones de servicio y realiza otro mantenimiento. La plataforma Azure controla estas actividades automáticamente, lo que garantiza que el mantenimiento sea transparente y sin problemas. No se espera ningún tiempo de inactividad durante los eventos de mantenimiento a menos que se le haya informado a través del mantenimiento planeado de Azure Service Health.

Acuerdo de nivel de servicio

El acuerdo de nivel de servicio (SLA) para Azure servicios describe la disponibilidad esperada de cada servicio y las condiciones que la solución debe cumplir para lograr esa expectativa de disponibilidad. Para obtener más información, consulte Acuerdos de Nivel de Servicio para servicios en línea.

Contenido relacionado

- ¿Qué es Azure Web PubSub servicio?
- Replicación geográfica en Azure Web PubSub
- Resiliencia y recuperación ante desastres en Azure Web PubSub Service
- Escalar una instancia del Servicio Web PubSub de Azure
- Escalar automáticamente las unidades de un servicio de Azure Web PubSub
- Confiabilidad en Azure

ⓘ **Nota:** El autor ha creado este artículo con ayuda de inteligencia artificial. Más información

---

Last updated on 30/04/2026

¿Qué son las regiones de Azure?

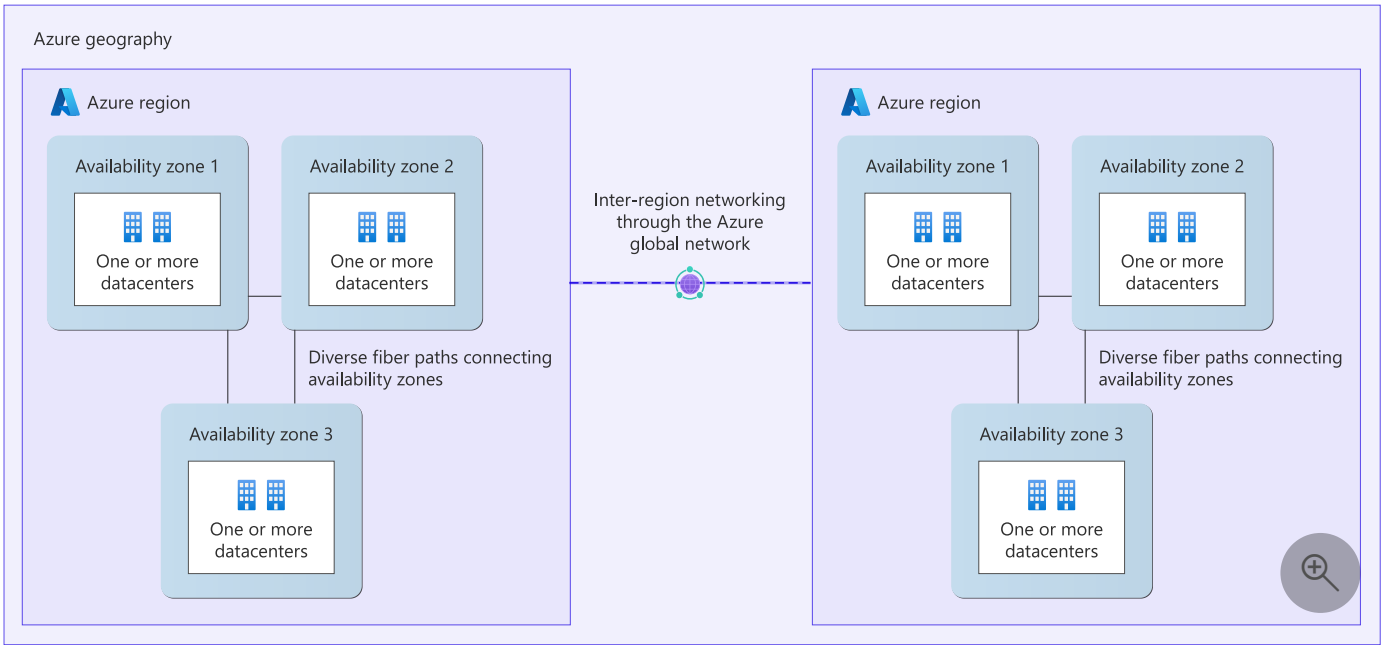
https://learn-video.azurefd.net/vod/player?id=d36b5b2d-8bd2-43df-a796-b0c77b2f82fc&locale=es-es&embedUrl=%2Fazure%2Freliability%2Fregions-overview

Azure proporciona más de 70 regiones globalmente. Las regiones se encuentran en muchas *zonas geográficas* diferentes. Cada geografía representa un límite de residencia de datos, por ejemplo, el Estados Unidos o Europa, y puede contener una o varias regiones. Cada región es un conjunto de instalaciones físicas con centros de datos e infraestructura de red.

Las regiones proporcionan determinados tipos de opciones de resistencia. Muchas regiones proporcionan zonas de disponibilidad y algunas tienen una región emparejada, mientras que otras regiones no están emparejadas. Al elegir una región para los servicios, es importante prestar atención a las opciones de resistencia disponibles en esa región. Este artículo le ayuda a comprender Azure regiones y proporciona información general sobre las opciones de resistencia que admiten algunas Azure regiones, al tiempo que ofrece vínculos a información más detallada sobre cada tema.

Entender las regiones y geografías de Azure

Una región de Azure consta de uno o varios centros de datos, conectados por una conexión de red de alta capacidad, tolerante a errores y baja latencia. Los centros de datos de Azure se encuentran normalmente dentro de un área metropolitana grande.



Cada región se encuentra dentro de una única *geografía* que actúa como límite fijo de residencia de datos. Si tiene requisitos de residencia de datos, es importante que seleccione las regiones dentro de la geografía requerida. Cada geografía tiene al menos una región equipada con zonas de disponibilidad. Para obtener una lista de todas las zonas geográficas de Azure, consulte Azure geographies.

> **Nota:**
>
> La mayoría de las regiones están disponibles para todos los clientes de Azure. Sin embargo, algunas regiones pertenecen a zonas *geográficas de nube soberanas* . Los clientes de áreas geográficas específicas con estrictas regulaciones de residencia de datos pueden acceder a estas regiones. Las regiones de nube soberana funcionan de la misma manera que otras regiones, pero a menudo tienen servicios y características limitados. Para obtener más ejemplos de disponibilidad limitada de servicios en regiones de nube soberana, consulte Comparación entre Azure Government y Azure global o Disponibilidad del servicio y hojas de ruta de Microsoft Azure operado por 21Vianet.

Lista de regiones

Para obtener una lista de regiones de Azure, consulte List of Azure regions. Si desea obtener más detalles sobre las regiones, incluida la residencia de datos y el cumplimiento normativo, consulte mapa de centros de datos de Microsoft.

Elección de regiones de Azure

Al diseñar una solución, debe elegir las Azure regiones que desea usar. Las regiones que elija pueden afectar a la resistencia de la solución, así como a muchos otros aspectos de la solución.

Al seleccionar regiones, tenga en cuenta los siguientes factores:

- **Latencia**. Seleccione las regiones que están geográficamente cerca de los usuarios para reducir la latencia. Por ejemplo, si los usuarios están en los Estados Unidos, puede seleccionar una región en los Estados Unidos o Canadá.
- **Zonas de disponibilidad**. Seleccione las regiones que admiten zonas de disponibilidad para proporcionar redundancia y aislamiento de errores. Asegúrese de distribuir los recursos entre varias zonas de disponibilidad de la región.
- **Residencia de datos:** Asegúrese de que las regiones que seleccione se encuentren dentro de un límite de residencia de datos que requiera su organización.

Para obtener información más detallada sobre cómo seleccionar regiones, consulte Select Azure regions.

Opciones de resistencia regional

Aunque todas las regiones de Azure proporcionan servicios de alta calidad, como la optimización de residencia y latencia de datos, pueden diferir en los tipos de opciones de resistencia que admiten.

En esta sección se resumen dos opciones de resistencia que podrían estar disponibles en las regiones que elija.

Zonas de disponibilidad

Muchas Azure regiones proporcionan zonas de disponibilidad. Las zonas de disponibilidad son conjuntos independientes de centros de datos que contienen conexiones de red, refrigeración y alimentación aisladas. Las zonas de disponibilidad se encuentran físicamente lo suficientemente cerca como para proporcionar una red de baja latencia, pero lo suficientemente lejos como para proporcionar aislamiento de errores de tales cosas como tormentas y interrupciones de energía aisladas. La mayoría de los servicios Azure proporcionan compatibilidad integrada con zonas de disponibilidad y puede decidir cómo usarlas para satisfacer sus necesidades. Al diseñar una solución de Azure, debe usar zonas de disponibilidad para proporcionar redundancia y aislamiento de errores.

Para obtener más información sobre las zonas de disponibilidad, consulte ¿Qué son las zonas de disponibilidad?.

Regiones emparejadas y no emparejadas

Microsoft empareja algunas regiones de Azure con otra región de Azure para formar *pares de áreas*. No puedes elegir tus propios pares. Algunos servicios Azure usan pares de regiones para admitir la replicación geográfica y la redundancia geográfica. Algunos también usan pares de regiones para admitir aspectos de la recuperación ante desastres, en el caso poco probable de que una región experimente un error catastrófico e irrecuperable.

Muchas regiones más recientes no están emparejadas y, en su lugar, usan zonas de disponibilidad como medio principal de redundancia. Muchos servicios Azure admiten redundancia geográfica tanto si las regiones están emparejadas como si no, y puede diseñar una solución altamente resistente si usa regiones emparejadas, regiones no emparejadas o una combinación de ambas.

Para obtener más información sobre las regiones emparejadas y no emparejadas y cómo usarlas, consulte Azure pares de regiones y regiones no emparejadas.

Uso de varias regiones de Azure

Es habitual usar varias regiones de Azure, ya sea emparejadas o no emparejadas, al diseñar una solución. Mediante el uso de varias regiones, puede aumentar la resistencia de la carga de trabajo a muchos tipos de errores y tiene muchas opciones para la recuperación ante desastres. Además, algunos servicios de Azure están disponibles en regiones específicas, por lo que mediante el diseño de una solución de varias regiones puede aprovechar la naturaleza global y distribuida de la nube.

Algunos servicios de Azure proporcionan capacidades integradas para la replicación y la conmutación por error entre regiones. Otros pueden requerir que diseñe e implemente sus propios enfoques de varias regiones. Para obtener una lista de las funcionalidades de varias regiones de servicio, consulte Azure services que admiten varias regiones. Para obtener información detallada sobre cada servicio, consulte su guía de confiabilidad.

Si usa varias regiones juntas, debe tener en cuenta los inconvenientes entre los siguientes factores:

- **Aislamiento físico:** Tenga en cuenta si debe usar regiones que están geográficamente distantes entre sí. Cuanto mayor sea la distancia, mayor será la resistencia en el caso de un desastre natural importante en una de las regiones. Para obtener información sobre la ciudad o el estado en el que se encuentra una región, consulte List of Azure regions and Microsoft Datacenters Map.

- **Latencia:** Al seleccionar regiones aisladas físicamente, aumenta la latencia de las conexiones de red entre esas regiones. La latencia puede afectar al diseño de una solución de varias regiones y puede restringir los tipos de replicación geográfica y redundancia geográfica que puede usar. Para más información sobre la latencia entre regiones de Azure, consulte estadísticas de latencia de ida y vuelta de red Azure. Para obtener más información sobre cómo seleccionar regiones, consulte Recomendaciones para usar zonas de disponibilidad y regiones.

- **Residencia de datos:** Asegúrese de que las regiones que seleccione se encuentren dentro de un límite de residencia de datos que requiera su organización. Para obtener más información, consulte Confiabilidad y soberanía en Azure.

Servicios no regionales

Puede implementar instancias de la mayoría de los servicios de Azure en una región específica. Sin embargo, algunos servicios no están vinculados a una sola región de Azure. Es importante reconocer cómo funcionan los servicios *no regionales* en el caso de un error regional y tenerlos en cuenta al diseñar sus soluciones y plan de continuidad empresarial. Microsoft implementa servicios no regionales en dos o más regiones. Si se produce un fallo regional, la instancia del servicio en una región operativa puede seguir atendiendo las solicitudes.

Algunos servicios no regionales permiten especificar una región o geografía en la que se almacenan los datos. Por ejemplo, con Microsoft Entra ID, puede seleccionar el área geográfica de los datos, como Europa o Norteamérica. Algunos servicios no regionales también admiten la especificación de la región en la que se implementan determinados componentes, como la región en la que se van a residir las máquinas virtuales Azure Virtual Desktop.

Para obtener más información y una lista de servicios no regionales, consulte Azure servicios no regionales.

Recursos relacionados

Para obtener más información sobre los servicios de Azure disponibles en cada región, consulte Products disponibles por región.

ⓘ**Nota:** El autor ha creado este artículo con ayuda de inteligencia artificial. Más información

---

Last updated on 29/06/2026

Azure regiones emparejadas y no emparejadas

En este artículo se describe cómo Azure usa pares de regiones y regiones no emparejadas.

Azure regiones son independientes entre sí. Sin embargo, Microsoft asocia algunas Azure regiones con otra región, donde ambas regiones suelen estar en la misma geografía. Juntas, las áreas forman un *par de áreas*. Un pequeño número de servicios Azure usan estos pares de regiones para admitir la replicación geográfica y la redundancia geográfica. Los pares también admiten algunos aspectos de la recuperación ante desastres en el improbable caso de que una región experimente un error catastrófico e irrecuperable.

Sin embargo, muchas regiones no están emparejadas y, en su lugar, usan zonas de disponibilidad como medio principal de redundancia. Además, muchos servicios de Azure admiten redundancia geográfica si las regiones están emparejadas o no.

Puede diseñar una solución altamente resistente si usa regiones emparejadas, regiones no emparejadas o una combinación.

Regiones emparejadas

Algunos servicios Azure usan regiones emparejadas para crear su estrategia de replicación geográfica multiregión y redundancia geográfica. Por ejemplo, Azure almacenamiento con redundancia geográfica (GRS) puede replicar automáticamente los datos en una región emparejada.

Si está en una región emparejada, el uso de su par como región secundaria proporciona varias ventajas:

- **Secuencia de recuperación de regiones**. En el improbable caso de una interrupción en toda la geografía, una región de cada par de regiones tiene prioridad para la recuperación. Los componentes que se implementan en regiones emparejadas usan una de las regiones como región prioritaria para la recuperación.
- **Actualización secuencial**. Azure se esfuerza por escalonar cualquier actualización planeada del sistema entre pares de regiones. Este enfoque minimiza el impacto de errores o errores lógicos en el caso poco frecuente de una actualización errónea y evita el tiempo de

inactividad en las soluciones diseñadas para usar regiones emparejadas juntas para la resistencia.

- **Residencia de datos**. Para cumplir con los requisitos de residencia de datos, casi todas las regiones residen dentro de la misma geografía que su par. Para obtener información sobre las excepciones, consulte la lista de regiones de Azure.

> **Importante**
>
> La implementación de recursos en una región en un par no hace que sean automáticamente más resistentes, ni proporciona alta disponibilidad automática, funcionalidades de recuperación ante desastres ni conmutación por error. Desarrolle sus propios planes de alta disponibilidad y recuperación ante desastres, independientemente de si usa regiones emparejadas o no.
>
> Incluso aunque se configuren características de servicio para usar pares de regiones, no confíe en la conmutación por error administrada por Microsoft entre esos pares como enfoque de recuperación ante desastres principal. Por ejemplo, la conmutación por error administrada por Microsoft de las cuentas de almacenamiento habilitadas para GRS solo se realiza en situaciones catastróficas y después de repetidos intentos de recuperación erróneos.

No está limitado a usar servicios en una sola región ni en el par de regiones de su región. Aunque un servicio Azure puede depender de un par regional específico para algunas de sus funcionalidades de confiabilidad, puede hospedar los servicios en cualquier región que satisfaga sus necesidades empresariales. Por ejemplo, una solución de Azure puede usar Azure Storage en la región Centro de Canadá con almacenamiento GRS para replicar datos en la región emparejada, Este de Canadá, mientras utiliza recursos de proceso de Azure ubicados en Este de EE. UU. y recursos de Azure OpenAI ubicados en Oeste de EE. UU.

Para ver una lista de regiones que incluye todos los pares de regiones, consulte Lista de regiones de Azure.

Regiones emparejadas asimétricamente

La mayoría de los pares de regiones son *simétricas*, lo que significa que cada región está emparejada bidireccionalmente con otra región. Por ejemplo, Oeste de EE. UU. está emparejado con Este de EE. UU. y Este de EE. UU. se empareja con Oeste de EE. UU.

Los pares de regiones asimétricas implican regiones que no están emparejadas bidireccionalmente. En la lista siguiente se incluyen pares de regiones asimétricas públicas:

- Sur de Brasil está emparejado con Centro-Sur de Estados Unidos, que no se encuentra dentro de la geografía de Brasil. Centro-sur de EE. UU. no está emparejado con el Sur de Brasil.
- La India Occidental está emparejada con el Sur de la India, pero la India del Sur está emparejada con Centro de la India.
- India centro-sur está emparejada con India central, pero India central está emparejada con India meridional.
- La región Oeste de EE. UU. 3 está emparejada en una dirección con Este de EE. UU. La región Este de EE UU está emparejada bidireccionalmente con Oeste de EE. UU.

Para obtener una lista de regiones que incluya todos los pares asimétricos de regiones, consulte pares de regiones de Azure.

lista de pares de regiones de Azure

En las siguientes tablas se enumeran las regiones de Azure y la región emparejada correspondiente de cada una, organizadas por zona geográfica. Las regiones sin un par muestran **N/A** en la columna región emparejada. Para obtener un conjunto más completo de información sobre cada región, vea Lista de regiones de Azure.

> **Nota:**
>
> El icono indica una región de acceso restringido. Para solicitar acceso, consulte proceso de solicitud de acceso a la región de Azure.

Todos

Expandir tabla

Region	Región emparejada
Centro de Australia	Centro de Australia 2
Centro de Australia 2	Centro de Australia
Australia East	Australia Southeast

Region	Región emparejada
Australia Southeast	Australia East
Este de Austria	N/A
Centro de Bélgica	N/A
Sur de Brasil	Centro-sur de EE. UU.
Sudeste de Brasil	Sur de Brasil
Canada Central	Este de Canadá
Este de Canadá	Canada Central
Centro de la India	South India
Central US	Este de EE. UU. 2
Chile Central	N/A
Este de Dinamarca	N/A
Este de Asia	Sudeste asiático
East US	West US
Este de EE. UU. 2	Central US
Centro de Francia	Sur de Francia
Sur de Francia	Centro de Francia
Norte de Alemania	Centro-oeste de Alemania
Centro-oeste de Alemania	Norte de Alemania
Centro-sur de la India	Centro de la India
Indonesia Central	N/A
Israel Central	N/A
Norte de Italia	N/A
Japón Oriental	Oeste de Japón
Oeste de Japón	Japón Oriental
Centro de Corea del Sur	Corea del Sur
Corea del Sur	Centro de Corea del Sur

Region	Región emparejada
Malaysia West	N/A
Centro de México	N/A
Norte de Nueva Zelanda	N/A
Centro-Norte de EE. UU	Centro-sur de EE. UU.
Norte de Europa	West Europe
Este de Noruega	Oeste de Noruega
Oeste de Noruega	Este de Noruega
Centro de Polonia	N/A
Qatar Central	N/A
Norte de Sudáfrica	Oeste de Sudáfrica
Oeste de Sudáfrica	Norte de Sudáfrica
Centro-sur de EE. UU.	Centro-Norte de EE. UU
South India	Centro de la India
Sudeste asiático	Este de Asia
Centro de España	N/A
Centro de Suecia	Sur de Suecia
Norte de Suiza	Oeste de Suiza
Oeste de Suiza	Norte de Suiza
Centro de Emiratos Árabes Unidos	Norte de Emiratos Árabes Unidos
Norte de Emiratos Árabes Unidos	Centro de Emiratos Árabes Unidos
UK South	UK West
UK West	UK South
Centro-oeste de EE. UU.	Oeste de EE. UU. 2
West Europe	Norte de Europa
West India	South India
West US	East US

Region	Región emparejada
Oeste de EE. UU. 2	Centro-oeste de EE. UU.
Oeste de EE. UU. 3	East US

Regiones no emparejadas

Azure continúa expandiéndose globalmente. Muchas de las regiones más recientes proporcionan varias zonas de disponibilidad para lograr una mayor resistencia y no tienen un par de regiones.

Muchos servicios Azure admiten la replicación geográfica y la redundancia geográfica entre cualquier conjunto arbitrario de regiones y no dependen de pares de regiones. Otros pueden requerir que diseñe e implemente sus propios enfoques de varias regiones. Para obtener una lista de las funcionalidades de varias regiones de servicio, incluidas las que funcionan entre regiones no emparejadas, consulte Azure servicios que admiten varias regiones. Para obtener información detallada sobre cada servicio, consulte su guía de confiabilidad.

Expandir tabla

Americas	Europe	Oriente Medio	Asia Pacífico
Chile Central	Este de Austria	Israel Central	Indonesia Central
Centro de México	Centro de Bélgica	Qatar Central	Malaysia West
	Este de Dinamarca		Norte de Nueva Zelanda
	Norte de Italia		
	Centro de Polonia		
	Centro de España		

Contenido relacionado

- ¿Qué son las regiones de Azure?
- Lista de regiones de Azure
- Guía de confiabilidad

ⓘ**Nota:** El autor ha creado este artículo con ayuda de inteligencia artificial. Más información

Last updated on 26/06/2026

Azure servicios que admiten varias regiones

En este artículo se proporciona una lista de referencia de Azure servicios. Algunos servicios ofrecen funcionalidades integradas de varias regiones, como la replicación geográfica, la copia de seguridad entre regiones, la conmutación por error o la conectividad entre regiones de Azure. Algunas funcionalidades funcionan automáticamente, mientras que otras requieren una configuración explícita.

Para cada servicio que ofrece compatibilidad integrada con varias regiones, en la tabla se muestra la siguiente información:

- **Tipo de compatibilidad con varias regiones:** La funcionalidad específica que proporciona el servicio, con un vínculo a la documentación que describe la funcionalidad con más detalle. Incluso cuando un servicio admite funcionalidades de varias regiones, revise la documentación de la funcionalidad para comprender las limitaciones o los requisitos de configuración.
- **Regiones emparejadas** y **regiones no emparejadas:** indica si la funcionalidad funciona entre regiones de ese tipo. Para obtener más información sobre las regiones emparejadas, consulte Azure pares de regiones y regiones no emparejadas. Algunas filas incluyen marcadores de nota al pie que apuntan a advertencias enumeradas debajo de la tabla. Los servicios no regionales muestran una marca de verificación en ambas columnas porque no están vinculadas a una región específica.
- **Descripción:** Un breve resumen de cómo se comporta la funcionalidad.

Los servicios sin compatibilidad integrada requieren que diseñe su propia solución de varias regiones, incluida la replicación, la conmutación por error y otros aspectos de la recuperación. Para ver los servicios que no proporcionan compatibilidad integrada con varias regiones, seleccione la pestaña **Sin compatibilidad integrada con varias regiones** .

Lista de servicios de Azure

Compatibilidad integrada con varias regiones

Expandir tabla

Service	Tipo de soporte técnico de varias regiones	Regiones emparejadas	Regiones no emparejadas	Descripción
Azure API Management	Implementación de varias regiones	✓	✓	Replica el componente de puerta de enlace en varias regiones para la optimización de latencia y la conmutación automática por error.
Azure App Configuration	Geo-replication	✓	✓	Replica los datos de configuración entre regiones con la funcionalidad de lectura y escritura activa-activa.
Azure App Service	Copias de seguridad personalizadas	✓ [1]	✓ [1]	Realiza una copia de seguridad de los datos de la aplicación en una cuenta de almacenamiento, que puede configurar con almacenamiento con redundancia geográfica (en regiones emparejadas) o replicación de objetos para la protección multiregión.
Azure App Service Environment	Copias de seguridad personalizadas	✓ [1]	✓ [1]	Realiza una copia de seguridad de los datos de la aplicación en una cuenta de almacenamiento, que puede configurar con almacenamiento con redundancia geográfica (en regiones emparejadas) o replicación de objetos para la protección multiregión.

Service	Tipo de soporte técnico de varias regiones	Regiones emparejadas	Regiones no emparejadas	Descripción
Azure Backup	Almacenamiento con redundancia geográfica y restauración entre regiones	✓		Replica los datos de copia de seguridad en la región emparejada para la recuperación ante desastres.
Azure Blob Storage	Azure Storage Mover	✓	✓	Copia los datos del blob en otra cuenta de almacenamiento, que puede estar en otra región.
Azure Blob Storage	Almacenamiento con redundancia geográfica (GRS)	✓		Replica los datos de forma asincrónica en la región emparejada con conmutación por error administrada por el cliente o Microsoft administrada.
Azure Blob Storage	Replicación de objetos	✓	✓	Copia los blobs en bloques de forma asincrónica entre una cuenta de almacenamiento de origen y una cuenta de destino, que puede estar en otra región.
Azure Container Registry	Geo-replication	✓	✓	Replica imágenes y artefactos de contenedor entre regiones con funcionalidad de lectura y escritura activas/escritura.
Azure Cosmos DB	Varias regiones de lectura con una sola región de escritura	✓	✓	Replica datos en varias regiones de lectura, lo que mejora la disponibilidad de lectura y la latencia global, con conmutación por error opcional para escrituras.

Service	Tipo de soporte técnico de varias regiones	Regiones emparejadas	Regiones no emparejadas	Descripción
Azure Cosmos DB	Varias regiones de escritura	✓	✓	Replica los datos entre regiones con funcionalidad de escritura múltiple activa-activa y resolución automática de conflictos.
Azure Cosmos DB	Copia de seguridad periódica	✓		Realiza copias de seguridad de los datos automáticamente a intervalos regulares y usa la replicación grS de almacenamiento para mantener copias en la región emparejada de forma predeterminada.
Azure Data Factory	Microsoft conmutación por error administrada en una región emparejada	✓ [2]		Replica los metadatos en la región emparejada. Microsoft puede iniciar la conmutación por error en la región emparejada durante interrupciones prolongadas. En regiones no emparejadas, puede diseñar su propia solución multiregión para Azure Data Factory.
Azure Data Lake Storage Gen2 (espacios de nombres jerárquicos)	Almacenamiento con redundancia geográfica (GRS)	✓		Replica los datos de forma asincrónica en la región emparejada con conmutación por error administrada por el cliente o Microsoft administrada.

Service	Tipo de soporte técnico de varias regiones	Regiones emparejadas	Regiones no emparejadas	Descripción
Azure Database for MySQL	Réplicas de lectura entre regiones	✓	✓	Replica los datos de forma asincrónica para leer réplicas en otras regiones con conmutación por error manual.
Azure Database for PostgreSQL	Réplicas de lectura entre regiones	✓	✓	Replica los datos de forma asincrónica para leer réplicas en otras regiones con puntos de conexión virtuales para la conmutación por error.
Azure DDoS Protection	Protección multiregión	✓	✓	Protege las direcciones IP públicas en varias regiones de un único plan de protección de red DDoS que se implementa en una región.
Azure Registro de dispositivos	Microsoft conmutación por error administrada en una región emparejada	✓		Replica los datos en la región emparejada. Microsoft puede iniciar la conmutación por error durante interrupciones prolongadas.
Azure DNS	Servicio no regional	✓	✓	Opera la infraestructura en varias regiones de Azure.
Azure Event Grid	Recuperación ante desastres geográfica de metadatos	✓		Replica los metadatos de forma sincrónica en la región emparejada. Los datos de eventos no se replican. En regiones no

Service	Tipo de soporte técnico de varias regiones	Regiones emparejadas	Regiones no emparejadas	Descripción
				emparejadas, puede diseñar su propia solución multiregión para Azure Event Grid.
Azure Event Hubs	Geo-replication	✓	✓	Replica los metadatos y los datos de eventos entre regiones con opciones sincrónicas o asincrónicas.
Azure Event Hubs	Recuperación ante desastres geográfica de metadatos	✓	✓	Replica solo metadatos (entidades, configuración). Los datos de eventos no se replican.
Azure Files	Almacenamiento con redundancia geográfica (GRS)	✓		Replica los datos de forma asincrónica en la región emparejada con conmutación por error administrada por el cliente o Microsoft administrada. Puede designar su propia solución de varias regiones para Azure Files, incluidos los recursos compartidos de archivos Premium o cuando se usan regiones no emparejadas.
Azure Front Door	Servicio no regional	✓	✓	Opera la infraestructura en varias regiones de Azure.
Azure IoT Hub	Microsoft conmutación por error administrada en una región emparejada	✓		Replica los datos de forma asincrónica en la región emparejada con Microsoft conmutación por error iniciada por el cliente

Service	Tipo de soporte técnico de varias regiones	Regiones emparejadas	Regiones no emparejadas	Descripción
				o administrada por el cliente. En regiones no emparejadas, puede diseñar su propia solución multiregión para Azure IoT Hub.
Azure Key Vault	Microsoft conmutación por error administrada en una región emparejada	✓ 2		Replica los datos de forma asincrónica en la región emparejada con la conmutación por error administrada por Microsoft. En regiones no emparejadas, puede diseñar su propia solución multiregión para Azure Key Vault.
Azure Key Vault Managed HSM	Replicación de varias regiones	✓ 2	✓ 2	Replica los datos en dos regiones con enrutamiento basado en Traffic Manager.
Azure Kubernetes Service (AKS)	Azure Kubernetes Fleet Manager	✓	✓	Administra varios clústeres de AKS entre regiones con orquestación centralizada y enrutamiento de cargas de trabajo de varios clústeres.
Azure Kubernetes Service (AKS)	Backup	✓		Realiza copias de seguridad de aplicaciones y datos en contenedores en el clúster de AKS, con compatibilidad con la restauración en una región emparejada mediante el nivel de almacén.

Service	Tipo de soporte técnico de varias regiones	Regiones emparejadas	Regiones no emparejadas	Descripción
Azure Load Balancer	Balanceador de carga global	✓ [2]	✓ [2]	En el caso de los equilibradores de carga públicos, el nivel Global proporciona una única dirección IP de difusión que enruta el tráfico al equilibrador de carga regional más cercano con conmutación automática por error.
Azure Managed Redis	Replicación geográfica activa	✓	✓	Se replica en varias regiones activas y activas con conmutación por error manual.
Azure Monitor Logs	Replicación del área de trabajo	✓ [2]	✓ [2]	Replica los registros y la configuración en una región secundaria dentro del mismo grupo de regiones con conmutación manual.
Azure NetApp Files	Replicación entre regiones	✓ [2]	✓ [2]	Replica los volúmenes de forma asincrónica entre pares de regiones seleccionadas Microsoft, incluidos algunos pares no estándar.
Azure Notification Hubs	Recuperación ante desastres geográfica de metadatos	✓	✓ [2]	Replica los metadatos en una región secundaria. Los datos de registro no se replican.
Azure Queue Storage	Almacenamiento con redundancia geográfica (GRS)	✓		Replica los datos de forma asincrónica en la región emparejada con conmutación por error administrada por

Service	Tipo de soporte técnico de varias regiones	Regiones emparejadas	Regiones no emparejadas	Descripción
				el cliente o Microsoft administrada. En regiones no emparejadas, puede diseñar su propia solución multiregión para Azure Queue Storage.
Azure Service Bus	Replicación geográfica:	✓	✓	Replica los metadatos y los datos de mensajes entre regiones con opciones sincrónicas o asincrónicas.
Azure Service Bus	Recuperación de metadatos Geo-Disaster	✓	✓	Replica solo metadatos (entidades, configuración). Los datos del mensaje no se replican.
Azure SignalR Service	Geo-replication	✓	✓	Replica entre regiones con enrutamiento basado en Traffic Manager y conmutación automática por error.
Azure Site Recovery	Replicación de máquinas virtuales entre regiones	✓	✓	Replica máquinas virtuales de forma asincrónica en una región secundaria.
Azure SQL Database	Replicación geográfica activa	✓	✓	Replica una base de datos única de forma asincrónica en una región secundaria con conmutación por error manual.
Azure SQL Database	Copias de seguridad automatizadas	✓		Realiza copias de seguridad de bases de datos automáticamente a intervalos regulares y

Service	Tipo de soporte técnico de varias regiones	Regiones emparejadas	Regiones no emparejadas	Descripción
				almacena copias en la región emparejada mediante el almacenamiento con redundancia geográfica.
Azure SQL Database	Grupos de conmutación por error	✓	✓	Replica grupos de bases de datos con conmutación automática por error y redirección de punto de conexión transparente.
Azure SQL Managed Instance	Copias de seguridad automatizadas	✓		Realiza copias de seguridad de instancias administradas automáticamente a intervalos regulares y almacena copias en la región emparejada mediante el almacenamiento con redundancia geográfica.
Azure SQL Managed Instance	Grupos de conmutación por error	✓	✓	Replica los datos de instancia en una región secundaria con conmutación por error administrada por el cliente o Microsoft administrada por el cliente.
Azure Storage Actions	Conmutación por error automatizada	✓		Continúa ejecutando tareas futuras en la región emparejada cuando la cuenta de almacenamiento de destino usa GRS.

Service	Tipo de soporte técnico de varias regiones	Regiones emparejadas	Regiones no emparejadas	Descripción
Azure Storage Mover	recuperación ante desastres iniciada Azure	✓		Microsoft puede iniciar la conmutación por error del servicio Storage Mover en la región emparejada durante interrupciones prolongadas. Independientemente de dónde se implemente Storage Mover, puede copiar entre cualquier región de origen y destino admitida.
Azure Table Storage	Almacenamiento con redundancia geográfica (GRS)	✓		Replica los datos de forma asincrónica en la región emparejada con conmutación por error administrada por el cliente o Microsoft administrada. En regiones no emparejadas, puede diseñar su propia solución de varias regiones para Azure Table Storage.
Azure Traffic Manager	Servicio no regional	✓	✓	Opera la infraestructura en varias regiones de Azure.
Azure Virtual Machines	Azure Site Recovery	✓ [1]	✓ [1]	Replica máquinas virtuales de forma asincrónica en una región secundaria.
Azure Web PubSub	Geo-replication	✓	✓	Replica entre regiones con enrutamiento basado en Traffic Manager y

Service	Tipo de soporte técnico de varias regiones	Regiones emparejadas	Regiones no emparejadas	Descripción
				conmutación automática por error.
Microsoft Entra ID	Almacenamiento de datos ubicados geográficamente y en todo el mundo	✓	✓	Opera la infraestructura en varias regiones de Azure.
Microsoft Fabric	Recuperación ante desastres entre regiones	✓		Proporciona recuperación ante desastres entre regiones en regiones emparejadas.
Microsoft Foundry Models	Implementaciones de zona de datos y globales	✓	✓	Admite implementaciones globales, de zona de datos y de modelo de una sola región.

[1] La compatibilidad con varias regiones para estos servicios puede requerir que configure varios servicios Azure.

[2] Estos servicios tienen restricciones de selección de regiones para las funcionalidades de varias regiones. Revise cuidadosamente sus guías de confiabilidad para confirmar las combinaciones de regiones admitidas y los requisitos previos relacionados antes de diseñar la arquitectura.

Contenido relacionado

- guías de confiabilidad del servicio Azure
- Servicios de Azure que admiten zonas de disponibilidad
- Servicios de Azure no regionales
- Lista de regiones de Azure
- Disponibilidad del servicio por categoría
- Azure Well-Architected Framework: estrategias de arquitectura para usar regiones y zonas de disponibilidad

Last updated on 02/06/2026

Soluciones de varias regiones en regiones no emparejadas

Algunos servicios de Azure admiten la redundancia geográfica y la replicación geográfica en regiones de Azure, ven cuando esas regiones no están emparejadas.

En este artículo se enumeran algunos de los servicios y las posibles configuraciones para las soluciones de varias regiones que no requieren regiones emparejadas. Para obtener una lista de las funcionalidades de varias regiones de servicio, consulte Azure services que admiten varias regiones. Para obtener información detallada sobre cada servicio, consulte su guía de confiabilidad.

Búsqueda de Azure AI

Para obtener información sobre cómo crear soluciones de varias regiones mediante Búsqueda de Azure AI, consulte Confiabilidad en la búsqueda de IA.

Azure API Management

Para obtener información sobre cómo crear soluciones de varias regiones mediante Azure API Management, consulte Confiabilidad en API Management.

Azure App Service

Para obtener información sobre cómo crear soluciones de varias regiones mediante Azure App Service, consulte Confiabilidad en App Service.

Azure Container Registry

Para obtener información sobre cómo crear soluciones de varias regiones mediante Azure Container Registry, consulte Confiabilidad en Container Registry.

Azure Cosmos DB

Para obtener información sobre cómo crear soluciones de varias regiones mediante Azure Cosmos DB, consulte Confiabilidad en Azure Cosmos DB.

Base de Datos Azure para MySQL

Para obtener información sobre cómo crear soluciones de varias regiones mediante Azure Database for MySQL, consulte Confiabilidad en Azure Database for MySQL.

Azure Database for PostgreSQL

Para obtener información sobre cómo crear soluciones de varias regiones mediante Azure Database for PostgreSQL, consulte Confiabilidad en Azure Database for PostgreSQL.

Azure Data Factory

Para obtener información sobre cómo crear soluciones de varias regiones mediante Azure Data Factory, consulte Confiabilidad en Azure Data Factory.

Azure Event Grid

Para obtener información sobre cómo crear soluciones de varias regiones mediante Azure Event Grid, consulte Confiabilidad en Azure Event Grid.

Azure IoT Hub

Para obtener información sobre cómo crear soluciones de varias regiones mediante Azure IoT Hub, consulte Confiabilidad en IoT Hub.

Azure Kubernetes Service (AKS)

Para obtener información sobre cómo crear soluciones de varias regiones mediante Azure Kubernetes Service (AKS), consulte Confiabilidad en AKS.

registros de Azure Monitor

Para obtener información sobre cómo crear soluciones de varias regiones mediante registros de Azure Monitor, consulte Confiabilidad en registros de Azure Monitor.

Azure Service Bus

Azure Service Bus puede proporcionar resistencia regional sin dependencia de los pares de regiones. Para obtener más información, consulte Reliability in Service Bus.

Azure SQL Database

Para obtener información sobre cómo crear soluciones de varias regiones mediante Azure SQL Database, consulte Confiabilidad en Azure SQL Database.

Azure SQL Managed Instance

Para obtener información sobre cómo crear soluciones de varias regiones mediante Azure SQL Managed Instance, consulte Confiabilidad en Azure SQL Managed Instance.

Azure Storage

Para lograr la replicación geográfica en regiones no emparejadas:

- **Para Azure Blob Storage (almacenamiento de objetos):** para obtener información sobre cómo crear soluciones de varias regiones mediante Azure Blob Storage, consulte Confiabilidad en Blob Storage.
- **Para obtener Azure NetApp Files:** Para obtener información sobre cómo crear soluciones de varias regiones mediante Azure NetApp Files, consulte Confiabilidad en Azure NetApp Files.
- **Para obtener Azure Files:** Para obtener información sobre cómo crear soluciones de varias regiones mediante Azure Files, consulte Confiabilidad en Azure Files.
- **Para Queue Storage:** Para obtener información sobre cómo crear soluciones de varias regiones mediante Azure Queue Storage, consulte Confiabilidad en Queue Storage.
- **Para obtener Azure Table Storage:** Para obtener información sobre cómo crear soluciones de varias regiones mediante Azure Table Storage, consulte Confiabilidad en Azure Table Storage.

Azure Virtual Desktop

Para la replicación geográfica en regiones no emparejadas para Azure Virtual Desktop, hay que considerar las máquinas virtuales del host de sesión y el almacenamiento para perfiles de usuario, aplicaciones y datos. Microsoft administra el plano de control de Azure Virtual Desktop, que se distribuye globalmente y tiene alta disponibilidad.

- En el caso de los hosts de sesión, puede implementar máquinas virtuales en varias regiones en un escenario activo-activo o replicarlas entre regiones mediante Azure Site Recovery en

un escenario activo-pasivo.

- Para obtener almacenamiento, consulte Azure Storage.

Para obtener más información, consulte Multiregion business continuity and disaster recovery (BCDR) for Azure Virtual Desktop and Azure Virtual Desktop service architecture and resilience.

Azure Virtual Machines

Para obtener información sobre cómo crear soluciones de varias regiones mediante Azure Virtual Machines, consulte Confiabilidad en Azure Virtual Machines.

Pasos siguientes

- servicios Azure con zonas de disponibilidad
- Lista de regiones de Azure
- Guía de confiabilidad

ⓘ**Nota:** El autor ha creado este artículo con ayuda de inteligencia artificial. Más información

---

Last updated on 13/08/2026

Azure servicios no regionales

Los servicios Azure no regionales son servicios que no se implementan en una región de Azure específica. En este artículo se explica la diferencia entre los servicios no regionales globales y geográficos, y se proporciona una lista de servicios no regionales que indican si cada uno es un servicio global o geográfico.

Tipos de servicios no regionales

Hay dos tipos de servicios no regionales:

- **Global services** se implementan en muchas regiones de Azure en todo el mundo. Si se produce un error regional, una instancia de otra región puede seguir funcionando. Con los servicios globales, no es necesario diseñar específicamente para interrupciones en toda la región, ya que estos servicios son resistentes a errores regionales.
- servicios **Geographic** se implementan en un área geográfica y no en una región de Azure específica. Dentro de esa área geográfica, se pueden usar una o varias regiones de Azure para atender solicitudes de cliente. Para comprender las áreas geográficas y aprender a diseñar una solución resistente con un servicio geográfico, consulte la documentación de ese servicio.

Por lo general, la mayoría de los servicios no regionales son globales o geográficos, pero no ambos. Por ejemplo, Azure Front Door un servicio global, mientras que Azure DevOps es un servicio geográfico. Sin embargo, Microsoft Entra ID es un servicio que puede ser global o geográfico, en función de cómo lo configure.

Alojamiento de datos para servicios no localizados regionalmente

Para la mayoría de los servicios no regionales que almacenan los datos de los clientes, puede seleccionar una región específica para el almacenamiento de datos. Para obtener más información, consulte la residencia de datos en Azure.

Componentes de servicio dependientes regionalmente

Cuando se usa un servicio no regional, no se elige una región de Azure específica para el propio servicio. Sin embargo, es posible que tenga que elegir una región para los componentes dependientes que use con el servicio. Por ejemplo, si usa Azure Virtual Desktop (AVD), que es un servicio no regional, puede elegir la región Azure donde reside el grupo de hosts (máquinas virtuales).

Lista de servicios no regionales

Microsoft publica una lista de servicios, que incluye servicios no regionales. La tabla siguiente también contiene la lista de servicios no regionales, pero especifica si cada servicio es *global* o *geográfico*, y proporciona notas adicionales sobre la residencia y configuración de datos.

Expandir tabla

Producto	Global	Geográfico	Notas
Azure AD B2C	✓		
Azure Advisor	✓		
Servicio de Bot de Azure AI	✓		
Azure Blueprints	✓		
Azure Cloud Shell	✓		
Azure Communication Services	✓		Algunos datos se almacenan en un área geográfica.
Azure Content Delivery Network	✓		
Azure DevOps		✓	
Azure DNS zonas privadas	✓		
Zonas públicas de Azure DNS	✓		
Azure Front Door	✓		
Azure Information Protection de Microsoft (un servicio de protección de información en la nube)	✓		
Azure Lighthouse	✓		
grupos de administración de Azure	✓		

Producto	Global	Geográfico	Notas
Azure Maps	✓		
Azure Migrate	✓		
Azure Peering Service	✓		
diagnósticos de rendimiento de Azure	✓		
Azure Policy	✓		
portal de Azure	✓		
Azure Resource Graph	✓		
Azure Stack Edge		✓	
Azure Static Web Apps	✓		
suscripciones de Azure	✓		
Azure Traffic Manager	✓		
Azure Virtual Desktop	✓		Algunos datos se almacenan en un área geográfica. Los grupos de hosts se encuentran en una región específica.
Microsoft Cost Management	✓		
Microsoft Defender para la nube	✓		
Microsoft Defender for Identity	✓		
Microsoft Defender para IoT	✓		
Microsoft Entra ID	✓	✓	Seleccione la ubicación durante la creación de un arrendatario de Microsoft Entra, que puede ser tanto geográfica como global. Consulte Microsoft Entra ID y residencia de datos.
Microsoft Graph	✓		
Microsoft Intune	✓		
Microsoft Sentinel		✓	

Contenido relacionado

- Residencia de Datos en Azure

- Lista de regiones de Azure

ⓘ**Nota:** El autor ha creado este artículo con ayuda de inteligencia artificial. Más información

---

Last updated on 26/08/2026

lista de regiones de Azure

Azure proporciona la superficie global más amplia de cualquier proveedor de nube y está abriendo rápidamente nuevas regiones. Este artículo contiene una lista de todas las regiones de la nube pública de Azure, el estado de la región emparejada, la ubicación física, la geografía y la compatibilidad con zonas de disponibilidad .

Leyenda

Expandir tabla

Símbolo	Description
	La región estará disponible próximamente. Para obtener más información sobre las zonas de disponibilidad y el soporte técnico de servicios disponibles en esta región, póngase en contacto con su representante de clientes o ventas de Microsoft. Para ver las próximas regiones que admiten zonas de disponibilidad, consulte Azure geographies .
	El acceso a esta región está restringido para admitir escenarios de clientes específicos, como la recuperación ante desastres dentro de un área geográfica específica. Para solicitar acceso a una región restringida para la suscripción de Azure, consulte Proceso de solicitud de acceso a la región de Azure.

Nota:

Incluso cuando una región proporciona zonas de disponibilidad, es posible que algunos servicios no las admitan en esa región. Consulte las guías de confiabilidad del servicio Azure para obtener información sobre la compatibilidad regional con los servicios habilitados para zonas de disponibilidad.

lista de regiones de Azure

Todos

Expandir tabla

Región	Soporte para zonas de disponibilidad	Región emparejada	Ubicación física	Geografía	Nombre programático
Centro de Australia		Centro de Australia 2	Canberra	Australia	australiacentral
Centro de Australia 2		Centro de Australia	Canberra	Australia	australiacentral2
Australia East	✓	Australia Southeast	Nueva Gales del Sur	Australia	australiaeast
Australia Southeast		Australia East	Victoria	Australia	australiasoutheast
Este de Austria	✓	N/A	Viena	Austria	austriaeast
Centro de Bélgica	✓	N/A	Brussels	Bélgica	belgiumcentral
Sur de Brasil	✓	Centro-sur de EE. UU.	Estado de Sao Paulo	Brasil	brazilsouth
Sudeste de Brasil		Sur de Brasil	Río	Brasil	brazilsoutheast
Canada Central	✓	Este de Canadá	Toronto	Canadá	canadacentral
Este de Canadá		Canada Central	Quebec	Canadá	canadaeast
Centro de la India	✓	South India	Pune	India	centralindia
Central US	✓	Este de EE. UU. 2	Iowa	Estados Unidos	centralus
Centro de Chile	✓	N/A	Santiago	Chile	chilecentral
Este de Dinamarca	✓	N/A	Copenhague	Dinamarca	denmarkeast
Este de Asia	✓	Sudeste Asiático	RAE de Hong Kong	Asia Pacífico	eastasia
East US	✓	West US	Virginia	Estados	eastus

Región	Soporte para zonas de disponibilidad	Región emparejada	Ubicación física	Geografía	Nombre programático
				Unidos	
Este de EE. UU. 2	✓	Central US	Virginia	Estados Unidos	eastus2
Centro de Francia	✓	Sur de Francia	París	Francia	francecentral
Sur de Francia		Centro de Francia	Marsella	Francia	francesouth
Norte de Alemania		Centro-oeste de Alemania	Berlín	Alemania	norte de Alemania
Centro-oeste de Alemania	✓	Norte de Alemania	Fráncfort	Alemania	germanywestcentral
Centro de Indonesia	✓	N/A	Yakarta	Indonesia	indonesiacentral
Israel Central	✓	N/A	Israel	Israel	israelcentral
Norte de Italia	✓	N/A	Milán	Italia	italynorth
Japón Oriental	✓	Oeste de Japón	Tokio, Saitama	Japón	japaneast
Oeste de Japón	✓	Japón Oriental	Osaka	Japón	japón occidente
Centro de Corea del Sur	✓	Corea del Sur	Seúl	Corea	koreacentral
Corea del Sur		Centro de Corea del Sur	Busán	Corea	koreasouth
Oeste de Malasia	✓	N/A	Kuala Lumpur	Malasia	malaysiawest
Centro de México	✓	N/A	Estado de Querétaro	México	mexico central

Región	Soporte para zonas de disponibilidad	Región emparejada	Ubicación física	Geografía	Nombre programático
Norte de Nueva Zelanda	✓	N/A	Auckland	Nueva Zelanda	nuevazelandanorte
Centro-norte de EE. UU.		Centro-sur de EE. UU.	Illinois	Estados Unidos	northcentralus
Norte de Europa	✓	West Europe	Irlanda	Europe	northeurope
Norway East	✓	Oeste de Noruega	Noruega	Noruega	norwayeast
Oeste de Noruega		Norway East	Noruega	Noruega	Noruega Oeste
Centro de Polonia	✓	N/A	Varsovia	Polonia	polandcentral
Qatar Central	✓	N/A	Doha	Qatar	qatarcentral
Norte de Sudáfrica	✓	Oeste de Sudáfrica	Johannesburgo	Sudáfrica	southafricanorth
Oeste de Sudáfrica		Norte de Sudáfrica	Ciudad del Cabo	Sudáfrica	southafricawest
Centro-sur de EE. UU.	✓	Centro-norte de EE. UU.	Texas	Estados Unidos	southcentralus
South India		Centro de la India	Chennai	India	southindia
Sudeste Asiático	✓	Este de Asia	Singapur	Asia Pacífico	southeastasia
Centro de España	✓	N/A	Madrid	España	spaincentral
Centro de Suecia	✓	Sur de Suecia	Gävle	Suecia	swedencentral
Norte de Suiza	✓	Oeste de Suiza	Zúrich	Suiza	switzerlandnorth

Región	Soporte para zonas de disponibilidad	Región emparejada	Ubicación física	Geografía	Nombre programático
Oeste de Suiza		Norte de Suiza	Ginebra	Suiza	switzerlandwest
Centro de Emiratos Árabes Unidos		Norte de Emiratos Árabes Unidos	Abu Dabi	Emiratos Árabes Unidos	uaecentral
Norte de Emiratos Árabes Unidos	✓	Centro de Emiratos Árabes Unidos	Dubai	Emiratos Árabes Unidos	uaenorth
UK South	✓	UK West	Londres	Reino Unido	uksouth
UK West		UK South	Cardiff	Reino Unido	ukwest
Centro-oeste de EE. UU.		Oeste de EE. UU. 2	Wyoming	Estados Unidos	westcentralus
West Europe	✓	Norte de Europa	Países Bajos	Europe	westeurope
West India		South India	Mumbai	India	westindia
West US		East US	California	Estados Unidos	westus
Oeste de EE. UU. 2	✓	Centro-oeste de EE. UU.	Washington	Estados Unidos	westus2
Oeste de EE. UU. 3	✓	East US	Fénix	Estados Unidos	westus3

Azure nombres de región programáticos

Para recuperar una lista de regiones, incluidos los nombres que se pueden usar para la programación y el scripting, puede usar los métodos siguientes:

- CLI de Azure: az account list-locations
- Azure PowerShell- Get-AzLocation

- Azure Resource Manager API REST

Regiones de nube soberana

Para obtener información sobre las regiones de las nubes soberanas de Azure, consulte los siguientes recursos:

- Regiones de Azure para la Administración Pública de Ee. UU.
- Azure en regiones de China

Contenido relacionado

- ¿Qué son las regiones de Azure?
- Pares de regiones de Azure y regiones que no están emparejadas
- Microsoft se compromete a expandir Azure zonas de disponibilidad a más regiones

---

Last updated on 21/05/2026

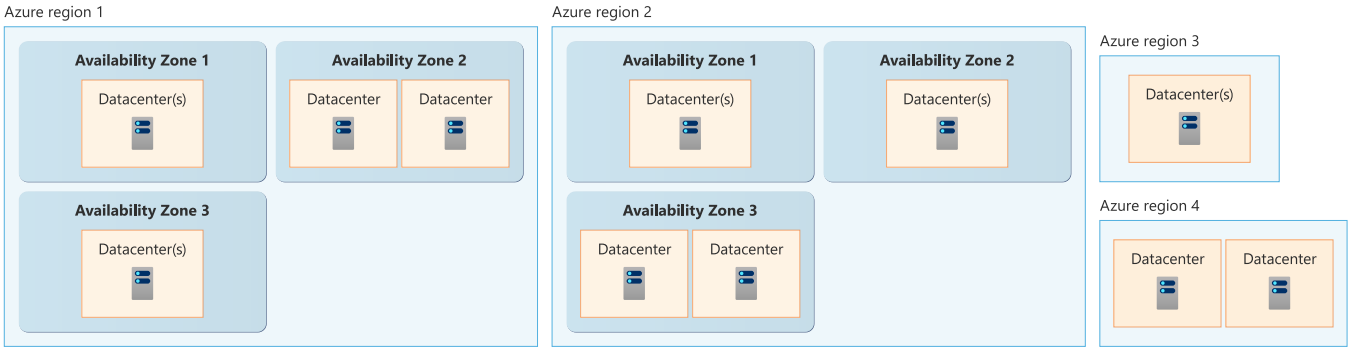
¿Qué son las zonas de disponibilidad?

https://learn-video.azurefd.net/vod/player?id=d36b5b2d-8bd2-43df-a796-b0c77b2f82fc&locale=es-es&embedUrl=%2Fazure%2Freliability%2Favailability-zones-overview

Muchas regiones de Azure proporcionan zonas *availability*, que son grupos de centros de datos separados dentro de una región. Cada zona de disponibilidad tiene una infraestructura independiente de alimentación, refrigeración y red, de modo que, si una zona experimenta una interrupción, el resto de zonas poseen los servicios regionales, la capacidad y la alta disponibilidad necesarios para respaldarla. Algunos servicios Azure usan automáticamente varias zonas, mientras que otras requieren que configure la implementación de varias zonas.

Las zonas de disponibilidad no solo protegen contra interrupciones a gran escala que afectan a toda una zona. También proporcionan resiliencia frente a fallos de menor alcance, como un rack de servidores o un fallo de clúster dentro de una zona. En estos escenarios, las cargas de trabajo pueden seguir ejecutándose en otras zonas, incluso si las partes de la zona afectada permanecen operativas.

En el diagrama siguiente se muestran varios ejemplos de regiones de Azure. Las regiones 1 y 2 admiten zonas de disponibilidad, mientras que las regiones 3 y 4 no tienen zonas de disponibilidad. Algunas zonas tienen un centro de datos y otros tienen varios.



> **Sugerencia**
>
> Para ver qué regiones admiten zonas de disponibilidad, consulte **List de regiones de Azure**.

Centros de datos y zonas de disponibilidad

Una zona de disponibilidad es una *agrupación lógica* de uno o varios centros de datos físicamente independientes dentro de una región. Cada zona de disponibilidad está integrada de una manera que si algo va mal en uno (como una interrupción de energía o un problema de red), los demás siguen funcionando. Un único centro de datos no ofrece este nivel de protección por sí solo.

Las zonas de disponibilidad suelen estar separadas por varios kilómetros, y normalmente están dentro de 100 kilómetros entre sí. Esto proporciona conectividad de baja latencia al tiempo que reduce la probabilidad de interrupciones locales o eventos meteorológicos que afectan simultáneamente a varias zonas.

Las ubicaciones del centro de datos se seleccionan mediante criterios rigurosos de evaluación de riesgos de vulnerabilidades. Este proceso identifica todos los riesgos significativos específicos del centro de datos y tiene en cuenta los riesgos compartidos entre las zonas de disponibilidad.

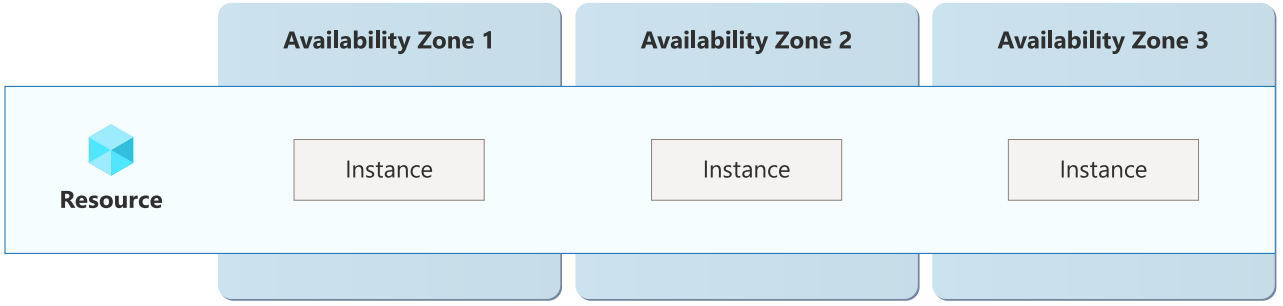
Tipos de compatibilidad con zonas de disponibilidad

Los servicios de Azure pueden proporcionar diferentes tipos de soporte de zona de disponibilidad para sus recursos, y la mayoría de los servicios admiten la implementación de recursos *con redundancia de zona* o *zonales*.

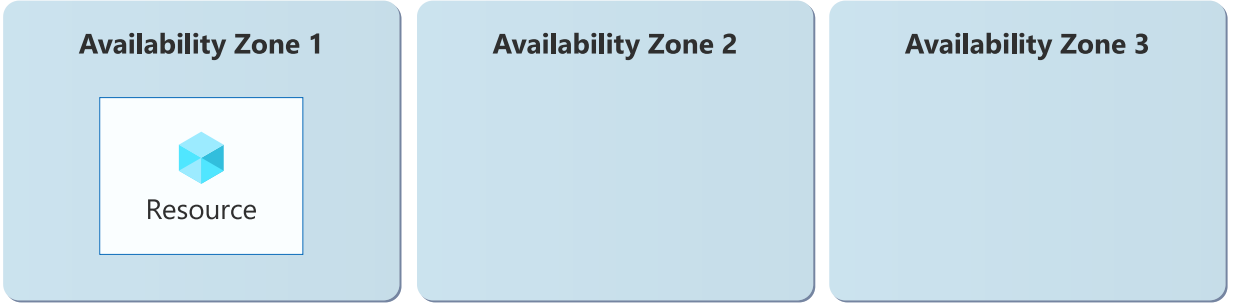
- **Recursos con redundancia** de zona: los recursos con redundancia de zona se replican o distribuyen entre varias zonas de disponibilidad por el servicio. Por ejemplo: los servicios de datos con redundancia de zona replican los datos en múltiples zonas para que un error en una zona no afecte a la alta disponibilidad de los datos. Algunos servicios son automáticamente con redundancia de zona en regiones admitidas, mientras que otros servicios requieren que configure el recurso para que tenga redundancia de zona.

  Para la mayoría de los servicios, Microsoft selecciona las zonas que usan los recursos, pero a veces puede seleccionar el conjunto de zonas. Los servicios pueden usar un número o subconjunto diferente de zonas en una región en función de su diseño. Un recurso es resistente a un error de zona siempre que use dos o más zonas.

  Con implementaciones con redundancia de zona, Microsoft administra la propagación de solicitudes entre zonas y la replicación de datos entre zonas. Si se produce una interrupción en una zona de disponibilidad, Microsoft gestiona el failover a otra zona automáticamente.



- **Recursos zonales**: un recurso zonal se implementa en una sola zona de disponibilidad que seleccione usted mismo.



Las implementaciones zonales no proporcionan automáticamente resistencia a interrupciones de zona de disponibilidad. Sin embargo, los recursos zonales están aislados de fallos en otras zonas. También pueden ayudarle a lograr requisitos de rendimiento o latencia inusualmente estrictos. Por ejemplo, para una carga de trabajo de chatty creada con máquinas virtuales, puede optar por implementar varias máquinas virtuales en la misma zona para reducir la latencia de las conexiones entre ellas.

Para que los recursos zonales sean resistentes a las interrupciones de zona de disponibilidad, debe diseñar una arquitectura con recursos independientes en varias zonas de disponibilidad dentro de la región. Microsoft no administra el proceso por usted. Si se produce una interrupción en una zona de disponibilidad, usted es responsable de la conmutación por error a otra zona.

Al configurar un recurso para que sea redundante de zona o si usa varias instancias de un recurso zonal en distintas zonas de disponibilidad, el recurso se considera *resistente a la zona*: es decir, es resistente a la interrupción de una sola zona de disponibilidad. Para más información sobre cómo usar implementaciones zonales y mantener la resistencia de zona, consulte Recursos zonales y resistencia de zona.

Al diseñar la estrategia de confiabilidad, asegúrese de comprender cómo cada servicio de la carga de trabajo admite zonas de disponibilidad. Por ejemplo, algunos servicios pueden tener requisitos adicionales para satisfacer la compatibilidad con zonas de disponibilidad, como determinados niveles o SKU. Las guías de confiabilidad contienen detalles de estos requisitos.

Si un recurso no está configurado para usar zonas de disponibilidad, ya sea que la región que utiliza no admite zonas o debido a sus opciones de configuración, se denomina una implementación *no zonal* o *regional*. Azure pueden colocar recursos no zonales en todas las zonas de la región. No decide qué recursos se asignan a qué zonas. Si alguna zona de disponibilidad de la región experimenta una interrupción, los recursos no zonales podrían estar en la zona afectada y podrían experimentar tiempo de inactividad.

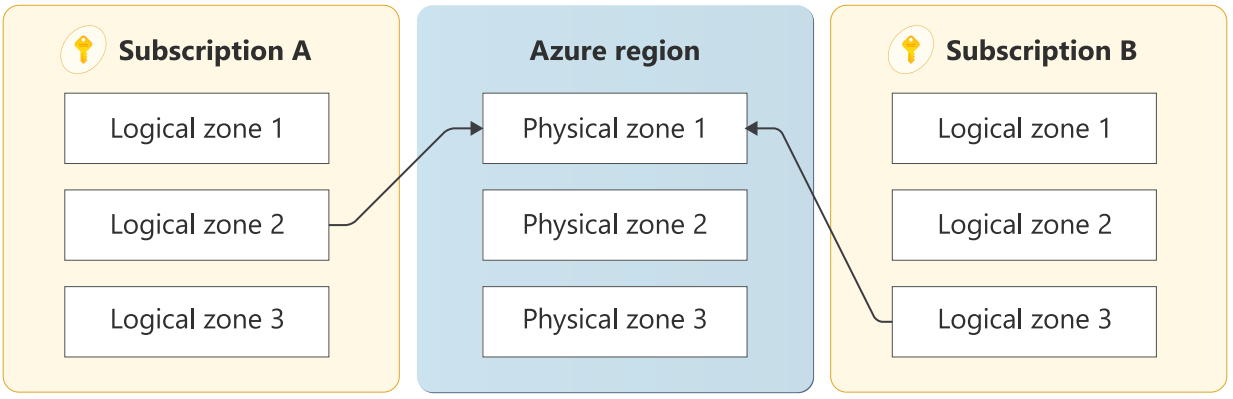
Configuración de recursos para la compatibilidad con zonas de disponibilidad

Cada servicio tiene su propio método para configurar la compatibilidad con zonas de disponibilidad y algunos servicios no requieren ninguna configuración. Para obtener información sobre cómo cada servicio admite zonas de disponibilidad y cómo configurar esa compatibilidad, consulte Azure guías de confiabilidad por servicio.

Zonas de disponibilidad físicas y lógicas

Cada centro de datos se asigna a una zona física. Las zonas físicas se asignan a zonas lógicas de la suscripción de Azure y es posible que diferentes suscripciones tengan un orden de asignación diferente. Las suscripciones de Azure se asignan automáticamente a su correspondencia en el momento en que se crea la suscripción. Debido a esto, la asignación de zona para una suscripción podría ser diferente para otras suscripciones.

Por ejemplo, la suscripción A puede tener la zona 1 física asignada a la zona 2 lógica, mientras que la suscripción B tiene la zona 1 física asignada a la zona 3 lógica:



Para comprender la asignación entre zonas lógicas y físicas de la suscripción, use el CLI de Azure o Azure PowerShell o las API de Azure Resource Manager.

CLI de Azure | API

Use el comando az account list-locations :

Azure CLI

```
az account list-locations \
    --query "[?availabilityZoneMappings].{availabilityZoneMappings:
availabilityZoneMappings, displayName: displayName, name: name}"
```

Zonas de disponibilidad y actualizaciones de Azure

Para cada región, Microsoft pretende implementar actualizaciones en Azure servicios dentro de una sola zona de disponibilidad a la vez. Este enfoque reduce el impacto que podrían tener las actualizaciones en una carga de trabajo activa, ya que la carga de trabajo puede seguir ejecutándose en otras zonas mientras la actualización está en proceso. Para aprovechar las actualizaciones de zona secuenciadas, la carga de trabajo ya debe estar configurada para ejecutarse en varias zonas. Para obtener más información sobre cómo Azure implementa las actualizaciones, consulte Advancing safe deployment practices .

Redes entre zonas

Dentro de cada región, las zonas de disponibilidad se conectan a través de una red de alto rendimiento. Microsoft se esfuerza por lograr una comunicación entre zonas con latencia de ida y vuelta de menos de 2 milisegundos aproximadamente. La baja latencia permite una comunicación de alto rendimiento dentro de una región y la replicación sincrónica de datos en varias zonas de disponibilidad. La latencia de destino hace referencia a la latencia de los vínculos

de red. Según el protocolo de comunicación que use y los saltos de red necesarios para cualquier flujo de red específico, la latencia que observe puede ser diferente.

En la mayoría de las cargas de trabajo, puede distribuir componentes de la solución entre zonas de disponibilidad sin un efecto notable en el rendimiento. Si tiene una carga de trabajo con un alto grado de sensibilidad a la latencia entre zonas, es importante probar la latencia entre las zonas de disponibilidad seleccionadas con los protocolos y la configuración reales. Para reducir el tráfico entre zonas, es posible usar implementaciones zonales, pero de forma óptima, debe usar varias zonas de disponibilidad en el plan de estrategia de confiabilidad. Para más información sobre cómo usar implementaciones zonales y mantener la resistencia de zona, consulte Recursos zonales y resistencia de zona.

Azure no cobra por la transferencia de datos entre zonas de disponibilidad de la misma región, tanto si usa direcciones IP públicas como privadas.

Guía de arquitectura de zona de disponibilidad

Para lograr cargas de trabajo confiables:

- Las cargas de trabajo de producción deberán configurarse para usar varias zonas de disponibilidad en caso de que la región en la que se encuentren admita las zonas de disponibilidad.
- En el caso de las cargas de trabajo críticas, debe tener en cuenta una solución que se *tanto* varias regiones como varias zonas.

En el caso de las cargas de trabajo restringidas a una sola región por requisitos de residencia o soberanía de datos, la implementación en varias zonas de disponibilidad es la forma recomendada principal de maximizar la disponibilidad sin mover datos fuera de la región. Una configuración resistente a zonas mantiene las réplicas dentro del límite regional mientras se mantiene la continuidad si se produce un error en una sola zona.

Las zonas de disponibilidad no protegen contra una interrupción de toda la región. Si los requisitos de cumplimiento no permiten usar una región de Azure secundaria, planee la redundancia en la región junto con procedimientos de copia de seguridad y restauración, como mediante Azure Backup y otros enfoques de copia de seguridad. Establezca objetivos de recuperación que reflejen el tiempo necesario para recuperarse de una interrupción en toda la región.

Para obtener información más detallada sobre cómo usar regiones y zonas de disponibilidad en una arquitectura de solución, consulte Recomendaciones para usar zonas de disponibilidad y

regiones.

Regiones con zonas de disponibilidad

Las regiones de la tabla siguiente admiten zonas de disponibilidad. Para obtener información más completa sobre cada región, consulte Lista de regiones de Azure.

Expandir tabla

Americas	Europe	Oriente Medio	Africa	Asia Pacífico
Sur de Brasil	Este de Austria	Israel Central	Norte de Sudáfrica	Este de Australia
Centro de Canadá	Centro de Bélgica	Qatar Central		Central India
Centro de EE. UU.	Este de Dinamarca	UAE North		East Asia
Chile Central	France Central			Centro-sur de la India
Este de EE. UU.	Centro-oeste de Alemania			Indonesia Central
Este de EE. UU. 2	Norte de Italia			Japan East
Centro de México	North Europe			Japan West
Centro-sur de EE. UU.	Norway East			Korea Central
Oeste de EE. UU. 2	Poland Central			Malaysia West
Oeste de EE. UU. 3	Spain Central			Norte de Nueva Zelanda
	Sweden Central			Southeast Asia
	Switzerland North			
	Sur de Reino Unido			
	West Europe			

Contenido relacionado

- ¿Qué son las regiones de Azure?

- Lista de regiones de Azure
- servicios Azure con zonas de disponibilidad
- Recomendaciones para el uso de zonas de disponibilidad y regiones

ⓘ**Nota:** El autor ha creado este artículo con ayuda de inteligencia artificial. Más información

---

Last updated on 26/06/2026

Azure servicios que admiten zonas de disponibilidad

Azure aumenta continuamente el número de servicios que admiten zonas availability, incluidas las ofertas con redundancia de zona y zona. Este artículo enumera los servicios que admiten zonas de disponibilidad y el tipo de ofertas que admiten.

Algunos servicios de Azure son *nonregional*, lo que significa que no implementa el servicio en una región de Azure específica ni configura la compatibilidad con zonas de disponibilidad. Para obtener más información, vea Non regional Azure services.

> **Importante**
>
> Algunos servicios pueden admitir zonas de disponibilidad, pero requieren configuraciones específicas para habilitarlas en los recursos. Por ejemplo, algunos servicios pueden admitir zonas de disponibilidad solo para regiones o niveles específicos. Para obtener más información sobre los requisitos de un servicio para la compatibilidad con zonas de disponibilidad, seleccione ese servicio en la siguiente tabla.

En la tabla siguiente se enumeran los servicios de Azure zonales y con redundancia de zona. Algunos servicios admiten ambos tipos de implementaciones. Para obtener más información sobre los servicios zonales y con redundancia de zona y cómo funcionan, consulte Tipos de compatibilidad con zonas de disponibilidad.

Expandir tabla

Service	Con redundancia de zona	Zonal
Búsqueda de Azure AI	✓	
Azure Centro de API	✓	
Azure API Management	✓	✓
Azure App Configuration	✓	
Azure App Service	✓	

Service	Con redundancia de zona	Zonal
Azure App Service: App Service Environment	✓	
Azure Application Gateway v2	✓	✓
Azure Automation	✓	
Azure Backup	✓	
Azure Bastion	✓	✓
Azure Batch	✓	
Azure Blob Storage	✓	
Azure Chaos Studio	✓	
Azure Compute Gallery	✓	
Azure Container Apps	✓	
Azure Container Instances		✓
Azure Container Registry	✓	
Azure Cosmos DB	✓	
Azure Data Explorer	✓	
Azure Data Factory	✓	
Azure Data Lake Storage Gen2	✓	
Azure Database for MySQL	✓	✓
Azure Database for PostgreSQL	✓	✓
Azure Databricks	✓	
Azure DDoS Protection	✓	
Host dedicado de Azure		✓
Azure Registro de dispositivos	✓	

Service	Con redundancia de zona	Zonal
Azure Disk Storage	✓	✓
Azure DNS Resolución privada	✓	
Azure DNS zonas privadas	✓	
Azure Elastic SAN	✓	
Azure Event Grid	✓	
Azure Event Hubs	✓	
Azure ExpressRoute gateway	✓	
Azure Files	✓	✓
Azure Firewall	✓	✓
Azure Firewall Manager	✓	
Azure Functions	✓	
Azure HDInsight		✓
Azure IoT Hub	✓	
Azure IoT Hub servicio de aprovisionamiento de dispositivos	✓	
Azure Key Vault	✓	
Azure Kubernetes Service (AKS)	✓	✓
Azure Load Balancer	✓	✓
Azure Logic Apps	✓	
Azure Managed Grafana	✓	
Azure Instancia administrada para Apache Cassandra	✓	
Azure Managed Redis	✓	

Service	Con redundancia de zona	Zonal
Azure Monitor Logs	✓	
Azure Monitor: Application Insights	✓	
Azure NAT Gateway	✓	✓
Azure NetApp Files		✓
Azure Network Watcher	✓	
Azure Network Watcher: Análisis de tráfico	✓	
Azure Notification Hubs	✓	
Azure Private Link service	✓	
Azure Queue Storage	✓	
Red Hat OpenShift en Azure	✓	
Azure Route Server	✓	
Azure Service Bus	✓	
Azure Service Fabric	✓	✓
Azure SignalR Service	✓	
Azure Site Recovery	✓	
Azure SQL Database	✓	
Azure SQL Managed Instance	✓	
Azure Stream Analytics	✓	
Azure Table Storage	✓	
Azure Ultra Disk Storage		✓
Conjuntos de escalado de máquinas virtuales de Azure	✓	✓

Service	Con redundancia de zona	Zonal
Azure Virtual Machines		✓
Azure Virtual Network	✓	
Azure Virtual WAN: ExpressRoute	✓	
Azure Virtual WAN: puertas de enlace vpn de punto a sitio	✓	
Azure Virtual WAN: puertas de enlace vpn de sitio a sitio	✓	
Azure VMware Solution	✓	✓
Azure VPN Gateway	✓	
Azure Web PubSub	✓	
Microsoft Entra Domain Services	✓	
Microsoft Fabric	✓	
Microsoft Foundry Agent Service	✓	
Power BI Embedded	✓	
Azure dirección IP pública	✓	✓
SQL Server en Virtual Machines	✓	

Contenido relacionado

- resistencia de zona Enable para cargas de trabajo de Azure
- Azure servicios que admiten varias regiones
- Servicios de Azure no regionales
- Lista de regiones de Azure
- Disponibilidad del servicio por categoría
- Azure Well-Architected Framework: estrategias de arquitectura para usar regiones y zonas de disponibilidad

ⓘ**Nota:** El autor ha creado este artículo con ayuda de inteligencia artificial. Más información

Last updated on 15/06/2026

Recursos zonales y resistencia de zona

En Azure, un recurso *zonal* es un recurso que está anclado a una sola zona. Dado que un recurso zonal está en una sola zona de disponibilidad, no es resistente a la zona. Si la zona que contiene el recurso tiene un problema, es probable que el recurso experimente tiempo de inactividad.

Algunos servicios de Azure requieren o permiten implementar recursos zonales. Puede optar por implementar un recurso de forma zonal debido a consideraciones de latencia o requisitos de servicio específicos. Puede anclar recursos individuales o conjuntos de recursos relacionados a una sola zona.

En este artículo se describen los escenarios en los que puede optar por implementar recursos zonales en lugar de recursos con redundancia de zona. También resalta las consideraciones y responsabilidades necesarias para asegurarse de que la solución sigue siendo resistente a las interrupciones de zona.

Tipos de implementación de recursos

En Azure, solo algunos tipos de implementación proporcionan resistencia de zona. En la tabla siguiente se comparan tres tipos de implementación de recursos y se describen su resistencia de zona, distribución de zona, opciones de configuración y recomendaciones.

Expandir tabla

Tipo de implementación de recursos	Soporte de resiliencia de zona	Distribución de zona	Cómo establecer la configuración	Recomendación
Zone-redundant	Siempre resistente por zona	Los recursos con redundancia de zona se distribuyen entre varias zonas y son resistentes a los errores de zona. Si se produce un error en una zona, el servicio puede seguir funcionando en otras zonas.	Algunos recursos con redundancia de zona proporcionan redundancia de zona automática entre zonas de disponibilidad, mientras que otros recursos requieren que habilite manualmente la redundancia de zona. Compruebe la guía de confiabilidad del servicio para ver	Use siempre los recursos con redundancia de zona siempre que sea posible, especialmente en las implementaciones de producción.

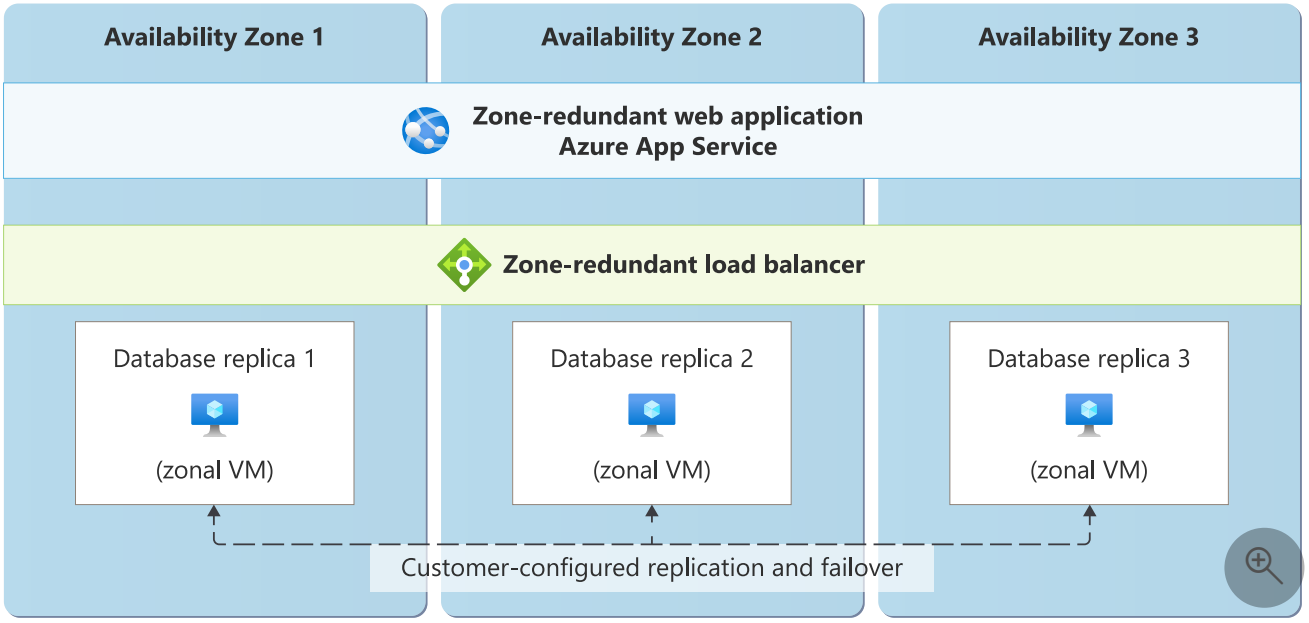
Tipo de implementación de recursos	Soporte de resiliencia de zona	Distribución de zona	Cómo establecer la configuración	Recomendación
			qué requiere su servicio para habilitar la resistencia.	
Zonal	No automático. Es su responsabilidad habilitar la resistencia de zona si lo decide. Los recursos zonales están aislados de errores en otras zonas, pero un error de su propia zona puede provocar tiempo de inactividad.	Usted selecciona la zona para el recurso.	Si tiene varios recursos que deben *estar alineados con zonas* (colocados en la misma zona), debe configurar la misma zona en *cada* recurso.	Use solo los recursos zonales cuando haya una necesidad clara. Para que la solución sea resistente a la zona, es su responsabilidad diseñar e implementar una solución de varias zonas.
Nozonal (regional)	Ninguno	Si la región proporciona compatibilidad con zonas de disponibilidad, Azure podría usar cualquier zona de la región.	No hay ninguna configuración de zona disponible para los recursos nozonales.	Dado que los recursos no zonales no se pueden hacer resistentes a la zona, evite las implementaciones no zonales para todas las cargas de trabajo de producción de las regiones que tienen zonas de disponibilidad.

Para obtener más información sobre las zonas de disponibilidad y las implementaciones de recursos, consulte Zonas de disponibilidad.

Cargas de trabajo que combinan recursos con redundancia de zona y zonales

Muchas cargas de trabajo combinan recursos zonales y con redundancia de zona. Por ejemplo, la carga de trabajo podría incluir un conjunto de máquinas virtuales zonales para el nivel de

base de datos, un servidor web con redundancia de zona hospedado en Azure App Service y un equilibrador de carga con redundancia de zona para enviar tráfico a las máquinas virtuales de base de datos.



Al combinar recursos zonales y con redundancia de zona en una tarea, tenga en cuenta cómo se comportan cada uno de los recursos y la solución global si una zona de disponibilidad tiene un problema. Normalmente, los servicios con redundancia de zona se recuperan automáticamente de interrupciones de zona con una pérdida de datos mínima o sin pérdida de datos, y Microsoft administra todo el proceso. En el caso de los recursos zonales, tiene la responsabilidad de configurar la conmutación por error automatizada o de realizar actividades de recuperación manual. Para obtener información sobre cómo se comporta cada servicio durante escenarios de zona baja, comprenda sus responsabilidades frente a las responsabilidades de Microsoft y supervise el estado de los servicios durante los eventos de zona abajo, consulte la guía de confiabilidad del servicio.

Cuándo usar una implementación zonal

Use recursos zonales solo cuando haya una necesidad clara. Entre los motivos típicos de una implementación de una sola zona se incluyen los casos en los que un recurso debe ser zonal, un servicio solo está disponible en una zona específica o una carga de trabajo es altamente sensible a la latencia entre zonas.

> ⓘ **Importante**

Algunos servicios de Azure permiten elegir entre implementaciones zonales y de zona redundante. Si no tiene una razón sólida para usar una implementación zonal, use una implementación con redundancia de zona.

Recursos que requieren implementaciones zonales

Algunos servicios de Azure solo admiten implementaciones zonales y no proporcionan implementaciones con redundancia de zona.

Las máquinas virtuales son un recurso zonal. Puede usar conjuntos de escalado de máquinas virtuales para crear conjuntos de máquinas virtuales. Los conjuntos de escalado de máquinas virtuales se pueden ampliar por zonas, lo que significa que las máquinas virtuales del conjunto se distribuyen entre varias zonas. Los conjuntos de escalado son una buena manera de lograr la resiliencia de zona para muchas cargas de trabajo que se basan en máquinas virtuales.

> **Sugerencia**
>
> Si despliega varias máquinas virtuales que realizan funciones similares, recomendamos que utilice conjuntos de escalado que abarquen zonas en lugar de máquinas virtuales de instancia única que despliegue de manera individual.

Otro ejemplo es Azure NetApp Files, que admite la implementación de volúmenes en una sola zona. El servicio también proporciona una manera de replicar entre varios volúmenes zonales.

Algunos servicios proporcionan opciones que solo están disponibles en zonas específicas. Por ejemplo, los tipos de máquina virtual específicos que usan unidades de procesamiento de gráficos avanzados (GPU) pueden estar disponibles solo en zonas específicas dentro de una región, lo que significa que no se pueden implementar en varias zonas. Para comprobar qué regiones y zonas admiten los tipos de máquina virtual que necesita, use los siguientes recursos:

- Para comprobar los tipos de máquina virtual disponibles en cada región, consulte Productos disponibles por región.
- Para comprobar los tamaños y tipos de máquina virtual admitidos dentro de cada zona de una región específica, consulte Comprobación de la disponibilidad de la SKU de máquina virtual.

Si el tipo de máquina virtual que necesita solo está disponible en una sola zona dentro de la región que usa, puede considerar una implementación zonal para esa máquina virtual y, a continuación, encontrar otras formas de hacer que la máquina virtual sea resistente a las

interrupciones de zona. Pero debe seguir asegurándose de que las otras partes de la solución son resistentes a la zona.

Para más información, consulte Servicios de Azure que admiten zonas de disponibilidad.

Latencia entre zonas

Si tiene una carga de trabajo que es inusualmente sensible a la latencia, puede usar recursos zonales en lugar de recursos con redundancia de zona, incluso si un servicio admite implementaciones con redundancia de zona.

Una red de baja latencia conecta zonas de disponibilidad, con latencia de ida y vuelta entre zonas normalmente debajo de dos milisegundos. Para la mayoría de las cargas de trabajo, la latencia entre zonas no es un problema. Las ventajas de resistencia de la propagación de recursos entre zonas de disponibilidad son más importantes que el impacto mínimo en el rendimiento del envío de tráfico entre zonas. Pero algunas cargas de trabajo son muy sensibles a la latencia entre zonas. Estas cargas de trabajo pueden incluir los siguientes escenarios:

- **Aplicaciones locales heredadas:** Algunas cargas de trabajo heredadas pueden contener aplicaciones diseñadas originalmente para un entorno local. Estas cargas de trabajo asumen que los componentes, como las bases de datos y otras aplicaciones y servicios, se colocan en el mismo host o en proximidad física cercana.
- **Replicación sincrónica de gran escala:** Las aplicaciones y bases de datos con estado realizan ocasionalmente un gran número de escrituras mediante la replicación sincrónica. La replicación sincrónica significa que los datos se escriben en varias *réplicas* antes de que la operación de escritura se considere completada. La distribución de réplicas entre zonas de disponibilidad mejora la resistencia, pero cuando se usa la replicación sincrónica, la latencia entre zonas puede aumentar la latencia de escritura de la carga de trabajo. Esta mayor latencia no suele ser significativa, pero debido a cómo se diseñan algunas aplicaciones, a veces puede resultar problemática a gran escala.

> **ⓘ Importante**
>
> Es inusual que las cargas de trabajo sean sensibles a la latencia entre zonas. No suponga que la carga de trabajo se ve afectada a menos que pruebe la latencia de la carga de trabajo y las necesidades específicas.

Si sospecha que la latencia entre zonas afecta a la carga de trabajo, pruebe su impacto en un entorno realista siguiendo estos pasos para la carga de trabajo específica:

1. **Defina los requisitos de rendimiento aceptables.** El tráfico entre zonas agrega una pequeña cantidad de latencia, pero es insignificante para la mayoría de las cargas de trabajo. Defina cómo debe ser un rendimiento aceptable para *su* carga de trabajo.

2. **Ejecute una prueba de rendimiento dentro de una sola zona de disponibilidad.** Establezca un conjunto de métricas de rendimiento de línea base.

   > **Importante**
   >
   > Pruebe la carga de trabajo, incluidas las aplicaciones, los protocolos, la configuración y la región de Azure. Use una carga realista. Los puntos de referencia y las pruebas sintéticas no son suficientes porque no muestran cómo se comporta realmente la solución.

3. **Habilite la replicación entre zonas.** En función de los componentes que use, puede habilitar la redundancia de zona o mover réplicas entre zonas.

4. **Vuelva a ejecutar pruebas de rendimiento.** Recopile las mismas métricas recopiladas anteriormente.

5. **Compare el impacto en el rendimiento con respecto a sus requisitos.** Use sus requisitos y los datos de rendimiento para tomar una decisión informada sobre el equilibrio entre latencia y resistencia a las interrupciones de zona.

   Si la prueba demuestra que la latencia es inaceptablemente alta para la carga de trabajo, considere la posibilidad de realizar las siguientes acciones:

   - ✓ Intente usar otro conjunto de zonas. Puede haber una ligera variabilidad en la latencia entre diferentes zonas porque pueden tener diferentes distancias físicas entre sí.

     > **Sugerencia**
     >
     > Si realiza pruebas entre suscripciones de Azure, revise el mapeo de zonas lógicas a físicas para asegurarse de probar los conjuntos de zonas físicas que anticipa.

   - ✓ Si hay otra región de Azure que satisfaga sus necesidades generales para la residencia de datos y otros factores, pruebe a usar varias zonas de esa región.

   - ✓ Considere si puede rediseñar la aplicación para minimizar la comunicación entre zonas requerida. Por ejemplo, puede consolidar varias operaciones de base de datos

pequeñas en una sola operación. Este enfoque puede reducir el impacto de la latencia en la carga de trabajo.

Si ninguna de estas acciones ayuda, considere la posibilidad de ejecutar la carga de trabajo o componentes específicos dentro de una sola zona de disponibilidad mediante máquinas virtuales zonales y otros servicios de Azure compatibles. A continuación, se encarga de asegurar que los componentes zonales sean resilientes frente a las interrupciones de zona. Revise el resto de este artículo para comprender sus responsabilidades y algunos enfoques que se deben tener en cuenta.

Sus responsabilidades para una implementación zonal

Un recurso zonal está en riesgo de tiempo de inactividad cuando su zona de disponibilidad experimenta una interrupción. Cuando implemente un recurso zonal, usted es responsable de hacer que su carga de trabajo sea resistente a fallos de nivel de zona.

> ⓘ **Importante**
>
> Los recursos zonales **no** son inherentemente resistentes a los errores de zona. Debe diseñar formas de mitigar el riesgo de una falla de zona mediante el desarrollo de un plan que incluya escenarios de caída de zona.

Para que los recursos zonales sean resistentes a la zona, tenga en cuenta las siguientes responsabilidades:

- **Implementación y configuración de varios recursos:** Implemente recursos zonales independientes manualmente en diferentes zonas o regiones. Determine cómo mantener la configuración coherente en cada recurso. El uso de la infraestructura como código (IaC) es un procedimiento recomendado porque permite la implementación rápida de varios recursos idénticos.

- **Enrutamiento y distribución del tráfico:** Debe seleccionar un componente del equilibrador de carga, asegurarse de que es resistente a la zona y configurarlo para enviar tráfico entre los recursos de diferentes zonas. Normalmente, se configura la directiva de enrutamiento (como activa-activa o activa-pasiva), comprobaciones de estado automatizadas y procesos de conmutación por error. Para obtener más información, consulte Opciones de equilibrio de carga.

- **Replicación o copia de seguridad de datos:** En el caso de los recursos con estado, es responsable de proteger los datos que almacenan y asegurarse de que se mantienen de

forma segura en varias zonas. Un enfoque común consiste en configurar la replicación en otra instancia de servicio en una zona de disponibilidad diferente. En algunas situaciones, puede confiar en copias de seguridad en su lugar. Pero las copias de seguridad requieren un tiempo de recuperación más largo durante un error de zona, lo que requiere que tenga un *objetivo de tiempo de recuperación mayor (RTO)*. También producen más pérdida de datos, lo que requiere un objetivo de *punto de recuperación (RPO)* mayor.

- **Implementación del proceso de detección y respuesta de errores de zona:** Debe determinar cómo supervisar el estado de los recursos zonales, definir las condiciones que los marcan como incorrectos y desencadenar acciones de respuesta como restaurar operaciones en otra zona o región.

- **Procesos de recuperación de zona:** Una vez que la zona se recupere, usted es responsable de las acciones de recuperación necesarias, como volver a los recursos de la zona principal.

Enfoques comunes para la resiliencia de la implementación zonal

Para tomar decisiones fundamentadas sobre cómo lograr la resistencia de zona para los recursos zonales, tenga en cuenta los siguientes factores:

- **Revise toda la carga de trabajo.** Comprenda cómo se comporta cada componente durante los eventos de zona descendente, incluidos los recursos con redundancia de zona, zonales y no regionales. Use la guía de confiabilidad de cada servicio para obtener información sobre cómo funciona el servicio durante escenarios de zona baja y cómo supervisar el estado de los servicios para eventos de zona descendente.

- **Comprenda la pérdida de datos permitida durante un error de zona.** El RPO especifica la cantidad de pérdida de datos que puede aceptar.

  Muchos recursos con redundancia de zona de Azure proporcionan un RPO de cero para errores de zona, lo que significa que no se produce ninguna pérdida de datos. Normalmente logran este RPO mediante la replicación sincrónica de todos los cambios entre zonas.

  Al planear una implementación zonal, debe asegurarse de que puede satisfacer los requisitos de RPO de su carga de trabajo en caso de fallo de una zona.

- **Comprenda el tiempo de inactividad permitido durante un fallo de zona.** El RTO especifica cuánto tiempo de inactividad puede aceptar usted.

Los recursos con redundancia de zona de Azure suelen proporcionar un RTO muy bajo para los errores de zona y normalmente requieren solo unos segundos de tiempo de inactividad.

Al planificar una implementación zonal, debe asegurarse de que puede satisfacer los requisitos de RTO de su carga de trabajo. Si tiene un RTO bajo, es posible que tenga que confiar en procesos de detección y recuperación automatizados. Un RTO mayor proporciona más flexibilidad para los procesos de respuesta.

- **Comprenda el costo.** Normalmente, los recursos zonales se facturan individualmente, por lo que la implementación de varios recursos zonales puede aumentar el costo de los recursos.

Diseño de una implementación zonal para la resiliencia

Al diseñar la implementación zonal para lograr resistencia, tenga en cuenta si usa zonas de disponibilidad para lograr *alta disponibilidad* o *recuperación ante desastres*. La distinción entre estos conceptos se basa en los requisitos de RTO y RPO.

Si tiene un requisito de RTO bajo y RPO bajo, debe tratar las zonas de disponibilidad como una construcción de *alta disponibilidad* . Pero si las cifras de RTO y RPO son más elevadas, podría optar por tratar las zonas de disponibilidad como un marco de *recuperación ante desastres*. Para obtener más información, consulte Continuidad empresarial, alta disponibilidad y recuperación ante desastres. El nivel de carga de trabajo puede ayudarle a determinar los requisitos y las acciones necesarias.

Diseño para lograr alta disponibilidad

Considere la posibilidad de implementar su propia arquitectura de alta disponibilidad en varias zonas. Una arquitectura de alta disponibilidad requiere la replicación de datos automatizada y frecuente entre los componentes que se implementan en varias zonas y la conmutación automática por fallo entre ellos si se produce un error de zona.

Algunas aplicaciones que se implementan en máquinas virtuales zonales proporcionan compatibilidad integrada con alta disponibilidad, por ejemplo, teniendo en cuenta la réplica. Por ejemplo, si usa SQL Server en máquinas virtuales de Azure, *los grupos de disponibilidad* proporcionan funcionalidades de enrutamiento de tráfico y conmutación por error. Puede seleccionar si desea usar la replicación sincrónica o asincrónica. Para más información, consulte Continuidad empresarial, alta disponibilidad y recuperación ante desastres para SQL Server en máquinas virtuales de Azure.

Diseño para la recuperación ante desastres

La recuperación ante desastres difiere de la alta disponibilidad, ya que un mayor tiempo de inactividad y pérdida de datos son aceptables en un escenario de desastre. RTO y RPO se miden normalmente en horas o más.

Un plan de recuperación ante desastres le ayuda a prepararse para diferentes escenarios y define cómo responder mediante una combinación de procesos automatizados y manuales.

Los siguientes enfoques de recuperación ante desastres pueden ayudar al planear una implementación zonal:

- **Recuperación ante desastres de zona a zona de Azure Site Recovery:** Este enfoque es útil cuando se necesita la replicación asincrónica de nivel de disco entre máquinas virtuales en distintas zonas. Para más información, consulte Habilitación de la recuperación ante desastres de máquinas virtuales de Azure entre zonas de disponibilidad.

- **Recuperación ante desastres de la región a región de Site Recovery:** Site Recovery admite la recuperación ante desastres de región a región y se basa en la replicación asincrónica. Este enfoque le permite conmutar por error a una zona de una región diferente de Azure en lugar de a otra zona de la región primaria. Para más información, consulte Replicación de máquinas virtuales de Azure en otra región de Azure.

- **Recuperación ante desastres basada en copia de seguridad:** Si la solución puede tolerar un RTO elevado y un RPO elevado, considere la posibilidad de usar copias de seguridad como una estrategia de recuperación ante desastres. Si la zona experimenta una interrupción, puede restaurar las copias de seguridad en otra zona o región. También debe considerar si crea previamente los demás recursos de Azure en su solución o si los crea durante el proceso de conmutación por error.

  En una arquitectura zonal, a menudo es responsable de almacenar y replicar esas copias de seguridad.

  Azure Backup es un servicio de copia de seguridad administrado ampliamente usado. Admite copias de seguridad con redundancia de zona y copias de seguridad con replicación geográfica entre regiones de Azure emparejadas. Algunas aplicaciones, como SQL Server en máquinas virtuales de Azure, también incluyen características integradas de copia de seguridad específicas de la aplicación.

Paso siguiente

- Recomendaciones para usar zonas de disponibilidad y regiones en Azure Well-Architected Framework

---

Last updated on 22/01/2026

Habilitación de la resistencia de zona para cargas de trabajo de Azure

Para que las aplicaciones sean más resistentes a los errores de hardware relacionados con la zona, las interrupciones de red y los desastres naturales, es importante diseñar las cargas de trabajo de Azure para la resistencia de zona. Al distribuir recursos entre varias zonas de disponibilidad dentro de una región, se reduce el riesgo de una interrupción de una sola zona que afecta a los servicios críticos.

Aunque es un procedimiento recomendado abordar la resistencia de zona durante la planeación inicial y la implementación de cargas de trabajo, es habitual convertir cargas de trabajo no resistentes existentes a configuraciones resistentes a zonas. En general, el procesamiento de la habilitación de la resistencia de zona para cargas de trabajo existentes es sencillo y Microsoft sigue simplificando el proceso. Sin embargo, cualquier cambio en la carga de trabajo puede introducir una cantidad de riesgo. Una vez que comprenda los riesgos implicados, podrá evaluar y priorizar qué cargas de trabajo y servicios dentro de esas cargas de trabajo son más importantes para su negocio y, después, aplicar resistencia de zona a los recursos más impactantes primero.

En este artículo se describen las consideraciones clave para habilitar la resistencia de zona en las cargas de trabajo de Azure. También le ayuda a planear e implementar una transición correcta a una arquitectura más resistente.

> **Sugerencia**
>
> Si actualmente está en proceso de diseñar sus cargas de trabajo o tiene previsto realizar una revisión de diseño de sus cargas de trabajo actuales, es importante que siga las recomendaciones para el diseño para la redundancia en el Azure Well-Architected Framework (WAF). La guía de recomendaciones de WAF le ayuda a diseñar la redundancia de la carga de trabajo en varios niveles, con un enfoque en flujos de trabajo críticos. Para facilitar la adopción de zonas de disponibilidad, también describe estrategias como despliegues en múltiples regiones y matrices de despliegue.

¿Qué es la resistencia de zona?

Los servicios de Azure pueden volverse resistentes a las interrupciones de zona de disponibilidad de dos maneras principales:

- **Servicios con redundancia de zona:** Muchos servicios de Azure admiten redundancia de zona. Estos servicios replican automáticamente los datos entre zonas de disponibilidad,

distribuyen las solicitudes entrantes y conmutan por error a distintas zonas durante un error de zona. Cada servicio admite estas funcionalidades de una manera que tenga sentido para ese servicio específico. Algunos servicios tienen redundancia de zona de forma predeterminada, mientras que es posible que otros servicios necesiten configurar la redundancia de zona.

- **Servicios de zona:** Algunos servicios de Azure son zonales, lo que significa que se pueden anclar a una zona de disponibilidad específica. Para lograr resistencia de zona mediante un servicio zonal, implemente instancias independientes del servicio en varias zonas de disponibilidad. También es posible que tenga que administrar la distribución del tráfico, la replicación de datos y la conmutación por error entre las instancias.

Algunos servicios se pueden implementar en una configuración con redundancia de zona o zonal. En la mayoría de los casos, es mejor implementar servicios con redundancia de zona cuando pueda.

Para más información, consulte Tipos de soporte de zonas de disponibilidad.

Procedimiento de habilitación de zona

Siga estos pasos para revisar sistemáticamente las cargas de trabajo de Azure, priorizarlas por resistencia de zona y habilitar la resistencia de zona para cada componente.

Prerrequisitos

Antes de empezar, realice las acciones siguientes:

- **Identifique cada carga de trabajo.** Una *carga de trabajo* hace referencia a una colección de recursos de aplicación, datos e infraestructura auxiliar que funcionan conjuntamente para lograr resultados empresariales definidos. Para más información sobre las cargas de trabajo y cómo definirlas, consulte Cargas de trabajo del Marco de buena arquitectura.

- **Priorice los flujos de usuario y sistema de cada carga de trabajo.** Comprenda las rutas críticas y las dependencias de sus cargas de trabajo para determinar qué componentes deben hacerse resilientes a nivel de zona primero. Para obtener más información sobre cómo usar el análisis de flujo crítico para priorizar los flujos de trabajo, consulte Prioridad de las cargas de trabajo para la resistencia de zona.

- **Asigne una clasificación de importancia crítica a cada carga de trabajo y flujo.** Esta clasificación le ayuda a comprender el impacto de una posible interrupción en su negocio y guía sus decisiones sobre qué cargas de trabajo priorizar para la resistencia de la zona.

Tenga en cuenta también la cantidad de tiempo de inactividad aceptable mientras vuelve a configurar las cargas de trabajo.

Puede usar una taxonomía para clasificar las cargas de trabajo en función de su importancia. Este enfoque le ayuda a centrar sus esfuerzos en los servicios más importantes.

Considere la taxonomía de ejemplo siguiente para clasificar las cargas de trabajo.

Expandir tabla

Tipo de carga de trabajo	Description	Efecto de la interrupción
Crítico para la misión	Flujos críticos y cargas de trabajo que deben ser altamente confiables, siempre disponibles, resistentes a errores y operativos	Cualquier interrupción de las funciones esenciales corre inmediatamente riesgos catastróficos de daño empresarial o introduce riesgos para la vida humana.
Crítico para la empresa	Flujos y cargas de trabajo esenciales que operan funciones empresariales importantes	La interrupción corre algunos riesgos de pérdida financiera o daño de marca.
Operaciones empresariales	Contribuye a la eficacia de las operaciones empresariales, pero fuera de la línea directa de servicio a los clientes	Puede tolerar algún nivel de interrupción.
Administrativo	Flujos de producción internos y cargas de trabajo no alineadas con las operaciones empresariales	Puede tolerar interrupciones.

Para obtener más información sobre cómo clasificar las cargas de trabajo según la clasificación de importancia crítica, consulte Asignación de una clasificación de importancia a cada flujo.

- **Compruebe que las regiones en las que residen los recursos de Azure admiten zonas de disponibilidad.** Consulte la lista de regiones de Azure. Si una región no admite zonas de disponibilidad, considere la posibilidad de reubicar los recursos en una región que sí lo haga. Para obtener más información, consulte Move Azure recursos entre grupos de recursos, suscripciones o regiones.

Paso 1: Priorizar los servicios de Azure para la resistencia de zona

Después de determinar qué flujos de carga de trabajo son más críticos para su empresa, puede centrarse en los servicios de Azure de los que dependen esos flujos. Algunos servicios Azure son más críticos para las aplicaciones que otras. Priorice estos servicios para ayudar a garantizar que las aplicaciones permanezcan disponibles y resistentes si se produce un error de zona.

Utilice las instrucciones siguientes para priorizar los grupos de servicios de Azure en función de su criticidad para las cargas de trabajo. Tenga en cuenta los requisitos empresariales y la arquitectura de aplicaciones específicos al determinar la prioridad de los servicios para la resistencia de zona.

1. **Comience con los servicios de red.** Las cargas de trabajo tienden a compartir servicios de red, por lo que un aumento de la resistencia de estos servicios puede mejorar la resistencia de varias cargas de trabajo a la vez.

   Muchos servicios de red principales son con redundancia de zona de forma automática, pero debe centrarse en componentes como Azure ExpressRoute Gateway, Azure VPN Gateway, Azure Application Gateway, Azure Load Balancer y Azure Firewall.

2. **Mejore la resistencia del almacenamiento de datos operativos.** Los almacenes de datos operativos contienen datos valiosos que suelen usar varias cargas de trabajo, por lo que mejorar la disponibilidad de esos almacenes de datos puede ayudar a muchas cargas de trabajo.

   Para la resistencia del almacenamiento de datos operativo, céntrese en servicios como Azure SQL Database, Azure SQL Managed Instance, Azure Storage, Azure Data Lake Storage, Azure Cosmos DB, Azure Servidor flexible de PostgreSQL, Azure servidor flexible de MySQL y Azure Managed Redis.

3. **Priorizar los servicios de computación.** Estos servicios suelen ser fáciles de replicar y distribuir entre zonas porque no tienen estado.

   Los servicios de proceso incluyen Azure Virtual Machines, Azure Virtual Machine Scale Sets, Azure Kubernetes Service (AKS), Azure App Service, App Service Environment, Azure Functions, Azure Service Fabric y Azure Container Apps.

4. **Revise los recursos críticos empresariales restantes que utilizan sus flujos críticos.** Es posible que estos recursos no sean tan críticos como los recursos enumerados anteriormente, pero siguen desempeñando un papel en la funcionalidad de la aplicación y debe considerarlos para la resistencia de zona.

5. **Revise el resto de los recursos operativos de la empresa.** Tome decisiones informadas acerca de si hacerlos resilientes a zonas. Esta revisión incluye servicios que podrían no

relacionarse directamente con las cargas de trabajo críticas, pero que siguen contribuyendo al rendimiento y la confiabilidad generales de las aplicaciones.

Paso 2: Evaluación de los enfoques de configuración de zona

Después de priorizar las cargas de trabajo y los servicios de Azure, identifique el enfoque necesario para habilitar la compatibilidad con zonas de disponibilidad para cada servicio y comprenda lo que debe hacer para configurar la resistencia de la zona.

Cada guía del servicio de confiabilidad de Azure proporciona una sección que describe cómo habilitar la resistencia de zona para ese servicio. Esta sección le ayuda a comprender el esfuerzo necesario para que cada zona de servicio sea resistente para que pueda planear su estrategia en consecuencia. Para obtener más información sobre un servicio específico, consulte guías de servicio de confiabilidad de Azure.

Use la tabla de configuración zone para comprender rápidamente los enfoques de los servicios comunes de Azure.

> **ⓘ Importante**
>
> Si la carga de trabajo incluye componentes implementados en una configuración zonal (o de una sola zona), planee hacer que estos componentes sean resistentes a las interrupciones de zona. Un enfoque común consiste en implementar instancias independientes en otra zona de disponibilidad y cambiar entre ellas si es necesario.

Paso 3: Probar la latencia

Al hacer que las cargas de trabajo sean resistentes a la zona, considere la posibilidad de tener en cuenta la latencia entre las zonas de disponibilidad. En ocasiones, algunos sistemas heredados no pueden tolerar la pequeña cantidad de latencia adicional que introduce el tráfico entre zonas, especialmente cuando se habilita la replicación sincrónica dentro del nivel de datos. Si sospecha que la latencia entre zonas podría afectar a la carga de trabajo, ejecute pruebas antes y después de habilitar la resistencia de zona. Para obtener más información sobre cómo la latencia entre zonas podría afectar a la aplicación y los enfoques para mitigar los problemas de latencia entre zonas, consulte Recursos zonales y resistencia de zona.

Enfoques de configuración de zona para servicios de Azure

Cada servicio Azure admite un tipo específico de compatibilidad con zonas de disponibilidad, que se basa en el uso previsto del servicio y la arquitectura interna. Si tiene un recurso que no está configurado para usar zonas de disponibilidad (o un recurso *no zonal*), es posible que quiera volver a configurarlo con compatibilidad con la zona de disponibilidad. La guía de confiabilidad de ese servicio proporciona instrucciones o vínculos a instrucciones de configuración de zona de disponibilidad.

En esta sección se proporciona información general sobre los distintos tipos de enfoques de configuración de zona y qué enfoque admite cada servicio.

> **ⓘ Importante**
>
> Al habilitar *la redundancia de zona* en un recurso, ese recurso se vuelve automáticamente resistente a los errores de zona. Cuando se usa una configuración *zonal* para anclar el recurso a una zona de disponibilidad específica, el recurso no es automáticamente redundante de zona. Debe hacer que sea resiliente a un fallo de zona. En el caso de los servicios zonales, este artículo refleja la complejidad y el costo del anclaje a una zona. Para más información sobre los pasos adicionales para lograr la resistencia de zona, consulte la **guía de confiabilidad del servicio**.

La tabla de configuración zone enumera el enfoque de configuración de zona compatible para muchos servicios Azure y contiene un vínculo a cada guía de confiabilidad de ese servicio. La guía de confiabilidad proporciona información sobre cómo configurar recursos de servicio no zonales para habilitar la compatibilidad con zonas de disponibilidad.

En la tabla siguiente se describen los enfoques de configuración de zona comunes, incluido el nivel de esfuerzo y el tiempo de inactividad necesarios para habilitar las zonas de disponibilidad. Algunos servicios tienen enfoques diferentes enumerados debido a la forma en que funcionan.

Expandir tabla

Enfoque	Description	Nivel de esfuerzo típico	Podría requerir tiempo de inactividad
Siempre resistente por zona	El servicio es resistente a la zona de forma predeterminada en regiones que admiten zonas de disponibilidad. No se requiere ninguna acción.	Ninguno	No
Habilitación	Se requieren cambios mínimos de configuración, como habilitar la redundancia de zona en la configuración. El proceso no afecta a la	Low	No

Enfoque	Description	Nivel de esfuerzo típico	Podría requerir tiempo de inactividad
	disponibilidad, pero tenga en cuenta los efectos en el costo o el rendimiento.		
Modificación	Es probable que requiera algunos cambios de configuración, como volver a implementar recursos dependientes o modificar la configuración de red.	Mediana	Sí
Reimplementación	Se requieren cambios significativos, como volver a implementar recursos completos, aplicaciones o servicios, o migrar datos a nuevos servicios.	High	Sí

Comprenda el costo de habilitar la compatibilidad con zonas de disponibilidad para un servicio. Para muchos servicios, la habilitación de zonas de disponibilidad no agrega costos. Pero algunos servicios requieren un nivel específico, un número específico de unidades de capacidad o ambos. Otros servicios cobran tarifas diferentes cuando se usan zonas de disponibilidad. En la tabla de la sección siguiente se muestra el impacto típico en los costos de cada servicio.

> **Nota:**
>
> La información de este artículo resume el enfoque típico para habilitar la compatibilidad con zonas de disponibilidad y describe el impacto típico en los costos. Sin embargo, algunos factores pueden afectar a cómo funciona para su solución específica. Por ejemplo, algunos servicios se muestran como *siempre resistentes a la zona*, pero esta designación solo se aplica en regiones específicas o para niveles específicos del servicio. Use estas tablas como punto de partida, pero revise los demás recursos mencionados para comprender los detalles específicos.

Servicios de Azure según el enfoque de configuración por zona

En la tabla siguiente se resume la compatibilidad de zona de disponibilidad para muchos servicios Azure y se proporciona un enfoque, incluido el impacto en los costos, para habilitar la compatibilidad con zonas de disponibilidad para cada servicio.

Expandir tabla

Service	Puede ser redundante por zona	Puede ser zonal	Enfoque típico de configuración de zona	Impacto típico en los costos
Azure AI Search	✓		Siempre resistente por zona	N/A
Azure API Management	✓	✓	Modificación	Nivel mínimo necesario
Azure App Configuration	✓		Siempre resistente por zona	N/A
Azure App Service	✓		Habilitación	Nivel mínimo y recuento de instancias necesarios
Azure App Service: App Service Environment	✓		Habilitación	Número mínimo de instancias requerido
Azure Application Gateway	✓	✓	Siempre resistente por zona	N/A
Azure Backup	✓		Reimplementación	Aumento moderado del costo
Azure Bastion	✓	✓	Reimplementación	Sin impacto en el costo
Azure Batch	✓		Reimplementación	Sin impacto en el costo del mismo número de máquinas virtuales (VM)
Azure Blob Storage	✓		Habilitación	Aumento moderado del costo
Azure Cache for Redis: Enterprise	✓		Reimplementación	Sin impacto en el costo
Azure Cache for Redis: Estándar y Premium	✓		Habilitación	Nivel mínimo necesario
Azure Container Apps	✓		Reimplementación	Número mínimo de réplicas necesario
Azure Container Instances		✓	Reimplementación	Sin impacto en el costo
Azure Container Registry	✓		Siempre resistente por zona	N/A

Service	Puede ser redundante por zona	Puede ser zonal	Enfoque típico de configuración de zona	Impacto típico en los costos
Azure Cosmos DB para NoSQL	✓		Modificación	Ninguno si usa escrituras de escalado automático o de varias regiones
Azure Data Factory	✓		Siempre resistente por zona	N/A
Azure Data Lake Storage	✓		Habilitación	Aumento moderado del costo
Azure Database for MySQL: servidor flexible	✓	✓	Reimplementación	Requiere la instancia principal y de alta disponibilidad (HA)
Azure Database for PostgreSQL: servidor flexible	✓	✓	Habilitación	Requiere la instancia principal y la de alta disponibilidad
Azure Databricks	✓		Habilitación	Sin impacto en los costos para el mismo número de máquinas virtuales; aumento moderado del costo para el almacenamiento
Azure Disk Storage (discos administrados)	✓	✓	Habilitación	Aumento moderado del costo
SAN elástico de Azure	✓		Reimplementación	Aumento moderado del costo
Azure Event Hubs: nivel dedicado	✓		Siempre resistente por zona	Unidades de capacidad mínimas (RU) necesarias
Azure Event Hubs: todos los demás niveles	✓		Siempre resistente por zona	N/A
puerta de enlace de Azure ExpressRoute	✓	✓	Modificación	Depende del nivel
Archivos de Azure	✓		Habilitación	Aumento moderado del costo
Azure Firewall	✓	✓	Nuevos firewalls: resilientes por zonas de forma predeterminada	Sin impacto en el costo

Service	Puede ser redundante por zona	Puede ser zonal	Enfoque típico de configuración de zona	Impacto típico en los costos
			Firewalls no zonales existentes: modificación (migración automática en curso)	
Azure Functions: plan dedicado	✓		Habilitación	Número mínimo de instancias requerido
Azure Functions: Plan de consumo flexible	✓		Habilitación	Número mínimo de instancias requerido
Azure Functions: Plan Premium	✓		Reimplementación	Número mínimo de instancias requerido
Azure HDInsight		✓	Reimplementación	Sin impacto en el costo para el mismo número de nodos
Azure IoT Hub	✓		Siempre resistente por zona	N/A
Azure Key Vault	✓		Siempre resistente por zona	N/A
Azure Kubernetes Service (AKS)	✓		Reimplementación	Sin impacto en el costo
Equilibrador de carga de Azure	✓	✓	Modificación	Sin impacto en el costo
Azure Logic Apps: nivel de consumo	✓		Siempre resistente por zona	N/A
Azure Logic Apps: nivel Estándar	✓		Reimplementación	Nivel mínimo y recuento de instancias necesarios
Azure Managed Grafana	✓		Volver a implementar	Aumento moderado del costo
Azure Managed Redis	✓		Habilitación	Requiere la instancia principal y la de alta disponibilidad
Azure Monitor: Log Analytics	✓		Siempre resistente por zona	

Service	Puede ser redundante por zona	Puede ser zonal	Enfoque típico de configuración de zona	Impacto típico en los costos
Puerta de enlace NAT de Azure	✓	✓	Reimplementación	Sin impacto en el costo
Azure NetApp Files		✓	Reimplementación	Depende de la configuración de replicación
servicio Azure Private Link	✓		Siempre resistente por zona	N/A
Azure Queue Storage	✓		Habilitación	Aumento moderado del costo
Azure Service Bus	✓		Siempre resistente por zona	N/A
Azure Service Fabric	✓	✓	Reimplementación	Sin impacto en el costo del mismo número de máquinas virtuales
Azure Site Recovery	✓		Reimplementación	Sin impacto en el costo de Site Recovery, aumento moderado del costo para el almacenamiento de réplicas
Azure SQL Database: nivel Crítico para la empresa	✓		Habilitación	Sin impacto en el costo
Azure SQL Database: nivel de uso general	✓		Habilitación	Aumento moderado del costo
Azure SQL Database: Nivel hiperescala	✓		Reimplementación	Número mínimo de réplicas necesario
Azure SQL Database: nivel Premium	✓		Habilitación	Sin impacto en el costo
Instancia administrada de Azure SQL	✓		Habilitación	Aumento moderado del costo
Azure Stream Analytics	✓		Siempre resistente por zona	N/A

Service	Puede ser redundante por zona	Puede ser zonal	Enfoque típico de configuración de zona	Impacto típico en los costos
Azure Table Storage (almacenamiento en tablas de Azure)	✓		Habilitación	Aumento moderado del costo
Conjuntos de escalado de máquinas virtuales de Azure	✓	✓	Reimplementación	Sin impacto en el costo del mismo número de máquinas virtuales
Máquinas virtuales de Azure		✓	Reimplementación	Sin impacto en el costo del mismo número de máquinas virtuales
Azure Virtual Network	✓		Siempre resistente por zona	N/A
Azure VMware Solution	✓	✓	Reimplementación	Sin impacto en el costo para el mismo número de nodos
Dirección IP pública	✓	✓	Siempre resistente por zona	N/A

Recursos relacionados

- guías de confiabilidad de Azure
- Servicios de Azure que admiten zonas de disponibilidad
- Lista de regiones de Azure

Last updated on 13/03/2026

Seguridad en Azure AI Search

Azure AI Search es una infraestructura de búsqueda escalable que indexa contenido heterogéneo y permite la recuperación a través de API, aplicaciones y agentes de IA. Es adecuado para escenarios de búsqueda empresarial y experiencias de clientes con tecnología de inteligencia artificial que requieren generación dinámica de contenido a través de modelos de finalización de chat. Como servicio de Azure, AI Search proporciona una variedad de funcionalidades para satisfacer sus requisitos de confiabilidad.

Cuando se usa Azure, la confiabilidad es una responsabilidad compartida. Microsoft proporciona una variedad de funcionalidades para admitir resistencia y recuperación. Es responsable de comprender cómo funcionan esas funcionalidades dentro de todos los servicios que usa y de seleccionar las funcionalidades que necesita para cumplir los objetivos empresariales y los objetivos de tiempo de actividad.

En este artículo se describe cómo hacer que Azure AI Search sea resistente a una variedad de posibles interrupciones y problemas, incluidos errores transitorios, interrupciones de zona de disponibilidad, interrupciones de regiones y mantenimiento del servicio. También se describe cómo puede usar copias de seguridad para recuperarse de otros tipos de problemas y se resalta cierta información clave sobre el acuerdo de nivel de servicio (SLA) de Azure AI Search.

Recomendaciones de implementación de producción para la confiabilidad

En el caso de las cargas de trabajo de producción, se recomienda que usted:

- ✓ Use un nivel facturable que tenga al menos dos réplicas. Esta configuración hace que su servicio de búsqueda sea más resistente a errores transitorios y operaciones de mantenimiento. También cumple el acuerdo de nivel de servicio (SLA) para la búsqueda de IA. El Acuerdo de Nivel de Servicio requiere dos réplicas para cargas de trabajo de solo lectura y tres o más réplicas para cargas de trabajo de lectura y escritura.
- ✓ No use el nivel Gratis para su uso en producción. Ai Search no proporciona un Acuerdo de Nivel de Servicio para el nivel Gratis, que está limitado a una réplica.

Introducción a la arquitectura de confiabilidad

Cuando se usa AI Search, se crea un *servicio de búsqueda*. Cada servicio de búsqueda admite muchos *índices de búsqueda* que almacenan el contenido que se puede buscar.

La búsqueda de IA no está diseñada como un almacén de datos principal. En su lugar, se usan *indexadores* para conectar el servicio de búsqueda a orígenes de datos externos. Un indexador

rastrea los datos de origen, invoca *aptitudes* que realizan el procesamiento y el enriquecimiento, y rellena el índice con las salidas de aptitud.

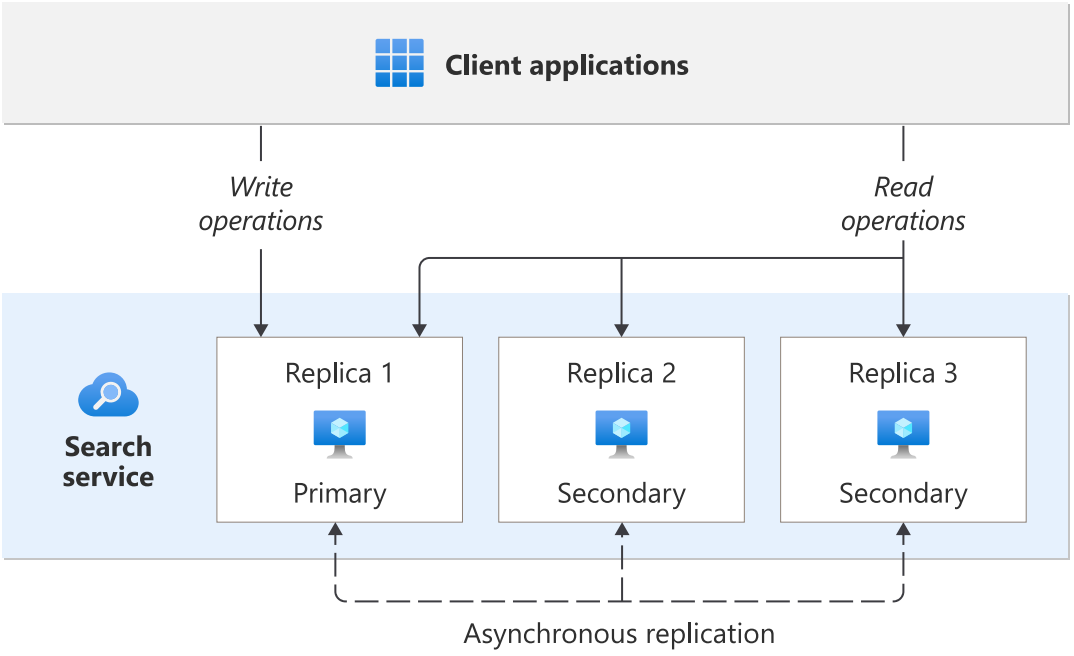
También configurará el número de *réplicas* para el servicio. En AI Search, una réplica es una copia del motor de búsqueda de tu servicio. Puede pensar en una réplica como si representara una sola máquina virtual (VM). Cada servicio Search puede tener entre 1 y 12 réplicas.

La adición de varias réplicas permite que AI Search pueda:

- Aumente la disponibilidad del servicio Search.
- Realice el mantenimiento en una réplica mientras las consultas continúan ejecutándose en otras réplicas.
- Controle cargas de trabajo de consulta e indexación más altas.
- Mejore la resistencia intentando aprovisionar réplicas en diferentes zonas de disponibilidad, si la región las admite.

Ai Search asigna automáticamente una réplica para que sea la *réplica principal*. Todas las operaciones de escritura se realizan en esa réplica. Las otras réplicas se usan para las operaciones de lectura.

En el diagrama siguiente se muestra cómo un servicio Search con tres réplicas se puede distribuir entre tres zonas de disponibilidad:



También puede configurar el número de *particiones*, que representan el almacenamiento que usan los índices de búsqueda.

Es importante comprender el impacto de agregar réplicas y particiones, ya que cada uno afecta al rendimiento de lectura y escritura de maneras diferentes. Para obtener más información sobre las réplicas y las particiones, consulte Estimación y administración de la capacidad de un servicio Search.

Resistencia a errores transitorios

Los errores transitorios son errores breves e intermitentes en los componentes. Se producen con frecuencia en un entorno distribuido como la nube y son una parte normal de las operaciones. Los errores transitorios se corrigen después de un breve período de tiempo. Es importante que las aplicaciones puedan controlar errores transitorios, normalmente mediante el reintento de solicitudes afectadas.

Todas las aplicaciones hospedadas en la nube deben seguir las instrucciones de control de errores transitorios de Azure cuando se comunican con cualquier API, bases de datos y otros componentes hospedados en la nube. Para obtener más información, consulte Recomendaciones para controlar errores transitorios.

Los indexadores de AI Search tienen un control de errores transitorios integrado. Si un origen de datos no está disponible brevemente, el indexador está diseñado para recuperarse y volver a intentarlo. Usa el seguimiento de cambios para reanudar la indexación del último documento indizado correctamente.

Los servicios de búsqueda pueden experimentar errores transitorios durante las operaciones de mantenimiento estándar y no programadas. Búsqueda de Azure AI no proporciona notificaciones anticipadas ni permite la programación de mantenimiento en momentos específicos. Aunque se hace todo lo posible por minimizar el tiempo de inactividad, incluso en los servicios de réplica única, pueden producirse breves interrupciones. Para mejorar la resistencia frente a estos errores transitorios, se recomienda usar dos o más réplicas.

Si construyes aplicaciones que interactúan con AI Search, estas deben manejar errores transitorios. Use una estrategia de reintento con retrocesos exponenciales para las operaciones de lectura y escritura.

Resistencia a errores de zona de disponibilidad

Las zonas de disponibilidad son grupos físicamente independientes de centros de datos dentro de una región de Azure. Cuando una zona falla, los servicios pueden transferirse a una de las zonas restantes.

AI Search tiene redundancia de zona, lo que significa que las réplicas se distribuyen a través de varias zonas de disponibilidad dentro de la región del servicio de búsqueda.

Al agregar dos o más réplicas al servicio, AI Search intenta colocar cada réplica en una zona de disponibilidad diferente. En el caso de los servicios que tienen más réplicas que las zonas disponibles, las réplicas se distribuyen entre zonas lo más uniformemente posible.

En el diagrama siguiente se muestra cómo se puede implementar un servicio de búsqueda de ejemplo con cuatro réplicas en tres zonas de disponibilidad:

> ⓘ **Importante**
>
> La búsqueda de IA no garantiza la ubicación exacta de las réplicas. La selección de ubicación está sujeta a restricciones de capacidad, operaciones de escalado y otros factores.

Requisitos

La redundancia de zona se habilita automáticamente cuando el servicio Search cumple todos los criterios siguientes:

- **Compatibilidad con regiones:** La compatibilidad con zonas de disponibilidad depende de la infraestructura y el almacenamiento. Para obtener una lista de las regiones admitidas, consulte Elección de una región para la búsqueda de IA.
- **Nivel:** El servicio debe estar en el nivel Básico o superior.
- **Número de réplicas:** El servicio debe tener al menos dos réplicas.

  > ⓘ **Nota:**
  >
  > Ai Search intenta distribuir réplicas entre varias zonas cuando tiene dos o más réplicas. Sin embargo, para las cargas de trabajo de lectura y escritura, debería usar tres o más réplicas para lograr el SLA de disponibilidad más alto posible.

Distribución de instancias a través de zonas

AI Search intenta colocar réplicas en diferentes zonas de disponibilidad. Sin embargo, en ocasiones hay situaciones en las que todas las réplicas de un servicio de búsqueda pueden colocarse en la misma zona de disponibilidad. Esta situación puede ocurrir cuando las réplicas se quitan del servicio, como cuando se *reduce horizontalmente* configurando el servicio para que use menos réplicas. La eliminación de réplicas no hace que las réplicas restantes se redistribuyan automáticamente entre las zonas de disponibilidad.

Para reducir la probabilidad de que todas las réplicas se coloquen en una sola zona de disponibilidad, puede desencadenar manualmente una operación de escalabilidad horizontal inmediatamente después de una operación de reducción horizontal. Por ejemplo, supongamos que el servicio Search tiene 10 réplicas y desea reducir horizontalmente a 7 réplicas. En lugar de realizar una sola operación de escalado, puede escalar temporalmente a 6 instancias y, a continuación, escalar inmediatamente a 7 instancias para desencadenar el reequilibrio de zona.

Cost

Cada servicio de búsqueda comienza con una réplica. La redundancia de zona requiere dos o más réplicas, lo que aumenta el costo de ejecutar el servicio. Para comprender las implicaciones de facturación de las réplicas, utilice la calculadora de precios.

Configurar soporte de zonas de disponibilidad

Si su servicio de búsqueda cumple los requisitos de redundancia de zona, no es necesaria ninguna configuración adicional. Siempre que sea posible, AI Search procura colocar las réplicas en diferentes zonas de disponibilidad.

Planeamiento y administración de capacidad

Para prepararse para un fallo en la zona de disponibilidad, considere sobreaprovisionar el número de réplicas. El aprovisionamiento excesivo permite al servicio de búsqueda tolerar cierta pérdida de capacidad y seguir funcionando sin un rendimiento degradado. Agregar réplicas durante una interrupción es difícil, por lo que el aprovisionamiento excesivo ayuda a garantizar que el servicio de búsqueda pueda controlar los volúmenes de solicitudes normales, incluso con una capacidad reducida. Para más información, consulte Administración de la capacidad mediante el sobreaprovisionamiento.

Comportamiento cuando todas las zonas están en buen estado

En esta sección se describe qué esperar cuando los servicios Search están configurados para la redundancia de zona y todas las zonas de disponibilidad están operativas.

- **Enrutamiento de tráfico entre zonas:** AI Search ejecuta el equilibrio de carga automático de todas las consultas y escrituras a través de todas las réplicas disponibles. AI Search puede enviar operaciones de lectura a cualquier réplica de cualquier zona de disponibilidad. Envía operaciones de escritura a una sola réplica principal que el servicio AI Search selecciona.

- **Replicación de datos entre zonas:** Los cambios en los datos se replican automáticamente entre réplicas entre zonas de disponibilidad. La replicación se produce de forma asincrónica, lo que significa que las escrituras se confirman en una réplica principal antes de que se repliquen en otras réplicas.

Comportamiento durante un fallo de zona

En esta sección se describe qué esperar cuando los servicios de búsqueda están configurados para la redundancia de zona y se produce una interrupción de zona de disponibilidad.

- **Detección y respuesta:** Ai Search es responsable de detectar un error en una zona de disponibilidad. No es necesario hacer nada para iniciar una conmutación por error de la zona.

- **Notificación**: Microsoft no le notifica automáticamente cuando una zona está inactiva. Sin embargo, puede usar Azure Resource Health para supervisar el estado de un recurso individual y puede configurar alertas de Resource Health para notificarle problemas. También puede usar Azure Service Health para comprender el estado general del servicio, incluidos los errores de zona, y puede configurar alertas de Service Health para notificarle problemas.

- **Solicitudes activas:** Se finalizan las solicitudes que las réplicas procesan en la zona con errores. Los clientes deben reintentar las solicitudes siguiendo las instrucciones para controlar errores transitorios.

- **Pérdida de datos esperada:** Si la zona de disponibilidad afectada solo contiene réplicas de lectura, no se espera ninguna pérdida de datos.

  Si se pierde la réplica principal porque estaba en la zona afectada, es posible que se pierdan las operaciones de escritura que aún no se hayan replicado.

- **Tiempo de inactividad esperado:** En la mayoría de las situaciones, no se espera que un error de zona provoque tiempo de inactividad en el servicio de búsqueda para las

operaciones de lectura porque las réplicas de lectura de otras zonas de disponibilidad siguen atendiendo solicitudes.

Si se pierde la réplica principal porque estaba en la zona afectada, AI Search promueve automáticamente otra réplica para convertirse en la nueva principal para que las operaciones de escritura se puedan reanudar. Normalmente, la promoción de la réplica tarda unos segundos en producirse. Durante este tiempo, es posible que las operaciones de escritura no se realicen correctamente. Asegúrese de que las aplicaciones están preparadas siguiendo las instrucciones de control de errores transitorios.

Sin embargo, hay algunas situaciones poco probables en las que todas las réplicas de tu servicio de búsqueda podrían encontrarse en una única zona de disponibilidad. En este escenario, puede experimentar una interrupción del servicio hasta que se recupere la zona. Para obtener más información y comprender una solución alternativa, consulte Distribución de instancias.

- **Reenrutamiento del tráfico:** Cuando se produce un error en una zona, AI Search detecta el error y enruta las solicitudes a réplicas activas en las zonas supervivientes. Si se pierde la réplica principal, se promueve otra réplica para que sea la nueva principal.

Recuperación de zona

Cuando se recupera la zona de disponibilidad, AI Search restaura automáticamente las operaciones normales y comienza a enrutar el tráfico a las réplicas disponibles en todas las zonas, incluida la zona recuperada.

Prueba de fallos de zona

AI Search administra el enrutamiento del tráfico para los servicios con redundancia de zona. No es necesario iniciar ni validar ningún proceso de error de zona.

Resistencia a errores en toda la región

AI Search es un servicio de una sola región. Si la región deja de estar disponible, su servicio de búsqueda también dejará de estar disponible.

Soluciones personalizadas de varias regiones para la resistencia

Opcionalmente, puede implementar varios servicios de búsqueda de IA en diferentes regiones. Es responsable de implementar y configurar servicios independientes en cada región. Si crea

una implementación idéntica en una región secundaria de Azure que usa una arquitectura de varias regiones, la aplicación será menos susceptible a un desastre de una sola región.

Cuando se sigue este enfoque, es necesario sincronizar los índices entre regiones para recuperar el último estado de la aplicación. También debe configurar las directivas de equilibrio de carga y conmutación por error.

Para optimizar el rendimiento de la solución general, busque oportunidades para realizar la indexación en réplicas de solo lectura de los orígenes de datos. Por ejemplo, algunos indexadores admiten la lectura desde las réplicas de lectura de un origen de datos distribuido geográficamente.

Para más información, consulte Implementaciones de varias regiones en Búsqueda de Azure AI.

Copias de seguridad y restauración

Dado que AI Search no es una solución de almacenamiento de datos principal, no proporciona opciones de copia de seguridad y restauración de autoservicio. Sin embargo, puede usar la muestra `index-backup-restore` para .NET o Python para hacer una copia de seguridad de la definición del índice y sus documentos en una serie de archivos JSON, que luego se usan para restaurar el índice.

Sin embargo, si elimina accidentalmente el índice y no tiene una copia de seguridad, puede volver a generar el índice. La regeneración implica volver a crear el índice en su servicio de búsqueda y luego volver a cargarlo recuperando los datos de su almacén de datos principal.

Acuerdo de nivel de servicio

El contrato de nivel de servicio (SLA) para los servicios de Azure describe la disponibilidad esperada de cada servicio y las condiciones que la solución deberá cumplir para lograr esa expectativa de disponibilidad. Para obtener más información, consulte Acuerdos de Nivel de Servicio para servicios en línea.

En AI Search, el Acuerdo de Nivel de Servicio de disponibilidad se aplica a los servicios de búsqueda que:

- Están configurados para usar un nivel facturable.
- Tener al menos dos réplicas para cargas de trabajo de solo lectura (consultas).
- Tener al menos tres réplicas para cargas de trabajo de lectura y escritura (consultas e indexación).

Contenido relacionado

- Confiabilidad en Azure
- Límites de servicio en AI Search
- Elección de una región para la búsqueda de IA
- Elección de un plan de tarifa para AI Search
- Planeamiento o adición de capacidad en la búsqueda de IA

ⓘ **Nota:** El autor ha creado este artículo con ayuda de inteligencia artificial. Más información

---

Last updated on 22/01/2026

Confiabilidad en Azure API Management

Azure API Management es un servicio totalmente administrado que ayuda a las organizaciones a publicar, proteger, transformar, mantener y supervisar las API. Como servicio de Azure, API Management proporciona una variedad de funcionalidades para satisfacer los requisitos de confiabilidad.

Cuando se usa Azure, la confiabilidad es una responsabilidad compartida. Microsoft proporciona una variedad de funcionalidades para admitir resistencia y recuperación. Es responsable de comprender cómo funcionan esas funcionalidades dentro de todos los servicios que usa y de seleccionar las funcionalidades que necesita para cumplir los objetivos empresariales y los objetivos de tiempo de actividad.

En este artículo se describe cómo hacer que API Management sea resistente a una variedad de posibles interrupciones y problemas, incluidos errores transitorios, interrupciones de zona de disponibilidad, interrupciones de regiones y mantenimiento del servicio. También se describe cómo puede usar copias de seguridad para recuperarse de otros tipos de problemas y se resalta cierta información clave sobre el acuerdo de nivel de servicio (SLA) de API Management.

Introducción a la arquitectura de confiabilidad

API Management usa una arquitectura basada en unidades de escalado para proporcionar redundancia y escalabilidad integradas. Al implementar una instancia de API Management, se configuran una o varias *unidades de escalado* o *unidades*. Cada unidad es una representación lógica de la capacidad que contiene los recursos de proceso necesarios para controlar las solicitudes de API.

Unidades

Cada unidad consta de dos recursos de proceso (máquinas virtuales o servidores similares, en función del nivel de servicio) que controlan las solicitudes de API juntas. Usted no ve estas máquinas virtuales ni otros servidores. La plataforma administra automáticamente su creación y monitorización del estado. Si se produce un error en un recurso de proceso, la unidad continúa funcionando pero con una capacidad reducida, lo que proporciona cierta protección de confiabilidad integrada.

Al configurar una instancia con dos o más unidades, las unidades disponibles funcionan juntas para procesar solicitudes y proporcionar equilibrio de carga automático. Si una de las unidades deja de estar disponible, las unidades restantes siguen controlando el tráfico, pero con capacidad potencialmente reducida.

Para obtener mayores niveles de confiabilidad, API Management admite la distribución de unidades entre zonas de disponibilidad dentro de una región y entre varias regiones.

> ⓘ **Nota:**
>
> API Management usa unidades para los componentes del gateway. Las unidades no son aplicables al portal para desarrolladores ni al plano de administración.

Niveles de servicio

Los niveles de servicio de API Management proporcionan distintos niveles de confiabilidad:

- **Nivel Premium (clásico):** Admite varias unidades que se pueden distribuir entre zonas de disponibilidad y regiones para lograr una máxima resistencia.
- **Nivel Premium v2**: admite varias unidades que se pueden distribuir entre zonas de disponibilidad. Actualmente no admite implementaciones de varias regiones.
- **Niveles Básico v2, Estándar y Estándar v2:** Todas admiten varias unidades dentro de un único centro de datos. No admiten zonas de disponibilidad ni implementaciones de varias regiones.
- **Nivel de desarrollador:** Solo admite una sola unidad y no proporciona compatibilidad con zonas de disponibilidad ni con varias regiones. Este nivel está diseñado para escenarios de desarrollo y pruebas. No es adecuado para cargas de trabajo de producción.
- **Nivel de consumo:** Tiene funcionalidades de resistencia integradas y es resistente a una serie de errores dentro de un único centro de datos de Azure. Sin embargo, el nivel Consumo no proporciona compatibilidad con zonas de disponibilidad ni implementaciones de varias regiones. Para comprender el tiempo de actividad esperado de una instancia de API Management de nivel de consumo, revise el acuerdo de nivel de servicio (SLA).

> ⓘ **Nota:**
>
> Los niveles Desarrollador y Premium de API Management admiten puertas de enlace autohospedada, que puede ejecutar en su propia infraestructura. Al usar puertas de enlace autohospedadas, es responsable de configurarlas para cumplir los requisitos de confiabilidad. Las puertas de enlace autohospedados están fuera del ámbito de este artículo.

Recomendaciones de implementación de producción

Azure Well-Architected Framework proporciona recomendaciones sobre confiabilidad, rendimiento, seguridad, costo y operaciones. Para comprender cómo estas áreas influyen entre sí y contribuyen a una solución confiable de API Management, consulte Procedimientos recomendados de arquitectura para API Management.

Resistencia a errores transitorios

Los errores transitorios son errores breves e intermitentes en los componentes. Se producen con frecuencia en un entorno distribuido como la nube y son una parte normal de las operaciones. Los errores transitorios se corrigen después de un breve período de tiempo. Es importante que las aplicaciones puedan controlar errores transitorios, normalmente mediante el reintento de solicitudes afectadas.

Todas las aplicaciones hospedadas en la nube deben seguir las instrucciones de control de errores transitorios de Azure cuando se comunican con cualquier API, bases de datos y otros componentes hospedados en la nube. Para obtener más información, consulte Recomendaciones para controlar errores transitorios.

Al usar API Management delante de una API, es posible que tenga que reintentar solicitudes que produzcan un error debido a errores transitorios. Para proteger la API de back-end de ser sobrecargada por demasiadas solicitudes, API Management proporciona directivas de reintento, límite de velocidad y cuota. También puede configurar las funcionalidades de equilibrio de carga y disyuntor mediante el uso de recursos de backend.

Resistencia a errores de zona de disponibilidad

Las zonas de disponibilidad son grupos físicamente independientes de centros de datos dentro de una región de Azure. Cuando una zona falla, los servicios pueden transferirse a una de las zonas restantes.

API Management proporciona dos tipos de compatibilidad con zonas de disponibilidad al implementar una instancia de API Management Premium (clásica) en una región admitida:

- **Automático** (recomendado): API Management proporciona compatibilidad con zonas de disponibilidad automáticas cuando no se especifican las zonas de disponibilidad que se van a usar.

- **Manual:** API Management proporciona compatibilidad con la zona de disponibilidad manual cuando se especifican explícitamente las zonas de disponibilidad que se van a usar.

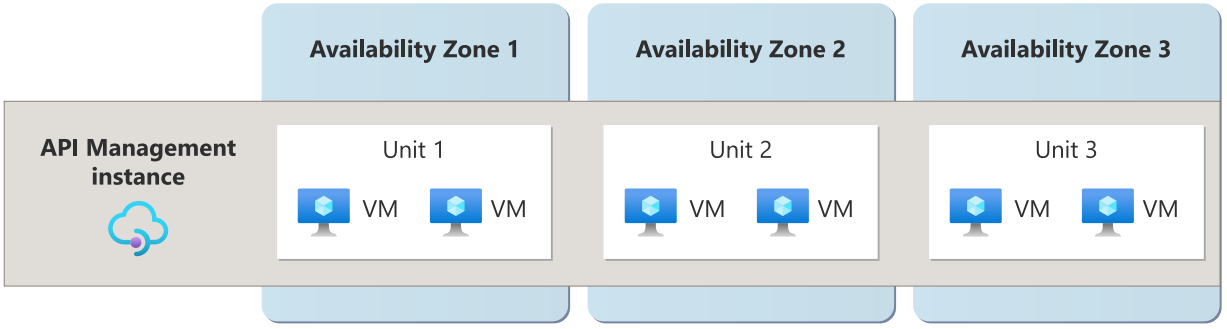
Con la compatibilidad con zonas de disponibilidad, API Management replica componentes de servicio entre zonas para lograr alta disponibilidad. En la región primaria, estos componentes incluyen la puerta de enlace (unidades de escalado), el plano de administración y el portal para desarrolladores. En regiones secundarias, solo se replican las unidades de puerta de enlace. Para más información sobre las regiones secundarias, consulte Resistencia a errores en toda la región.

Compatibilidad con zonas de disponibilidad automáticas

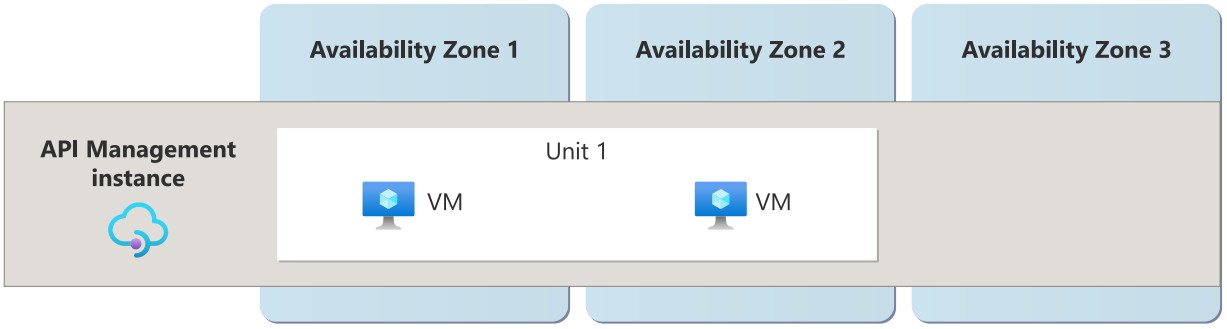
Puede usar la compatibilidad automática con zonas de disponibilidad para elegir una configuración de instancia de una sola unidad o de varias unidades con el fin de lograr la redundancia de zonas:

- **Configuración** de varias unidades (recomendada): si la instancia tiene dos o más unidades, API Management realiza un mejor intento de distribuir las unidades de la instancia entre las zonas de disponibilidad de la región. No puede determinar en qué zonas de disponibilidad se colocan las unidades. Implemente un mínimo de dos unidades, que se pueden distribuir entre dos zonas.

  En el diagrama siguiente se muestra una instancia de API Management con tres unidades configuradas para la compatibilidad automática con zonas de disponibilidad:



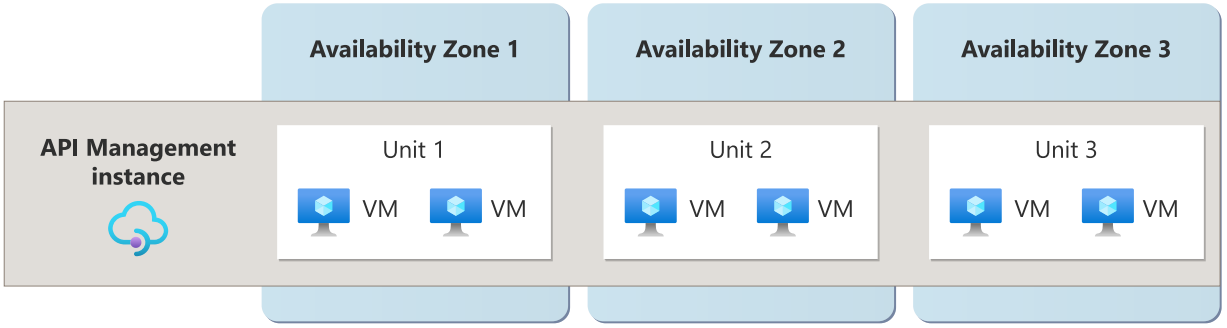
- **Configuración de una sola unidad:** Si la instancia tiene una sola unidad, los recursos de proceso subyacentes de la unidad se distribuyen a dos zonas de disponibilidad. No puede determinar en qué zonas de disponibilidad se colocan los recursos de proceso de la unidad.



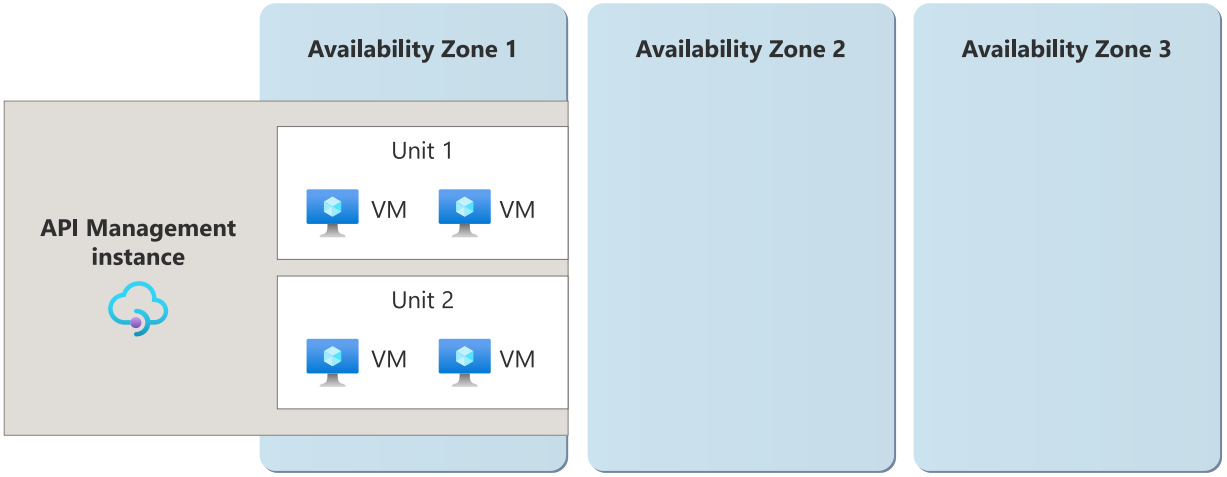
Compatibilidad con zonas de disponibilidad configuradas manualmente

Si desea seleccionar explícitamente las zonas de disponibilidad que se van a usar, puede elegir entre configuraciones con redundancia de zona y zonales:

- **Redundancia de zona:** Configure manualmente la redundancia de zona para una instancia de API Management en una región compatible con el fin de proporcionar redundancia a los componentes del servicio. Al seleccionar dos o más zonas de disponibilidad que se van a usar, Azure replica automáticamente los componentes de servicio en las zonas seleccionadas.



- **Zonal:** Los componentes del servicio API Management se implementan en una sola zona que se selecciona dentro de una región de Azure. Todas las unidades se colocan en la misma zona de disponibilidad.



> **Importante**
>
> Ancle a una sola zona de disponibilidad solo si la latencia entre zonas es demasiado alta para sus necesidades y después de comprobar que la latencia no cumple sus requisitos. Por sí mismo, una instancia zonal no proporciona resistencia a una interrupción de zona de disponibilidad. Para mejorar la resistencia de una implementación de API Management zonal, debe implementar explícitamente instancias independientes en varias zonas de disponibilidad y configurar el enrutamiento y la conmutación por error del tráfico.

Requisitos

- **Compatibilidad con regiones:** API Management admite zonas de disponibilidad para los niveles Premium (clásico) y Premium v2 en regiones en las que el nivel api Management está disponible y la región admite zonas de disponibilidad.
- **Requisito de nivel:** Debe usar el nivel Premium (clásico) o Premium v2 para configurar la compatibilidad con zonas de disponibilidad. API Management no admite actualmente zonas de disponibilidad en los niveles consumo clásico, desarrollador, básico y estándar o en los niveles Básico v2 y Estándar v2. Para ver las opciones de actualización, consulte Actualización y escalado de una instancia de API Management.

Consideraciones

- **Número de unidades para instancias con redundancia de zona:** Si configura manualmente la redundancia de zona para una instancia, también debe configurar una serie de unidades de API Management que se pueden distribuir uniformemente en todas las zonas de disponibilidad seleccionadas. Por ejemplo, si selecciona dos zonas, debe configurar al menos dos unidades. En su lugar, puede configurar cuatro unidades u otras

varias de dos unidades. Si selecciona tres zonas de disponibilidad, debe configurar tres unidades, seis unidades u otro múltiplo de tres unidades.

Si usa la compatibilidad con la zona de disponibilidad automática, no es necesario usar un número específico de unidades. Las unidades que implemente se distribuyen entre las zonas de disponibilidad de una manera más adecuada. Para obtener redundancia máxima de zona, use al menos dos unidades para que una interrupción de zona de disponibilidad no afecte al rendimiento de la puerta de enlace.

Para determinar el número de unidades que proporcionan el rendimiento necesario de su puerta de enlace, use las métricas de capacidad y sus propias pruebas. Para más información sobre el escalado y la actualización de la instancia de servicio, consulte Actualización y escalado de una instancia de API Management.

- **Escalado automático:** Si configura manualmente zonas de disponibilidad en una instancia de API Management configurada con el escalado automático, es posible que tenga que ajustar la configuración de escalado automático después de la configuración. En este caso, el número de unidades de API Management en reglas y límites de escalado automático debe ser un múltiplo del número de zonas. Si usa la compatibilidad con la zona de disponibilidad automática, no es necesario ajustar la configuración de escalado automático.

- **Requisitos de dirección IP:** Cuando habilita la compatibilidad con zonas de disponibilidad en una instancia de API Management implementada en una red virtual externa o interna, debe especificar un recurso de dirección IP pública para que lo utilice la instancia. En una red virtual interna, la dirección IP pública solo se usa para las operaciones de administración, no para las solicitudes de API. Para más información, consulte Direcciones IP en API Management.

Cost

Independientemente de la configuración de la zona de disponibilidad, si agrega más unidades, incurrirá en más costos. Para más información, consulte Precios de API Management.

Configurar soporte de zonas de disponibilidad

En esta sección se explica cómo configurar la compatibilidad con zonas de disponibilidad para la instancia de API Management. Para obtener más información, consulte Habilitación de la compatibilidad con zonas de disponibilidad en instancias de API Management.

- **Cree una instancia de API Management que admita zonas de disponibilidad:** Al crear una instancia de API Management Premium (clásica) en una región que admita zonas de

disponibilidad, la instancia admite zonas de disponibilidad de forma predeterminada. Puede seleccionar la compatibilidad automática con zonas de disponibilidad o configurar manualmente el soporte zonal o el soporte con redundancia de zona.

> ⓘ **Nota:**
>
> Cuando selecciona qué zonas de disponibilidad usar, en realidad está seleccionando la *zona de disponibilidad lógica*. Si implementa otros componentes de la carga de trabajo en una suscripción de Azure diferente, podrían usar un número de zona de disponibilidad lógica distinto para acceder a la misma zona de disponibilidad física. Para obtener más información, consulte zonas de disponibilidad física y lógica.

- **Habilite o vuelva a configurar la compatibilidad con la zona de disponibilidad:** Puede cambiar la configuración de zona de disponibilidad de una instancia de API Management, incluida la adición de zonas de disponibilidad y el traslado de una instancia zonal entre zonas de disponibilidad. Para obtener información sobre cómo configurar la compatibilidad con zonas de disponibilidad en una instancia de API Management, consulte Habilitar la compatibilidad con zonas de disponibilidad en instancias de API Management. Ninguna de las opciones de configuración requiere tiempo de inactividad.

  Al cambiar la configuración de la zona de disponibilidad, los cambios pueden tardar entre 15 y 45 minutos o más en aplicarse. La puerta de enlace de API Management puede seguir controlando las solicitudes de API durante ese tiempo.

  Cambiar la configuración de la zona de disponibilidad desencadena un cambio de dirección IP pública y privada.

Planeamiento y administración de capacidad

En un escenario de reducción de zona, no hay ninguna garantía de que las solicitudes de más capacidad en otra zona de disponibilidad tengan éxito. El reposición de unidades perdidas se produce de forma más adecuada. Si necesita capacidad garantizada cuando se produce un error en una zona de disponibilidad, cree y configure la instancia de API Management para que tenga en cuenta la pérdida de una zona mediante la realización de todas las acciones siguientes:

- Aprovisionamiento excesivo de las unidades de la instancia de API Management.
- Use la configuración de zona de disponibilidad automática o redundante por zonas.

Para más información, consulte Administración de la capacidad con sobreaprovisionamiento.

Use métricas de capacidad y sus propias pruebas para determinar el número de unidades que proporcionan el rendimiento de la puerta de enlace necesaria. Para más información sobre cómo escalar y actualizar la instancia de servicio, consulte Actualización y escalado de una instancia de API Management.

Comportamiento cuando todas las zonas están en buen estado

En esta sección se describe qué esperar cuando las instancias de API Management están configuradas con compatibilidad con zonas de disponibilidad y todas las zonas de disponibilidad están operativas.

- **Enrutamiento de tráfico entre zonas:** Durante las operaciones normales, el tráfico se enruta entre todas las unidades de API Management disponibles en todas las zonas de disponibilidad seleccionadas.
- **Replicación de datos entre zonas:** API Management almacena y replica los siguientes datos.
  - *La configuración de la puerta de enlace*, como las API y las definiciones de directiva, se sincroniza periódicamente entre las zonas de disponibilidad que seleccione para la instancia. La propagación de actualizaciones entre las zonas de disponibilidad normalmente tarda menos de 10 segundos.
  - *Datos de la memoria caché interna*, si usa la memoria caché interna que proporciona API Management. Las entradas de caché se distribuyen entre zonas de disponibilidad. La caché interna es volátil y no se garantiza que los datos se conserven. Considere la posibilidad de usar una caché externa si necesita conservar los datos almacenados en caché.
  - *Contadores de límite de velocidad*, si usa las funcionalidades de limitación de velocidad que proporciona API Management. Los contadores de límite de velocidad se replican de forma asíncrona entre las zonas de disponibilidad que seleccione para la instancia.

Comportamiento durante un fallo de zona

En esta sección se describe qué esperar cuando las instancias de API Management están configuradas con compatibilidad con la zona de disponibilidad y hay una interrupción de zona de disponibilidad.

- **Detección y respuesta:** La responsabilidad de la detección y la respuesta depende de la configuración de zona de disponibilidad que usa la instancia.

- *Automático y con redundancia de zona:* En el caso de las instancias configuradas para utilizar la compatibilidad automática con zonas de disponibilidad o configuradas manualmente para utilizar la redundancia de zona, la plataforma API Management se encarga de detectar los fallos en una zona de disponibilidad y responder a ellos. No es necesario hacer nada para iniciar una conmutación por error de la zona.
- *Zonal:* para que las instancias configuradas sean zonales, debe detectar la pérdida de una zona de disponibilidad e iniciar una conmutación por error a una instancia secundaria que cree en otra zona de disponibilidad.

- **Solicitudes activas:** Cuando una zona de disponibilidad no está disponible, las solicitudes en curso que están conectadas a una unidad de API Management en la zona de disponibilidad defectuosa se finalizan y deben reintentarse.
- **Notificación**: Microsoft no le notifica automáticamente cuando una zona está inactiva. Sin embargo, puede usar Azure Resource Health para supervisar el estado de un recurso individual y puede configurar alertas de Resource Health para notificarle problemas. También puede usar Azure Service Health para comprender el estado general del servicio, incluidos los errores de zona, y puede configurar alertas de Service Health para notificarle problemas.
- **Pérdida de datos esperada:** API Management almacena los siguientes datos.
  - *Cambios en la configuración de la puerta de enlace*, y que se replican en cada zona de disponibilidad seleccionada en aproximadamente 10 segundos. Si se produce una interrupción de una zona de disponibilidad, puede perder los cambios de configuración que no se replican.
  - *Datos de la memoria caché interna*, si usa la característica de caché interna. La caché interna es volátil y no se garantiza que los datos se conserven. Durante una interrupción del servicio en una zona de disponibilidad, es posible que se pierda parte o la totalidad de los datos almacenados en caché. Considere la posibilidad de usar una caché externa si necesita conservar los datos almacenados en caché.
  - *Contadores de límite de velocidad*, si usa la característica de límite de velocidad. Durante una interrupción de zona de disponibilidad, es posible que los contadores de límite de frecuencia no estén actualizados en las zonas supervivientes.
- **Tiempo de inactividad esperado:** El tiempo de inactividad esperado depende de la configuración de zona de disponibilidad que usa la instancia.
  - *Automático:* Puede esperar que las instancias que usen compatibilidad con zonas de disponibilidad automáticas no tengan tiempo de inactividad durante una interrupción

de la zona de disponibilidad. Las unidades de la zona o zonas no afectadas siguen funcionando.

También puede esperar instancias que usen compatibilidad automática con zonas de disponibilidad, pero que tengan una sola unidad, sin tiempo de inactividad. En este caso, API Management distribuye los recursos de proceso subyacentes de la unidad a dos zonas. El recurso de la zona no afectada sigue funcionando.

- *Con redundancia de zona:* Puede esperar que las instancias con redundancia de zona no sufran interrupciones durante una interrupción de la zona de disponibilidad.
- *Zonal:* En el caso de las instancias zonales, cuando una zona no está disponible, la instancia no está disponible hasta que se recupere la zona de disponibilidad.

- **Reenrutamiento del tráfico:** El comportamiento de reenrutamiento del tráfico depende de la configuración de zona de disponibilidad que usa la instancia.
  - *Automático y con redundancia de zona:* En el caso de las instancias que están configuradas para utilizar la compatibilidad automática con zonas de disponibilidad o que están configuradas manualmente para utilizar la redundancia de zona, cuando una zona no está disponible, ninguna de las unidades de la zona afectada estará disponible. Puede optar por escalar la instancia para agregar más unidades.
  - *Zonal:* En el caso de las instancias zonales, cuando una zona no está disponible, la instancia no está disponible. Si tiene una instancia secundaria en otra zona de disponibilidad, es responsable de redirigir el tráfico a esa instancia secundaria.

Recuperación de zona

- **Automático y con redundancia de zona:** En el caso de las instancias que están configuradas para utilizar la compatibilidad automática con zonas de disponibilidad o que están configuradas manualmente para utilizar la redundancia de zona, cuando la zona de disponibilidad se recupera, API Management restaura automáticamente las unidades de la zona de disponibilidad y redirige el tráfico entre sus unidades con normalidad.
- **Zonal:** En el caso de las instancias zonales, usted es responsable de redirigir el tráfico a la instancia en la zona de disponibilidad original una vez que la zona de disponibilidad se haya recuperado.

Prueba de fallos de zona

- **Automático y con redundancia de zona:** En el caso de las instancias que están configuradas para utilizar la compatibilidad automática con zonas de disponibilidad o que están configuradas manualmente para utilizar la redundancia de zona, la plataforma API Management administra el enrutamiento del tráfico, la conmutación por error y la conmutación por recuperación. Esta característica está totalmente administrada, por lo que no es necesario iniciar ni validar los procesos de error de zona de disponibilidad.

- **Zonal:** En el caso de las instancias zonales, no se puede simular una interrupción de la zona de disponibilidad que contiene la instancia de API Management. Sin embargo, puede configurar manualmente puertas de enlace ascendentes o equilibradores de carga para redirigir el tráfico a una instancia diferente en una zona de disponibilidad diferente.

Resistencia a errores en toda la región

Mediante una implementación de varias regiones, puede agregar puertas de enlace de API regionales a una instancia de API Management existente en una o varias regiones de Azure admitidas. La implementación en varias regiones ayuda a reducir la latencia de las solicitudes que perciben los consumidores de API distribuidos geográficamente. Una implementación de varias regiones también mejora la disponibilidad del servicio si una región se queda sin conexión.

> ⓘ **Importante**
>
> Las implementaciones de varias regiones solo se admiten en el nivel Premium (clásico) de API Management.

Implementación multirregión administrada por Microsoft

Al agregar una región, configure:

- Número de unidades que hospeda esa región.

- Resistencia a errores de zona de disponibilidad, si esa región proporciona zonas de disponibilidad.

- La configuración de red virtual en la región agregada, si las redes están configuradas en la región o regiones existentes.

Requisitos

- **Compatibilidad con regiones:** Puede crear implementaciones de varias regiones en el nivel Premium (clásico) con cualquier región de Azure que admita API Management. Para ver qué regiones admiten implementaciones de varias regiones, consulte Disponibilidad del producto por región.

- **Requisito de nivel:** Debe usar el nivel Premium (clásico) para configurar la compatibilidad con varias regiones. Para actualizar la instancia al nivel Premium (clásico), consulte Actualización al nivel Premium.

Consideraciones

- **Solo puerta de enlace:** Únicamente el componente de puerta de enlace de su instancia de gestión de API se replica en múltiples regiones. El plano de administración y el portal para desarrolladores de la instancia permanecen hospedados solo en la región primaria donde implementó originalmente el servicio.

- **Requisitos de red:** Si desea configurar una ubicación secundaria para su instancia de API Management cuando se implementa (inyecta) en una red virtual, la red virtual y la región de la subred deben coincidir con la ubicación secundaria que configure. Si agrega, quita o habilita la zona de disponibilidad en la región primaria, o si cambia la subred de la región primaria, cambia la dirección IP virtual (VIP) de la instancia de API Management. Para obtener más información, consulte Cambios en las direcciones IP. Sin embargo, si se agrega una región secundaria, el VIP de la región principal no cambia, ya que cada región tiene su propio VIP privado.

- **Nombres del sistema de nombres de dominio (DNS):** La puerta de enlace de cada región, incluida la región primaria, tiene un nombre DNS regional que sigue el patrón de dirección URL de `https://<service-name>-<region>-01.regional.azure-api.net`, por ejemplo `https://contoso-westus2-01.regional.azure-api.net`.

Cost

Agregar regiones implica costos. Para más información, consulte Precios de API Management.

Configuración de la compatibilidad con varias regiones

Para configurar la compatibilidad con varias regiones en una instancia de API Management, consulte Implementar una instancia de API Management en varias regiones de Azure.

Para quitar una región de una instancia de API Management, consulte Eliminación de una región del servicio API Management.

Planeamiento y administración de capacidad

En un escenario de caída regional, no hay garantía de que las solicitudes de más capacidad en otra región se realicen con éxito. Si necesita capacidad garantizada cuando se produce un error en una región, debe crear y configurar la instancia de API Management para que tenga en cuenta la pérdida de una región. Puede hacerlo mediante el aprovisionamiento excesivo de la capacidad de la instancia de API Management. Para más información sobre el principio de sobreaprovisionamiento, consulte Gestionar la capacidad con sobreaprovisionamiento.

En las implementaciones de varias regiones, el escalado automático solo se aplica a la región primaria. Las regiones secundarias requieren ajustes de escalado manual o herramientas personalizadas que controle.

Comportamiento cuando todas las regiones están en buen estado

En esta sección se describe qué se puede esperar cuando las instancias de API Management se configuran con compatibilidad con varias regiones y todas las regiones están operativas.

- **Enrutamiento del tráfico entre regiones:** API Management enruta automáticamente las solicitudes recibidas a una puerta de enlace regional. Una solicitud se enruta a la puerta de enlace regional con la latencia más baja del cliente. Si necesita usar un enfoque de enrutamiento diferente, puede configurar su propio Administrador de tráfico con reglas de enrutamiento personalizadas. Para obtener más información, consulte Uso del enrutamiento personalizado en las puertas de enlace regionales de API Management.

  Cuando una solicitud llega a una puerta de enlace regional de API Management, se enruta a la API backend, a menos que una directiva devuelva una respuesta directamente desde la puerta de enlace, como una respuesta almacenada en caché o un código de error. En una solución de varias regiones, debe tener cuidado de enrutar a una instancia de la API de back-end que cumpla los requisitos de rendimiento. Para obtener más información, consulte Enrutamiento de llamadas API a servicios back-end regionales.

- **Replicación de datos entre regiones:** La configuración de puerta de enlace, como las API y las definiciones de directiva, se sincronizan periódicamente entre las regiones primarias y secundarias que agregue. La propagación de actualizaciones a las puertas de enlace regionales normalmente tarda menos de 10 segundos.

  Los contadores de límite de velocidad y los datos de la caché interna son específicos de la región, por lo que no se replican entre regiones.

Comportamiento durante una falla de región

En esta sección se describe qué puede ocurrir cuando las instancias de API Management se configuran con compatibilidad con varias regiones y se produce una interrupción del servicio en una de las regiones que utiliza.

- **Detección y respuesta:** La API Management se encarga de detectar un error en una región y de realizar automáticamente una conmutación por error a una puerta de enlace en una de las otras regiones que usted configure.

- **Solicitudes activas:** Las solicitudes activas que se están procesando en la región defectuosa se pueden quitar y el cliente debe reintentarlas.

- **Pérdida de datos esperada:** API Management no almacena datos, excepto los contadores de configuración, caché y límite de velocidad.

  Los cambios de configuración se replican en cada región en aproximadamente 10 segundos. Si se produce una interrupción del servicio en su región principal, es posible que pierda los cambios de configuración que no se hayan replicado.

  Los contadores de límite de velocidad y los datos de la caché interna son específicos de la región, por lo que no se replican entre regiones.

- **Tiempo de inactividad esperado:** No se espera ningún tiempo de inactividad de puerta de enlace.

  Si la región primaria se queda sin conexión, el plano de administración de API Management y el portal para desarrolladores no estarán disponibles. Sin embargo, las regiones secundarias siguen atendiendo solicitudes de API mediante la configuración de puerta de enlace más reciente.

- **Reenrutamiento del tráfico:** Si una región se queda sin conexión, las solicitudes de API se enrutan automáticamente evitando la región fallida hacia la siguiente puerta de enlace más cercana.

Recuperación de regiones

Cuando se recupera la región primaria, API Management restaura automáticamente las unidades de la región y vuelve a enrutar el tráfico entre las unidades.

Prueba de fallos de región

Para estar preparado para interrupciones inesperadas de regiones, realice pruebas periódicas de sus respuestas a los fallos de la región. Puede simular algunos aspectos de un error de región deshabilitando el enrutamiento a una puerta de enlace regional.

Copias de seguridad y restauración

API Management no almacena la mayoría de los datos en tiempo de ejecución. Sin embargo, puede realizar una copia de seguridad de la configuración del servicio API Management. También puede usar operaciones de copia de seguridad y restauración para replicar las configuraciones del servicio API Management entre entornos operativos, como el desarrollo y el almacenamiento provisional.

> **Importante**
>
> En un procedimiento de copia de seguridad, se incluyen datos en tiempo de ejecución, como usuarios y suscripciones, que podrían no ser siempre deseables.

La copia de seguridad se admite en los niveles Desarrollador, Básico, Estándar y Premium.

Para obtener más información, consulte Cómo implementar la recuperación ante desastres mediante la copia de seguridad y restauración del servicio en API Management.

Para realizar copias de seguridad o restaurar algunos componentes o recursos del servicio, también puede considerar opciones administradas por el cliente, como las herramientas APIOps y las soluciones de infraestructura como código (IaC).

Resistencia al mantenimiento del servicio

API Management realiza actualizaciones de servicio periódicas y otras formas de mantenimiento.

En los niveles Básico, Estándar y Premium (clásico), puede personalizar cuándo en el proceso de actualización la instancia recibe una actualización. Si necesita validar el efecto de las actualizaciones en la carga de trabajo, considere la posibilidad de configurar una instancia de prueba para recibir actualizaciones al principio de un ciclo de actualización y establecer la instancia de producción para recibir actualizaciones en un momento posterior al ciclo. También puede especificar una ventana de mantenimiento, que es la hora del día en la que desea que la instancia aplique las actualizaciones del servicio.

Para obtener más información, consulte Configurar los ajustes de actualización del servicio para sus instancias de API Management.

Acuerdo de nivel de servicio

El contrato de nivel de servicio (SLA) para los servicios de Azure describe la disponibilidad esperada de cada servicio y las condiciones que la solución deberá cumplir para lograr esa expectativa de disponibilidad. Para obtener más información, consulte Acuerdos de Nivel de Servicio para servicios en línea.

Al implementar una instancia de API Management en varias zonas de disponibilidad o regiones, aumenta el porcentaje de tiempo de actividad definido en el Acuerdo de Nivel de Servicio.

El servicio proporciona su propio Acuerdo de Nivel de Servicio, pero también debe tener en cuenta la confiabilidad prevista de otros componentes de carga de trabajo, como los back-end de API.

Contenido relacionado

- Recuperación ante desastres y continuidad empresarial para API Management
- Uso de zonas de disponibilidad en API Management
- Implementación en varias regiones de API Management
- Supervisar API Management
- Planeamiento de la capacidad de API Management
- Procedimientos recomendados de arquitectura para API Management

ⓘ **Nota:** El autor ha creado este artículo con ayuda de inteligencia artificial. Más información

---

Last updated on 22/01/2026

Confiabilidad en Azure App Configuration

Azure App Configuration almacena y administra centralmente las opciones de configuración de la aplicación y las marcas de características, que reemplazan los archivos de configuración incrustados directamente dentro de las aplicaciones. Con este enfoque, puede actualizar dinámicamente los valores de configuración, realizar un seguimiento del historial de versiones y mantener un registro de los cambios de configuración a lo largo del tiempo. La disponibilidad y confiabilidad de App Configuration son consideraciones importantes porque el comportamiento de la aplicación puede depender directamente del acceso a los datos de configuración en tiempo de ejecución.

Al usar Azure, relibilidad es una responsabilidad compartida. Microsoft proporciona una variedad de capacidades para admitir resiliencia y recuperación. Es responsable de comprender cómo funcionan esas funcionalidades dentro de todos los servicios que usa y de seleccionar las funcionalidades que necesita para cumplir los objetivos empresariales y los objetivos de tiempo de actividad.

En este artículo se describe la arquitectura de confiabilidad de App Configuration y cómo el servicio está diseñado para permanecer disponible durante errores transitorios, errores de zona de disponibilidad y interrupciones de región.

Recomendaciones de implementación de producción para la confiabilidad

Para la mayoría de las implementaciones de producción de App Configuration, tenga en cuenta las siguientes recomendaciones:

- ✓ **SKU:** Use la SKU Estándar o Premium.
- ✓ **Eliminación temporal y protección de purga:** Active la eliminación temporal y la protección de purga para ayudar a evitar la eliminación de datos.
- ✓ **Para escenarios críticos:** Use la SKU Premium y configure la réplica incluida para replicar en varias regiones. Este enfoque mejora la alta disponibilidad y la resistencia a las interrupciones de la región.

Para obtener una lista de los procedimientos recomendados y la configuración de las cargas de trabajo de producción, consulte Compilación de aplicaciones que tienen alta resistencia.

Introducción a la arquitectura de confiabilidad

Al implementar App Configuration, se implementa una *tienda*. Tu tienda contiene varios tipos de configuración que la aplicación puede usar, incluidas las claves y los valores, y las banderas de características. El servicio también incluye funcionalidades integradas para organizar, proteger, control de versiones y implementar de forma segura los cambios de configuración en todos los entornos. Para obtener más información, consulte ¿Qué es App Configuration?

App Configuration es un servicio totalmente administrado. Microsoft es responsable de realizar el mantenimiento en el servicio y almacenar y administrar la configuración.

Al crear aplicaciones cliente que se conectan a App Configuration, opcionalmente puede usar App Configuration con Azure Front Door (versión preliminar) para habilitar el caché y la aceleración global del tráfico. Esta configuración presenta otras consideraciones para la replicación geográfica, que se resaltan en este artículo cuando corresponda.

Resistencia a errores transitorios

Los errores transitorios son errores breves e intermitentes en los componentes. Se producen con frecuencia en un entorno distribuido como la nube y son una parte normal de las operaciones. Los errores transitorios se corrigen después de un breve período de tiempo. Es importante que las aplicaciones puedan controlar errores transitorios, normalmente mediante el reintento de solicitudes afectadas.

Todas las aplicaciones hospedadas en la nube deben seguir las instrucciones de control de errores transitorios Azure cuando se comunican con cualquier API, bases de datos y otros componentes hospedados en la nube. Para obtener más información, consulte Recomendaciones para controlar errores transitorios.

Al usar App Configuration, tenga en cuenta los siguientes procedimientos recomendados para minimizar el efecto de los errores transitorios en el acceso a la configuración, especialmente dentro de las rutas de acceso de código críticas.

- **Proveedores de configuración:** Use proveedores de App Configuration, que tienen integradas funcionalidades de reintento, almacenamiento en caché y otras capacidades de resiliencia.

- **SDK de Azure:** Utilice los SDK de App Configuration si su aplicación necesita enviar solicitudes de escritura. Los SDKs reintentan automáticamente en las respuestas de código de estado HTTP 429 y otros errores transitorios.

- **Lógica de reintento:** Incluya lógica de reintento en clientes personalizados si no puede utilizar proveedores o SDK de App Configuration. El `retry-after-ms` encabezado de la respuesta proporciona un tiempo de espera sugerido en milisegundos antes de que el cliente vuelva a intentar la solicitud.

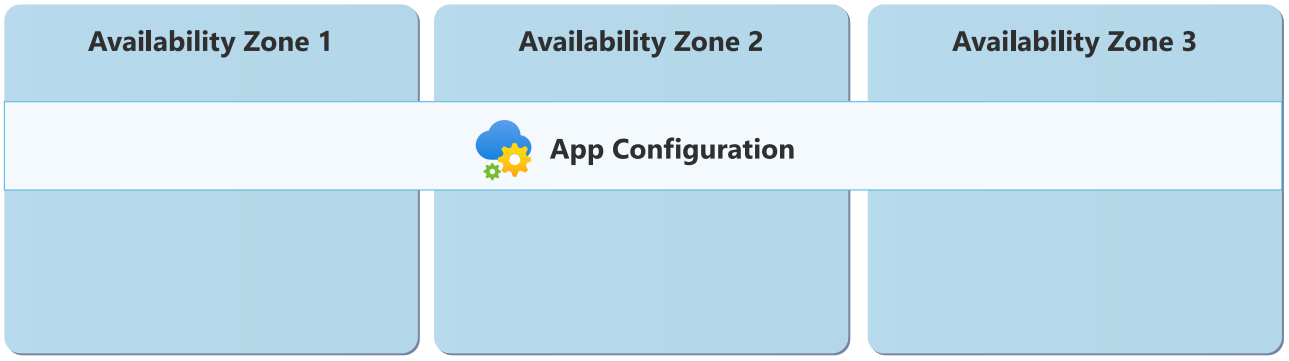
- **Almacenamiento en caché:** Configure los ajustes de caché en la memoria cuando sea posible para reducir las solicitudes directas a su tienda.

Para obtener otras instrucciones de configuración de aplicaciones, consulte Preguntas más frecuentes sobre App Configuration.

Resistencia a errores de zona de disponibilidad

Availability zones son grupos de centros de datos físicamente independientes dentro de una región de Azure. Cuando una zona falla, los servicios pueden transferirse a una de las zonas restantes.

App Configuration proporciona automáticamente redundancia de zona en regiones que admiten zonas de disponibilidad. Esta redundancia proporciona alta disponibilidad dentro de una región sin necesidad de ninguna configuración específica.



Cuando una zona de disponibilidad deja de estar disponible, App Configuration redirige automáticamente las solicitudes a otras zonas de disponibilidad correctas para garantizar una alta disponibilidad.

Requisitos

Compatibilidad con regiones: Los almacenes implementados en las siguientes regiones son automáticamente redundantes de zona.

Expandir tabla

Americas	Europa	Oriente Medio	África	Asia Pacífico
Sur de Brasil	Centro de Francia	Israel Central		Australia East
Canada Central	Centro-oeste de Alemania	Qatar Central		Centro de la India
Central US	Norte de Italia	Norte de Emiratos Árabes Unidos		Norte de China 3
East US	Norte de Europa			Este de Asia
Este de EE. UU. 2	Norway East			Japón Oriental
Centro de México	Centro de Polonia			Centro de Corea del Sur
Centro-sur de EE. UU.	Centro de España			Sudeste de Asia
Gobierno de EE. UU. - Virginia	Centro de Suecia			
Oeste de EE. UU. 2	Norte de Suiza			
Oeste de EE. UU. 3	Sur de Reino Unido			
	Oeste de Europa			

Cost

No hay ningún costo adicional para la redundancia de zona para App Configuration.

Configurar soporte de zonas de disponibilidad

Microsoft configura automáticamente la redundancia de zona para un almacén cuando se encuentra en a región que admite zonas de disponibilidad.

Si App Configuration agrega compatibilidad con zonas de disponibilidad a una región existente, no es necesario realizar ninguna acción para beneficiarse de la compatibilidad con la zona de disponibilidad. La tienda se beneficia de la compatibilidad con la zona de disponibilidad que está disponible para los almacenes de App Configuration en la región.

Comportamiento cuando todas las zonas están en buen estado

En esta sección describimos lo que puede esperar cuando tiene un almacén de App Configuration con redundancia de zona y todas las zonas están operativas.

- **Operación entre zonas:** App Configuration administra automáticamente el enrutamiento de tráfico entre zonas de disponibilidad. Durante las operaciones normales, distribuye de forma transparente las solicitudes entre zonas.
- **Replicación de datos entre zonas:** En las regiones que admiten zonas, App Configuration replica de forma sincrónica los datos entre zonas de disponibilidad. Esta replicación garantiza que la configuración siga siendo coherente y disponible incluso si una zona deja de estar disponible.

Comportamiento durante un fallo de zona

En esta sección se describe qué cabe esperar al usar un almacén de App Configuration con redundancia de zona y se produce una interrupción en una de las zonas.

- **Detección y respuesta:** El servicio App Configuration detecta errores de zona y responde automáticamente a ellos. No es necesario realizar ninguna acción ante un fallo de zona.
- **Notification:** Microsoft no le notifica automáticamente cuando una zona está inactiva. Sin embargo, puede usar Azure Service Health para comprender el estado general del servicio, incluidos los errores de zona, y puede configurar alertas de Service Health para notificarle problemas.
- **Solicitudes activas:** durante un error de zona, es posible que la zona afectada no controle las solicitudes en curso, lo que requiere que las aplicaciones cliente vuelvan a intentarlas. Las aplicaciones cliente deben seguir los procedimientos transitorios de control de errores para asegurarse de que pueden reintentar solicitudes si se produce un error de zona.
- **Pérdida de datos esperada:** No se espera ninguna pérdida de datos durante un error de zona debido a la replicación sincrónica entre zonas.
- **Tiempo de inactividad esperado:** No se espera ningún tiempo de inactividad.
- **Redistribución:** App Configuration vuelve a enrutar automáticamente el tráfico de la zona afectada a zonas correctas sin necesidad de intervención del cliente.

Recuperación de zona

Cuando se recupera una zona no disponible anteriormente, App Configuration restaura automáticamente las operaciones normales en todas las zonas de disponibilidad. No es necesario realizar ninguna acción para recuperar de una falla de zona.

Prueba de fallos de zona

La plataforma App Configuration gestiona el enrutamiento del tráfico, la conmutación ante fallos y la recuperación de zona para almacenes con redundancia de zona. Microsoft administra completamente este proceso, por lo que no es necesario validar los procesos de error de zona de disponibilidad.

Resistencia a errores en toda la región

App Configuration proporciona capacidades nativas de geo-replicación para admitir la resiliencia durante interrupciones regionales. La replicación geográfica permite que los datos de configuración se repliquen entre regiones como una característica de servicio administrado.

Replicación geográfica

Con la replicación geográfica, puede replicar un almacén en varias regiones de Azure. Cada almacén puede tener varias *réplicas* en regiones diferentes. El almacén original también es una réplica. Esta funcionalidad ayuda a proteger las aplicaciones frente a interrupciones en toda la región.

Requisitos

- **Region support:** Puede crear réplicas en cualquier región de Azure compatible con App Configuration, incluso si las regiones no son regiones emparejadas de Azure.
- **Nivel:** El almacén de configuración debe usar un nivel admitido para habilitar la replicación geográfica. Para obtener más información, consulte Habilitación de la replicación geográfica.

Consideraciones

Al habilitar la replicación geográfica, tenga en cuenta los siguientes factores:

- **Réplicas redundantes por zona:** Cualquier réplica creada en una región donde App Configuration admita zonas de disponibilidad será automáticamente redundante por zona.
- **Azure Front Door:** Para habilitar la entrega de configuración con redundancia geográfica con Azure Front Door, configure réplicas de App Configuration como orígenes dentro de un grupo de origen. Los orígenes configurados correctamente son necesarios para que Azure Front Door realice el enrutamiento basado en la salud, el equilibrio de carga y la

conmutación automática por error entre regiones. Para obtener más información, consulte Métodos de enrutamiento del tráfico al origen.

Cost

Cada región con replicación geográfica se factura por separado según los precios del plan y la región correspondientes. No se aplican cargos de salida de datos para la replicación entre regiones. Para obtener más información sobre los precios, consulte Precios de App Configuration.

Configuración de la compatibilidad con varias regiones

Para configurar la replicación para un almacén de configuración recién creado, consulte Habilitación de la replicación geográfica.

Comportamiento cuando todas las regiones están en buen estado

En esta sección se describe qué esperar al configurar un almacén de App Configuration para la replicación geográfica y todas las regiones están operativas.

- **Operación entre regiones:** Cada réplica es direccionable individualmente y tiene su propio nombre del sistema de nombres de dominio (DNS). Todas las réplicas pueden aceptar operaciones de lectura y escritura.

  App Configuration no enruta automáticamente el tráfico entre regiones. Cuando se usan proveedores de configuración de App Configuration, la aplicación puede usar opcionalmente la detección automática de réplicas. Alternativamente, puede especificar una lista priorizada de réplicas, y App Configuration selecciona la primera réplica saludable. Este enfoque permite a la aplicación controlar qué réplica usa.

  > **Nota:**
  >
  > Si usa Azure Front Door, el comportamiento del enrutamiento del tráfico es diferente. Para obtener más información, consulte Conmutación por error y equilibrio de carga.

- **Replicación de datos entre regiones:** Los datos se replican de forma asincrónica y finalmente son coherentes. Puede usar la métrica de latencia de replicación en Azure Monitor para supervisar la latencia de replicación actual entre réplicas.

Comportamiento durante una falla de región

En esta sección se describe qué se puede esperar al configurar una tienda de configuración de aplicaciones para la replicación geográfica, y cuando se produce una interrupción en una de las regiones de réplica.

- **Detección y respuesta:** Microsoft es responsable de detectar errores de región o réplica e iniciar procesos de recuperación.

  Cuando se usan proveedores de configuración de App Configuration con detección automática de réplicas o una lista de varias réplicas, la aplicación detecta automáticamente errores y realiza una conmutación en caso de error a una réplica sana.

  Si no usa proveedores de App Configuration, es responsable de cambiar su aplicación a una réplica saludable.

- **Notification:** Microsoft no le notifica automáticamente cuando una región está inactiva. Sin embargo, puede usar Azure Service Health para comprender el estado general del servicio, incluidos los errores de región, y puede configurar alertas de Service Health para notificarle problemas.

- **Solicitudes activas:** Las solicitudes activas contra una réplica en la región pueden fallar. Las aplicaciones cliente deben reintentar las solicitudes en una réplica diferente.

- **Pérdida de datos esperada:** Si se produce un error en una réplica, es posible que los cambios recientes realizados en esa réplica aún no se repliquen en otras réplicas. Esos cambios pueden permanecer no disponibles hasta que se recupere la réplica. Para calcular la posible pérdida de datos, supervise la métrica de latencia de replicación en Azure Monitor.

- **Tiempo de inactividad esperado:** Cuando una réplica deja de estar disponible, permanece sin conexión hasta que se recupera su región. Otras réplicas siguen gestionando las solicitudes. Las aplicaciones pueden experimentar un breve tiempo de inactividad mientras detectan el error y cambian a una réplica correcta. La duración depende de la rapidez con la que cada aplicación realiza esta detección y conmutación por error.

- **Redistribución:** Las aplicaciones deben enrutar el tráfico a una réplica saludable cuando se produce un error.

  Si utiliza proveedores de configuración de App Configuration, estos gestionan automáticamente la selección de réplica y la conmutación por error.

Si coloca Azure Front Door delante del almacenamiento de datos y configura el grupo de origen con varias réplicas como orígenes para la conmutación por error, Azure Front Door vuelve a enrutar automáticamente las solicitudes a una réplica en buen estado.

Recuperación de regiones

Una vez que la región se recupera, App Configuration sincroniza la réplica con las demás réplicas sin intervención del usuario.

Usted es responsable de volver a configurar su aplicación para enrutar el tráfico de regreso a la instancia de la región recuperada. Las aplicaciones que utilizan App Configuration reanudan automáticamente el uso de la réplica.

Prueba de fallos de región

No se puede simular una conmutación por error de réplica directamente en App Configuration. Sin embargo, dado que las aplicaciones controlan la selección de réplica, puede probar el comportamiento de la conmutación por error forzando la aplicación a un estado en el que debe cambiar las réplicas.

Para validar el comportamiento de failover de réplica de su aplicación, puede introducir un error de conectividad controlado en un entorno de no producción y observar la respuesta de la aplicación.

Un enfoque consiste en usar la máquina local u otro entorno en el que tenga acceso administrativo. Siga estos pasos:

1. Active el registro exhaustivo para el SDK de Azure. En .NET, use la clase `AzureEventSourceListener` para configurar un registrador. Para obtener más información, vea el documento sobre supervisión y registro.

2. Configure manualmente el `hosts` archivo para que las solicitudes al almacén de App Configuration se enruten a una dirección IP que no pueda recibirlos, como `127.0.0.1` (localhost).

   > ⚠ **Advertencia**
   >
   > Este paso bloquea eficazmente el acceso desde tu equipo a tu almacén de Configuración de la Aplicación. Siga estos pasos solo en un entorno que no sea de producción.

3. Supervise los registros de un mensaje similar al ejemplo siguiente:

   Consola

   ```
   [Warning] Microsoft-Extensions-Configuration-AzureAppConfiguration-Refresh:
   Failed to get configuration settings from endpoint
   'https://myappconfigstore.azconfig.io'. Failing over to endpoint
   https://myappconfigstore-eus.azconfig.io'.
   ```

   Este mensaje indica que la aplicación conmutó con éxito para usar otra réplica de su almacén.

4. Después de completar la prueba, deshaga los cambios en el archivo `hosts`.

Copias de seguridad y restauración

Puede usar App Configuration para exportar datos de configuración desde un almacén y usarlos como parte de una estrategia de copia de seguridad más amplia.

Para la mayoría de las soluciones, no debe confiar exclusivamente en copias de seguridad. En su lugar, utilice las otras capacidades descritas en esta guía para apoyar los requisitos de resiliencia. Sin embargo, las copias de seguridad protegen contra algunos riesgos que otros enfoques no. Para más información, consulte ¿Qué son la redundancia, la replicación y la copia de seguridad?.

Resistencia a la eliminación accidental

App Configuration proporciona dos características clave de recuperación para evitar la eliminación accidental o malintencionada:

- **Eliminación temporal:** Al activar la eliminación temporal, puede recuperar almacenes y configuraciones eliminados durante un período de retención configurable. Las funciones de borrado suave actúan como una papelera de reciclaje para los recursos de App Configuration.

- **Protección de purga:** Al activar la protección de purga, el servicio evita la eliminación permanente del almacén y su configuración hasta que transcurre el período de retención. Esta protección evita que los actores malintencionados destruyan permanentemente la configuración.

Use ambas características para entornos de producción. Para más información, consulte Eliminación temporal y protección contra purgas.

Resistencia al mantenimiento del servicio

Microsoft realiza periódicamente actualizaciones de servicio y otro mantenimiento. El servicio controla estas actividades automáticamente, lo que hace que el mantenimiento sea transparente y sin problemas. No se espera ningún tiempo de inactividad durante los eventos de mantenimiento a menos que Azure Service Health proporcione un aviso de mantenimiento planeado.

Resistencia a problemas de configuración

Los cambios de configuración incorrectos o accidentales pueden provocar tiempo de inactividad de la aplicación. Utiliza instantáneas de configuración para desplegar los cambios en la configuración de forma segura. Supervise el estado de la aplicación después de cualquier cambio de configuración y vuelva a la última instantánea de configuración correcta conocida si los cambios presentan un problema.

Acuerdo de nivel de servicio

El acuerdo de nivel de servicio (SLA) para Azure servicios describe la disponibilidad esperada de cada servicio y las condiciones que la solución debe cumplir para lograr esa expectativa de disponibilidad. Para obtener más información, vea SLAs for servicios en línea.

Contenido relacionado

- Confiabilidad en Azure
- Documentación de Configuración de la aplicación

ⓘ**Nota:** El autor ha creado este artículo con ayuda de inteligencia artificial. Más información

---

Last updated on 15/04/2026

Confiabilidad en Azure App Service

Azure App Service es un servicio basado en HTTP para hospedar aplicaciones web, API REST y back-end móvil. App Service se integra con Microsoft Azure para proporcionar seguridad, equilibrio de carga, escalado automático y administración automatizada para aplicaciones. Como servicio de Azure, App Service proporciona una variedad de funcionalidades para satisfacer los requisitos de confiabilidad.

Cuando se usa Azure, la confiabilidad es una responsabilidad compartida. Microsoft proporciona una variedad de funcionalidades para admitir resistencia y recuperación. Es responsable de comprender cómo funcionan esas funcionalidades dentro de todos los servicios que usa y de seleccionar las funcionalidades que necesita para cumplir los objetivos empresariales y los objetivos de tiempo de actividad.

En este artículo se describe cómo hacer que App Service sea resistente a una variedad de posibles interrupciones y problemas, incluidos errores transitorios, interrupciones de zona de disponibilidad, interrupciones de regiones y mantenimiento del servicio. También se describe cómo puede usar copias de seguridad para recuperarse de otros tipos de problemas y se resalta cierta información clave sobre el acuerdo de nivel de servicio (SLA) de App Service.

> **Nota:**
>
> Si busca información sobre el soporte de confiabilidad en App Service Environment, consulte Confiabilidad en App Service Environment.

Recomendaciones de implementación de producción

Azure Well-Architected Framework proporciona recomendaciones sobre confiabilidad, rendimiento, seguridad, costo y operaciones. Para comprender cómo estas áreas influyen entre sí y contribuir a una solución confiable de App Service, consulte Procedimientos recomendados de arquitectura para App Service (Web Apps) en Azure Well-Architected Framework.

Introducción a la arquitectura de confiabilidad

Al crear una aplicación web de App Service, se especifica el plan de App Service que ejecuta la aplicación.

Un plan de App Service define un conjunto de recursos de proceso que ejecutan las aplicaciones web. Todas las Web Apps deben ejecutarse dentro de un plan. Puede escalar un

plan para ejecutarse en varias *instancias* de máquina virtual, también denominadas *trabajadores*. Estas instancias proporcionan los recursos de proceso que ejecutan el código de su aplicación. Un único plan de App Service puede hospedar varias aplicaciones. Todas las aplicaciones se ejecutan en el mismo conjunto compartido de instancias de máquina virtual.

App Service ofrece las siguientes características de redundancia:

- **Distribución entre dominios de error:** En el nivel de plataforma, Azure distribuye automáticamente las instancias de máquina virtual del plan de App Service entre dominios de error dentro de la región de Azure. Esta distribución minimiza el riesgo de errores de hardware localizados mediante la agrupación de máquinas virtuales que comparten una fuente de alimentación común y un conmutador de red.

- **Distribución entre zonas de disponibilidad:** Si habilita la redundancia de zona en un plan de App Service compatible, Azure distribuye las instancias entre zonas de disponibilidad dentro de la región. Esta configuración proporciona una mayor resistencia si se produce una interrupción de zona. Para obtener más información sobre la redundancia de zonas, consulte Compatibilidad con zonas de disponibilidad.

- **Escalado de aplicaciones:** Al configurar el plan de App Service para ejecutar varias instancias de máquina virtual, todas las aplicaciones del plan se ejecutan en todas las instancias de forma predeterminada. Si configura su plan para el escalado automático, todas las aplicaciones se escalarán juntas según la configuración de escalado automático. Sin embargo, puede personalizar el número de instancias de plan que ejecutan una aplicación específica mediante el escalado por aplicación.

- **Unidades de escalado:** Internamente, App Service se ejecuta en una infraestructura de plataforma denominada *unidades de escalado*, también conocida como *stamps* o *espacios web*. Una unidad de escalado incluye todos los componentes necesarios para hospedar y ejecutar App Service, incluido el proceso, el almacenamiento, las redes y el equilibrio de carga. Azure administra las unidades de escalado para garantizar una distribución equilibrada de cargas de trabajo, realizar un mantenimiento rutinario y mantener la confiabilidad general de la plataforma.

  Es posible que algunas funcionalidades solo se apliquen a unidades de escalado específicas. Por ejemplo, algunas unidades de escalado de App Service pueden admitir la redundancia de zona, mientras que otras unidades de escala de la misma región no.

Resistencia a errores transitorios

Los errores transitorios son errores breves e intermitentes en los componentes. Se producen con frecuencia en un entorno distribuido como la nube y son una parte normal de las

operaciones. Los errores transitorios se corrigen después de un breve período de tiempo. Es importante que las aplicaciones puedan controlar errores transitorios, normalmente mediante el reintento de solicitudes afectadas.

Todas las aplicaciones hospedadas en la nube deben seguir las instrucciones de control de errores transitorios de Azure cuando se comunican con cualquier API, bases de datos y otros componentes hospedados en la nube. Para obtener más información, consulte Recomendaciones para controlar errores transitorios.

Los SDK proporcionados por Microsoft suelen controlar errores transitorios. Dado que hospeda sus propias aplicaciones en App Service, siga los pasos necesarios para reducir la posibilidad de errores transitorios:

- **Implemente varias instancias en el plan.** App Service realiza actualizaciones automatizadas y otras formas de mantenimiento en instancias del plan. Si una instancia se vuelve incorrecta, el servicio puede reemplazar automáticamente esa instancia por una nueva instancia correcta. Durante el proceso de reemplazo, puede haber un breve período cuando la instancia anterior no está disponible y una nueva instancia no está lista para atender el tráfico. Para mitigar estos efectos, implemente varias instancias del plan de App Service.

- **Uso de ranuras de implementación.** Las ranuras de implementación de App Service permiten implementaciones sin tiempo de inactividad de las aplicaciones. Use ranuras de implementación para minimizar el efecto de las implementaciones y los cambios de configuración para los usuarios. Las ranuras de implementación también reducen la probabilidad de que se reinicie la aplicación. Reiniciar la aplicación produce un error transitorio.

- **Evite escalar o reducir verticalmente.** Estas operaciones modifican la CPU, la memoria y otros recursos asignados a cada instancia, y pueden provocar el reinicio de la aplicación empresarial. En su lugar, seleccione un nivel y un tamaño de instancia que cumplan los requisitos de rendimiento bajo una carga típica. Para escalar y reducir horizontalmente, agregue y elimine instancias de forma dinámica para administrar los cambios en el volumen de tráfico.

Resistencia a errores de zona de disponibilidad

Las zonas de disponibilidad son grupos físicamente independientes de centros de datos dentro de una región de Azure. Cuando una zona falla, los servicios pueden transferirse a una de las zonas restantes.

En el caso de los **niveles Premium v2 a v4**, puede configurar App Service como *redundante de zona*, lo que significa que los recursos se distribuyen entre varias zonas de disponibilidad. La distribución entre varias zonas ayuda a las cargas de trabajo de producción a lograr resistencia y confiabilidad. Al configurar la redundancia de zona en los planes de App Service, todas las aplicaciones que utilizan el plan pasan a ser redundantes en la zona.

Requisitos

Para habilitar la redundancia de zona, debe cumplir los siguientes requisitos:

- **Compatibilidad con regiones:** En el caso de los planes de App Service **Premium v2 y v3** , la redundancia de zona se admite en cualquier región que admita zonas de disponibilidad.
- **Tipo de plan:** Use tipos de plan Premium v2 a v4.

  > ⓘ **Importante**
  >
  > Para habilitar la redundancia de zona para los planes **de App Service Premium v4** , debe **confirmar que la región deseada admite planes v4** y que **admite zonas de disponibilidad**.

- **Número mínimo de instancias:** Implemente un mínimo de dos instancias en el plan.
- **Unidad de escalado:** La aplicación debe implementarse en una unidad de escalado que admita zonas de disponibilidad. No controla directamente la unidad de escalado que utiliza el plan. En su lugar, al crear un plan de App Service, el plan se asigna a una unidad de escalado en función del grupo de recursos del propio plan. Para determinar si la unidad de escalado del plan de App Service admite redundancia de zona, consulte Comprobar la compatibilidad con redundancia de zona para un plan de App Service.

  Si tu plan de App Service está en una unidad de escalado que no admite redundancia de zona, no puedes habilitar la redundancia de zona en tu plan. En su lugar, debe reimplementar tus aplicaciones en un nuevo plan en una unidad de escala diferente.

Distribución de instancias a través de zonas

Al crear un plan de App Service con redundancia de zona, Azure distribuye las instancias del plan entre zonas de disponibilidad de la región. Esta distribución garantiza que sus aplicaciones sigan estando disponibles incluso si una zona sufre una interrupción del servicio.

La distribución de instancias en una implementación con redundancia de zona sigue reglas específicas. Estas reglas también se aplican a medida que la aplicación se escala y se reduce horizontalmente:

- **Instancias mínimas:** El plan de App Service debe tener un mínimo de dos instancias para la redundancia de zona.
- **Zonas de disponibilidad máxima compatibles con el plan:** Azure determina el número de zonas de disponibilidad que puede usar el plan, lo que se conoce como *maximumNumberOfZones*. Para ver el número de zonas de disponibilidad que puede usar su plan específico, consulte Comprobar la compatibilidad con la redundancia de zonas para un plan de App Service.
- **Distribución de instancias:** Cuando se habilita la redundancia de zona, Azure distribuye instancias de plan entre varias zonas de disponibilidad automáticamente. La distribución se basa en las reglas siguientes:
  - Si el número de instancias supera el *maximumNumberOfZones* y divide uniformemente, Azure distribuye las instancias uniformemente entre zonas.
  - Si el número de instancias no se divide de manera uniforme, Azure distribuye las instancias restantes entre las zonas restantes.
  - Cuando la plataforma de App Service asigna instancias para un plan de App Service con redundancia de zona, usa el equilibrio de zona de mejor esfuerzo que proporcionan los conjuntos de escalado de máquinas virtuales de Azure subyacentes. Un plan se equilibra si cada zona tiene el mismo número de máquinas virtuales o difiere en una instancia de todas las demás zonas. Para obtener más información, consulte Equilibrio de zona.
- **Ubicación de zona física:** Puede ver la zona de disponibilidad física que se usa para cada una de las instancias del plan de App Service. Para más información, consulte Visualización de zonas físicas para un plan de App Service.

Consideraciones

En el caso de los planes **Premium v2 a v4**, una interrupción de zona de disponibilidad podría afectar a algunos aspectos de Azure App Service, aunque la aplicación siga sirviendo tráfico. Estos comportamientos incluyen el escalado del plan de App Service, la creación de aplicaciones, la configuración de la aplicación y la publicación de aplicaciones.

Al habilitar la redundancia de zona en su plan App Service **Premium v2 a v4**, también mejora la resistencia durante las actualizaciones de la plataforma. Para obtener más información,

consulte Resistencia al mantenimiento del servicio.

En los planes de App Service que no están configurados como redundantes por zona, las instancias de máquina virtual (VM) subyacentes no son resistentes a los errores de la zona de disponibilidad. Pueden experimentar tiempo de inactividad durante una interrupción en cualquier zona de esa región.

Cost

Cuando use los **planes Premium v2 a v4 de App Service**, habilitar las zonas de disponibilidad no supone ningún coste adicional si tiene dos o más instancias. Los cargos se basan en la SKU de su plan de App Service, la capacidad que especifique y las instancias a las que escale según sus criterios de escalado automático.

Si habilita zonas de disponibilidad pero especifica una capacidad de menos de dos, la plataforma aplica un recuento mínimo de instancias de dos. La plataforma le cobra por esas dos instancias.

Configurar soporte de zonas de disponibilidad

- **Cree un nuevo plan de App Service con redundancia de zona.** Para más información, consulte Crear un nuevo plan de App Service que incluya redundancia de zona.
- **Habilite o deshabilite la redundancia de zona en un plan de App Service existente.** Para obtener más información, consulte Configurar la redundancia de zona para un plan de App Service existente.

Planeamiento y administración de capacidad

Para prepararse para errores de zona de disponibilidad, considere la posibilidad de *sobreaprovisionar* la capacidad del plan de App Service. Este enfoque permite que la solución tolere cierta pérdida de capacidad y siga funcionando sin que se vea afectado su rendimiento. Para obtener más información, consulte Administración de la capacidad mediante el aprovisionamiento excesivo.

Comportamiento cuando todas las zonas están en buen estado

En la siguiente lista se describe qué esperar cuando los planes de App Service están configurados para la redundancia de zona y todas las zonas de disponibilidad están operativas:

- **Enrutamiento del tráfico entre zonas:** Durante el funcionamiento normal, el tráfico se enruta entre todas las instancias disponibles del plan de App Service en todas las zonas de disponibilidad.

- **Replicación de datos entre zonas:** Durante las operaciones normales, cualquier estado almacenado en el sistema de archivos de la aplicación se almacena en el almacenamiento con redundancia de zona y se replica sincrónicamente entre zonas de disponibilidad.

Comportamiento durante un fallo de zona

Una interrupción de la zona de disponibilidad puede afectar a algunos aspectos de App Service, aunque la aplicación siga sirviendo al tráfico. Estos comportamientos incluyen el escalado del plan de App Service, la creación de aplicaciones, la configuración de la aplicación y la publicación de aplicaciones.

La siguiente lista describe qué puede ocurrir cuando los planes de App Service están configurados para la redundancia de zona y una o varias zonas de disponibilidad no están disponibles:

- **Detección y respuesta:** La plataforma de App Service detecta automáticamente errores en una zona de disponibilidad e inicia una respuesta. No se requiere ninguna intervención manual para iniciar una conmutación por error de zona.

- **Notificación**: Microsoft no le notifica automáticamente cuando una zona está inactiva. Sin embargo, puede usar Azure Resource Health para supervisar el estado de un recurso individual y puede configurar alertas de Resource Health para notificarle problemas. También puede usar Azure Service Health para comprender el estado general del servicio, incluidos los errores de zona, y puede configurar alertas de Service Health para notificarle problemas.

- **Solicitudes activas:** Las solicitudes en curso que se conectan a una instancia del plan de App Service en la zona de disponibilidad defectuosa se finalizan. Reintentar esas solicitudes.

- **Enrutamiento del tráfico:**App Service detecta las instancias perdidas de esa zona e intenta encontrar nuevas instancias de reemplazo Después de que App Service encuentre reemplazos, distribuye el tráfico entre las nuevas instancias según sea necesario.

  Si se configura el ajuste automático y se determina que se necesitan más instancias, se solicitan instancias a App Service. El comportamiento de escalabilidad automática funciona independientemente del comportamiento de la plataforma de App Service. Por lo tanto, la especificación del número de instancias no tiene por qué ser un múltiplo de

dos. Para más información, consulte Escalado vertical de una aplicación en App Service y Información general sobre el escalado automático.

> **Importante**
>
> Azure no garantiza que las solicitudes de más instancias se realicen correctamente en un escenario de zona descendente. La plataforma intenta reponer instancias perdidas de forma óptima. Si necesita capacidad garantizada durante un error de zona de disponibilidad, cree y configure sus planes de App Service para tener en cuenta la pérdida de zona mediante el **sobreaprovisionamiento de la capacidad**.

- **Comportamientos fuera del tiempo de ejecución:** Las aplicaciones en un plan de App Service con redundancia de zona continúan ejecutándose y prestando servicio al tráfico incluso si una zona de disponibilidad sufre una interrupción. Sin embargo, los comportamientos que no son de ejecución pueden verse afectados durante una interrupción de la zona de disponibilidad. Estos comportamientos incluyen el escalado del plan de App Service, la creación de aplicaciones, la configuración de la aplicación y la publicación de aplicaciones.

Recuperación de zona

Cuando se recupera la zona de disponibilidad, App Service crea automáticamente instancias en la zona de disponibilidad recuperada, quita las instancias temporales creadas en las otras zonas de disponibilidad y enruta el tráfico entre las instancias como de costumbre.

Prueba de fallos de zona

La plataforma de App Service administra el enrutamiento del tráfico, la conmutación por error y la conmutación por recuperación para planes de App Service con redundancia de zona. Esta característica está totalmente administrada, por lo que no es necesario iniciar ni validar los procesos de error de zona de disponibilidad.

Resistencia a errores en toda la región

App Service es un servicio de una sola región. Si la región deja de estar disponible, la aplicación tampoco está disponible.

Soluciones personalizadas de varias regiones para la resistencia

Para reducir el riesgo de un error de una sola región que afecte a la aplicación, puede implementar planes en varias regiones. Los pasos siguientes ayudan a reforzar la resistencia:

- Despliegue su aplicación en los planes de cada región.
- Configure directivas de equilibrio de carga y conmutación por error.
- Replique los datos entre regiones para que pueda recuperar el último estado de la aplicación.

Considere los siguientes recursos relacionados:

- Arquitectura de referencia: aplicación web de alta disponibilidad en múltiples regiones
- Enfoques que se deben tener en cuenta
- Tutorial: Crear una aplicación de alta disponibilidad en varias regiones en App Service

Copias de seguridad y restauración

Cuando use el nivel Básico o superior, puede hacer una copia de seguridad de sus aplicaciones de App Service en un archivo usando las funciones de copia de seguridad y restauración de App Service.

Estas funcionalidades resultan útiles cuando resulta complicado volver a implementar el código o cuando se almacena el estado en el disco. La mayoría de las soluciones no deberían depender exclusivamente de las copias de seguridad. En su lugar, use las otras capacidades de esta guía para satisfacer sus requisitos de resistencia. Sin embargo, las copias de seguridad protegen contra algunos riesgos que otros enfoques no.

> ⓘ **Importante**
>
> A partir del **31 de marzo de 2028**, las copias de seguridad personalizadas de Azure App Service **ya no admiten la copia de seguridad de bases de datos vinculadas**. Consulte Desuso de copias de seguridad de bases de datos vinculadas para obtener más información.
>
> En su lugar, use las herramientas nativas de copia de seguridad y restauración de la base de datos vinculada. Para más información, consulte Copia de seguridad y restauración de la aplicación en App Service.

Resistencia al mantenimiento del servicio

App Service realiza actualizaciones de servicio periódicas y otras tareas de mantenimiento. Para mantener la capacidad esperada durante una actualización, la plataforma agrega

automáticamente instancias adicionales del plan de App Service durante el proceso de actualización.

Habilitación de la redundancia de zona. Al habilitar la redundancia de zona en su plan de App Service, también mejora la resiliencia durante las actualizaciones de la plataforma. *Los dominios de actualización* constan de colecciones de máquinas virtuales que se desconectan durante una actualización y se asignan a zonas de disponibilidad. La implementación de varias instancias en su plan de App Service y la habilitación de la redundancia de zona para su plan agregan una capa adicional de resistencia en caso de que una instancia o zona sea incorrecta durante una actualización.

Para obtener más información, consulte Mantenimiento planeado rutinario para App Service y Mantenimiento rutinario para App Service, reinicios y tiempo de inactividad.

Acuerdo de nivel de servicio

El contrato de nivel de servicio (SLA) para los servicios de Azure describe la disponibilidad esperada de cada servicio y las condiciones que la solución deberá cumplir para lograr esa expectativa de disponibilidad. Para obtener más información, consulte Acuerdos de Nivel de Servicio para servicios en línea.

Al implementar un plan de App Service con redundancia de zona, aumenta el porcentaje de tiempo de actividad definido en el Acuerdo de Nivel de Servicio.

Contenido relacionado

- Confiabilidad en Azure
- Confiabilidad en App Service Environment

Last updated on 22/01/2026

Confiabilidad en Azure Application Gateway v2

Azure Application Gateway es un equilibrador de carga de tráfico web que puede usar para administrar el tráfico a las aplicaciones web. Proporciona funciones avanzadas como el escalado automático, la redundancia de zona, las direcciones IP virtuales estáticas (VIP) y la integración del firewall de aplicaciones web (WAF) para ofrecer servicios de entrega de aplicaciones de alta disponibilidad y seguros.

Cuando se usa Azure, la confiabilidad es una responsabilidad compartida. Microsoft proporciona una variedad de funcionalidades para admitir resistencia y recuperación. Es responsable de comprender cómo funcionan esas funcionalidades dentro de todos los servicios que usa y de seleccionar las funcionalidades que necesita para cumplir los objetivos empresariales y los objetivos de tiempo de actividad.

En este artículo se describe cómo hacer que Application Gateway v2 sea resistente a una variedad de posibles interrupciones y problemas, incluidos errores transitorios, interrupciones de zona de disponibilidad y interrupciones en regiones. También se resalta cierta información clave sobre el acuerdo de nivel de servicio (SLA) de Application Gateway v2.

> **ⓘ Importante**
>
> La confiabilidad de la solución general depende de la configuración de los servidores back-end a los que Application Gateway enruta el tráfico. En función de la solución, pueden ser máquinas virtuales (VM) de Azure, Microsoft Azure Virtual Machines de Azure, Azure App Services o puntos de conexión externos.
>
> Los servidores back-end no están en el ámbito de este artículo, pero sus configuraciones de disponibilidad afectan directamente a la resistencia de la aplicación. Revise las guías de confiabilidad de todos los servicios de Azure de la solución para comprender cómo admite cada uno los requisitos de confiabilidad. Al asegurarse de que los servidores de back-end también están configurados para alta disponibilidad y redundancia de zona, puede lograr una confiabilidad de un extremo a otro para la aplicación.

Recomendaciones de implementación de producción

Para obtener información sobre cómo implementar Application Gateway para admitir los requisitos de confiabilidad de la solución y cómo afecta la confiabilidad a otros aspectos de la

arquitectura, consulte Procedimientos recomendados de arquitectura para Application Gateway en Azure Well-Architected Framework.

Introducción a la arquitectura de confiabilidad

Application Gateway es un servicio administrado. Es importante comprender algunos elementos clave de la arquitectura de servicio para que pueda tomar decisiones de confiabilidad informadas.

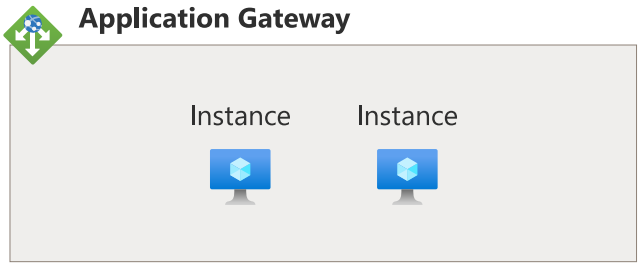
- **Puerta de Enlace:** Al implementar Application Gateway, se implementa una *puerta de enlace*, que es el recurso de Azure con el que trabaja.

- **Instancias:** Una *instancia* es una sola unidad de la puerta de enlace. Una puerta de enlace tiene varias instancias. Cada instancia tiene su propia dirección IP privada, controla el tráfico y enruta las solicitudes.

  > **Importante**
  >
  > Para lograr una alta disponibilidad, cada puerta de enlace siempre se crea con un mínimo de dos instancias. Sin embargo, Azure Portal puede indicar que la puerta de enlace tiene una sola instancia, aunque realmente tenga dos.

  La plataforma administra automáticamente la creación de instancias, la supervisión del estado y la sustitución de instancias incorrectas. También administra la eliminación correcta de instancias durante los eventos de escalado horizontal. Este proceso se conoce como *drenaje de conexiones*.

  En el diagrama siguiente se muestra una puerta de enlace con dos instancias:



Para distribuir instancias entre varias zonas de disponibilidad y, por tanto, aumentar la redundancia y la disponibilidad durante los errores del centro de datos, puede habilitar la redundancia de zona.

- **Escalado:** un aspecto importante de la confiabilidad de una puerta de enlace es cómo se escala para satisfacer la demanda de tráfico. Si una puerta de enlace no se escala

correctamente, el tráfico no puede fluir y la aplicación puede experimentar tiempo de inactividad. Puede configurar Application Gateway para que use uno de los siguientes modos de escalado.

  - *Escalado automático:* ajusta automáticamente el número de instancias dentro de un intervalo que especifique. El escalado automático escala el número de instancias en función de la demanda de tráfico actual.
  - *Escalado manual:* requiere que especifique un número exacto de instancias. Es responsable de seleccionar un recuento de instancias que cumpla la demanda de tráfico.

  Una *unidad de capacidad* representa una cantidad de capacidad que la puerta de enlace puede procesar. Una unidad de capacidad es una medida sintética que incorpora el tráfico, el número de conexiones y los recursos de proceso. Cada instancia puede controlar al menos 10 unidades de capacidad. Para más información, vea Escalado de Application Gateway v2 y WAF v2.

- **Sondeos de estado:** Application Gateway usa sondeos de estado para supervisar continuamente los servidores back-end, como los servidores de aplicaciones individuales. El tráfico se puede redirigir automáticamente a servidores de back-end correctos cuando se detectan servidores incorrectos.

Resistencia a errores transitorios

Los errores transitorios son errores breves e intermitentes en los componentes. Se producen con frecuencia en un entorno distribuido como la nube y son una parte normal de las operaciones. Los errores transitorios se corrigen después de un breve período de tiempo. Es importante que las aplicaciones puedan controlar errores transitorios, normalmente mediante el reintento de solicitudes afectadas.

Todas las aplicaciones hospedadas en la nube deben seguir las instrucciones de control de errores transitorios de Azure cuando se comunican con cualquier API, bases de datos y otros componentes hospedados en la nube. Para obtener más información, consulte Recomendaciones para controlar errores transitorios.

Al usar Application Gateway, tenga en cuenta los procedimientos recomendados siguientes:

- **Implementación de la lógica de reintento.** Los clientes deben implementar mecanismos de reintento adecuados para errores de conexión transitorios.
- **Configure sondeos de estado con tolerancia.** Configure los sondeos de estado a fin de permitir un período de gracia para errores transitorios. Los sondeos de estado se pueden

configurar con un *umbral incorrecto*, que especifica el número de intentos de conexión con error consecutivos que deben desencadenar que el servidor de back-end se marque como incorrecto. El valor predeterminado de tres garantiza que los errores transitorios de los servidores back-end no desencadenen Application Gateway para marcar innecesariamente el servidor como incorrecto.

Resistencia a errores de zona de disponibilidad

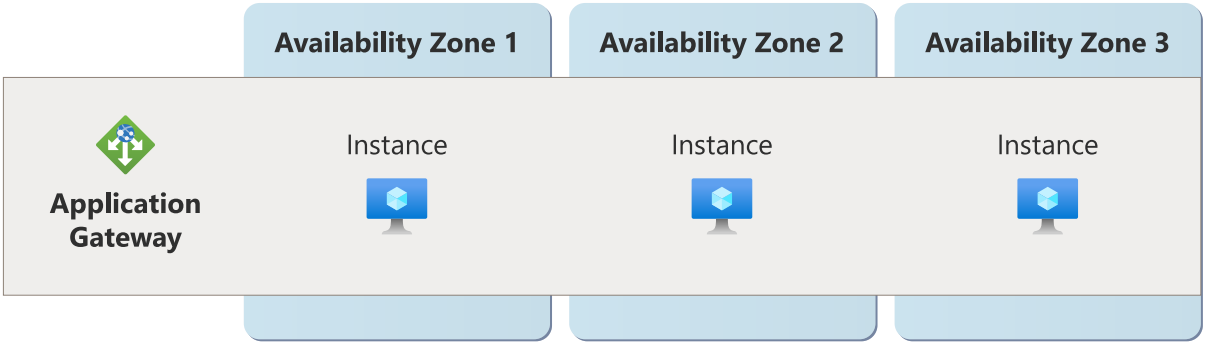
Las zonas de disponibilidad son grupos físicamente independientes de centros de datos dentro de una región de Azure. Cuando una zona falla, los servicios pueden transferirse a una de las zonas restantes.

Application Gateway proporciona dos tipos de compatibilidad con zonas de disponibilidad al implementar una puerta de enlace de Standard_v2 o WAF_v2 en una región admitida.

- *Con redundancia de zona:* Azure distribuye automáticamente las instancias en dos o más zonas de disponibilidad.

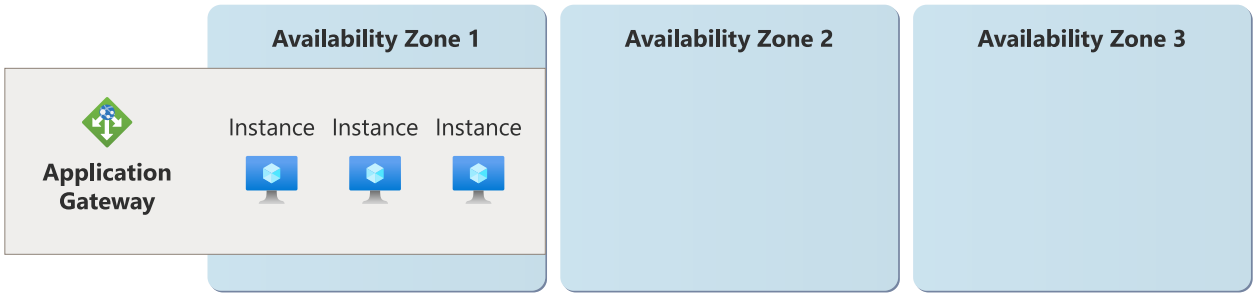
  Al implementar un nuevo gateway, es redundante en zona por defecto.

  En el diagrama siguiente se muestra una puerta de enlace con redundancia de zona con tres instancias distribuidas entre tres zonas de disponibilidad:



- *Zonal:* Azure implementa todas las instancias de Application Gateway en una sola zona que seleccione en la región de Azure elegida.

  En el diagrama siguiente se muestra una puerta de enlace zonal con tres instancias que se implementan en la misma zona de disponibilidad:



> ⓘ **Importante**
>
> La asignación a una sola zona de disponibilidad solo se recomienda cuando la latencia entre zonas es demasiado alta para sus necesidades y cuando se comprueba que la latencia no cumple sus requisitos. Por sí sola, una puerta de enlace zonal no proporciona resiliencia ante una interrupción del servicio en una zona de disponibilidad. Para mejorar la resistencia de una implementación de Application Gateway zonal, debe implementar explícitamente puertas de enlace independientes en varias zonas de disponibilidad y configurar el enrutamiento de tráfico y la conmutación por error.

Requisitos

- **Compatibilidad con regiones:** Application Gateway admite zonas de disponibilidad para los niveles de Standard_v2 y WAF_v2 en todas las regiones de Azure que admiten zonas de disponibilidad.
- **SKU:** Debe usar la Standard_v2 o WAF_v2 SKU para habilitar la compatibilidad con zonas de disponibilidad.
- **Recuento de instancias:** La puerta de enlace debe tener al menos dos instancias.

> ⓘ **Nota:**
>
> Todas las puertas de enlace tienen un mínimo de dos instancias para garantizar alta disponibilidad. Incluso si Azure Portal indica que la puerta de enlace tiene una sola instancia, internamente siempre se crea con un mínimo de dos instancias.

Consideraciones

Las puertas de enlace con redundancia de zona se distribuyen entre dos o más zonas de disponibilidad en la región. Por ejemplo, en una región que tiene tres zonas de disponibilidad, una implementación de Application Gateway v2 con redundancia de zona tiene instancias en al

menos dos de esas zonas. En función de la capacidad regional y las decisiones de plataforma, las instancias se pueden distribuir entre dos zonas o tres zonas.

Cost

La compatibilidad con la zona de disponibilidad para Application Gateway v2 no incurre en cargos adicionales más allá de los precios de la unidad de capacidad estándar. Para obtener más información sobre los precios, consulte Descripción de los precios de Application Gateway y WAF y Application Gateway.

Configurar soporte de zonas de disponibilidad

En esta sección se explica cómo configurar la compatibilidad con zonas de disponibilidad para las puertas de enlace.

- **Cree una nueva puerta de enlace con compatibilidad con la zona de disponibilidad.** El enfoque que se usa para configurar zonas de disponibilidad depende de si desea crear una puerta de enlace con redundancia de zona o zonal.
  - *Con redundancia de zona:* las puertas de enlace se crean como con redundancia de zona de manera predeterminada. Las instancias se distribuyen entre varias zonas de disponibilidad y pueden usar dos o más zonas en la región.

    Para implementar una nueva puerta de enlace, consulte Inicio rápido: tráfico web directo mediante Application Gateway: Azure Portal.

    Si usa la CLI de Azure, Azure PowerShell, Bicep, plantillas de Azure Resource Manager (plantillas de ARM) o Terraform, puede especificar opcionalmente las zonas de disponibilidad en las que implementar la puerta de enlace. Puede implementar una puerta de enlace con redundancia de zona si especifica dos o más zonas. Sin embargo, se recomienda omitir la lista de zonas para que la puerta de enlace pueda usar todas las zonas de disponibilidad, a menos que tenga una razón específica para no usar una zona específica.
  - *Zonal:* Puede implementar una puerta de enlace zonal mediante las siguientes herramientas.
    - *CLI de Azure:* debe seleccionar explícitamente zonas mediante el parámetro `--zones` en el comando `az network application-gateway create`. Para anclar la puerta de enlace a una sola zona, especifique el número de zona lógica.
    - *Azure PowerShell:* use el parámetro `-Zone` en el comando `New-AzApplicationGateway`. Para anclar la puerta de enlace a una sola zona, especifique el

número de zona lógica.

- *Plantillas de Bicep y ARM:* Configure la `zones` propiedad en la definición de recursos. Para anclar la puerta de enlace a una sola zona, especifique el número de zona lógica.

> ⓘ **Nota:**
>
> Cuando selecciona qué zonas de disponibilidad usar, en realidad está seleccionando la *zona de disponibilidad lógica*. Si implementa otros componentes de la carga de trabajo en una suscripción de Azure diferente, podrían usar un número de zona de disponibilidad lógica distinto para acceder a la misma zona de disponibilidad física. Para obtener más información, consulte zonas de disponibilidad física y lógica.

- **Cambie la configuración de zona de disponibilidad de una instancia de Application Gateway v2 existente.** Todas las puertas de enlace tienen redundancia de zona a menos que se configuren para que sean zonales.

  Si necesita pasar de una puerta de enlace con redundancia de zona a una configuración zonal, debe implementar una nueva puerta de enlace.

- **Deshabilite la compatibilidad con la zona de disponibilidad.** No se puede deshabilitar la compatibilidad con la zona de disponibilidad. Todas las puertas de enlace de regiones que admiten zonas de disponibilidad deben ser con redundancia de zona o zonales.

Planeamiento y administración de capacidad

Cuando planifique fallos de zona para una puerta de enlace de aplicaciones redundante por zonas o varias puertas de enlace zonales implementadas en varias zonas, tenga en cuenta que las instancias de las zonas supervivientes pueden experimentar un aumento de la carga a medida que se redistribuye el tráfico. Las conexiones pueden experimentar interrupciones breves que pueden durar unos segundos.

Para administrar la capacidad de forma eficaz, realice las siguientes acciones:

- **Use el escalado automático.** Configure el escalado automático con los recuentos máximos de instancias adecuados para controlar la posible redistribución del tráfico durante las interrupciones de zona.

  Si usa el escalado manual, a fin de prepararse para el error de zona de disponibilidad, considere la posibilidad de *sobreaprovisionar* el número de instancias de la puerta de

enlace. El aprovisionamiento excesivo permite a la solución tolerar cierto grado de pérdida de capacidad, al tiempo que sigue funcionando sin un rendimiento degradado. Para más información, consulte Administración de la capacidad con sobreaprovisionamiento.

- **Responder a los cambios en los patrones de tráfico.** Supervise las métricas de capacidad y ajuste los parámetros de escalado en función de los patrones de tráfico y los requisitos de rendimiento.

Comportamiento cuando todas las zonas están en buen estado

En la siguiente sección se describe qué se puede esperar cuando Application Gateway v2 se configura con compatibilidad con zonas de disponibilidad y todas las zonas de disponibilidad están operativas.

- **Enrutamiento de tráfico entre zonas**: Application Gateway distribuye automáticamente las solicitudes entrantes entre instancias de todas las zonas que usa la puerta de enlace. Esta configuración activa-activa garantiza un rendimiento y una distribución de carga óptimos en condiciones de funcionamiento normales.

- **Administración de instancias:** la plataforma administra automáticamente la colocación de instancias en las zonas que usa la puerta de enlace. Reemplaza las instancias con errores y mantiene el recuento de instancias configurados. El seguimiento de estado garantiza que solo las instancias correctas reciban tráfico.

Comportamiento durante un fallo de zona

En la siguiente sección se describe qué ocurre cuando Application Gateway v2 se configura con compatibilidad con zonas de disponibilidad y una o varias zonas de disponibilidad no están disponibles.

- **Detección y respuesta:** la responsabilidad de la detección y la respuesta depende de la configuración de zona de disponibilidad que usa la puerta de enlace.
  - *Con redundancia de zona:* Microsoft administra la detección de errores de zona e inicia automáticamente la conmutación por error. No se requiere ninguna acción del cliente.
  - *Zonal:* debe detectar la pérdida de una zona de disponibilidad e iniciar una conmutación por error a una puerta de enlace secundaria que cree en otra zona de disponibilidad.

- **Notificación**: Microsoft no le notifica automáticamente cuando una zona está inactiva. Sin embargo, puede usar Azure Resource Health para supervisar el estado de un recurso individual y puede configurar alertas de Resource Health para notificarle problemas. También puede usar Azure Service Health para comprender el estado general del servicio, incluidos los errores de zona, y puede configurar alertas de Service Health para notificarle problemas.

- **Solicitudes activas:** durante una interrupción de zona, se finalizan las solicitudes que procesan las instancias de esa zona. Los clientes deben reintentar las solicitudes siguiendo las instrucciones para controlar errores transitorios.

- **Pérdida de datos esperada:** no se espera que los errores de zona causen pérdida de datos porque Application Gateway es un servicio sin estado.

- **Tiempo de inactividad esperado:** El tiempo de inactividad que puede esperar depende de la configuración de zona de disponibilidad que usa la puerta de enlace.

  - *Con redundancia de zona:* Durante las interrupciones de zona, las conexiones pueden experimentar interrupciones breves que suelen durar unos segundos a medida que se redistribuye el tráfico.

  - *Zonal:* cuando una zona no está disponible, la puerta de enlace no está disponible hasta que se recupere la zona de disponibilidad.

- **Administración de instancias:** el comportamiento de administración de instancias depende de la configuración de zona de disponibilidad que usa la puerta de enlace.

  - *Con redundancia de zona:* la plataforma intenta mantener la capacidad de la puerta de enlace mediante la creación de instancias temporales en otras zonas de disponibilidad.

    Internamente, Application Gateway usa conjuntos de escalado de máquinas virtuales, que realizan el equilibrio de zona más indicado. Debido a este comportamiento, es posible que las operaciones de escalado no se produzcan cuando la capacidad no se pueda dividir uniformemente entre zonas (+/- 1 instancia).

  - *Zonal:* es responsable de crear instancias en zonas correctas si las necesita.

- **Re enrutamiento del tráfico:** El comportamiento se re direcciona del tráfico depende de la configuración de zona de disponibilidad que usa la puerta de enlace.

  - *Con redundancia de zona:* Application Gateway redistribuye inmediatamente el tráfico a las instancias de zonas correctas, incluidas las que se crean temporalmente.

  - *Zonal:* cuando una zona no está disponible, la puerta de enlace no está disponible. Si tiene una puerta de enlace secundaria en otra zona de disponibilidad, es responsable

de redirigir el tráfico a esa puerta de enlace secundaria.

Recuperación de zona

El comportamiento de recuperación de zona depende de la configuración de zona de disponibilidad que usa la puerta de enlace:

- *Con redundancia de zona:* Cuando se recupera la zona de disponibilidad afectada, Application Gateway realiza automáticamente las siguientes acciones:
  - Restaura instancias en la zona recuperada
  - Quita las instancias temporales que se hayan creado en otras zonas durante la interrupción
  - Vuelve a la distribución normal del tráfico en todas las zonas disponibles
- *Zonal:* es responsable de redirigir el tráfico a la puerta de enlace en la zona de disponibilidad original después de que se recupere la zona de disponibilidad.

Prueba de fallos de zona

Las opciones para probar errores de zona dependen de la configuración de zona de disponibilidad que usa la puerta de enlace.

- *Con redundancia de zona:* La plataforma Application Gateway administra completamente el enrutamiento del tráfico, la conmutación por error y la conmutación por recuperación para las puertas de enlace con redundancia de zona. Como Microsoft administra esta característica, no es necesario iniciar ni validar los procesos de error de zona de disponibilidad. La plataforma controla todos los escenarios de error de zona de forma transparente.
- *Zonal:* puede simular algunos aspectos del error de una zona de disponibilidad si detiene explícitamente una puerta de enlace. Al detener la puerta de enlace, puede probar cómo otros sistemas y equilibradores de carga controlan una interrupción en la puerta de enlace. Para más información, consulte Cómo detener e iniciar Application Gateway.

Resistencia a errores en toda la región

Application Gateway v2 es un servicio de una sola región. Si la región deja de estar disponible, la puerta de enlace tampoco está disponible.

Soluciones personalizadas de varias regiones para la resistencia

Para lograr la resistencia de varias regiones mediante Application Gateway v2, debe implementar puertas de enlace independientes en cada región deseada e implementar la administración del tráfico entre regiones. Es responsable de implementar y configurar cada una de las puertas de enlace, así como el enrutamiento del tráfico y la conmutación por error. Considere los siguientes puntos:

- Configure reglas y directivas coherentes de Application Gateway entre regiones. Puede definir la infraestructura como código (IaC) mediante herramientas como Bicep o Terraform para simplificar las implementaciones y configuraciones entre regiones.
- Implemente una solución de equilibrio de carga global que pueda enviar tráfico entre las puertas de enlace regionales. Los servicios globales de equilibrio de carga en Azure son Azure Traffic Manager y Azure Front Door. Cada servicio enruta el tráfico en función de las comprobaciones de estado, la proximidad geográfica o las métricas de rendimiento. Azure Front Door también proporciona una variedad de otras funcionalidades, incluida la protección contra ataques por denegación de servicio distribuido (DDoS), las funcionalidades de WAF y las reglas avanzadas y las características de enrutamiento.
- Más allá de la puerta de enlace, considere la posibilidad de replicar aplicaciones de back-end y datos entre regiones. Consulte las guías de confiabilidad de cada servicio de Azure para comprender los enfoques de implementación de varias regiones.

Para obtener un enfoque de ejemplo, consulte Uso de Application Gateway con Traffic Manager.

Copias de seguridad y restauración

Application Gateway v2 es un servicio sin estado que no requiere operaciones tradicionales de copia de seguridad y restauración. Todos los datos de configuración se almacenan en Resource Manager y se pueden volver a implementar mediante enfoques IaC, como archivos de Bicep o plantillas de ARM.

Para la administración de la configuración y la recuperación ante desastres, debe realizar las siguientes acciones:

- Defina la configuración de la implementación de Application Gateway mediante archivos de Bicep o plantillas de ARM, o bien exporte la configuración de una puerta de enlace existente.

- Almacene certificados de seguridad de la capa de transporte (TLS) en Azure Key Vault para la administración y replicación seguras.
- Directivas, reglas y configuraciones personalizadas de control de versiones y documentos.
- Implemente canalizaciones de implementación automatizadas para el aprovisionamiento coherente de la puerta de enlace.

Para la mayoría de las soluciones, no debe confiar exclusivamente en copias de seguridad. En su lugar, utilice las otras capacidades descritas en esta guía para apoyar los requisitos de resiliencia. Sin embargo, las copias de seguridad protegen contra algunos riesgos que otros enfoques no. Para más información, consulte ¿Qué son la redundancia, la replicación y la copia de seguridad?.

Resistencia al mantenimiento del servicio

Application Gateway v2 realiza actualizaciones de servicio periódicas y otras tareas de mantenimiento. Para mantener la capacidad esperada durante una actualización, la plataforma agrega automáticamente instancias adicionales de la puerta de enlace durante el proceso de actualización. Sin embargo, debe asegurarse de que la subred de la puerta de enlace tenga suficiente espacio de direcciones IP libre para que se creen las instancias temporales. Para obtener más información, consulte ¿Cómo controla Application Gateway el mantenimiento rutinario?.

Acuerdo de nivel de servicio

El contrato de nivel de servicio (SLA) para los servicios de Azure describe la disponibilidad esperada de cada servicio y las condiciones que la solución deberá cumplir para lograr esa expectativa de disponibilidad. Para obtener más información, consulte Acuerdos de Nivel de Servicio para servicios en línea.

Para que una puerta de enlace sea apta para el Acuerdo de Nivel de Servicio de disponibilidad de Application Gateway, debe cumplir los criterios siguientes:

- Debe configurarse correctamente para realizar el equilibrio de carga HTTP.
- Debe tener redundancia de zona o debe configurarse para usar la escalabilidad automática.

Contenido relacionado

- Introducción a la confiabilidad de Azure

ⓘ **Nota:** El autor ha creado este artículo con ayuda de inteligencia artificial. Más información

---

Last updated on 22/01/2026

Confiabilidad en Azure Blob Storage

Azure Blob Storage es una solución de almacenamiento de objetos para la nube desde Microsoft. Se diseñó para almacenar grandes cantidades de datos no estructurados, como texto, datos binarios, documentos, archivos multimedia y copias de seguridad de aplicaciones. Como servicio básico de Azure Storage, Blob Storage proporciona varias características de confiabilidad para asegurarse de que los datos permanezcan disponibles y duren durante eventos planeados y no planeados.

Cuando se usa Azure, la confiabilidad es una responsabilidad compartida. Microsoft proporciona una variedad de funcionalidades para admitir resistencia y recuperación. Es responsable de comprender cómo funcionan esas funcionalidades dentro de todos los servicios que usa y de seleccionar las funcionalidades que necesita para cumplir los objetivos empresariales y los objetivos de tiempo de actividad.

En este artículo se describe cómo hacer que Blob Storage sea resistente a una variedad de posibles interrupciones y problemas, incluidos errores transitorios, interrupciones de zona de disponibilidad y interrupciones de regiones. También se describe cómo puede usar copias de seguridad para recuperarse de otros tipos de problemas y se resalta cierta información clave sobre el acuerdo de nivel de servicio (SLA) de Blob Storage.

> ⓘ **Nota:**
>
> Blob Storage forma parte de la plataforma Azure Storage. Algunas de las capacidades de Blob Storage son comunes en muchos servicios de Azure Storage. En este artículo, se usa *Azure Storage* para hacer referencia a estas características.

Recomendaciones de implementación de producción

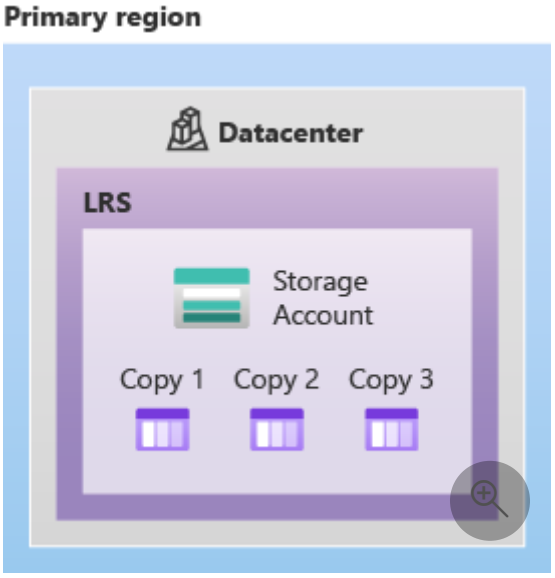
Para obtener información sobre cómo implementar Blob Storage para admitir los requisitos de confiabilidad de la solución y cómo afecta la confiabilidad a otros aspectos de la arquitectura, consulte Procedimientos recomendados de arquitectura para Blob Storage en el marco de arquitectura recomendada de Azure.

Introducción a la arquitectura de confiabilidad

Azure Storage proporciona varias opciones de redundancia para ayudarle a proteger los datos frente a distintos tipos de errores. Cada opción proporciona un nivel específico de redundancia

de datos, por lo que es posible elegir el nivel que mejor se adapte a los requisitos de la aplicación.

El almacenamiento con redundancia local (LRS) replica los datos de las cuentas de almacenamiento en una o varias zonas de disponibilidad de Azure ubicadas en la región primaria de su elección. Aunque no hay ninguna opción para elegir la zona de disponibilidad preferida, Azure puede mover o expandir cuentas LRS entre zonas para mejorar el equilibrio de carga. No hay ninguna garantía de que los datos se distribuirán entre zonas. Para obtener más información sobre las zonas de disponibilidad, consulte ¿Qué son las zonas de disponibilidad?.



El almacenamiento con redundancia de zona (ZRS), el almacenamiento con redundancia geográfica (GRS) y el almacenamiento con redundancia de zona geográfica (GZRS) proporcionan protecciones adicionales. En este artículo se describen estas opciones al detalle.

Resistencia a errores transitorios

Los errores transitorios son errores breves e intermitentes en los componentes. Se producen con frecuencia en un entorno distribuido como la nube y son una parte normal de las operaciones. Los errores transitorios se corrigen después de un breve período de tiempo. Es importante que las aplicaciones puedan controlar errores transitorios, normalmente mediante el reintento de solicitudes afectadas.

Todas las aplicaciones hospedadas en la nube deben seguir las instrucciones de control de errores transitorios de Azure cuando se comunican con cualquier API, bases de datos y otros componentes hospedados en la nube. Para obtener más información, consulte Recomendaciones para controlar errores transitorios.

Para administrar de forma eficaz los errores transitorios al usar Blob Storage, implemente las siguientes recomendaciones:

- **Utilice las bibliotecas cliente de Azure Storage**, que incluyen directivas de reintento integradas con retroceso exponencial y fluctuación. Los SDK de .NET, Java, Python y JavaScript controlan automáticamente los reintentos en caso de fallos transitorios. Para obtener más información sobre las opciones de configuración de reintentos, consulte Implementación de directivas de reintentos con .NET.

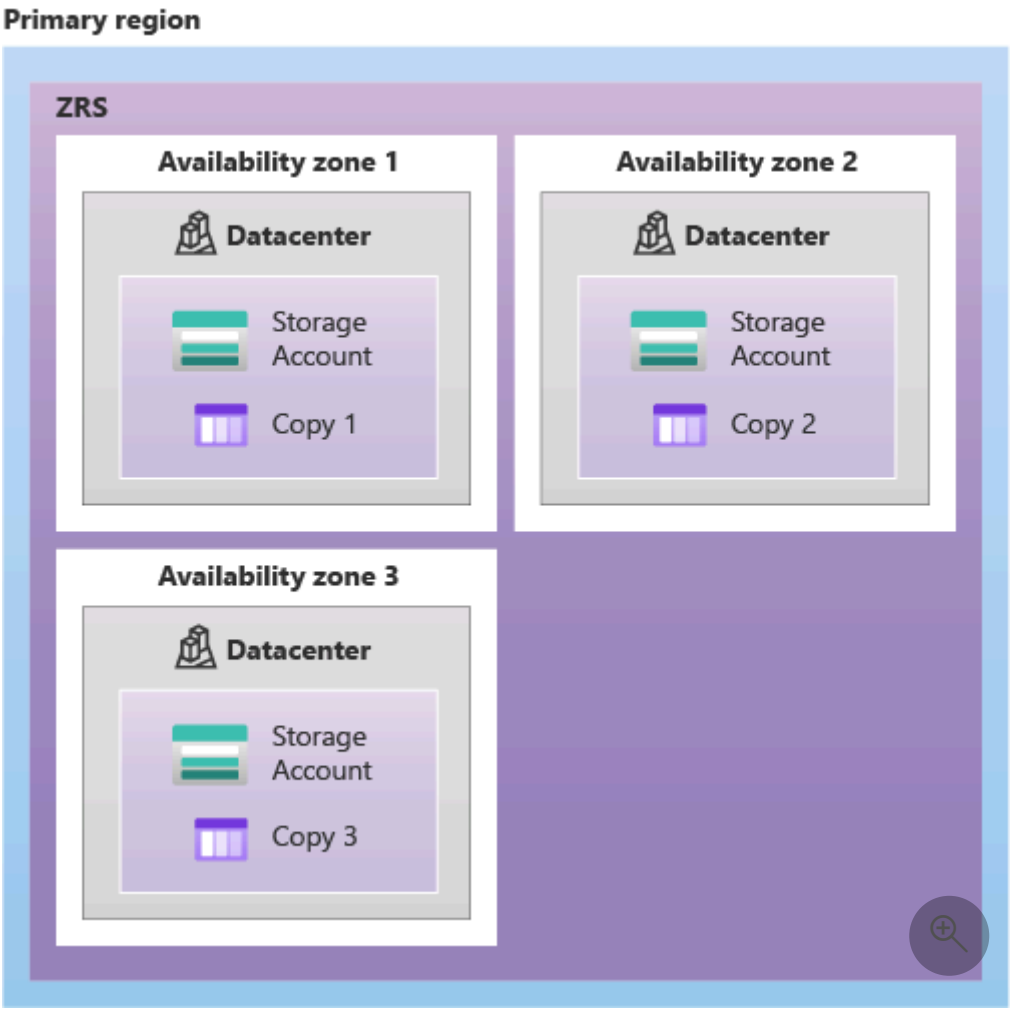
- **Configure los valores de tiempo de espera adecuados** para sus operaciones con blobs en función del tamaño de los blobs y las condiciones de la red. Los blobs más grandes requieren tiempos de espera más largos, pero las operaciones más pequeñas pueden usar valores más cortos para detectar errores rápidamente.

Resistencia a errores de zona de disponibilidad

Las zonas de disponibilidad son grupos físicamente independientes de centros de datos dentro de una región de Azure. Cuando una zona falla, los servicios pueden transferirse a una de las zonas restantes.

Blob Storage proporciona una sólida compatibilidad con zonas de disponibilidad a través de configuraciones de ZRS que distribuyen automáticamente los datos entre varias zonas de disponibilidad dentro de una región. A diferencia del almacenamiento con redundancia local (LRS), ZRS garantiza que Azure replique de forma sincrónica los datos del blob en varias zonas de disponibilidad. ZRS garantiza que sus datos sigan estando accesibles incluso si se produce un corte en una zona.

La redundancia de zona se habilita en el nivel de la cuenta de almacenamiento y se aplica a todos los contenedores de blobs dentro de esa cuenta. No se pueden establecer distintos niveles de redundancia para contenedores individuales. La configuración de redundancia se aplica a toda la cuenta de almacenamiento. Cuando una zona de disponibilidad experimente una interrupción, Azure Storage enrutará automáticamente las solicitudes a zonas correctas sin necesidad de intervención de usted o de la aplicación.



Requisitos

- **Compatibilidad con regiones:** Puede implementar cuentas de Azure Storage con redundancia de zona en cualquier región que admita zonas de disponibilidad.
- **Tipos de cuenta de almacenamiento:** La redundancia de zona está disponible tanto para los tipos de cuentas de almacenamiento v2 de uso general estándar como para las cuentas de almacenamiento Blob en bloques Premium. Los blobs en bloques, los blobs en anexos y los blobs en páginas admiten configuraciones con redundancia de zona, pero el tipo de cuenta de almacenamiento que se use determinará qué funcionalidades estarán disponibles. Para obtener más información, consulte Tipos de cuentas de almacenamiento compatibles.

Cost

Cuando se habilita el almacenamiento con redundancia de zona (ZRS), se cobra una tarifa diferente a la del almacenamiento redundante local (LRS) debido a la replicación adicional y los gastos generales de almacenamiento.

Para más información, consulte Precios de Blob Storage.

Configurar soporte de zonas de disponibilidad

- **Crear una cuenta de Blob Storage con redundancia de zona.** Para crear una cuenta de almacenamiento con ZRS, consulte Creación de cuentas de almacenamiento y seleccione **ZRS**, **almacenamiento con redundancia de zona geográfica (GZRS)** o **almacenamiento con redundancia geográfica con acceso de lectura (RA-GZRS)** como opción de redundancia durante la creación de la cuenta.

- **Cambiar el tipo de replicación.** Para obtener información sobre cómo cambiar una cuenta de almacenamiento existente a almacenamiento con redundancia de zona (ZRS) y sobre las opciones y requisitos de configuración, consulte Cambiar la forma en que se replica una cuenta de almacenamiento.

- **Desactivar la redundancia de zona.** Convierta las cuentas de ZRS de nuevo a una configuración no zonal, como el almacenamiento con redundancia local (LRS), utilizando el mismo proceso de cambio de configuración de redundancia.

Comportamiento cuando todas las zonas están en buen estado

En esta sección se describe qué se puede esperar cuando se configura una cuenta de almacenamiento de blobs para la redundancia de zona y todas las zonas de disponibilidad están operativas.

- **Enrutamiento del tráfico entre zonas:** Azure Storage con almacenamiento con redundancia de zona (ZRS) distribuye automáticamente las solicitudes entre los clústeres de almacenamiento en varias zonas de disponibilidad. La distribución del tráfico es transparente para las aplicaciones y no requiere ninguna configuración del lado cliente.

- **Replicación de datos entre zonas:** todas las operaciones de escritura en ZRS se replican de forma sincrónica en todas las zonas de disponibilidad dentro de la región. Al cargar o modificar datos, la operación no se considera completa hasta que los datos se hayan replicado correctamente en todas las zonas de disponibilidad. Esta replicación sincrónica garantiza una fuerte coherencia y cero pérdida de datos durante los errores de zona.

Comportamiento durante un fallo de zona

En esta sección se describe qué puede ocurrir cuando se configura una cuenta de almacenamiento de blobs para ZRS y se produce una interrupción en la zona de disponibilidad.

- **Detección y respuesta:** Microsoft detecta automáticamente errores de zona e inicia procesos de recuperación. No se requiere ninguna acción del cliente para las cuentas de

almacenamiento con redundancia de zona (ZRS). Si una zona deja de estar disponible, Azure lleva a cabo actualizaciones de red como el repunto del sistema de nombres de dominio (DNS).

- **Notificación**: Microsoft no le notifica automáticamente cuando una zona está inactiva. Sin embargo, puede usar Azure Resource Health para supervisar el estado de un recurso individual y puede configurar alertas de Resource Health para notificarle problemas. También puede usar Azure Service Health para comprender el estado general del servicio, incluidos los errores de zona, y puede configurar alertas de Service Health para notificarle problemas.
- **Solicitudes activas:** Es posible que las solicitudes en curso se descarten durante el proceso de recuperación y se deban reintentar. Las aplicaciones deben implementar lógica de reintento para controlar estas interrupciones temporales.
- **Pérdida de datos esperada:** No se produce ninguna pérdida de datos durante los errores de zona porque los datos se replican sincrónicamente en varias zonas antes de completar las operaciones de escritura.
- **Tiempo de inactividad previsto:** es posible que se produzca un breve periodo de inactividad, normalmente de unos segundos, durante la recuperación automática, ya que el tráfico se redirige a zonas correctas. Cuando diseñe aplicaciones para ZRS, siga las prácticas de manejo de fallos transitorios, incluyendo la implementación de directivas de reintento con retroceso exponencial.
- **Reenrutamiento del tráfico:** si una zona de disponibilidad se quedase sin conexión, Azure iniciará los cambios de red, como la reconfiguración del punto de conexión del sistema de nombres de dominio (DNS). Estas actualizaciones garantizan que el tráfico se redireccione a las zonas de disponibilidad correctas restantes. El servicio mantiene toda la funcionalidad mediante las zonas supervivientes y no requiere intervención del cliente.

Recuperación de zona

Cuando se recupera la zona de disponibilidad con errores, Azure Storage restaura automáticamente las operaciones normales en todas las zonas de disponibilidad. El servicio garantiza automáticamente la coherencia de los datos mediante la sincronización de las operaciones que se produjeron durante el período de interrupción.

Prueba de fallos de zona

El usar el almacenamiento con redundancia de zona (ZRS), Azure Storage administrará automáticamente la replicación, el enrutamiento del tráfico y las respuestas ante caídas de

zonas. Dado que esta característica está totalmente administrada, no es necesario iniciar ni validar los procesos de error de zona de disponibilidad.

Resistencia a errores en toda la región

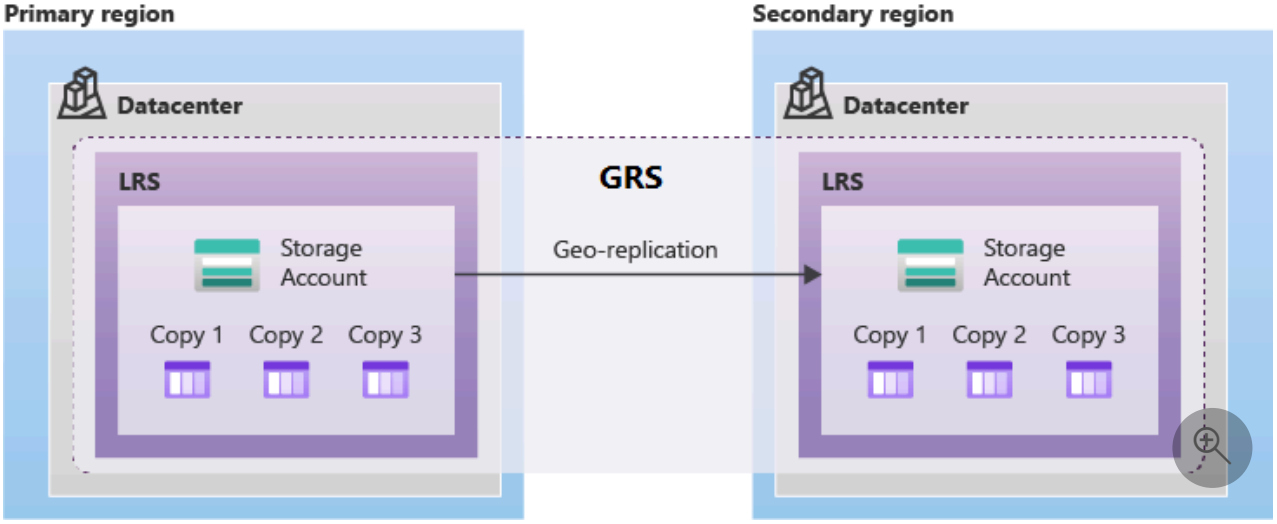
Azure Storage, que incluye Azure Blob Storage, Azure Files, Azure Table Storage y Azure Queue Storage, ofrece una amplia gama de capacidades de redundancia geográfica y conmutación por error para adaptarse a diferentes requisitos.

> ⓘ **Importante**
>
> El GRS solo funciona dentro de regiones emparejadas de Azure. Si la región de la cuenta de almacenamiento no está emparejada, considere la posibilidad de usar las soluciones personalizadas de varias regiones para lograr resistencia.

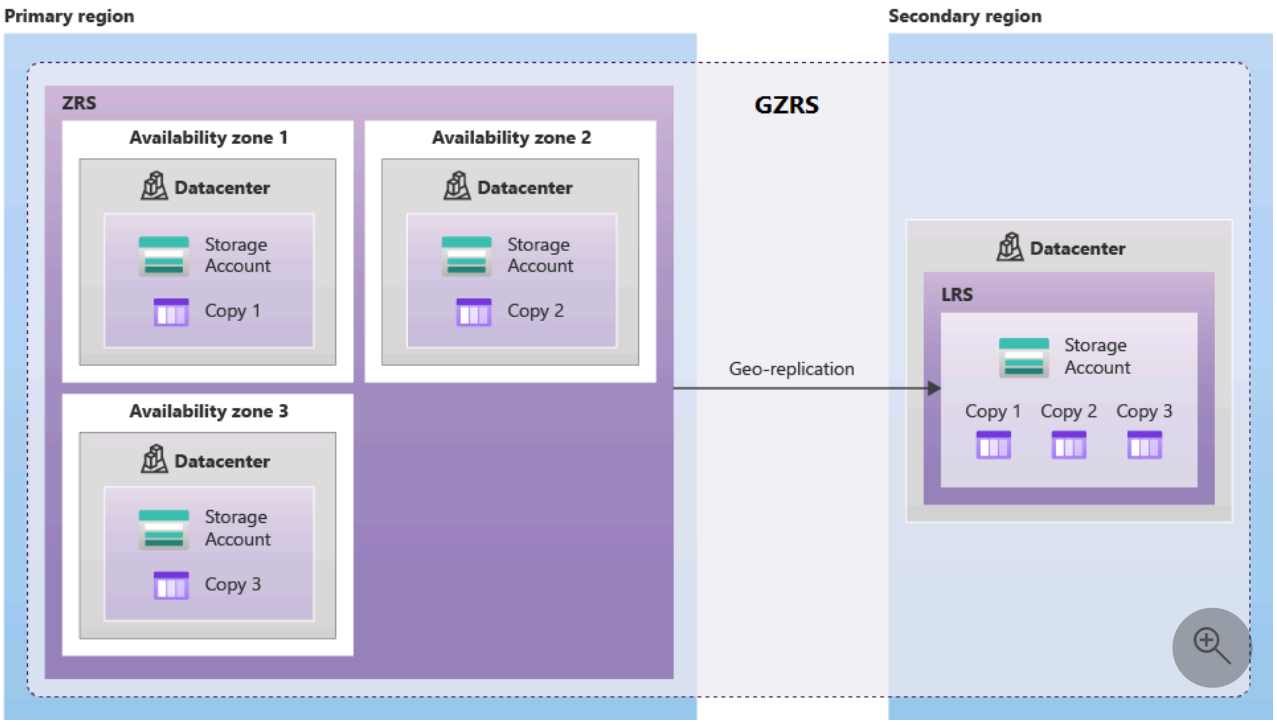
Almacenamiento con redundancia geográfica para regiones emparejadas

Azure Storage ofrece varios tipos de GRS en regiones emparejadas. Independientemente del tipo de GRS que utilice, los datos de la región secundaria siempre se replicarán mediante almacenamiento con redundancia local (LRS). Este enfoque proporciona protección contra fallos de hardware dentro de la región secundaria.

- GRS proporciona soporte para conmutaciones por error planificadas y no planificadas a la región emparejada de Azure cuando se produce una interrupción en la región principal. GRS replica de forma asincrónica los datos de la región primaria a la región emparejada.



- El almacenamiento con redundancia de zona geográfica (GZRS) replica los datos en varias zonas de disponibilidad de la región principal y en la región emparejada.



- El almacenamiento con redundancia geográfica con acceso de lectura (RA-GRS) y el almacenamiento con redundancia de zona geográfica con acceso de lectura (RA-GZRS) amplían el almacenamiento con redundancia geográfica (GRS) y el almacenamiento con redundancia de zona geográfica (GZRS), con la ventaja adicional del acceso de lectura al punto de conexión secundario. Estas opciones son idóneas y están diseñadas para obtener alta disponibilidad en aplicaciones críticas para la empresa. En el improbable caso de que el punto de conexión principal experimente una interrupción, las aplicaciones configuradas para el acceso de lectura a la región secundaria pueden seguir funcionando.

Tipos de conmutación por error

Azure Storage admite tres tipos de conmutación por error para distintos escenarios.

- **Conmutación por error no planeada administrada por el cliente:** es responsable de iniciar la recuperación si hubiera un error de almacenamiento en toda la región principal.

- **Conmutación por error planeada administrada por el cliente:** Usted es responsable de iniciar la recuperación si otra parte de su solución sufre un fallo en la región primaria y necesita trasladar toda su solución a la región secundaria. Use una conmutación por error planeada cuando el almacenamiento permanezca operativo en la región primaria, pero debe realizar una conmutación por error de toda la solución a una región secundaria, como para los simulacros de recuperación ante desastres diseñados para garantizar el cumplimiento de los requisitos de auditoría.

- **Conmutación por error administrada por Microsoft:** en circunstancias excepcionales, Microsoft podría iniciar la conmutación por error para todas las cuentas de almacenamiento con redundancia geográfica (GRS) de una región. Sin embargo, la

conmutación por error administrada por Microsoft es un último recurso y se espera que solo se realice después de un período prolongado de interrupción. No se debe confiar en el failover administrado por Microsoft.

Las cuentas de GRS pueden utilizar cualquiera de estos tipos de conmutación por error. No es necesario preconfigurar una cuenta de almacenamiento para usar ninguno de los tipos de conmutación por error con antelación.

Requisitos

- **Compatibilidad con regiones:** Las configuraciones con redundancia geográfica de Azure Storage usan regiones emparejadas de Azure para la replicación de regiones secundarias. La región secundaria se determina automáticamente en función de la selección de la región principal y no se puede personalizar. Para obtener una lista completa de las regiones emparejadas de Azure, consulte Lista de regiones de Azure.

  Si la región de la cuenta de almacenamiento no está emparejada, considere la posibilidad de usar las soluciones personalizadas de varias regiones para lograr resistencia.

- **Tipos de cuenta de almacenamiento:** El almacenamiento con redundancia geográfica (GRS) y la conmutación por error y la conmutación por recuperación iniciadas por el cliente están disponibles en todas las regiones emparejadas de Azure que admiten cuentas de Azure Storage de uso general v2.

Consideraciones

Al implementar Blob Storage en varias regiones, tenga en cuenta los siguientes factores clave:

- **Latencia de replicación asincrónica:** la replicación de datos en la región secundaria es asincrónica, lo que significa que hay un retraso entre cuando los datos se escriben en la región primaria y cuando estarán disponibles en la región secundaria. Este retraso podría provocar una posible pérdida de datos si se produjese un fallo en la región principal antes de que se repliquen los datos recientes. La pérdida de datos se mide mediante el objetivo de punto de recuperación (RPO). Espere que el retraso de replicación sea inferior a 15 minutos, pero este tiempo es una estimación y no se garantiza.

  Compruebe la propiedad Última hora de sincronización para saber cuántos datos se podrían perder si su cuenta de almacenamiento sufriese una conmutación por error no planificada.

- **Acceso a la región secundaria:** con las configuraciones de almacenamiento con redundancia geográfica (GRS) y almacenamiento con redundancia de zona geográfica

(GZRS), la región secundaria no será accesible a las lecturas hasta que se produzca una conmutación por error.

Las configuraciones de almacenamiento con redundancia geográfica con acceso de lectura (RA-GRS) y almacenamiento con redundancia de zona geográfica con acceso de lectura (RA-GZRS) proporcionan acceso de lectura a la región secundaria durante las operaciones normales, pero debido a la latencia de replicación asincrónica, podrían devolver datos ligeramente obsoletos.

- **Limitaciones de las características:** algunas características de Azure Storage no son compatibles o tienen limitaciones cuando se utiliza almacenamiento con redundancia geográfica (GRS) o conmutación por error administrada por el cliente. Revise la compatibilidad de las característica antes de implementar la redundancia geográfica.

Cost

Las configuraciones de cuentas de Azure Storage multirregionales conllevan costes adicionales por la replicación entre regiones y el almacenamiento en la región secundaria. La transferencia de datos entre regiones de Azure se cobra en función de las tarifas estándar de ancho de banda entre regiones.

Para más información, consulte Precios de Blob Storage.

Configuración de la compatibilidad con varias regiones

- **Cree una nueva cuenta de almacenamiento con redundancia geográfica (GRS).** Para crear una cuenta de GRS, consulte Creación de una cuenta de almacenamiento y seleccione GRS, almacenamiento con redundancia geográfica con acceso de lectura (RA-GRS), almacenamiento con redundancia de zona geográfica (GZRS) o almacenamiento con redundancia de zona geográfica con acceso de lectura (RA-GZRS) durante la creación de la cuenta.

- **Habilite la redundancia geográfica en una cuenta de almacenamiento existente.** Para convertir una cuenta de almacenamiento existente en almacenamiento con redundancia geográfica (GRS), consulte Cambio de cómo se replica una cuenta de almacenamiento.

> ⚠ **Advertencia**
>
> Una vez que su cuenta se haya reconfigurado para la redundancia geográfica, es posible que transcurra un tiempo considerable antes de que los datos existentes en la nueva región principal se copien por completo a la nueva región secundaria.

Para evitar una pérdida importante de datos, compruebe el valor de la **propiedad Última hora de sincronización** antes de iniciar una conmutación por error no planificada. Para evaluar la posible pérdida de datos, compare la última hora de sincronización con la última hora en que se escribieron los datos en la nueva región principal.

- **Deshabilite la redundancia geográfica.** Convierta las cuentas de GRS en configuraciones de una sola región, como el almacenamiento con redundancia local (LRS) o el almacenamiento con redundancia local (ZRS) con el mismo proceso de cambio de configuración de redundancia.

Comportamiento cuando todas las regiones están en buen estado

En esta sección se describe qué esperar cuando una cuenta de almacenamiento está configurada para redundancia geográfica y todas las regiones están operativas.

- **Enrutamiento del tráfico entre regiones:** Azure Storage utiliza un enfoque activo-pasivo en el que todas las operaciones de escritura y la mayoría de las operaciones de lectura se dirigen a la región principal.

  En las configuraciones de almacenamiento con redundancia geográfica con acceso de lectura (RA-GRS) y almacenamiento con redundancia de zona geográfica con acceso de lectura (RA-GZRS), las aplicaciones pueden leer opcionalmente desde la región secundaria accediendo al punto de conexión secundario. Este enfoque requiere una configuración explícita de la aplicación y no es automática. Además, debido al retraso de replicación asincrónica, los datos de la región secundaria podrían estar ligeramente obsoletos.

- **Replicación de datos entre regiones:** las operaciones de escritura se confirman primero en la región principal utilizando los siguientes tipos de redundancia configurados:
  - Almacenamiento con redundancia local (LRS) para almacenamiento con redundancia geográfica (GRS) y RA-GRS
  - Almacenamiento con redundancia de zona (ZRS) para el almacenamiento con redundancia de zona geográfica (GZRS) y RA-GZRS

  Tras completarse con éxito en la región principal, los datos se replican de forma asíncrona en la región secundaria, donde se almacenan mediante LRS.

  La naturaleza asíncrona de la replicación entre regiones significa que suele haber un retraso entre el momento en que los datos se escriben en la región principal y el momento en que están disponibles en la región secundaria. Supervise el tiempo de replicación utilizando la propiedad Última hora de sincronización.

Comportamiento durante una falla de región

En esta sección se describe qué esperar cuando una cuenta de almacenamiento está configurada para la redundancia geográfica y hay una interrupción en la región primaria.

- **Conmutación por error administrada por el cliente (no planificada):** utilice una conmutación por error no planificada cuando el almacenamiento en la región principal no esté disponible.
  - **Detección y respuesta:** En el caso poco probable de que su cuenta de almacenamiento no esté disponible en la región primaria, puede considerar iniciar un failover no planeado administrado por el cliente. Para tomar esta decisión, tenga en cuenta los siguientes factores:
    - Si Azure Resource Health muestra problemas al acceder a la cuenta de almacenamiento en su región principal
    - Si Microsoft le recomienda realizar una conmutación por error a otra región

    > ⚠ **Advertencia**
    >
    > Una conmutación por error no planeada puede **provocar la pérdida de datos**. Antes de iniciar una conmutación por error administrada por el cliente, decida si la restauración del servicio justifica el riesgo de pérdida de datos.

  - **Notificación**: Microsoft no le notifica automáticamente cuando una región está inactiva. Sin embargo:
    - Puede usar Azure Resource Health para supervisar el estado de un recurso individual y puede configurar alertas de Resource Health para notificarle problemas.
    - Puede usar Azure Service Health para comprender el estado general del servicio, incluidos los errores de región, y puede configurar alertas de Service Health para notificarle problemas.
  - **Solicitudes activas:** Durante el proceso de conmutación por error, los puntos de conexión de la cuenta de almacenamiento principal y secundaria dejan de estar disponibles temporalmente para lecturas y escrituras. Las solicitudes activas podrían fallar, y las aplicaciones cliente deben intentar nuevamente después de que se complete la conmutación por error.
  - **Pérdida de datos prevista:** la pérdida de datos es habitual durante una conmutación por error no planificada debido al retraso en la replicación asíncrona, lo que significa que es posible que las escrituras recientes no se repliquen. Consulte la propiedad

Última hora de sincronización para saber cuántos datos se podrían perder durante una conmutación por error no planificada. La pérdida de datos esperada se conoce a menudo como objetivo de punto de recuperación (RPO). Normalmente, puede esperar que el RPO sea inferior a 15 minutos, pero ese tiempo no está garantizado.

  - **Tiempo de inactividad esperado:** La cantidad de tiempo de inactividad esperado se conoce a menudo como objetivo de tiempo de recuperación (RTO). La conmutación por error administrada por el cliente se completa normalmente en un plazo de hasta 60 minutos, dependiendo del tamaño y la complejidad de la cuenta.

  - **Reenrutamiento del tráfico:** A medida que se completa la conmutación por error, Azure actualiza automáticamente los puntos de conexión de la cuenta de almacenamiento para que no sea necesario volver a configurar las aplicaciones. Si su aplicación mantiene entradas del Sistema de nombres de dominio (DNS) almacenadas en caché, puede que sea necesario borrar la caché para asegurarse de que la aplicación envíe el tráfico a la nueva región principal.

  - **Configuración posterior a la conmutación por error:** una vez completada una conmutación por error no planificada, su cuenta de almacenamiento en la región de destino utilizará el nivel de almacenamiento con redundancia local (LRS). Si necesitase volver a replicar geográficamente los datos, deberá volver a habilitar el almacenamiento con redundancia geográfica (GRS) y esperar a que los datos se repliquen en la nueva región secundaria.

  Para obtener más información sobre cómo iniciar una conmutación por error administrada por el cliente, consulte Cómo funciona la conmutación por error administrada por el cliente (no planificada) e Iniciar una conmutación por error de una cuenta de almacenamiento.

- **Conmutación por error administrada por el cliente (planificada):** utilice una conmutación por error planificada cuando el almacenamiento siga operativo en la región principal, pero necesite conmutar por error toda su solución a una región secundaria por otro motivo. Por ejemplo, otro servicio de Azure podría estar experimentando un problema y debe cambiar al uso de una región secundaria para toda la solución. O bien, puede usar una conmutación por error planeada para realizar un simulacro de recuperación ante desastres (DR) con fines de cumplimiento y auditoría.

  - **Detección y respuesta:** Usted es responsable de decidir conmutar por error. Normalmente, debe tomar esta decisión si necesita realizar una conmutación entre regiones, aunque su cuenta de almacenamiento esté en buen estado. Por ejemplo, puede desencadenar una conmutación por error cuando se produce una interrupción importante en otro componente de la aplicación que no se puede recuperar en la región principal.

- **Notificación**: Microsoft no le notifica automáticamente cuando una región está inactiva. Sin embargo:
  - Puede usar Azure Resource Health para supervisar el estado de un recurso individual y puede configurar alertas de Resource Health para notificarle problemas.
  - Puede usar Azure Service Health para comprender el estado general del servicio, incluidos los errores de región, y puede configurar alertas de Service Health para notificarle problemas.

- **Solicitudes activas:** Durante el proceso de conmutación por error, los puntos de conexión de la cuenta de almacenamiento principal y secundaria dejan de estar disponibles temporalmente para lecturas y escrituras. Las solicitudes activas podrían fallar, y las aplicaciones cliente deben intentar nuevamente después de que se complete la conmutación por error.

- **Pérdida de datos esperada:** No se espera ninguna pérdida de datos porque el proceso de conmutación por error se completa solo después de sincronizar todos los datos, lo que da como resultado un RPO de cero.

- **Tiempo de inactividad esperado:** La conmutación por error normalmente se completa en un plazo de 60 minutos, lo que significa que el tiempo de recuperación objetivo (RTO) esperado es de 60 minutos, en función del tamaño y la complejidad de la cuenta. Durante el proceso de conmutación por error, los puntos de conexión de la cuenta de almacenamiento principal y secundaria dejan de estar disponibles temporalmente para lecturas y escrituras.

- **Reenrutamiento del tráfico:** A medida que se completa la conmutación por error, Azure actualiza automáticamente los puntos de conexión de la cuenta de almacenamiento para que no sea necesario volver a configurar las aplicaciones. Si su aplicación mantiene entradas DNS almacenadas en caché, puede que sea necesario borrar la caché para asegurarse de que la aplicación envíe tráfico a la nueva región principal.

- **Configuración posterior a la conmutación por error:** Una vez completada una conmutación por error planeada, la cuenta de almacenamiento de la región de destino sigue siendo replicada geográficamente y permanece en el nivel GRS.

Para obtener más información sobre cómo iniciar una conmutación por error administrada por el cliente, consulte Cómo funciona la conmutación por error administrada por el cliente (planificada) e Iniciar una conmutación por error de una cuenta de almacenamiento.

- **Conmutación por error administrada por Microsoft:** En el caso excepcional de un desastre importante en el que Microsoft determina que la región primaria es irrecuperable permanentemente, es posible que se inicie una conmutación automática por error a la región secundaria. Microsoft controla todo el proceso y no se requiere ninguna acción del cliente. El tiempo que transcurre antes de que ocurra un failover depende de la gravedad del desastre y del tiempo necesario para evaluar la situación.
  - **Notificación**: Microsoft no le notifica automáticamente cuando una región está inactiva. Sin embargo:
    - Puede usar Azure Resource Health para supervisar el estado de un recurso individual y puede configurar alertas de Resource Health para notificarle problemas.
    - Puede usar Azure Service Health para comprender el estado general del servicio, incluidos los errores de región, y puede configurar alertas de Service Health para notificarle problemas.

> ⓘ **Importante**
>
> Utilice las opciones de conmutación por error administradas por el cliente para desarrollar, probar e implementar tus planes de recuperación ante desastres. **No confíe en la conmutación por error administrada por Microsoft**, que solo se puede utilizar en circunstancias extremas. Es probable que se haya iniciado una conmutación por error administrada por Microsoft para toda una región. No se puede iniciar para cuentas de almacenamiento individuales, suscripciones o clientes. La conmutación por error puede producirse en momentos diferentes para distintos servicios de Azure. Recomendamos que utilice la conmutación por error administrada por el cliente.

Recuperación de regiones

El proceso de conmutación por recuperación difiere significativamente entre los escenarios de conmutación por error administrados por Microsoft y los administrados por el cliente.

- **Conmutación por error administrada por el cliente (no planeada):** después de una conmutación por error no planeada, la cuenta de almacenamiento se configura con almacenamiento con redundancia local (LRS). Para realizar una conmutación por error, es necesario restablecer la relación de almacenamiento con redundancia geográfica (GRS) y esperar a que se repliquen los datos.
- **Conmutación por error administrada por el cliente (planificada):** después de una conmutación por error planificada, la cuenta de almacenamiento seguirá estando

replicada geográficamente. Inicie otra conmutación por error administrada por el cliente para volver a la región principal original. Se aplican las mismas consideraciones de conmutación por error.

- **Conmutación por error administrada por Microsoft:** si Microsoft iniciase una conmutación por error, es probable que se haya producido un desastre importante en la región principal y que esta no se pueda recuperar. Cualquier calendario o plan de recuperación dependerá del alcance del desastre regional y de los esfuerzos de recuperación. Debe supervisar las comunicaciones de Azure Service Health para más información.

Prueba de fallos de región

Simule fallos regionales para probar sus procedimientos de recuperación ante desastres.

- **Pruebas de conmutación por error planificadas:** en el caso de las cuentas de almacenamiento con redundancia geográfica (GRS), es posible realizar operaciones de conmutación por error planificadas durante las ventanas de mantenimiento para probar el proceso completo de conmutación por error y conmutación por recuperación. La conmutación por error planificada no implica pérdida de datos, pero sí conlleva tiempo de inactividad tanto durante la conmutación por error como durante la conmutación por recuperación.

- **Pruebas de puntos de conexión secundarios:** para el almacenamiento con redundancia geográfica con acceso de lectura (RA-GRS) y las configuraciones de almacenamiento con redundancia de zona geográfica con acceso de lectura (RA-GZRS), pruebe periódicamente las operaciones de lectura en el punto de conexión secundario para asegurarse de que la aplicación pueda leer correctamente los datos de la región secundaria.

Soluciones personalizadas de varias regiones para la resistencia

Las funcionalidades de conmutación por error entre regiones de Azure Storage podrían no ser adecuadas por los siguientes motivos:

- La cuenta de almacenamiento está en una región no emparejada.

- Los objetivos de tiempo de actividad empresarial no se satisfacen por el tiempo de recuperación o la pérdida de datos que proporciona las opciones de conmutación por error integradas.

- Debe conmutar por error a una región que no esté emparejada con la región principal.
- Necesita una configuración activa o activa entre regiones.

En esta sección se proporciona información general de alto nivel de algunos enfoques que se deben tener en cuenta. Una introducción completa de las topologías de implementación de varias regiones para Azure Storage está fuera del ámbito de este artículo.

Implemente Azure Storage en varias regiones utilizando cuentas de almacenamiento independientes en cada región. Este enfoque ofrece flexibilidad en la selección de regiones, la posibilidad de utilizar regiones no emparejadas y un control más granular sobre el momento de la replicación y la coherencia de los datos. Al implementar varias cuentas de almacenamiento entre regiones, es necesario configurar la replicación de datos entre regiones, implementar directivas de equilibrio de carga y conmutación por error y garantizar la coherencia de los datos entre regiones.

La replicación de objetos proporciona una opción adicional para la replicación de datos entre regiones que proporciona copia asincrónica de blobs en bloques entre cuentas de almacenamiento. A diferencia de las opciones de GRS integradas que usan regiones emparejadas fijas, la replicación de objetos permite replicar datos entre cuentas de almacenamiento en cualquier región de Azure, incluyendo las regiones no emparejadas. Este enfoque le ofrece un control total sobre las regiones de origen y destino, las directivas de replicación y los contenedores y prefijos de blobs específicos que se van a replicar.

Configure la replicación de objetos para replicar todos los blobs de un contenedor o subconjuntos específicos en función de prefijos y etiquetas de blobs. La replicación es asincrónica y se produce en segundo plano. Puede configurar varias directivas de replicación e incluso encadenar la replicación entre varias cuentas de almacenamiento para crear topologías multirregionales sofisticadas.

La replicación de objetos no es compatible con todas las cuentas de almacenamiento. Por ejemplo, no funciona con cuentas de almacenamiento que usen espacios de nombres jerárquicos (también conocidos como *cuentas de Azure Data Lake Storage Gen2*).

Para obtener más información, consulte Replicación de objetos para blobs en bloques y Configuración de la replicación de objetos.

Copia de seguridad y recuperación

Blob Storage proporciona varios mecanismos de protección de datos que complementan la redundancia para estrategias de copia de seguridad completas. La redundancia integrada del servicio protege frente a errores de infraestructura, mientras que las funcionalidades de copia

de seguridad adicionales protegen contra actividades malintencionadas, daños y eliminación accidentales.

La restauración a un momento dado (PITR) permite restaurar los datos de blobs en bloques a un estado anterior dentro de un período de retención configurado de hasta 365 días. Microsoft administra completamente esta característica. También proporciona funcionalidades de recuperación pormenorizadas en el nivel de contenedor o blob. Los datos PITR se almacenan en la misma región que la cuenta de origen y son accesibles incluso durante interrupciones regionales si se usan configuraciones con redundancia geográfica.

El **control de versiones de blobs** mantiene automáticamente las versiones anteriores de blobs cuando se modifiquen o eliminen. Cada versión se almacena como un objeto independiente y se puede acceder a ella de forma independiente. Las versiones se almacenan en la misma región que el blob actual y siguen la misma configuración de redundancia que la cuenta de almacenamiento.

La eliminación temporal proporciona una red de seguridad para los blobs y contenedores eliminados accidentalmente, ya que conserva los datos eliminados durante un período configurable. Los datos eliminados temporalmente permanecen en la misma cuenta de almacenamiento y región, lo que hace que estén disponibles inmediatamente para la recuperación. En el caso de las cuentas con redundancia geográfica, los datos eliminados de forma temporal también se replican en la región secundaria.

Las **instantáneas de blobs** crean copias de solo lectura y a un momento dado de blobs que se pueden usar para escenarios de copia de seguridad y recuperación. Las instantáneas se almacenan en la misma cuenta de almacenamiento y siguen la misma configuración de redundancia y replicación geográfica que el blob base.

En el caso de los requisitos de copia de seguridad entre regiones, considere la posibilidad de usar **Azure Backup para blobs**, que proporciona administración centralizada de copias de seguridad y almacena datos de copia de seguridad en regiones diferentes de los datos de origen. Este servicio proporciona opciones de copia de seguridad operativas y de almacén que tienen directivas de retención configurables, así como funcionalidades de restauración. Para más información, consulte Introducción a Backup para blobs.

Para la mayoría de las soluciones, no debe confiar exclusivamente en copias de seguridad. En su lugar, utilice las otras capacidades descritas en esta guía para apoyar los requisitos de resiliencia. Sin embargo, las copias de seguridad protegen contra algunos riesgos que otros enfoques no. Para más información, consulte ¿Qué son la redundancia, la replicación y la copia de seguridad?.

Acuerdo de nivel de servicio

El acuerdo de nivel de servicio (SLA) para Azure Storage describe la disponibilidad esperada del servicio y las condiciones que se deben cumplir para lograr esa expectativa de disponibilidad. El contrato de nivel de servicio de disponibilidad al que tiene derecho dependerá del nivel de almacenamiento y del tipo de replicación que utilice. Para obtener más información, consulte los SLAs for Online Services.

Contenido relacionado

- Introducción a Blob Storage
- Redundancia de Azure Storage
- Recuperación ante desastres y failover de la cuenta de almacenamiento
- Uso de redundancia geográfica para diseñar aplicaciones de alta disponibilidad
- Cambio de la replicación de una cuenta de almacenamiento

ⓘ **Nota:** El autor ha creado este artículo con ayuda de inteligencia artificial. Más información

---

Last updated on 22/01/2026

Confiabilidad en Azure Backup

Azure Backup es un servicio de Azure integrado que protege de forma segura las cargas de trabajo locales y en la nube. La copia de seguridad puede escalar su protección en múltiples cargas de trabajo y proporciona integración nativa con las cargas de trabajo de Azure, incluidas las máquinas virtuales (VM) en Azure, SAP HANA en máquinas virtuales en Azure, SQL en máquinas virtuales en Azure, Azure Files, Azure Blob Storage, Azure Data Lake Storage, discos administrados de Azure, volúmenes de Azure Elastic SAN y Azure Kubernetes Service (AKS). No es necesario administrar la automatización ni la infraestructura, escribir scripts ni aprovisionar el almacenamiento.

Al usar Azure, relibilidad es una responsabilidad compartida. Microsoft proporciona una variedad de capacidades para admitir resiliencia y recuperación. Es responsable de comprender cómo funcionan esas funcionalidades dentro de todos los servicios que usa y de seleccionar las funcionalidades que necesita para cumplir los objetivos empresariales y los objetivos de tiempo de actividad.

En este artículo se describe cómo la copia de seguridad puede ser resistente a una variedad de posibles interrupciones y problemas, incluidos errores transitorios, interrupciones de zona de disponibilidad y interrupciones de regiones. También resalta cierta información clave sobre el acuerdo de nivel de servicio (SLA) de copia de seguridad.

> **Nota:**
>
> En este artículo se describe cómo el propio servicio Backup es resistente a varios problemas y cómo puede hacer que sea más resistente. No explica cómo usar backup para proteger las máquinas virtuales, los datos u otros recursos. Para obtener información sobre cómo usar Backup, consulte **Introducción a la copia de seguridad**.

Recomendaciones de implementación de producción para la confiabilidad

Para realizar copias de seguridad de las cargas de trabajo de producción, se recomienda configurar la bóveda de las siguientes maneras:

- ✔ Use el almacenamiento con redundancia de zona (ZRS) como nivel de redundancia mínimo para las copias de seguridad. ZRS replica las copias de seguridad en varias zonas de disponibilidad para poder restaurar las copias de seguridad durante una interrupción de la zona de disponibilidad.
- ✔ Si usa almacenamiento con redundancia geográfica (GRS) para replicar las copias de seguridad en una región de Azure emparejada, habilite la restauración entre regiones (CRR) para orígenes de datos admitidos. CRR le permite restaurar las copias de seguridad en la región emparejada en cualquier momento.

En las secciones siguientes de este artículo se proporcionan más detalles sobre estas configuraciones.

> ⓘ **Nota:**
>
> Estas recomendaciones de redundancia de almacenamiento se aplican a ubicaciones en las que se replican las copias de seguridad, no al servicio Backup o a los recursos de los que realiza una copia de seguridad. La protección de copia de seguridad y la redundancia de almacenamiento se complementan entre sí. Las copias de seguridad protegen contra la pérdida de datos y la redundancia protege frente a errores de infraestructura.

Para obtener una lista de otras recomendaciones para copia de seguridad, incluidas las recomendaciones centradas en la confiabilidad, consulte Copia de seguridad de cargas de trabajo en la nube y cargas de trabajo locales a la nube.

Introducción a la arquitectura de confiabilidad

En esta sección se describen algunos de los aspectos importantes de cómo funciona el servicio que es más relevante desde una perspectiva de confiabilidad. En la sección se presenta la arquitectura lógica, que incluye algunos de los recursos y características que se implementan y usan. También se describe la arquitectura física, que proporciona detalles sobre cómo funciona el servicio en segundo plano.

Arquitectura lógica

La copia de seguridad puede realizar copias de seguridad y restaurar una variedad de *orígenes de datos*. Las copias de seguridad se configuran de forma diferente en función del origen de datos con el que trabaje. Los orígenes de datos siguientes son comunes:

- máquinas virtuales de Azure
- Varias bases de datos
- cuentas de Blob Storage
- Clústeres de AKS
- Servidores locales a través del agente de Microsoft Azure Recovery Services (MARS)

La copia de seguridad almacena los datos de copia de seguridad en *almacenes*. Las bóvedas son entidades de almacenamiento en línea en Azure que contienen datos, como copias de seguridad, puntos de recuperación y políticas de copia de seguridad. Los almacenes de Recovery Services y los almacenes de Backup son dos tipos de almacenes. Puede usar uno o ambos tipos en función de lo que necesite proteger. Para obtener una lista de los orígenes de datos que admite cada tipo de almacén, consulte Preguntas más frecuentes sobre los almacenes admitidos para la copia de seguridad y restauración.

Los trabajos representan la actividad de copia de seguridad o restauración de los datos. Los trabajos de copia de seguridad incluyen operaciones programadas o bajo demanda que copian sus datos desde el origen a la bóveda. Los trabajos de restauración incluyen operaciones que recuperan los datos del almacenamiento de copia de seguridad en una ubicación de destino. Cada trabajo tiene un identificador único y un seguimiento de estado para que pueda supervisar el progreso y solucionar problemas que se producen durante las operaciones de copia de seguridad y restauración. Tú también creas *directivas de copia de seguridad* asociadas a trabajos. Las directivas especifican la configuración como la programación de copia de seguridad y cuánto tiempo desea conservar los datos.

Bóvedas almacenan las directivas de copia de seguridad y la configuración junto con los metadatos de los trabajos, permitiéndole realizar un seguimiento de los trabajos y solucionar problemas.

Arquitectura física

Microsoft administra la infraestructura principal del servicio Backup. Esta infraestructura es responsable de la administración y el funcionamiento del servicio, incluidos los trabajos de desencadenamiento y supervisión.

Los respaldos se almacenan en la bóveda. Las bóvedas se construyen sobre Azure Storage. Las bóvedas replican automáticamente los datos de copias de seguridad, y la durabilidad y resistencia de la copia de seguridad dependen de la redundancia de almacenamiento de la bóveda.

- Almacenamiento con redundancia de ubicación (LRS) replica los datos dentro de su almacén en una o varias zonas de disponibilidad de Azure ubicadas en la región primaria de su elección. No puede elegir la zona de disponibilidad preferida, pero Azure puede mover o expandir cuentas de LRS entre zonas para mejorar el equilibrio de carga. No se garantiza que los datos se repartan entre zonas. Para obtener más información, consulte Introducción a las zonas de disponibilidad.

- ZRS y GRS proporcionan protecciones adicionales. En este artículo se describen estas opciones al detalle.

> ⓘ **Nota:**
>
> Algunas fuentes de datos admiten copias de seguridad de *nivel operativo*, que almacenan datos en otra ubicación en lugar del almacén. Por ejemplo, la **copia de seguridad de discos administrados de Azure** y las **copias de seguridad de AKS** admiten copias de seguridad de nivel operativo, que se almacenan en instantáneas de disco. En este artículo no se describe el almacenamiento de copia de seguridad de nivel operativo, pero puede aplicar las instrucciones de resistencia de este artículo a operaciones y flujos de trabajo de copia de seguridad para estos tipos de copia de seguridad.

Resistencia a errores transitorios

Los errores transitorios son errores breves e intermitentes en los componentes. Se producen con frecuencia en un entorno distribuido como la nube y son una parte normal de las operaciones. Los errores transitorios se corrigen después de un breve período de tiempo. Es importante que las aplicaciones puedan controlar errores transitorios, normalmente mediante el reintento de solicitudes afectadas.

Todas las aplicaciones hospedadas en la nube deben seguir las instrucciones de control de errores transitorios Azure cuando se comunican con cualquier API, bases de datos y otros componentes hospedados en la nube. Para obtener más información, consulte Recomendaciones para controlar errores transitorios.

Cuando se usa Backup, los flujos de trabajo de copia de seguridad y restauración son resistentes a errores intermitentes. El servicio vuelve a intentarlo automáticamente cuando encuentra errores transitorios de red o interrupciones temporales del servicio. Tú no configuras ninguna lógica de reintento. Si experimenta errores repetidos, consulte Solución de problemas de las operaciones de administración de la bóveda de copias de seguridad.

Resistencia a errores de zona de disponibilidad

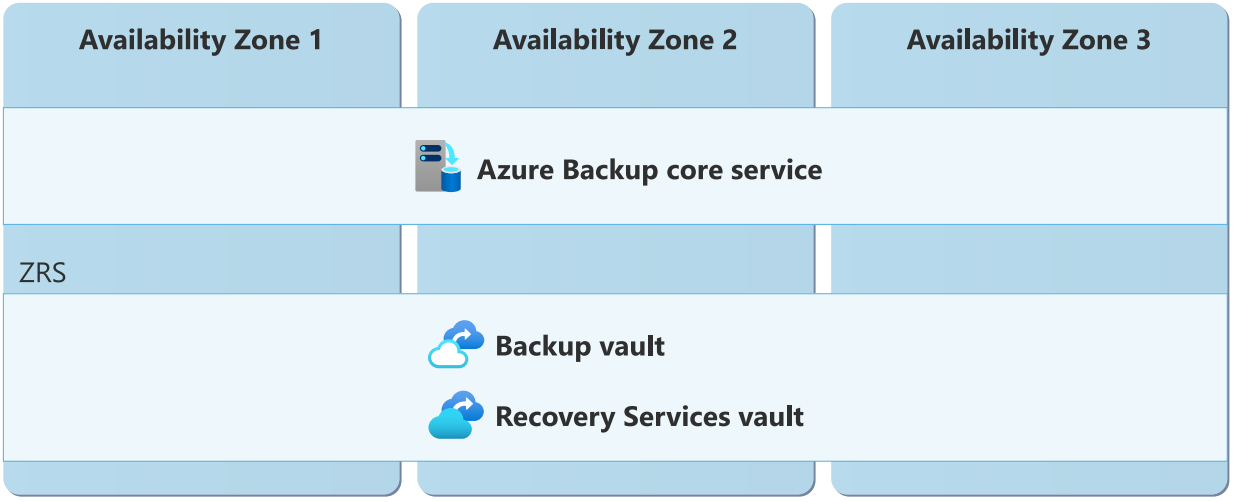
Availability zones son grupos de centros de datos físicamente independientes dentro de una región de Azure. Cuando una zona falla, los servicios pueden transferirse a una de las zonas restantes.

La copia de seguridad administra por separado la configuración de zona de disponibilidad del servicio y de los datos.

- **Servicio:** El servicio de Backup es automáticamente resistente a las zonas en regiones admitidas. Sin embargo, esta resistencia de zona integrada no se aplica a los datos de copia de seguridad.

- **Redundancia de almacenamiento de copia de seguridad:** Seleccione el nivel de redundancia que desea para los datos de copia de seguridad mediante la configuración del almacén de Recovery Services o el almacén de Backup. Si selecciona ZRS, las copias de los datos de copia de seguridad se almacenan automáticamente en varias zonas de disponibilidad en la región de Azure que use.

  Si no usa ZRS, los datos de copia de seguridad se consideran *no zonales* y pueden almacenarse en cualquier zona. Si alguna zona de la región tiene un problema, es posible que los datos de copia de seguridad no zonales no estén disponibles.



Requisitos

- **Compatibilidad con regiones:** El servicio es automáticamente resistente a zonas en todas las regiones que tienen zonas de disponibilidad. Los almacenes de ZRS se admiten en las mismas regiones.

- **Solo bóvedas nuevas:** Configure ZRS en su bóveda antes de la primera copia de seguridad.

Cost

Al habilitar ZRS para las copias de seguridad, se le cobra a una tarifa diferente de LRS debido a la sobrecarga adicional de replicación y almacenamiento. Para obtener más información, consulte Precios de copia de seguridad .

Configurar soporte de zonas de disponibilidad

- **Cree un nuevo almacén que use ZRS:** Configure la redundancia de almacenamiento al crear un almacén. Siga los diferentes pasos según el tipo de bóveda. Para obtener más información, consulte los artículos siguientes:
  - Crear y eliminar almacenes de copias de seguridad
  - Creación y configuración de un almacén de Recovery Services

- **Configuración de ZRS en almacenes existentes:** En el caso de los almacenes de Backup, configure la redundancia de almacenamiento al crear el almacén. Después de crear un almacén de Backup, la configuración está bloqueada y no se puede cambiar.

  En el caso de las bóvedas de Recovery Services, debe configurar la redundancia de almacenamiento antes de proteger cualquier carga de trabajo. Después de proteger una carga de trabajo, la configuración está bloqueada y no se puede cambiar.

  Puede crear una nueva bóveda configurada para usar ZRS y reasignar las cargas de trabajo a la nueva bóveda. Sin embargo, este enfoque requiere tiempo de inactividad. Para obtener más información, consulte Modificación de la configuración predeterminada. También es responsable de eliminar manualmente los puntos de recuperación existentes y otros datos porque las directivas de retención del almacén antiguo ya no se aplican. Para obtener más información, consulte Eliminación de un almacén de Backup o Eliminación de un almacén de Recovery Services.

Comportamiento cuando todas las zonas están en buen estado

En esta sección se describe qué esperar al configurar bóvedas para ZRS cuando todas las zonas están operativas.

- **Operación entre zonas:** Los trabajos de copia de seguridad se ejecutan en la infraestructura replicada entre zonas. Azure administra los trabajos de la infraestructura en cualquier zona.

- **Replicación de datos entre zonas:** ZRS replica los datos de copia de seguridad entre zonas. La replicación se produce de forma sincrónica, lo que significa que varias zonas reconocen cada operación de escritura antes de que se complete.

Comportamiento durante un fallo de zona

En esta sección se describe qué esperar al configurar bóvedas para ZRS cuando se produce una interrupción en una de las zonas.

- **Detección y respuesta:** Para el servicio de copia de seguridad, Microsoft es responsable de detectar fallos en las zonas de disponibilidad y responder. No es necesario hacer nada para iniciar una conmutación por error de la zona.

  > ⓘ **Importante**
  >
  > En el caso de los datos o recursos que no están disponibles debido a una falla de la zona, usted es responsable de detectar la falla y tomar medidas de recuperación, incluida la restauración de copias de seguridad en una zona operativa.

- **Notification**: Microsoft no le notifica automáticamente cuando una zona está inactiva. Sin embargo, puede usar Azure Resource Health para supervisar el estado de un recurso individual y puede configurar Resource Health alertas para notificarle problemas. También puede usar Azure Service Health para comprender el estado general del servicio, incluidos los errores de zona, y puede configurar alertas de Service Health para notificarle problemas.

- **Solicitudes activas:** El comportamiento de las tareas activas depende de qué zona falla.

  - Para cualquier origen de datos en la zona de disponibilidad fallida, el fallo de la zona hace que los orígenes de datos no estén disponibles. Los trabajos activos podrían pausarse o fallar.

  - En el caso de los orígenes de datos en zonas de disponibilidad correctas que ejecutan trabajos activos, una pequeña cantidad de tiempo de inactividad, normalmente unos segundos, puede producirse mientras la plataforma cambia a zonas de disponibilidad correctas para el servicio Backup.

- **Pérdida de datos esperada:** La cantidad esperada de pérdida de datos también se conoce como *objetivo de punto de recuperación (RPO)*. El RPO de los datos de copia de seguridad depende de varios factores, incluida la programación de las copias de seguridad. En general, para una interrupción de zona, no se espera ninguna pérdida de datos de copia de seguridad porque todos los datos se replican de manera sincrónica entre las zonas.

- **Tiempo de inactividad esperado:** La cantidad esperada de tiempo de inactividad también se conoce como *objetivo de tiempo de recuperación (RTO)*. El RTO es diferente para cada uno de los escenarios siguientes:

  - En el caso de los orígenes de datos de la zona de disponibilidad fallida, es posible que estos no estén disponibles hasta que la zona se recupere. Es posible que los trabajos de copia de seguridad no se ejecuten hasta que el origen de datos esté disponible de nuevo. El RTO no está definido.

  - En el caso de los orígenes de datos en zonas de disponibilidad correctas, una pequeña cantidad de tiempo de inactividad, normalmente unos segundos, puede producirse mientras la plataforma cambia a zonas de disponibilidad correctas para el servicio Backup.

- **Redistribución:** Las ejecuciones posteriores de trabajos usan automáticamente la infraestructura en zonas saludables siempre y cuando los orígenes de datos estén disponibles.

  Es responsable de restaurar la copia de seguridad en la infraestructura de una zona saludable y reconfigurar los balanceadores de carga, los clientes y otros sistemas para redirigir el tráfico a una infraestructura saludable en la nueva zona.

Recuperación de zona

Cuando se recupera la zona de disponibilidad, Backup restaura automáticamente las operaciones en la zona de disponibilidad y vuelve a enrutar el tráfico entre las zonas como es normal. Los trabajos siguen ejecutándose y los datos permanecen disponibles.

Prueba de fallos de zona

La plataforma de respaldo gestiona el enrutamiento del tráfico, la replicación de datos, la conmutación por error y la recuperación después de fallos. Esta característica está totalmente administrada, por lo que no es necesario iniciar ni validar los procesos de error de zona de disponibilidad.

Resistencia a errores en toda la región

Backup admite la redundancia geográfica y la conmutación por error a través de GRS y CRR.

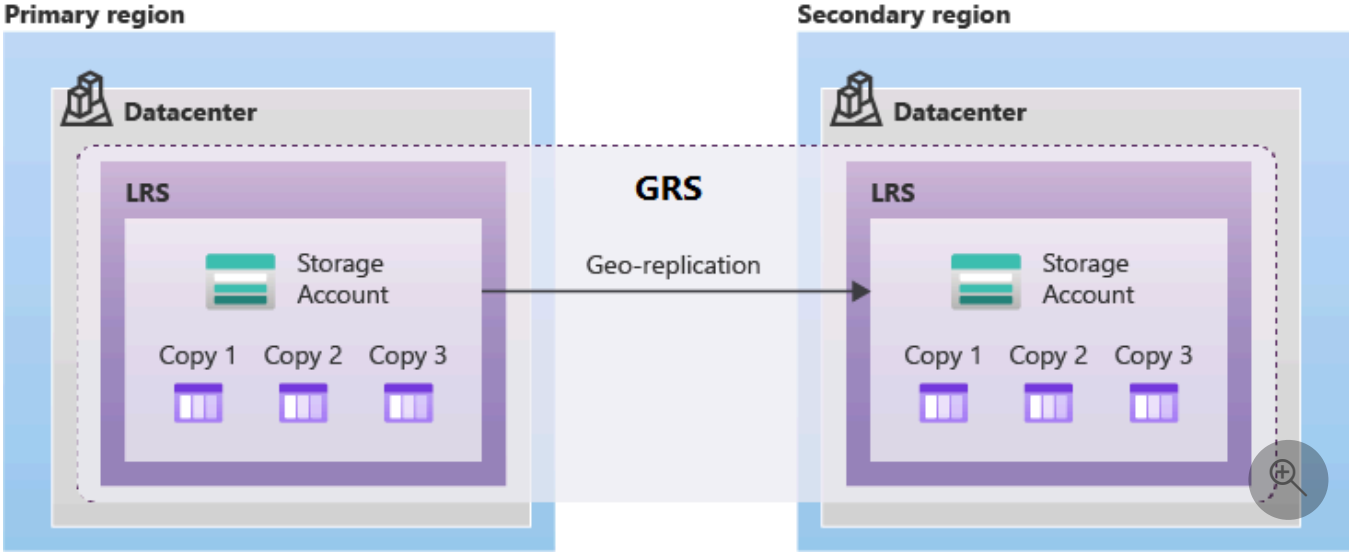
> ⓘ **Importante**
>
> GRS for Backup solo funciona en regiones Azure emparejadas.

Almacenamiento con redundancia geográfica y restauración entre regiones

Para lograr la redundancia regional de los datos de copia de seguridad, use Backup para replicar las copias de seguridad en una región emparejada Azure mediante GRS. GRS protege las copias de seguridad frente a interrupciones regionales.

La región en la que se implementa el almacén se llama *región primaria*. Los orígenes de datos deben encontrarse en la región primaria. No se pueden configurar copias de seguridad en una bóveda de otra región.

La región emparejada también se conoce como *región secundaria*.



Si no configura GRS y se produce una interrupción en la región de la bóveda, es posible que pueda acceder a la bóveda y ver los elementos de copia de seguridad. Sin embargo, sin redundancia regional, los datos de copia de seguridad subyacentes permanecen no disponibles para las operaciones de restauración.

Restauración cruzada de regiones

Al configurar GRS en un almacén, Microsoft pone a disposición las copias de seguridad en la región emparejada después de que se produzca una interrupción en la región primaria. Si el origen de datos admite CRR, puede restaurar desde puntos de recuperación de regiones secundarias incluso cuando no se produzca ninguna interrupción en la región primaria. CRR también le permite ejecutar simulacros para evaluar la resistencia frente a interrupciones regionales. Al activar CRR, Microsoft actualiza el almacenamiento de copia de seguridad de GRS al almacenamiento con redundancia geográfica con acceso de lectura (RA-GRS).

Requisitos

- **Soporte de región:** GRS para Backup solo funciona dentro de regiones emparejadas de Azure.
- **Solo bóvedas nuevas:** Debe configurar GRS en su bóveda antes de realizar la primera copia de seguridad.

Consideraciones

- **CRR:** Después de activar CRR, los elementos de copia de seguridad pueden tardar hasta 48 horas en estar disponibles en la región secundaria.

Cost

Las bóvedas GRS incurren en costos adicionales para la replicación entre regiones y para el almacenamiento en la región secundaria. La transferencia de datos entre regiones de Azure se cobra en función de las tasas estándar de ancho de banda entre regiones. La CRR se cobra a una tarifa distinta porque Microsoft actualiza el almacenamiento del cofre de GRS a RA-GRS. Para obtener más información, consulte Precios de copia de seguridad .

Configuración de la compatibilidad con varias regiones

- **Cree un nuevo almacén que use GRS y CRR:** Al crear un almacén, también debe configurar la redundancia de almacenamiento. Después de seleccionar GRS, opcionalmente puede habilitar CRR en el almacén. Los pasos que siga dependen del tipo de bóveda. Para obtener más información, consulte los artículos siguientes:
  - Crear y eliminar almacenes de copias de seguridad
  - Creación y configuración de un almacén de Recovery Services
- **Configuración de GRS y CRR en almacenes existentes:** En el caso de los almacenes de Backup, debe configurar la redundancia de almacenamiento al crear el almacén.

En el caso de las bóvedas de Recovery Services, debe configurar la redundancia de almacenamiento antes de proteger cualquier carga de trabajo. Una vez protegida una carga de trabajo, la configuración está bloqueada y no se puede cambiar.

Puede habilitar CRR en almacenes GRS existentes. Después de habilitar CRR, no se puede deshabilitar.

Comportamiento cuando todas las regiones están en buen estado

En esta sección se describe qué esperar al configurar bóvedas para el uso de GRS cuando todas las regiones están operativas.

- **Operación entre regiones:** Las copias de seguridad siempre se completan en la región primaria, que es la región donde están desplegados el almacén y el origen de datos.
- **Replicación de datos entre regiones:** Cuando configuras el vault para usar GRS, las copias de seguridad se confirman primero en la región primaria mediante LRS. Después de finalizar correctamente en la región primaria, los datos se replican de forma asincrónica en la región secundaria. La región secundaria usa LRS para almacenar datos. Los datos de copia de seguridad pueden tardar hasta 12 horas en replicarse desde la región primaria a la región secundaria.

Comportamiento durante una falla de región

En esta sección se describe qué esperar al configurar bóvedas para usar GRS en caso de que ocurra una interrupción en la región primaria.

- **Detección y respuesta:** Para los orígenes de datos que admiten CRR y donde CRR está habilitado en la bóveda, puede iniciar su propio CRR en la región emparejada en cualquier momento, incluido durante una caída de la región o un desastre. Es responsable de detectar la interrupción y realizar acciones de recuperación, incluida la restauración de copias de seguridad en una región correcta.

  En todos los demás escenarios, los datos replicados en la región secundaria están disponibles para restaurarse en la región secundaria solo si Azure declara un desastre en la región primaria. Microsoft es responsable de declarar un desastre. La cantidad de tiempo que se tarda en declarar un desastre depende de la gravedad del incidente y del tiempo necesario para evaluar la situación. Microsoft normalmente declara un desastre solo después de un período de tiempo prolongado.

- **Notification**: Microsoft no le notifica automáticamente cuando una región está inactiva. Sin embargo:
  - Puede usar Azure Resource Health para supervisar el estado de un recurso individual y puede configurar Resource Health alertas para notificarle problemas.
  - Puede usar Azure Service Health para comprender el estado general del servicio, incluidos los errores de región, y puede configurar alertas de Service Health para notificarle los problemas.
- **Pérdida de datos esperada:** El RPO de los datos de copia de seguridad depende de varios factores, incluida la programación de copia de seguridad. En términos generales, para una interrupción regional, se puede esperar hasta 36 horas de pérdida de datos porque el RPO en la región primaria es de 24 horas, y puede tardar hasta 12 horas en replicar los datos de copia de seguridad de la región primaria a la región secundaria.
- **Tiempo de inactividad esperado:** El RTO es diferente para cada uno de los escenarios siguientes:
  - Es posible que los orígenes de datos y otros recursos de la región con errores no estén disponibles hasta que se recupere la región, por lo que el RTO no está definido.
  - Es posible que la copia de seguridad no pueda realizar operaciones de copia de seguridad o restauración en la región afectada hasta que la región se recupere, por lo que el RTO no está definido.
  - Si usa CRR, el RTO para iniciar la restauración de las copias de seguridad ya replicadas en la región emparejada es cero. Si no usa CRR, el RTO depende del tiempo que tarda Microsoft en declarar un desastre en la región afectada.
- **Redistribución:** No se pueden ejecutar trabajos de copia de seguridad mientras la región primaria está sin conexión. Puede restaurar datos en la bóveda, pero no puede agregar nuevos.

  Es responsable de restaurar la copia de seguridad en la infraestructura de la región emparejada y de volver a configurar equilibradores de carga, clientes y otros sistemas para redirigir el tráfico a una infraestructura correcta en la región emparejada.

Recuperación de regiones

Cuando se recupera la región primaria, Backup restaura automáticamente las operaciones en la región. Los trabajos se reanudan y los datos permanecen disponibles.

Prueba de fallos de región

Puede usar CRR para realizar una operación de restauración en la región emparejada. Puede usar este enfoque para comprobar la restauración y otros procesos de recuperación.

Resistencia a la pérdida de datos de copia de seguridad

La copia de seguridad proporciona dos características clave de recuperación para evitar la eliminación accidental o malintencionada de los datos de copia de seguridad:

- **La eliminación temporal** permite recuperar los objetos y almacenes eliminados durante un período de retención configurable. De forma predeterminada, este período es de 14 días, pero puede editarlo. Piensa en la eliminación suave como una papelera de reciclaje para tus copias de seguridad y bóvedas. Para obtener más información, consulte Seguro por defecto con la eliminación suave de Backup.

- **Los almacenes inmutables** pueden ayudarle a proteger los datos de copia de seguridad bloqueando las operaciones que podrían provocar la pérdida de puntos de recuperación. Puede bloquear la configuración del almacén inmutable para que sea irreversible. También puede usar el almacenamiento de escribir una vez, leer muchas veces (WORM) para los respaldos, para evitar que usuarios malintencionados deshabiliten la inmutabilidad y eliminen copias de seguridad. Para obtener más información, consulte almacén inmutable para copia de seguridad.

Acuerdo de nivel de servicio

El acuerdo de nivel de servicio (SLA) para Azure servicios describe la disponibilidad esperada de cada servicio y las condiciones que la solución debe cumplir para lograr esa expectativa de disponibilidad. Para obtener más información, vea SLAs for servicios en línea.

El Acuerdo de Nivel de Servicio de copia de seguridad cubre la disponibilidad del servicio para las operaciones de copia de seguridad y restauración. Para estar cubierto por el Acuerdo de Nivel de Servicio, debe reintentar los trabajos de copia de seguridad o restauración con errores al menos una vez cada 30 minutos.

Contenido relacionado

- Confiabilidad en Azure
- Reliability in Azure Virtual Machines
- Reliability in Azure Disk Storage
- Fiabilidad en Azure Site Recovery

ⓘ **Nota:** El autor ha creado este artículo con ayuda de inteligencia artificial. Más información

---

Last updated on 14/04/2026

Fiabilidad en Azure Batch

En este artículo se describe el soporte de confiabilidad en Azure Batch. Describe cómo mejorar la resistencia intrar regional mediante zonas de disponibilidad, grupos de lotes y nodos de proceso para minimizar el tiempo de inactividad y la pérdida de datos. También se vincula a información sobre la recuperación entre regiones y la continuidad empresarial.

Soporte para zonas de disponibilidad

Las zonas de disponibilidad son grupos físicamente independientes de centros de datos dentro de una región de Azure. Cuando una zona falla, los servicios pueden transferirse a una de las zonas restantes.

Batch mantiene la paridad con Azure con la compatibilidad de las zonas de disponibilidad.

Prerrequisitos

- Para las cuentas de Batch del modo de suscripción de usuario, asegúrese de que la suscripción en la que va a crear el grupo no tenga una restricción de oferta de zona en la SKU de máquina virtual solicitada. Para ver si la suscripción no tiene restricciones, llame a la API de enumeración de SKU de recursos y compruebe el campo `ResourceSkuRestrictions`. Si existe una restricción de zona, puede enviar una incidencia de soporte técnico para quitar la restricción de zona.

- Como InfiniBand no admite la comunicación entre zonas, no puede crear un grupo con una directiva de zona si tiene habilitada la comunicación entre nodos y usa una SKU de máquina virtual que admite InfiniBand.

- Batch mantiene la paridad con Azure con la compatibilidad de las zonas de disponibilidad. Para usar la opción de zona, el grupo se debe crear en una región de Azure compatible con las zonas de disponibilidad.

- Para asignar el grupo de Batch en las zonas de disponibilidad, la región de Azure en la que se creó el grupo debe admitir la SKU de máquina virtual solicitada en más de una zona. Para validar que la región admite la SKU de máquina virtual solicitada en más de una zona, llame a la API de enumeración de SKU de recursos y compruebe el campo `locationInfo` de `resourceSku`. Asegúrese de que se admite más de una zona para la SKU de máquina virtual solicitada. También puede usar la CLI de Azure para enumerar todas las SKU de recursos disponibles con el siguiente comando:

  Azure CLI

```
az vm list-skus
```

Creación de un grupo de Azure Batch en zonas de disponibilidad

Para obtener ejemplos sobre cómo crear un grupo de Batch en zonas de disponibilidad, consulte Creación de un grupo de Azure Batch en zonas de disponibilidad.

Obtenga más información acerca de cómo crear cuentas de Batch con Azure Portal, la CLI de Azure, Powershell o Batch Management API.

Experiencia a nivel de zona

Durante la interrupción de una zona, los nodos de esa zona dejan de estar disponibles. Los nodos de ese mismo grupo de nodos de otras zonas no se ven afectados y siguen estando disponibles.

La cuenta de Azure Batch no reasigna ni crea nodos nuevos para compensar los nodos que han quedado fuera de servicio debido a la interrupción. Los usuarios deben agregar más nodos al grupo de nodos, que luego se asignan desde otras zonas en estado correcto.

Tolerancia a errores

Para prepararse para un posible error de zona de disponibilidad, debe sobreaprovisionar la capacidad del servicio para asegurarse de que la solución puede tolerar un tercio de la pérdida de capacidad y seguir funcionando sin degradar el rendimiento durante interrupciones en toda la zona. Como la plataforma distribuye las máquinas virtuales entre tres zonas y debe tener en cuenta al menos el error de una de ellas, multiplique el número máximo de instancias de carga de trabajo por un factor de zonas/(zonas-1) o 3/2. Por ejemplo, si la carga de trabajo máxima típica necesita cuatro instancias, debería aprovisionar seis instancias: (2/3 * 6 instancias) = 4 instancias.

Migración de zonas de disponibilidad

No se puede migrar un grupo de Batch existente a la compatibilidad con la zona de disponibilidad. Si desea recrear el grupo de Batch entre zonas de disponibilidad, consulte Creación de un grupo de Azure Batch entre zonas de disponibilidad.

Recuperación ante desastres entre regiones y continuidad empresarial

Azure Batch está disponible en todas las regiones de Azure. Sin embargo cuando se crea una cuenta de Batch, se debe asociar con una región específica. Todas las operaciones posteriores para dicha cuenta de Batch solo se aplican a esa región. Por ejemplo, se crean los grupos y las máquinas virtuales (VM) asociadas en la misma región que la cuenta de Batch.

Al diseñar una aplicación que usa Batch, debe considerar la posibilidad de que Batch no esté disponible en una región. Es posible que encuentre una situación poco frecuente en la que haya un problema con la región como un todo, con el servicio completo de Batch en la región o con su cuenta de Batch específica.

Si la aplicación o solución que usa Batch siempre debe estar disponible, se debe diseñar para conmutarse por error a otra región o para que la carga de trabajo siempre se divida en dos o más regiones. Ambos enfoques requieren al menos dos cuentas de Batch y que cada cuenta se encuentre en una región distinta.

La responsabilidad de configurar la recuperación ante desastres entre regiones con Azure Batch es suya. Si ejecuta varias cuentas de Batch en regiones específicas y aprovecha las zonas de disponibilidad, la aplicación puede cumplir los objetivos de recuperación ante desastres cuando una de las cuentas de Batch deje de estar disponible.

Al proporcionar la capacidad de conmutación por error a una región alternativa, se deben tener en cuenta todos los componentes de una solución. No basta con tener simplemente una segunda cuenta de Batch. Por ejemplo, en la mayoría de las aplicaciones de Batch, se requiere una cuenta de almacenamiento de Azure. La cuenta de almacenamiento y la cuenta de Batch deben estar en la misma región para lograr un rendimiento aceptable.

Tenga en cuenta lo siguiente a la hora de diseñar una solución con conmutación por error:

- Cree previamente todos los servicios necesarios en cada región, como la cuenta de Batch y la de almacenamiento. No se suele aplicar ningún cargo por tener cuentas creadas y los cargos se generan solo cuando se usa la cuenta o se almacenan datos.
- Asegúrese con tiempo de que se establecen las cuotas adecuadas para todas las cuentas de Batch de **suscripción de usuario** para asignar el número necesario de núcleos mediante la cuenta de Batch.
- Use scripts o plantillas para automatizar la implementación de la aplicación en una región.

- Mantenga actualizados los archivos binarios de aplicación y los datos de referencia en todas las regiones. Mantenerlos actualizados garantizará que la región se pueda conectar rápidamente sin necesidad de esperar a que los archivos se carguen y se implementen. Por ejemplo, considere el caso en el que se almacena una aplicación personalizada para instalar en los nodos del grupo y se hace referencia a ella mediante paquetes de aplicación de Batch. Cuando se publica una actualización de la aplicación, se debe cargar en cada cuenta de Batch y hacer referencia a ella mediante la configuración del grupo (o convertir la versión más reciente en la versión predeterminada).

- En la aplicación que realice llamadas a Batch, al almacenamiento y a cualquier otro servicio, cambie los clientes o la carga con facilidad a regiones distintas.

- Considere la posibilidad de cambiar con frecuencia a una región alternativa como parte del funcionamiento normal. Por ejemplo, si tiene dos implementaciones en regiones distintas, cambie a la región alternativa cada mes.

La duración del tiempo de recuperación ante un desastre depende de la configuración que elija. Batch es independiente respecto a si usa varias cuentas o una sola cuenta. En las configuraciones de tipo activo-activo, en las que dos instancias de Batch reciben tráfico simultáneamente, la recuperación ante desastres es más rápida que en el caso de una configuración de tipo activo-pasivo. La configuración que elija debe basarse en las necesidades empresariales (regiones diferentes, requisitos de latencia) y las consideraciones técnicas.

Recuperación ante desastres de una sola región

La forma de implementar la recuperación ante desastres en Batch es la misma, tanto si trabaja en una zona geográfica de una sola región o de varias regiones. Las únicas diferencias son las SKU que usa para el almacenamiento y si tiene previsto usar la misma cuenta de almacenamiento o una diferente en todas las regiones.

Pruebas de recuperación ante desastres

Debe realizar sus propias pruebas de recuperación ante desastres de la solución habilitada para Batch. Se considera un procedimiento recomendado para permitir cambiar de forma sencilla entre la carga de cliente y servicio en diferentes regiones.

Probar el plan de recuperación ante desastres para Batch puede ser tan sencillo como alternar cuentas de Batch. Por ejemplo, podría confiar en una sola cuenta de Batch en una región específica durante un día operativo. A continuación, el día siguiente, podría cambiar a una segunda cuenta de Batch en otra región. La recuperación ante desastres se administra principalmente en el lado cliente. Este enfoque de varias cuentas para la recuperación ante

desastres se encarga de las expectativas de RTO y RPO en zonas geográficas de una sola región o de varias regiones.

Capacidad y resistencia proactiva de la recuperación ante desastres

Microsoft y sus clientes operan bajo el modelo de responsabilidad compartida. Microsoft es responsable de la resistencia de la plataforma y la infraestructura. La responsabilidad de abordar la recuperación ante desastres para cualquier servicio específico que implemente y controle es suya. Para asegurarse de que la recuperación sea proactiva:

- Siempre debe implementar previamente recursos secundarios. La implementación previa de recursos secundarios es necesaria porque no hay ninguna garantía de capacidad en el momento del impacto para aquellos que no han asignado previamente dichos recursos.
- Cree previamente todos los servicios necesarios en cada región, como las cuentas de Batch y las cuentas de almacenamiento asociadas. No se aplica ningún cargo por crear nuevas cuentas; solo se generan cargos cuando se usa la cuenta o se almacenan datos.
- Asegúrese de establecer las cuotas adecuadas en todas las suscripciones de antemano, de forma que pueda asignar el número necesario de núcleos mediante la cuenta de Batch. Al igual que en otros servicios de Azure, existen límites en determinados recursos asociados con el servicio Batch. Muchos de ellos son cuotas predeterminadas que Azure aplica en el nivel de cuenta o suscripción. Tenga en cuenta estas cuotas al diseñar y escalar verticalmente sus cargas de trabajo de Batch.

> ⓘ **Nota:**
>
> Si planea ejecutar cargas de trabajo de producción en Batch, es posible que tenga que aumentar el valor predeterminado de una o varias de las cuotas. Para generar una cuota, puede solicitar un aumento de la cuota sin cargo alguno. Para obtener más información, consulte **Solicitud de un aumento de cuota**.

Storage

Debe configurar el almacenamiento de Batch para asegurarse de que se realiza una copia de seguridad de los datos entre regiones; la responsabilidad del cliente es el enfoque predeterminado. La mayoría de las soluciones de Batch usan Azure Storage para almacenar los archivos de recursos y los archivos de salida. Por ejemplo, las tareas de Batch (incluidas las tareas estándar, las de inicio, las de preparación de trabajos y las de liberación de trabajos) especifican normalmente archivos de recursos que residen en cuentas de almacenamiento. Las

cuentas de almacenamiento también almacenan los datos que se procesan y los datos de salida que se generan. Comprender la posible pérdida de datos en las regiones de las operaciones de servicio es una consideración importante. También debe confirmar si los datos se pueden reescribir o si son de solo lectura.

Batch admite los siguientes tipos de cuentas de almacenamiento de Azure:

- Cuentas de uso general v2 (GPv2)
- Cuentas de uso general v1 (GPv1)
- Cuentas de Blob Storage (actualmente admitidas para grupos en la configuración de máquina virtual)

Para más información sobre las cuentas de almacenamiento, vea Introducción a las cuentas de Azure Storage.

Puede asociar una cuenta de almacenamiento con su cuenta de Batch cuando crea la cuenta o llevar a cabo este paso en otro momento.

Si va a configurar una cuenta de almacenamiento independiente para cada región en la que está disponible el servicio, debe usar cuentas de almacenamiento con redundancia de zona (ZRS). Use cuentas de almacenamiento con redundancia de zona geográfica (GZRS) si va a usar la misma cuenta de almacenamiento en varias regiones emparejadas. En el caso de las zonas geográficas que contienen una sola región, debe crear una cuenta de almacenamiento con redundancia de zona (ZRS), ya que el almacenamiento con redundancia de zona geográfica no está disponible.

El planeamiento de capacidad es otra consideración importante respecto al almacenamiento y debe abordarse de forma proactiva. Al elegir una cuenta de almacenamiento, tenga en cuenta los requisitos de costo y rendimiento. Por ejemplo, las opciones de cuenta GPv2 y de Blob Storage admiten mayores límites de capacidad y escalabilidad si se compara con GPv1. (Póngase en contacto con el servicio de soporte técnico de Azure para solicitar un aumento en el límite de almacenamiento). Estas opciones de cuenta pueden mejorar el rendimiento de las soluciones de Batch que contienen un gran número de tareas en paralelo que se leen o escriben en la cuenta de almacenamiento.

Cuando una cuenta de almacenamiento está vinculada a una cuenta de Batch, imagínese que se trata de la cuenta de almacenamiento automático. Se requiere una cuenta de almacenamiento automático si tiene previsto usar la funcionalidad de paquetes de aplicación, ya que se usa para almacenar los archivos ZIP de los paquetes de aplicación. También se puede usar una cuenta de almacenamiento automático para los archivos de recursos de tarea; dado que la cuenta de almacenamiento automático ya está vinculada a la cuenta de Batch, esto evita la necesidad de que las direcciones URL de firma de acceso compartido (SAS) accedan a los archivos de recursos.

Pasos siguientes

- Confiabilidad en Azure

Last updated on 22/01/2026

Confiabilidad en Azure Container Apps

Azure Container Apps es un servicio de hospedaje de contenedores sin servidor totalmente administrado para implementar microservicios y aplicaciones en contenedor.

Cuando se usa Azure, la confiabilidad es una responsabilidad compartida. Microsoft proporciona una variedad de funcionalidades para admitir resistencia y recuperación. Es responsable de comprender cómo funcionan esas funcionalidades dentro de todos los servicios que usa y de seleccionar las funcionalidades que necesita para cumplir los objetivos empresariales y los objetivos de tiempo de actividad.

En este artículo se describe cómo hacer que Las aplicaciones de contenedor sean resistentes a varias posibles interrupciones y problemas, incluidos errores transitorios, interrupciones de zona de disponibilidad, interrupciones de regiones y mantenimiento del servicio. También se describe cómo usar copias de seguridad para recuperarse de otros tipos de problemas y se resalta información clave sobre el acuerdo de nivel de servicio (SLA) de Container Apps.

Recomendaciones de implementación de producción

Para obtener información sobre cómo implementar Container Apps para admitir los requisitos de confiabilidad de la solución y cómo afecta la confiabilidad a otros aspectos de la arquitectura, consulte Procedimientos recomendados de arquitectura para Aplicaciones de contenedor en Azure Well-Architected Framework.

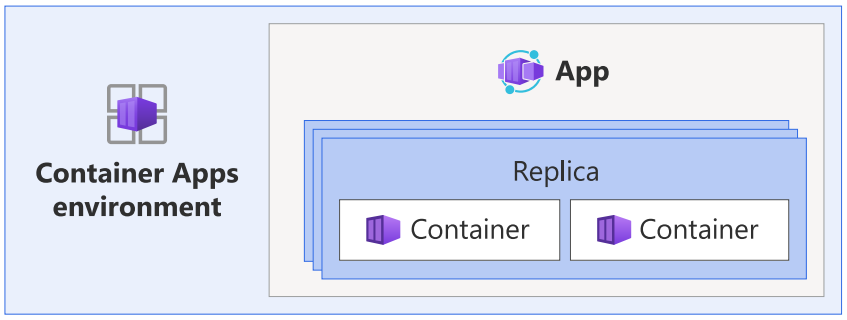
Introducción a la arquitectura de confiabilidad

Al usar Container Apps, implementará un *entorno* que actúa como unidad de implementación básica y define un límite seguro alrededor de un grupo de aplicaciones contenedoras. El entorno es donde se configuran los valores principales, incluida la compatibilidad de la zona de disponibilidad y la configuración de red. Los dos tipos de entornos son entornos de perfiles de carga de trabajo y entornos de solo consumo. Para obtener más información, consulte Estructuras de proceso y facturación en Container Apps.

Puede implementar varias *aplicaciones* en un único entorno. Cada aplicación ejecuta uno o varios *contenedores*. Un entorno también puede ejecutar uno o varios *trabajos*, que representan tareas no interactivas. Para obtener más información, consulte Contenedores en Aplicaciones de contenedor y trabajos en Aplicaciones de contenedor.

Cada aplicación tiene una o varias *réplicas*, que representan las instancias en ejecución de la aplicación. Puedes controlar cómo se escala la aplicación, incluido el número mínimo y máximo

de réplicas y cómo la aplicación agrega y quita réplicas dinámicamente. El planificador de la plataforma garantiza una distribución óptima entre hosts físicos al mismo tiempo que cumple con los requisitos mínimos de cantidad de réplicas. Para obtener más información, consulte Establecimiento de reglas de escalado en Container Apps.



Container Apps apoya la fiabilidad de las aplicaciones mediante diferentes capacidades.

- **Supervisión automática del estado:** El controlador de ingreso integrado equilibra automáticamente la carga del tráfico entre réplicas saludables. Si se produce un error en las comprobaciones de estado de una réplica o su infraestructura subyacente deja de estar disponible durante un tiempo prolongado, el servicio reinicia automáticamente los contenedores con errores o crea réplicas de reemplazo. También redistribuye el tráfico fuera de las réplicas no saludables y administra los reintentos de red en el clúster. Este proceso de recuperación automática no requiere ninguna intervención del cliente y mantiene el recuento de réplicas especificado. Para más información, consulte Sondeos de estado.

- **Resistencia de la aplicación a través de Dapr:** Container Apps proporciona una estrecha integración con Dapr, que es un marco que admite microservicios de nivel de producción y aplicaciones en contenedor. Dapr incluye características que ayudan a mejorar la resistencia, incluido el control de errores en otros servicios. Para obtener más información, consulte Microservicios con Aplicaciones de Contenedores.

- **Resistencia de la infraestructura para los componentes del sistema:** Esta resistencia incluye el plano de control, los controladores de entrada y el entorno de ejecución del contenedor. En las regiones que tienen zonas de disponibilidad, Container Apps proporciona redundancia de zona. Para obtener más información, consulte Resilience to availability zone failures (Resistencia a errores de zona de disponibilidad).

Resistencia a errores transitorios

Los errores transitorios son errores breves e intermitentes en los componentes. Se producen con frecuencia en un entorno distribuido como la nube y son una parte normal de las operaciones. Los errores transitorios se corrigen después de un breve período de tiempo. Es

importante que las aplicaciones puedan controlar errores transitorios, normalmente mediante el reintento de solicitudes afectadas.

Todas las aplicaciones hospedadas en la nube deben seguir las instrucciones de control de errores transitorios de Azure cuando se comunican con cualquier API, bases de datos y otros componentes hospedados en la nube. Para obtener más información, consulte Recomendaciones para controlar errores transitorios.

Container Apps gestiona automáticamente muchos errores transitorios a través de sus mecanismos de reintento a nivel de plataforma y supervisión de salud. Para asegurarse de que las aplicaciones son resistentes a errores transitorios, realice las siguientes acciones:

- **Configure sondeos de estado** que permitan a la plataforma detectar y responder a condiciones de error específicas de la aplicación. Establezca los umbrales de error y los valores de tiempo de espera adecuados en función de las características de inicio de la aplicación. Por ejemplo, para evitar reinicios prematuros del contenedor durante problemas temporales, use un umbral de fallos de 3 con un período de 10 segundos para los probes de vivacidad. Para más información, consulte Sondeos de estado.

- **Utilice directivas de resiliencia para el descubrimiento de servicios (versión preliminar)** para prevenir, detectar y recuperar proactivamente de los fallos en las solicitudes de servicio. Por ejemplo, cuando se usa una directiva de resistencia, cada solicitud entrante a la aplicación se puede reintentar automáticamente si hay un error transitorio que impide que la aplicación responda. Para más información, consulte Resistencia de detección de servicios (versión preliminar).

- **Implemente la lógica de reintento** en las aplicaciones para llamadas de servicio externas, conexiones de base de datos y solicitudes de API.

  Si la aplicación usa Dapr para integrarse con los servicios en la nube, use la resistencia de componentes de Dapr (versión preliminar) para configurar reintentos, tiempos de espera y disyuntores.

  Para otras dependencias, la aplicación debe controlar los errores transitorios. Use estrategias de retroceso exponencial y patrones de cortacircuitos al llamar a servicios externos para evitar fallos en cascada durante interrupciones en servicios posteriores. Las características integradas de detección y equilibrio de carga de Container Apps enrutan automáticamente el tráfico de las instancias con errores, pero las políticas de reintento a nivel de aplicación garantizan una gestión eficaz de manera elegante de los errores transitorios antes de que las comprobaciones de estado a nivel de plataforma desencadenen el reinicio del contenedor.

- **Diseñe trabajos para ser resistentes a errores transitorios**, incluidos los errores durante la ejecución del trabajo o en sus dependencias. Diseñe los trabajos para reanudarlos si se

reinician, o diseñe para la idempotencia de modo que se puedan ejecutar nuevamente de forma segura.

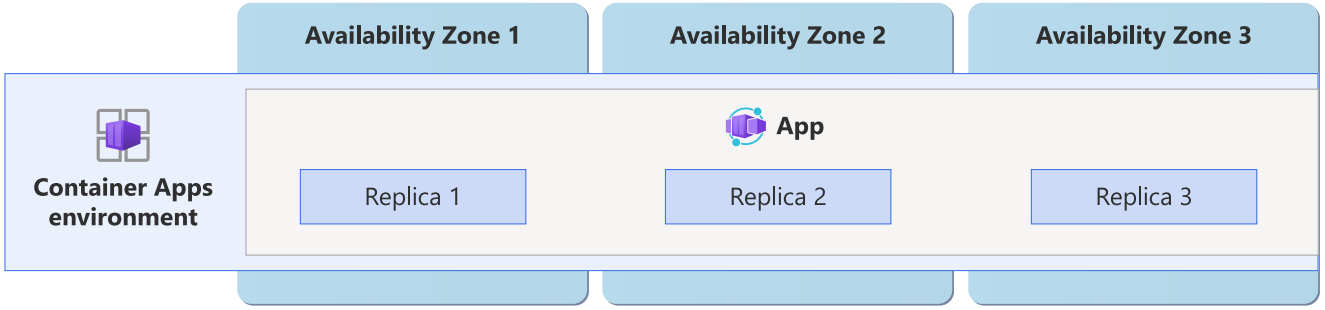
Resistencia a errores de zona de disponibilidad

Las zonas de disponibilidad son grupos físicamente independientes de centros de datos dentro de una región de Azure. Cuando una zona falla, los servicios pueden transferirse a una de las zonas restantes.

Al crear un entorno de Container Apps, puede habilitar la *redundancia de zona* para distribuir la infraestructura subyacente en varias zonas de disponibilidad en la región de Azure elegida. Container Apps programa automáticamente las réplicas de las aplicaciones entre zonas. Esta distribución se produce de forma transparente, lo que significa que no es necesario especificar la ubicación de zona para las réplicas individuales.

La redundancia de zona mejora la resistencia de la aplicación a errores de nivel de zona al garantizar que las réplicas de la aplicación contenedor se reparten entre varias zonas.

En el diagrama siguiente se muestra una aplicación contenedora con redundancia de zona con tres réplicas. Cada réplica se ejecuta en una zona de disponibilidad independiente.



Requisitos

- **Compruebe la compatibilidad con regiones.** La redundancia de zona está disponible en todas las regiones que admiten aplicaciones de contenedores y zonas de disponibilidad.

  Para ver qué regiones admiten zonas de disponibilidad, consulte Regiones de Azure con compatibilidad con zonas de disponibilidad.

  Para ver qué regiones admiten Container Apps, consulte Disponibilidad del producto por región.

- **Use perfiles de carga de trabajo.** La redundancia de zona está disponible para todos los planes de Aplicaciones de Contenedor, incluidos los perfiles de carga de trabajo de consumo y dedicado.

- **Habilite la redundancia de zona durante la creación del entorno.** Esta configuración no se puede cambiar después de crear el entorno.
- **Implemente un entorno de Container Apps en una red virtual.** La red virtual debe estar en una región que admita zonas de disponibilidad. Asegúrese de que la red virtual tiene una subred de tamaño adecuado. Los entornos de solo consumo necesitan una subred con un `/23` intervalo de enrutamiento de Inter-Domain sin clases (CIDR) o mayor, mientras que los entornos de perfiles de carga de trabajo requieren un `/27` intervalo CIDR o superior.
- **Establezca el número mínimo de réplicas en al menos dos para garantizar la distribución entre varias zonas de disponibilidad.** Considere la posibilidad de establecer un número mínimo de réplicas mayor si se aplica alguna de las condiciones siguientes:
  - La carga máxima esperada necesita más de dos réplicas.
  - Debe ser resistente a varias interrupciones de zona simultáneas.
  - Quieres minimizar el tiempo que esperas a que se creen nuevas réplicas en otras zonas durante una interrupción de zona.

Cost

No incurre en cargos adicionales más allá de los precios estándar de Container Apps al habilitar la redundancia de zona. Se pagan las mismas tarifas por los recursos de computación, las solicitudes y los segundos de vCore, independientemente de si la redundancia de zona está habilitada o no. Para más información, consulte Precios de Container Apps y facturación de Container Apps.

Configurar soporte de zonas de disponibilidad

- **Cree un entorno de Container Apps con zonas redundantes.** Para obtener instrucciones de implementación que abarcan Azure Portal, la CLI de Azure y Azure PowerShell, consulte Creación de una aplicación de contenedor con redundancia de zona.
- **Migre a una implementación redundante por zonas.** No se puede habilitar la redundancia de zona en un entorno de Container Apps existente. Para actualizar los entornos existentes que no tienen redundancia de zona, cree un nuevo entorno con redundancia de zona habilitada en una región admitida. A continuación, vuelva a implementar las aplicaciones de contenedor.
- **Desactivar la redundancia de zona.** La redundancia de zona no se puede deshabilitar después de habilitarla durante la creación del entorno. Si necesita una implementación

sin redundancia de zona, debe crear un nuevo entorno sin habilitar la opción de redundancia de zona ni implementarla en una región que no admita zonas de disponibilidad.

- **Compruebe la redundancia de zona.** Puede usar Azure Portal, la CLI de Azure y Azure PowerShell para comprobar el estado de redundancia de zona del entorno.

Planeamiento y administración de capacidad

Si una zona de disponibilidad deja de estar disponible, la plataforma Container Apps usa las reglas de escalado para decidir cuándo reemplazar las réplicas perdidas de esa zona. Es importante configurar correctamente las reglas de escalado para que el programador pueda tomar las decisiones de programación adecuadas.

Para configurar correctamente las reglas de escalado, siga estos principios:

- **Establezca un número mínimo de réplicas** que la aplicación pueda tolerar. Puede tardar un breve período de tiempo en reemplazar las réplicas perdidas porque la plataforma debe detectar que las réplicas antiguas han desaparecido. A continuación, las nuevas réplicas deben iniciarse y devolver un estado de sondeo de preparación correcto para poder recibir solicitudes entrantes. Si no puede tolerar ningún período con menos del número mínimo de réplicas que ha especificado, considere la posibilidad de sobreaprovisionar para mantener el rendimiento de la aplicación incluso si una zona deja de estar disponible.

- **Establezca las solicitudes de recursos y los límites** para ayudar al programador de Container Apps a tomar decisiones de ubicación óptimas entre zonas. Los requisitos de recursos infraespecificados pueden provocar errores de distribución o ubicación desiguales durante una carga alta.

Para obtener más información sobre las opciones de configuración, consulte Establecimiento de reglas de escalado.

Comportamiento cuando todas las zonas están en buen estado

En esta sección se describe qué esperar cuando los recursos de Container Apps están configurados para la redundancia de zona y todas las zonas de disponibilidad están operativas.

- **Enrutamiento de tráfico entre zonas:** Con Container Apps con redundancia de zonas, la plataforma funciona en un modelo activo-activo donde varias réplicas sirven tráfico simultáneamente. El controlador de entrada distribuye las solicitudes entrantes entre

todas las réplicas funcionales, independientemente de su zona, y utiliza el equilibrio de carga 'round robin' de forma predeterminada. Cada zona procesa solicitudes de forma independiente y la plataforma no da prioridad a ninguna zona específica para la distribución del tráfico. Los sondeos de salud se originan en todas las zonas para garantizar una evaluación precisa de la salud de cada réplica desde varias perspectivas.

- **Replicación de datos entre zonas:** Container Apps no replica los datos de la aplicación entre zonas porque está diseñado para cargas de trabajo sin estado. Los datos que la aplicación almacena en almacenamiento efímero, incluido el almacenamiento con ámbito de contenedor y el almacenamiento con ámbito de réplica, se eliminan cuando se cierra el contenedor o la réplica.

  Para los requisitos de datos con estado, monte un recurso compartido de archivos de Azure Files configurado para el almacenamiento con redundancia de zona o use otros servicios de Azure, como Azure Cosmos DB o Azure SQL Database, que proporcionan sus propias funcionalidades de replicación entre zonas.

  La plataforma solo replica los metadatos del plano de control, incluidas las configuraciones de la aplicación, las reglas de escalado y los secretos entre zonas para lograr una alta disponibilidad. Las imágenes de contenedor se extraen del registro de contenedor en cada zona según sea necesario cuando se crean réplicas.

Comportamiento durante un fallo de zona

En esta sección se describe qué esperar cuando los recursos de Container Apps están configurados para redundancia de zona y hay una interrupción en una zona de disponibilidad.

- **Detección y respuesta:** Azure detecta automáticamente errores de zona. Container Apps detiene inmediatamente la programación de nuevas réplicas en la zona fallida y comienza a redistribuir el tráfico a réplicas saludables en las zonas restantes. La plataforma gestiona todas las operaciones de conmutación por error automáticamente, sin necesidad de tu intervención.

- **Notificación:** Microsoft no le notifica automáticamente cuando una zona está inactiva. Sin embargo, puede usar Azure Service Health para comprender el estado general del servicio, incluidos los errores de zona, y puede configurar alertas de Service Health para notificarle problemas.

  También puede supervisar el estado de las aplicaciones a través de las métricas de Container Apps en Azure Monitor. Configure alertas en las caídas del recuento de réplicas y las tasas de error de solicitudes para recibir notificaciones inmediatas cuando se produzcan problemas relacionados con la zona.

- **Solicitudes activas:** Es posible que las solicitudes en vuelo a las réplicas de la zona fallida puedan eliminarse o experimenten tiempos de espera o errores de conexión. Las ejecuciones de trabajos en la zona afectada se anulan y se marcan como fallidas.

- **Pérdida de datos esperada:** No se produce ninguna pérdida de datos en el nivel de plataforma container Apps porque el servicio está diseñado para cargas de trabajo sin estado. Los datos almacenados en el almacenamiento efímero dentro de la zona de disponibilidad se pierden cuando finaliza la réplica y el almacenamiento efímero solo se debe usar para los datos temporales.

- **Tiempo de inactividad esperado:** Las aplicaciones experimentan un tiempo de inactividad mínimo o sin tiempo de inactividad durante los errores de zona. El impacto real depende de la configuración del sondeo de estado de la aplicación y del número de réplicas en zonas correctas. Asegúrese de que los clientes sigan las instrucciones de control de errores transitorios para minimizar los efectos.

  Los trabajos que se ejecutan en la zona afectada son abortados y se marcan como fallidos. Si necesita que un trabajo sea resistente a un error de zona, configure reintentos o configure paralelismo para que el trabajo ejecute varias copias de la misma ejecución. Para obtener más información, consulte Configuración avanzada del trabajo.

- **Reenrutamiento del tráfico:** Las sondas de estado del controlador de ingreso detectan rápidamente las réplicas inaccesibles y las eliminan del pool de equilibrio de carga. En función de la configuración del sondeo de estado de la aplicación, este proceso de conmutación por error suele producirse en unos 30 segundos. El tráfico entrante subsiguiente se distribuye entre las réplicas saludables restantes. Esta redirección de tráfico se produce de forma transparente a los clientes, que siguen usando la misma dirección URL de la aplicación.

  Si la afinidad de sesión está habilitada y una zona deja de funcionar, los clientes que se enrutaron previamente a réplicas de esa zona se enrutan a nuevas réplicas porque las réplicas anteriores ya no están disponibles. Se pierde cualquier estado asociado a las réplicas anteriores.

  No se iniciarán nuevas instancias de trabajos en la zona defectuosa.

- **Administración de instancias:** Es posible que se creen nuevas instancias de réplica en zonas saludables si las reglas de escalado automático se activan debido a un aumento en la carga.

Recuperación de zona

Cuando una zona de disponibilidad se recupera de un error, Container Apps vuelve a reinsertar automáticamente la zona en un servicio activo sin necesidad de intervención. Los sondeos de estado de la plataforma detectan cuándo la infraestructura de la zona recuperada está disponible y Container Apps comienza a programar nuevas réplicas en esa zona en función de la configuración de escalado. Las réplicas existentes en zonas saludables siguen atendiendo el tráfico durante el proceso de reintegración, lo que ayuda a evitar la interrupción del servicio.

Container Apps equilibra gradualmente la distribución de réplicas entre todas las zonas disponibles como parte de las operaciones de escalado estandarizados. Este reajuste automático ocurre cuando se crean réplicas en respuesta a eventos de escalado o cuando se reemplazan las réplicas defectuosas. La plataforma no fuerza la redistribución inmediata de réplicas correctas existentes, lo que evita reinicios innecesarios del contenedor y mantiene la estabilidad de la aplicación durante la recuperación.

Prueba de fallos de zona

La plataforma Container Apps administra el enrutamiento del tráfico, la conmutación por error y la conmutación por recuperación para aplicaciones de contenedor con redundancia de zona. Esta característica está totalmente administrada, por lo que no es necesario iniciar ni validar los procesos de error de zona de disponibilidad.

Para validar la resistencia de la aplicación a errores de zona, simule interrupciones de nivel de zona en el nivel de aplicación mediante enfoques de prueba controlados. Detenga o quite réplicas de zonas específicas mediante el escalado descendente de su aplicación y supervise cómo las réplicas restantes manejan la carga aumentada. Supervise las métricas clave durante las pruebas de resistencia, incluido el recuento de réplicas, las tasas de éxito de las solicitudes, los tiempos de respuesta y el comportamiento de escalado automático. Asegúrese de que el número mínimo de réplicas mantiene la disponibilidad del servicio cuando se eliminan las réplicas y compruebe que las reglas de escalado pueden manejar el aumento de carga en las réplicas restantes. Pruebe las configuraciones de sondeo de estado de salud fallando deliberadamente los puntos de conexión de estado para confirmar que la plataforma quita las instancias no saludables del ciclo de rotación durante los períodos de tiempo esperados.

Resistencia a errores en toda la región

Container Apps es un servicio de una sola región. Si la región deja de estar disponible, el entorno y las aplicaciones tampoco están disponibles.

Soluciones personalizadas de varias regiones para la resistencia

Para reducir el riesgo de un error de una sola región que afecte a la aplicación, puede implementar entornos en varias regiones. Los pasos siguientes ayudan a reforzar la resistencia:

- Implemente las aplicaciones en entornos de cada región. Cada entorno requiere su propia configuración de red virtual y los requisitos de subred se aplican de forma independiente a cada implementación regional. Las imágenes de contenedor deben estar disponibles en todas las regiones, lo que puede lograr mediante Azure Container Registry con la replicación geográfica habilitada.
- Configure las directivas de equilibrio de carga y conmutación por error usando servicios como Azure Front Door o Azure Traffic Manager.
- Replique los datos entre regiones para que pueda recuperar el último estado de la aplicación.

Copias de seguridad y restauración

Container Apps no proporciona funcionalidades de copia de seguridad integradas para las aplicaciones o los datos. Como plataforma de hospedaje de contenedores sin estado, Container Apps espera que las aplicaciones administren sus propias estrategias de persistencia y recuperación de datos a través de servicios externos. Los contenedores de aplicaciones y sus sistemas de archivos locales son efímeros y los datos almacenados localmente se pierden cuando las réplicas se reinician o se mueven.

Resistencia durante las actualizaciones de la aplicación

Use la administración de revisiones para implementar actualizaciones en la aplicación sin tiempo de inactividad. Puede crear nuevas revisiones con imágenes de contenedor actualizadas y realizar una transición mediante una estrategia de implementación azul-verde o cambiar gradualmente el tráfico mediante reglas de división de tráfico. Durante las actualizaciones de aplicaciones, la plataforma mantiene los recuentos mínimos de réplicas mediante la creación de nuevos contenedores antes de retirar los antiguos, lo que ayuda a evitar interrupciones del servicio.

Para obtener más información, consulte Actualización e implementación de cambios en Container Apps.

Resistencia al mantenimiento del servicio

Container Apps realiza el mantenimiento automático de la plataforma para aplicar actualizaciones de seguridad, implementar nuevas características y mejorar la confiabilidad del servicio. La plataforma usa actualizaciones graduales en dominios de error y zonas de disponibilidad para reducir la interrupción de las aplicaciones en ejecución. Durante las ventanas de mantenimiento, los contenedores siguen ejecutándose sin interrupción porque las actualizaciones se aplican a la infraestructura subyacente en fases.

Puede especificar sus propias ventanas de mantenimiento, que son períodos de tiempo que desea realizar el mantenimiento en las aplicaciones. Tenga en cuenta que las actualizaciones críticas pueden producirse fuera de las ventanas de mantenimiento. Para obtener más información, consulte Mantenimiento planeado de Container Apps.

Acuerdo de nivel de servicio

El contrato de nivel de servicio (SLA) para los servicios de Azure describe la disponibilidad esperada de cada servicio y las condiciones que la solución deberá cumplir para lograr esa expectativa de disponibilidad. Para obtener más información, consulte Acuerdos de Nivel de Servicio para servicios en línea.

El Acuerdo de Nivel de Servicio de disponibilidad para las Aplicaciones de Contenedor se basa en las reglas de escalado que establezca en las aplicaciones.

Contenido relacionado

- Introducción a Container Apps
- Sondas de estado en Container Apps
- Escalado automático en aplicaciones de contenedores
- Observabilidad en Aplicaciones de Contenedores
- Documentación sobre confiabilidad de Azure
- Guía de confiabilidad deWell-Architected Framework

ⓘ **Nota:** El autor ha creado este artículo con ayuda de inteligencia artificial. Más información

---

Last updated on 22/01/2026

Confiabilidad en Azure Container Instances

En este artículo se describe la compatibilidad con la confiabilidad en Azure Container Instances, que proporciona una manera sencilla de ejecutar contenedores de Linux o Windows en Azure, sin necesidad de administrar máquinas virtuales (VM) ni adoptar un servicio de nivel superior más complejo.

Al usar Azure, relibilidad es una responsabilidad compartida. Microsoft proporciona una variedad de funcionalidades para admitir resistencia y recuperación. Es responsable de comprender cómo funcionan esas funcionalidades dentro de todos los servicios que usa y de seleccionar las funcionalidades que necesita para cumplir los objetivos empresariales y los objetivos de tiempo de actividad.

En este artículo se describe cómo hacer que Azure Container Instances sean resistentes a una variedad de posibles interrupciones y problemas, incluidos errores transitorios, interrupciones de zona de disponibilidad y interrupciones de regiones. Resalta cierta información clave sobre el acuerdo de nivel de servicio (SLA) de Azure Container Instances.

Recomendaciones de implementación de producción para la confiabilidad

Para aumentar la confiabilidad de las aplicaciones de producción basadas en Container Instances, se recomienda realizar las siguientes acciones:

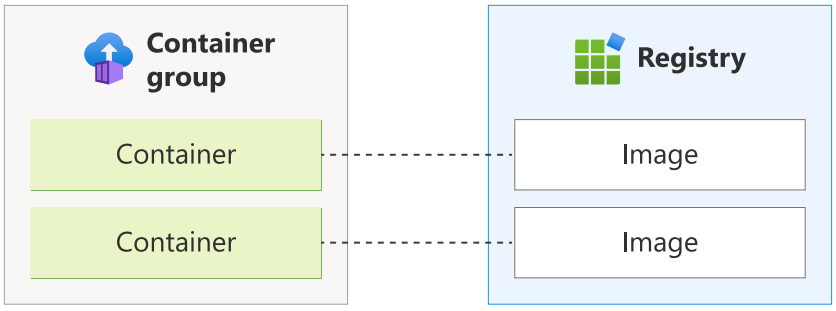
- ✓ Ejecute las aplicaciones en varias zonas de disponibilidad.
- ✓ Considere si también se deben ejecutar grupos de contenedores independientes en varias regiones.
- ✓ Use sondeos de actividad para detectar y automáticamente reiniciar contenedores no saludables.
- ✓ Utilice sondeos de preparación para esperar hasta que los contenedores estén listos antes de recibir tráfico.
- ✓ Use actualizaciones graduales para aplicar cambios progresivamente si usa NGroups. Este enfoque reduce la probabilidad de tiempo de inactividad debido a las actualizaciones.
- ✓ Revise las prácticas recomendadas y consideraciones para Container Instances.

Introducción a la arquitectura de confiabilidad

Para usar Container Instances, implemente un grupo *container*. Un grupo de contenedores contiene uno o varios *contenedores*. Cada contenedor se crea a partir de una imagen *container*, que se almacena en un registro como Azure Container Registry.

Todos los contenedores de un grupo de contenedores se implementan juntos como una sola unidad lógica y comparten la misma infraestructura física.

En el diagrama siguiente se muestra la relación entre grupos de contenedores, contenedores e imágenes.



Container Instances proporciona las siguientes características para administrar grupos de contenedores:

- NGroups (versión preliminar) proporciona un conjunto de funcionalidades para administrar varios grupos de contenedores relacionados. Al crear un NGroup, se define el número de grupos de contenedores para crear. Container Instances proporciona funcionalidades como lanzamientos de actualización automatizados y distribución de grupos de contenedores entre zonas de disponibilidad.
- Los grupos en espera crean un grupo de grupos de contenedores aprovisionados previamente que se pueden usar en respuesta al tráfico entrante. Los grupos en espera están concebidos para optimizar la creación de grupos de contenedores y no para aumentar la resiliencia.

Resistencia a errores transitorios

Los errores transitorios son errores breves e intermitentes en los componentes. Se producen con frecuencia en un entorno distribuido como la nube y son una parte normal de las operaciones. Los errores transitorios se corrigen después de un breve período de tiempo. Es importante que las aplicaciones puedan controlar errores transitorios, normalmente mediante el reintento de solicitudes afectadas.

Todas las aplicaciones hospedadas en la nube deben seguir las instrucciones de control de errores transitorios Azure cuando se comunican con cualquier API, bases de datos y otros componentes hospedados en la nube. Para obtener más información, consulte Recomendaciones para controlar errores transitorios.

Los SDK proporcionados por Microsoft suelen controlar errores transitorios. Dado que hospeda sus propias aplicaciones en Container Instances, siga los pasos necesarios para reducir la posibilidad de errores transitorios:

- Ejecute varios grupos de contenedores para cargas de trabajo importantes para asegurarse de que un error en un contenedor o grupo de contenedores no afecte a toda la aplicación.
- Construye el código de tu aplicación para que sea resistente a errores transitorios en los servicios a los que te conectas, por ejemplo, utilizando políticas de reintentos con estrategias de espera incremental.

Para obtener más información sobre otros errores que pueden producirse en tiempo de ejecución y cómo responder a ellos, consulte Problemas durante el tiempo de ejecución del grupo de contenedores.

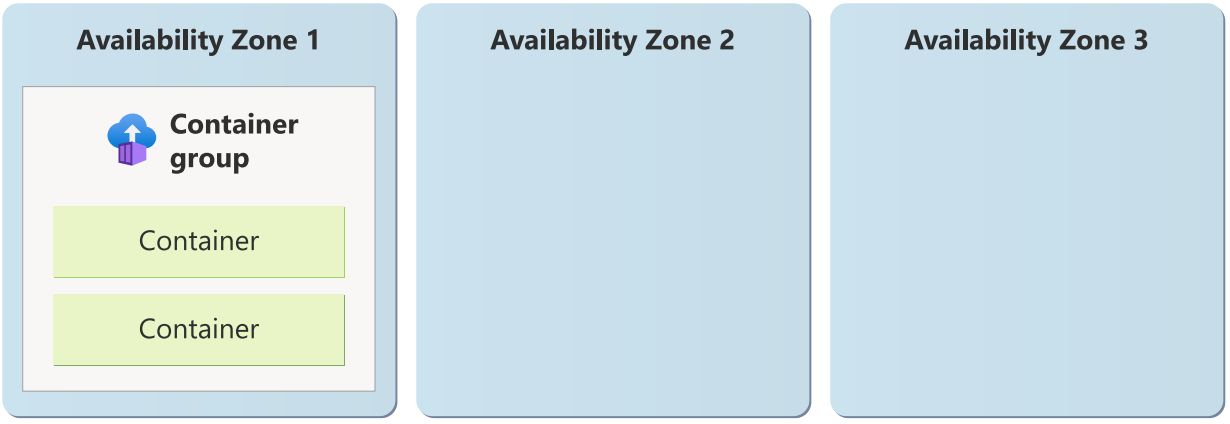
Resistencia a errores de zona de disponibilidad

Availability zones son grupos de centros de datos físicamente independientes dentro de una región de Azure. Cuando una zona falla, los servicios pueden transferirse a una de las zonas restantes.

Container Instances admite zonas de disponibilidad de diferentes maneras, en función de cómo implemente los grupos de contenedores:

- **Grupos de contenedores creados manualmente:** Un grupo de contenedores individual es un recurso *zonal* , lo que significa que se puede implementar en una sola zona de disponibilidad que seleccione. Todos los contenedores del grupo se despliegan en la misma zona de disponibilidad. Si esa zona de disponibilidad tiene una interrupción, el grupo de contenedores y todos sus contenedores podrían experimentar tiempo de inactividad.

  En el diagrama siguiente se muestra un grupo de contenedores que se ha implementado manualmente en la zona de disponibilidad 1:

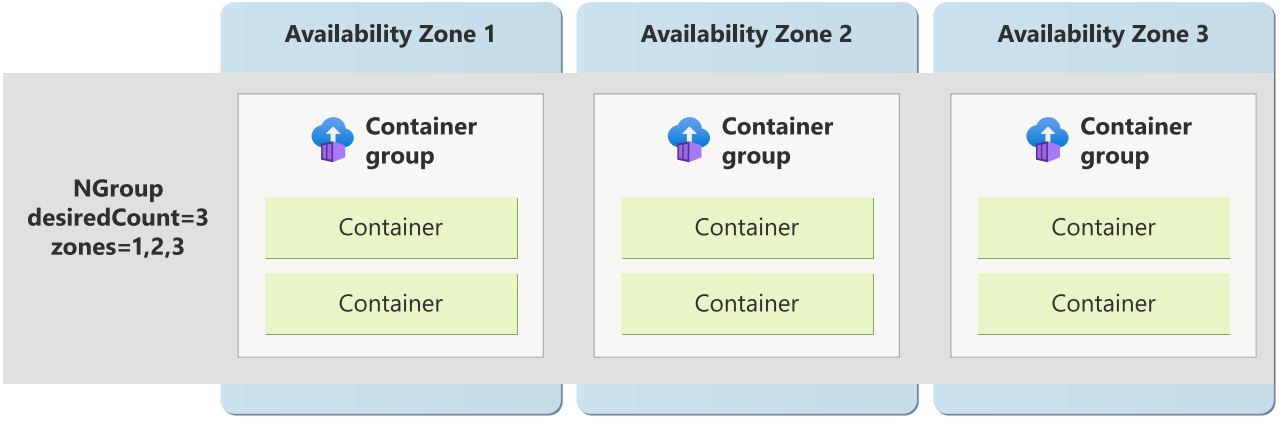


> ⓘ **Nota:**
>
> Para asegurarse de que la aplicación sigue ejecutándose cuando cualquier zona única de la región experimenta una interrupción, se recomienda crear un mínimo de dos grupos de contenedores en dos zonas de disponibilidad diferentes.

Si no especifica zonas de disponibilidad que se van a usar para el grupo de contenedores, es *no zonal* o *regional*, lo que significa que podría colocarse en cualquier zona de disponibilidad dentro de la región o dentro de la misma zona. Si alguna zona de disponibilidad de la región tiene un problema, el grupo de contenedores podría experimentar tiempo de inactividad.

- **NGroups:** Al implementar un NGroup, puede especificar una o varias zonas en las que implementarla. Si implementa un NGroup en dos o más zonas, se trata de un NGroup *con redundancia de zona* y una interrupción de una zona de disponibilidad solo provoca problemas para los grupos de contenedores dentro de la zona afectada.

  En el diagrama siguiente se muestra un NGroup que se implementa en tres zonas de disponibilidad:



Si no especifica zonas de disponibilidad para su NGroup, será no zonal y podría experimentar tiempo de inactividad si alguna zona de disponibilidad en la región presenta un problema.

- **Grupos en espera:** Al implementar un grupo en espera, puede especificar opcionalmente una o varias zonas. La plataforma puede solicitar contenedores en las zonas que seleccione.

  Sin embargo, los grupos en espera no tienen redundancia de zona ni resistencia de zona porque no hay garantía de que los contenedores se creen en varias zonas. Si se produce una interrupción de zona, es posible que todos los contenedores del grupo se coloquen en la zona afectada.

  Debido a que los grupos de respaldo no están diseñados para soportar la resiliencia ante fallos de zona, esta guía no describe el comportamiento detallado de los grupos de respaldo con zonas de disponibilidad.

  > ⓘ **Importante**
  >
  > Los grupos en espera no se diseñaron para tener resistencia de zona. No deben usarse para cargas de trabajo que requieran resistencia a errores de zona.

Requisitos

- **Compatibilidad con regiones:** Las implementaciones de grupos de contenedores zonales se admiten en todas las regiones con zonas de disponibilidad.
- Las implementaciones zonales están disponibles para Linux y Windows Server 2019 grupos de contenedores.
- Para seleccionar una zona de disponibilidad, debe usar la SKU estándar. Los grupos de contenedores zonales no están disponibles con la SKU confidencial.

Consideraciones

Los contenedores de acceso puntual no admiten zonas de disponibilidad y siempre son no zonales.

Cost

No hay ningún costo adicional para configurar zonas de disponibilidad para un grupo de contenedores.

Configurar soporte de zonas de disponibilidad

- **Cree grupos de contenedores con soporte para zonas de disponibilidad.** El enfoque que se usa para configurar zonas de disponibilidad depende de cómo cree grupos de contenedores.
  - *Grupos de contenedores creados manualmente:* Para crear un grupo de contenedores zonales en una zona específica, puede usar uno de los métodos siguientes:
    - El portal Azure
    - El CLI de Azure
    - Azure PowerShell
    - Bicep
    - Una plantilla de Azure Resource Manager (plantilla de ARM)
    - Terraform
  - *NGroups:* Puede implementar un NGroup con redundancia de zona mediante un archivo de Bicep o una plantilla de ARM y especificando varias zonas. Para obtener más información, consulte ejemplo de NGroups con zonas.
  - *Grupos en espera:* Puede implementar un grupo en espera que use zonas de disponibilidad especificando una o varias zonas al crear o actualizar el grupo. Sin embargo, es posible que los contenedores no se creen en varias zonas. Los pools de reserva no deben usarse para cargas de trabajo que requieran tolerancia a fallos de zona. Para obtener más información, consulte Crear un grupo en espera para Container Instances.
- **Habilitar soporte para zonas de disponibilidad en los recursos existentes.** El enfoque que se usa para configurar zonas de disponibilidad depende de cómo cree grupos de contenedores.
  - *Grupos de contenedores creados manualmente:* No se pueden habilitar zonas de disponibilidad en un grupo de contenedores nozonal existente. Debe eliminar el grupo de contenedores y crear un grupo de contenedores zonal.
  - *NGroups:* No se pueden habilitar zonas de disponibilidad en un NGroup no zonal existente. Debe eliminar el NGroup y crear un NGroup con redundancia zonal.

- *Grupos en espera:* No se pueden habilitar zonas de disponibilidad en un grupo de espera no zonal existente. Debe eliminar el grupo de contenedores y crear un nuevo grupo de espera que use varias zonas de disponibilidad.

- **Mueva grupos de contenedores entre zonas o deshabilite el soporte para zonas de disponibilidad.** El enfoque que se usa para modificar zonas de disponibilidad depende de cómo cree grupos de contenedores.

  - *Grupos de contenedores creados manualmente:* Para cambiar la zona de disponibilidad de un grupo de contenedores, debe eliminar el grupo de contenedores y crear otro grupo de contenedores con la nueva zona de disponibilidad. Para obtener información sobre cómo eliminar el grupo de contenedores, consulte los siguientes recursos:
    - El portal Azure
    - El CLI de Azure
    - Azure PowerShell
    - Bicep
    - Una plantilla de ARM
    - Terraform
    - La CLI de Docker

  - *NGroups:* Puede agregar zonas a un NGroup, pero no puede quitar zonas.

  - *Grupos en espera:* Puede agregar zonas a un grupo en espera, pero no puede quitar zonas.

Distribución de grupos de contenedores

La forma en que los grupos de contenedores se distribuyen entre zonas de disponibilidad depende de cómo implemente los grupos de contenedores.

- **Grupos de contenedores creados manualmente:** Es responsable de distribuir los grupos de contenedores creados manualmente en varias zonas de disponibilidad.

- **NGroups:** Durante las operaciones de escalado horizontal, NGroups elimina aleatoriamente instancias, lo que podría no mantener una distribución entre zonas de disponibilidad. Las operaciones de escalabilidad horizontal intentan reequilibrar la distribución entre zonas.

- **Grupos en espera:** Un grupo en espera puede crear contenedores en cualquiera de las zonas de disponibilidad que configure en el grupo. Sin embargo, es posible que los

contenedores no se creen en varias zonas. Los pools de reserva no deben usarse para cargas de trabajo que requieran tolerancia a fallos de zona.

Planeamiento y administración de capacidad

** Para prepararse para un fallo de zona de disponibilidad, considere el sobreaprovisionamiento del número de grupos de contenedores que despliegue. Este enfoque permite que la solución tolere cierta pérdida de capacidad y siga funcionando sin que se vea afectado su rendimiento. Para obtener más información, consulte Administración de la capacidad mediante el sobreaprovisionamiento.

El enfoque que se usa para sobreaprovisionar grupos de contenedores depende de cómo implemente los grupos de contenedores.

- **Grupos de contenedores creados manualmente:** Es responsable de planear la capacidad de los grupos de contenedores creados manualmente, incluido el planeamiento del número de grupos de contenedores que se van a implementar en cada zona.
- **NGroup:** Considere *la posibilidad de sobreaprovisionar* la capacidad de NGroup para tolerar la pérdida de una zona.
- **Grupos en espera:** Los grupos en espera no están diseñados para ser resistentes a los errores de zona. Considere la posibilidad de usar varios grupos en espera en diferentes zonas o usar NGroups.

Comportamiento cuando todas las zonas están en buen estado

En esta sección se describe qué esperar cuando los recursos de Container Instances están configurados para ser compatibles con zonas de disponibilidad y todas las zonas de disponibilidad están operativas.

- **Enrutamiento de tráfico entre zonas:** Usted es responsable de dirigir el tráfico a sus contenedores. Por ejemplo, puede usar Azure Application Gateway como puerta de enlace y equilibrador de carga para los grupos de contenedores.

  Si usa grupos de NGroup o grupos en espera, es responsable del equilibrio de carga en cada contenedor. También debe configurar el sistema de enrutamiento de tráfico para detectar el estado de cada grupo de contenedores y redirigir el tráfico a un grupo alternativo cuando sea necesario.

- **Replicación de datos entre zonas:** Los contenedores y los grupos de contenedores no tienen estado. Puede adjuntar su propio recurso compartido de archivos o conectarse a bases de datos u otros servicios de almacenamiento desde sus aplicaciones. Usted es responsable de asegurar que los recursos compartidos de archivos y los servicios de almacenamiento sean resilientes por zonas. Revise las guías de confiabilidad de cada servicio para comprender cómo hacer que cada zona de componentes sea resistente.

Comportamiento durante un fallo de zona

En esta sección se describe qué esperar cuando los recursos de Container Instances están configurados para admitir la zona de disponibilidad y se produce una interrupción en esta.

- **Detección y respuesta:** La responsabilidad de la detección de errores de zona y la respuesta asociada dependen de cómo implemente los grupos de contenedores.
  - *Grupos de contenedores creados manualmente:* debe detectar la pérdida de una zona de disponibilidad e iniciar una conmutación por error a un grupo de contenedores alternativo que ha creado en otra zona de disponibilidad.
  - *NGroups:* La plataforma Container Instances es responsable de detectar un error en una zona de disponibilidad y responder.

    Sin embargo, es responsable de garantizar que el tráfico se enrute a los contenedores en una zona saludable.
  - *Standby pools:* No se garantiza que la plataforma de Container Instances responda a errores de zona para pools en espera. Los pools de reserva no deben usarse para cargas de trabajo que requieran tolerancia a fallos de zona.
- **Notificación:** Microsoft no le notifica automáticamente cuando una zona está inactiva. Sin embargo, puede usar Azure Service Health para comprender el estado general del servicio, incluidos los errores de zona, y puede configurar alertas de Service Health para notificarle problemas.
- **Solicitudes activas:** Si se produce un error en una zona, es probable que todos los contenedores que se ejecutan en esa zona se detengan, incluido cualquier trabajo activo que controlen.
- **Pérdida de datos esperada:** Dado que los contenedores y los grupos de contenedores no tienen estado, no se espera ninguna pérdida de datos debido a un fallo de zona. Sin

embargo, es responsable de garantizar que cada componente de la carga de trabajo sea resistente a la zona, incluidos los servicios de almacenamiento y las bases de datos.

- **Tiempo de inactividad esperado:** El tiempo de inactividad que puede esperar de un error de zona depende de cómo implemente los grupos de contenedores.
  - *Grupos de contenedores creados manualmente:* En el caso de los grupos de contenedores zonales, cuando una zona no está disponible, el grupo de contenedores y sus contenedores no están disponibles hasta que se recupere la zona de disponibilidad.
  - *NGroups:* En el caso de NGroups, si una zona deja de funcionar, la aplicación sigue estando disponible porque los grupos de contenedores restantes de NGroups siguen ejecutándose en otras zonas. No se espera ningún tiempo de inactividad.
  - *Grupos en espera:* Los grupos en espera no proporcionan resistencia de zona. Si todos los grupos de contenedores del grupo en espera se encuentran en una sola zona de disponibilidad, es posible que todos los grupos de contenedores y sus contenedores queden inaccesibles hasta que se recupere la zona de disponibilidad.
- **Reenrutamiento del tráfico:** Dado que es responsable del enrutamiento del tráfico a los contenedores, también es responsable de volver a enrutar el tráfico si se produce un error en un grupo de contenedores debido a una interrupción de zona de disponibilidad.

Recuperación de zona

Una vez recuperada la zona, la plataforma Azure reinicia automáticamente los grupos de contenedores que se habían detenido. No se requiere ninguna acción del cliente.

Prueba de fallos de zona

No hay ninguna manera de simular una interrupción de la zona de disponibilidad que contiene el grupo de contenedores. Sin embargo, puede configurar manualmente puertas de enlace ascendentes o equilibradores de carga para redirigir el tráfico a otro grupo de contenedores en una zona de disponibilidad diferente.

Resistencia a errores en toda la región

Container Instances es un servicio de una sola región. Si la región deja de estar disponible, los grupos de contenedores y sus contenedores tampoco están disponibles.

Soluciones de varias regiones personalizadas para la resistencia

Opcionalmente, puede implementar grupos de contenedores independientes en varias regiones. Es responsable de implementar y configurar los grupos de contenedores en cada región. También debe configurar el equilibrio de carga mediante un servicio como Azure Traffic Manager o Azure Front Door. Usted es responsable de cualquier sincronización, conmutación por error y recuperación después de falla de datos.

Acuerdo de nivel de servicio

El acuerdo de nivel de servicio (SLA) para Azure servicios describe la disponibilidad esperada de cada servicio y las condiciones que la solución debe cumplir para lograr esa expectativa de disponibilidad. Para obtener más información, vea SLAs for servicios en línea.

Contenido relacionado

- Confiabilidad en Azure

ⓘ**Nota:** El autor ha creado este artículo con ayuda de inteligencia artificial. Más información

---

Last updated on 22/01/2026

Confiabilidad en Azure Container Registry

Azure Container Registry es un servicio de registro de contenedor administrado que se usa para almacenar y administrar las imágenes de contenedor privadas de Docker y los artefactos relacionados para las implementaciones de contenedor.

Cuando se usa Azure, la confiabilidad es una responsabilidad compartida. Microsoft proporciona una variedad de funcionalidades para admitir resistencia y recuperación. Es responsable de comprender cómo funcionan esas funcionalidades dentro de todos los servicios que usa y de seleccionar las funcionalidades que necesita para cumplir los objetivos empresariales y los objetivos de tiempo de actividad.

En este artículo se describe cómo hacer que Container Registry sea resistente a una variedad de posibles interrupciones y problemas, incluidos errores transitorios, interrupciones de zona de disponibilidad y interrupciones de regiones. También se describe cómo puede usar copias de seguridad para recuperarse de otros tipos de problemas y se resalta cierta información clave sobre el acuerdo de nivel de servicio (SLA) de Container Registry.

Recomendaciones de implementación de producción para la confiabilidad

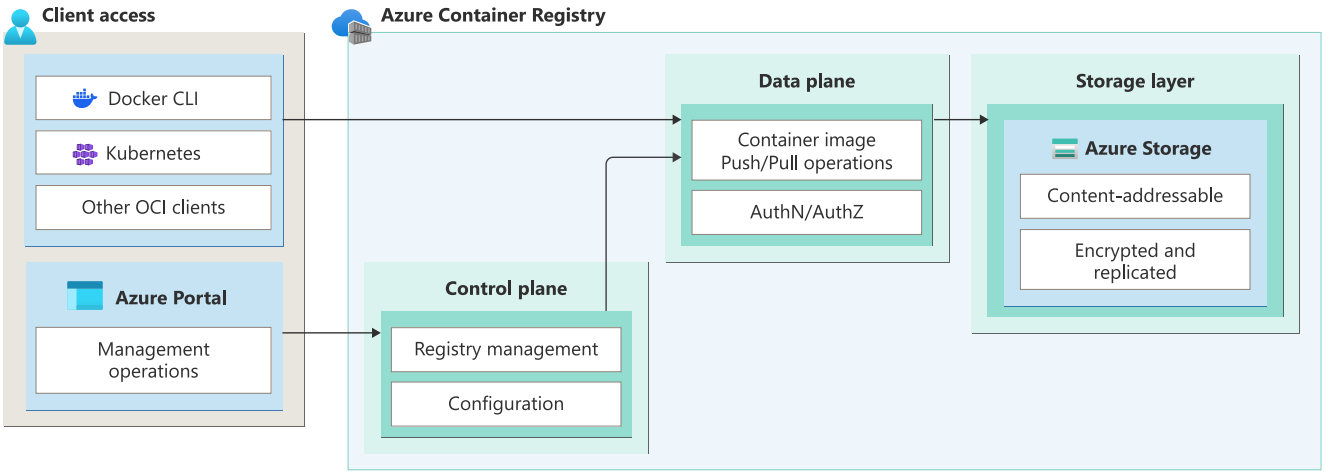
Para las cargas de trabajo de producción, le recomendamos que realice las siguientes acciones:

- Use el nivel Premium de Container Registry, que proporciona las características de confiabilidad más completas. El nivel Premium también proporciona límites de rendimiento más altos, características de seguridad mejoradas y funcionalidades avanzadas que son esenciales para las cargas de trabajo de contenedor de producción. Para más información sobre los niveles de servicio y las características, consulte Niveles de servicio de Container Registry.
- Aprovisione su registro de contenedores en una región que admita zonas de disponibilidad.
- En escenarios de varias regiones, configure la replicación geográfica para distribuir el registro en varias regiones en función de sus requisitos geográficos y de cumplimiento específicos.

Introducción a la arquitectura de confiabilidad

Container Registry se basa en la infraestructura distribuida de Azure para proporcionar alta disponibilidad y durabilidad de los datos. El servicio consta de varios componentes clave que

funcionan conjuntamente para garantizar la confiabilidad. En el siguiente diagrama se muestra la arquitectura de servicio principal.



- **El plano de control** es un componente de administración centralizado en la región principal para la configuración del registro, la configuración de autenticación y las directivas de replicación.
- **El plano de datos** es un servicio distribuido que controla las operaciones de inserción y extracción de imágenes de contenedor entre regiones y zonas de disponibilidad.
- **La capa de almacenamiento** es una solución de Azure Storage direccionable con contenido que conserva imágenes y artefactos de contenedor. Incluye desduplicación automática, cifrado en reposo y replicación integrada.

Microsoft es responsable de administrar la infraestructura subyacente de Container Registry, que incluye los siguientes tipos de mantenimiento:

- **Mantenimiento del plano de control** para la administración del registro
- **Mantenimiento del plano de datos** para las operaciones de imagen de contenedor en regiones y zonas de disponibilidad

Como cliente, es responsable de las siguientes acciones:

- **Resistencia de nivel de aplicación:** Implemente la lógica de reintento adecuada y el control de conmutación por error en las aplicaciones de contenedor y las plataformas de orquestación.
- **Configuración de resistencia de zona:** Asegúrese de que el registro de contenedor está implementado en una región que admita zonas de disponibilidad.
- *Configuración de replicación geográfica:* Elija las regiones adecuadas para la replicación geográfica en función de los requisitos de distribución geográfica, cumplimiento y rendimiento.

Container Registry también admite *tareas*, lo que puede ayudarle a automatizar las compilaciones de contenedor y las operaciones de mantenimiento. Las tareas se ejecutan en una infraestructura informática administrada por Microsoft y admiten desencadenadores manuales, basados en eventos o programados. Para más información, consulte Automatizar la creación y el mantenimiento de imágenes de contenedores mediante tareas de Container Registry.

> ⓘ **Nota:**
>
> Container Registry admite registros conectados, que son réplicas locales o remotas que se sincronizan con su Container Registry basado en la nube. Al usar registros conectados, es responsable de configurarlos para cumplir los requisitos de confiabilidad. Los registros conectados quedan fuera del ámbito de este artículo.

Resistencia a errores transitorios

Los errores transitorios son errores breves e intermitentes en los componentes. Se producen con frecuencia en un entorno distribuido como la nube y son una parte normal de las operaciones. Los errores transitorios se corrigen después de un breve período de tiempo. Es importante que las aplicaciones puedan controlar errores transitorios, normalmente mediante el reintento de solicitudes afectadas.

Todas las aplicaciones hospedadas en la nube deben seguir las instrucciones de control de errores transitorios de Azure cuando se comunican con cualquier API, bases de datos y otros componentes hospedados en la nube. Para obtener más información, consulte Recomendaciones para controlar errores transitorios.

Container Registry controla los errores transitorios internamente a través de varios mecanismos. El servicio implementa una lógica de reintento automático para las operaciones del registro y mantiene un grupo de conexiones para un uso eficiente de los recursos. Las operaciones de Container Registry están diseñadas para ser idempotentes, lo que permite reintentos seguros de las operaciones de envío y extracción. Las tareas administran automáticamente los fallos transitorios cuando realizan muchos tipos de operaciones.

En el caso de las aplicaciones cliente que usan Container Registry, implemente las directivas de reintento adecuadas con retroceso exponencial al realizar operaciones del Registro. Use el cliente oficial de Docker o los SDK de Container Registry, que incluyen mecanismos de reintento integrados para los errores transitorios más comunes.

Resistencia a errores de zona de disponibilidad

Las zonas de disponibilidad son grupos físicamente independientes de centros de datos dentro de una región de Azure. Cuando una zona falla, los servicios pueden transferirse a una de las zonas restantes.

La redundancia de zona protege su registro de contenedores contra errores de una sola zona distribuyendo los datos y las operaciones del registro entre varias zonas de disponibilidad dentro de la región. Las operaciones de extracción e inserción de imágenes de contenedor continúan funcionando durante las interrupciones de zona, con conmutación automática por error a zonas correctas.

La redundancia de zona está habilitada de forma predeterminada para todos los registros de regiones que admiten zonas de disponibilidad, lo que hace que los recursos sean más resistentes automáticamente y sin costo adicional. Esta mejora se aplica a todos los niveles de servicio, incluidos Básico y Estándar, y se ha aplicado a registros nuevos y existentes.

> ⓘ **Importante**
>
> Es posible que Azure Portal y otras herramientas aún no reflejen la actualización de redundancia de zona con precisión.
>
> La `zoneRedundancy` propiedad de la configuración del registro podría seguir apareciendo como `false`, pero la redundancia de zona está activa para todos los registros de las regiones admitidas.
>
> Estamos actualizando activamente las superficies del portal y de la API para reflejar este comportamiento predeterminado de forma más transparente. Todas las características habilitadas anteriormente siguen funcionando según lo previsto.

Consideraciones

- **Compatibilidad con regiones:** Los registros con redundancia de zona se pueden implementar en cualquier región que admita zonas de disponibilidad. Si su registro está en una región que no admite zonas de disponibilidad, para que sea redundante en zonas, debe crear un nuevo registro en una región que admita zonas de disponibilidad. A continuación, debe migrar las imágenes de contenedor mediante la creación de una canalización de transferencia o mediante la importación de imágenes de contenedor.

Consideraciones

- **Tareas:** Las tareas de Container Registry no admiten actualmente zonas de disponibilidad. La redundancia de zona se aplica al propio servicio del registro, pero no a las tareas ni a

sus operaciones.

- **Replicación geográfica:** Si el registro usa replicación geográfica, las réplicas creadas en regiones con zonas de disponibilidad son automáticamente redundantes a nivel de zona.

Cost

La redundancia de zona se incluye con registros de contenedor sin coste adicional.

Configurar soporte de zonas de disponibilidad

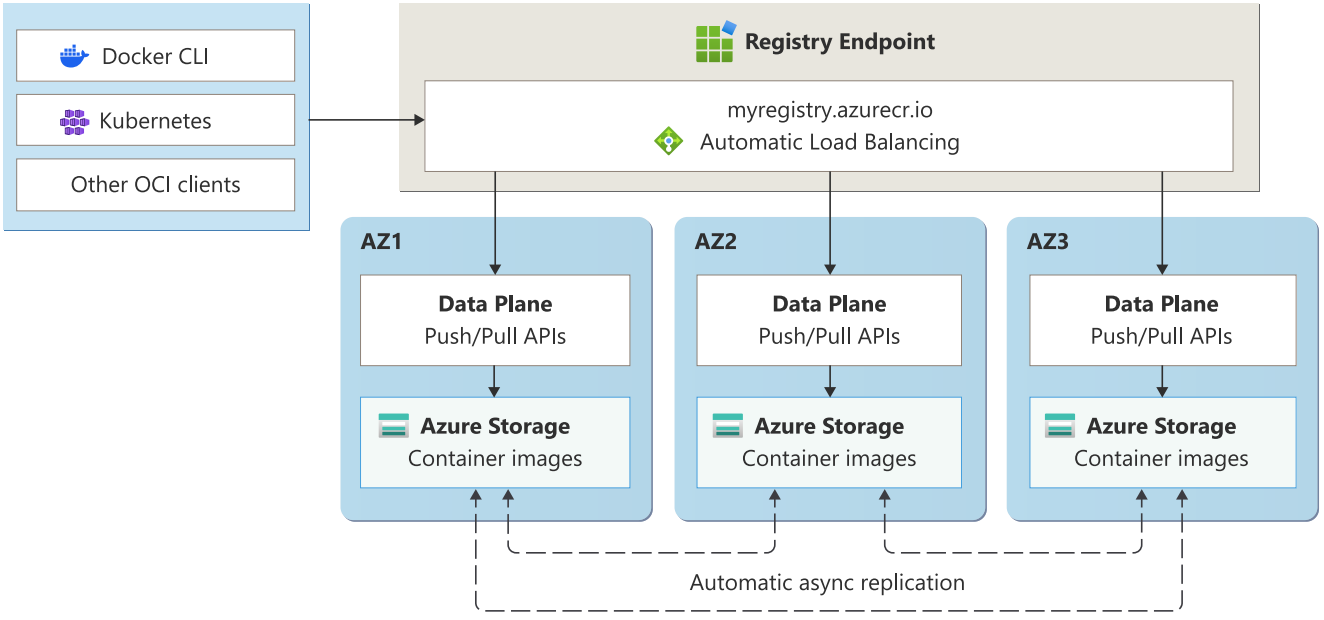
- **Cree un registro con redundancia de zona.** Cuando creas un registro en una región admitida, es automáticamente redundante a nivel de zona. Para más información sobre cómo crear un registro, consulte Creación de un registro con redundancia de zona en Container Registry.

- **Habilite la redundancia de zona en un registro existente.** Los registros que se encuentran en regiones con zonas de disponibilidad automáticamente tienen redundancia de zona. No es necesario habilitar la redundancia de zona.

- **Desactivar la redundancia de zona.** No se puede deshabilitar la redundancia de zona.

Comportamiento cuando todas las zonas están en buen estado

En esta sección se describe qué esperar cuando los recursos de Container Registry están configurados para la redundancia de zona y todas las zonas de disponibilidad están operativas.



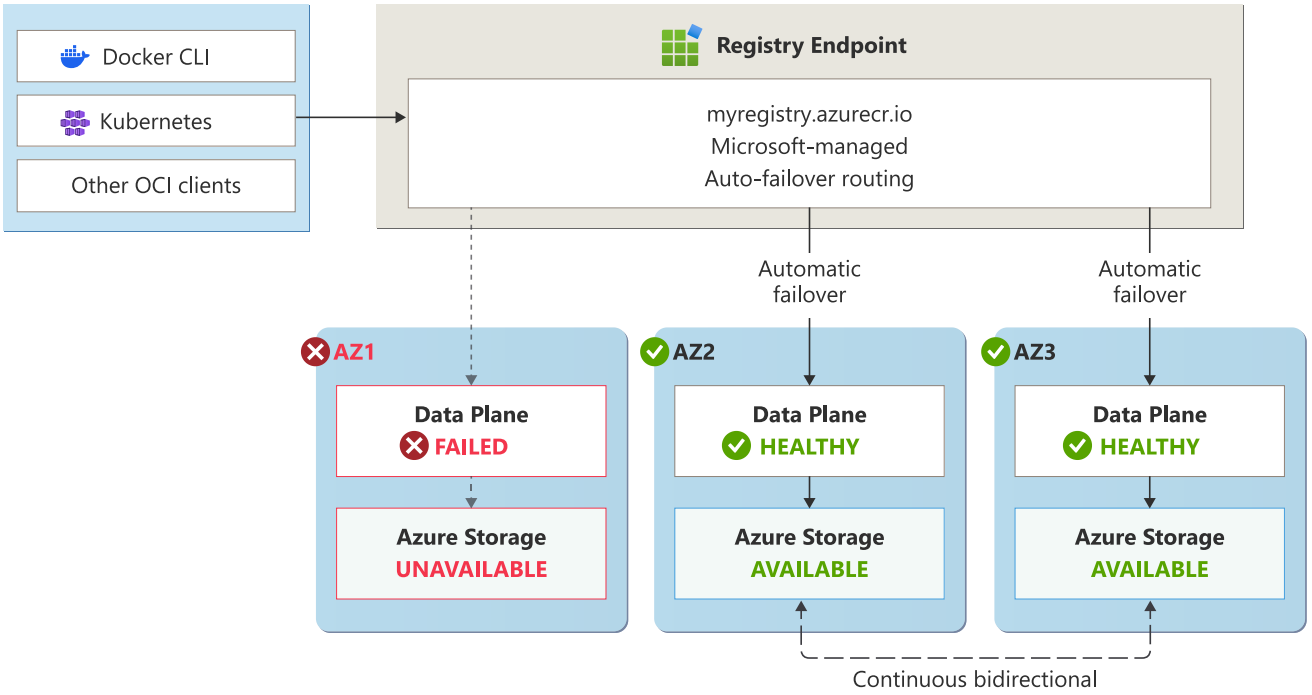
- **Enrutamiento de tráfico entre zonas:** Container Registry usa la funcionalidad de enrutamiento interno para distribuir automáticamente las operaciones del plano de datos en todas las zonas de disponibilidad dentro de una región. El servicio del Registro enruta automáticamente las solicitudes a zonas correctas sin necesidad de equilibradores de carga externos.

- **Replicación de datos entre zonas:** Los datos del Registro, incluidas las imágenes de contenedor, los manifiestos y los metadatos, se replican de forma asincrónica en varias zonas de disponibilidad. Los cambios se replican rápidamente entre zonas para mantener una alta disponibilidad y durabilidad de los datos. La replicación es asíncrona, pero normalmente se completa en cuestión de minutos, y todas las zonas permanecen disponibles para operaciones de lectura y escritura durante la replicación.

Comportamiento durante un fallo de zona

En esta sección se describe qué esperar cuando los recursos de Container Registry están configurados para la redundancia de zona y hay una interrupción de zona de disponibilidad.

Cuando una zona deja de estar disponible, Container Registry administra automáticamente el proceso de conmutación por error con un impacto mínimo en las operaciones del registro.



- **Detección y respuesta:** La plataforma Container Registry detecta automáticamente errores en una zona de disponibilidad e inicia una respuesta. El servicio enruta automáticamente el tráfico a las zonas correctas restantes. No se requiere ninguna intervención manual para iniciar una conmutación por error de zona.

- **Notificaciones:** Microsoft no le notifica automáticamente cuando una zona está inactiva. Sin embargo, puede usar Azure Service Health para comprender el estado general del servicio, incluidos los errores de zona, y puede configurar alertas de Service Health para notificarle problemas.

  También puede supervisar las métricas de disponibilidad del Registro en Azure Monitor.

- **Solicitudes activas:** Cuando una zona de disponibilidad no está disponible, se finalizan las solicitudes en curso conectadas a los recursos de la zona de disponibilidad errónea. Deben reintentarse.

- **Pérdida de datos esperada:** Las escrituras recientes realizadas en la zona defectuosa podrían no replicarse en otras regiones, lo que significa que podrían perderse hasta que la zona se recupere. Normalmente, se espera que la pérdida de datos sea inferior a 15 minutos, pero eso no está garantizado.

- **Tiempo de inactividad esperado:** Es posible que se produzca un breve periodo de inactividad durante la conmutación automática por error, ya que el tráfico se redirige a zonas correctas. Este tiempo de inactividad suele ser de unos segundos para la mayoría de las operaciones del registro. Se recomienda seguir los procedimientos recomendados de control de errores transitorios para minimizar el efecto de la conmutación por error de zona en las aplicaciones.

- **Reenrutamiento del tráfico:** La plataforma vuelve a enrutar automáticamente el tráfico a zonas correctas sin necesidad de realizar ningún cambio de configuración.

Recuperación de zona

Cuando se recupera la zona de disponibilidad afectada, Container Registry distribuye automáticamente las operaciones en todas las zonas disponibles, incluida la zona recuperada. El servicio reequilibra el tráfico y la distribución de datos sin necesidad de intervención manual ni provoca interrupciones del servicio.

Prueba de fallos de zona

La plataforma Container Registry administra el enrutamiento del tráfico, la conmutación por error y la conmutación por recuperación para registros redundantes por zona. Dado que esta característica está totalmente administrada, no es necesario iniciar ni validar los procesos de error de zona de disponibilidad.

Resistencia a errores en toda la región

Container Registry proporciona compatibilidad nativa con varias regiones a través de la replicación geográfica cuando el registro usa el nivel Premium. La replicación geográfica crea réplicas del Registro en varias regiones de su elección. La región en la que se implementa el recurso de registro se conoce como la *región principal*.

La replicación geográfica permite la resistencia a interrupciones regionales. Si su registro está replicado geográficamente y se produce una interrupción regional, los datos del registro seguirán estando disponibles desde las otras regiones que haya seleccionado. Si no habilita la replicación geográfica, es posible que sus datos no estén disponibles durante una interrupción del servicio en la región.

También puede utilizar la replicación geográfica para obtener acceso local a imágenes de contenedores dentro de esas regiones y reducir la latencia de las aplicaciones distribuidas a nivel mundial.

Al implementar Container Registry y habilitar la replicación geográfica, Microsoft usa Azure Traffic Manager para distribuir las solicitudes del plano de datos entre las réplicas y conmutar automáticamente por error entre réplicas si una réplica regional no está disponible.

La replicación geográfica de Container Registry no depende de las regiones emparejadas de Azure. Puede seleccionar cualquier combinación de regiones de Azure para la replicación en función de sus requisitos específicos de ubicación geográfica, rendimiento y cumplimiento normativo. Cada registro replicado geográficamente funciona como un punto de conexión de registro. Admite la mayoría de las operaciones del registro, incluyendo tareas de envío, extracción y administración de imágenes.

En esta sección se resume la información sobre la replicación geográfica en relación con la confiabilidad. Para más información, consulte Replicación geográfica en Container Registry.

Soporte para regiones

La replicación geográfica está disponible en todas las regiones de Azure en las que se admite el nivel Premium. Puede replicar en cualquier combinación de regiones, independientemente de si Azure empareja esas regiones.

Requisitos

Debe usar el nivel Premium para habilitar la replicación geográfica.

Consideraciones

- **Réplicas con redundancia de zona:** cualquier réplica que cree en una región con zonas de disponibilidad tiene redundancia de zona de manera automática.
- **Plano de control:** El plano de control se ejecuta en la región principal. Si la región principal no está disponible, las operaciones del plano de control no están disponibles y es posible que no pueda modificar la configuración del Registro.
- **Tareas:** Las tareas de Container Registry no admiten actualmente réplicas geográficas. Las tareas siempre se ejecutan en la región principal. Si la región principal no está disponible, la tarea no se ejecuta.

Cost

Cada región con replicación geográfica se factura por separado según los precios del plan Premium de la región correspondiente. Los cargos de salida también se aplican a la transferencia de datos entre regiones durante la replicación inicial y la sincronización continua.

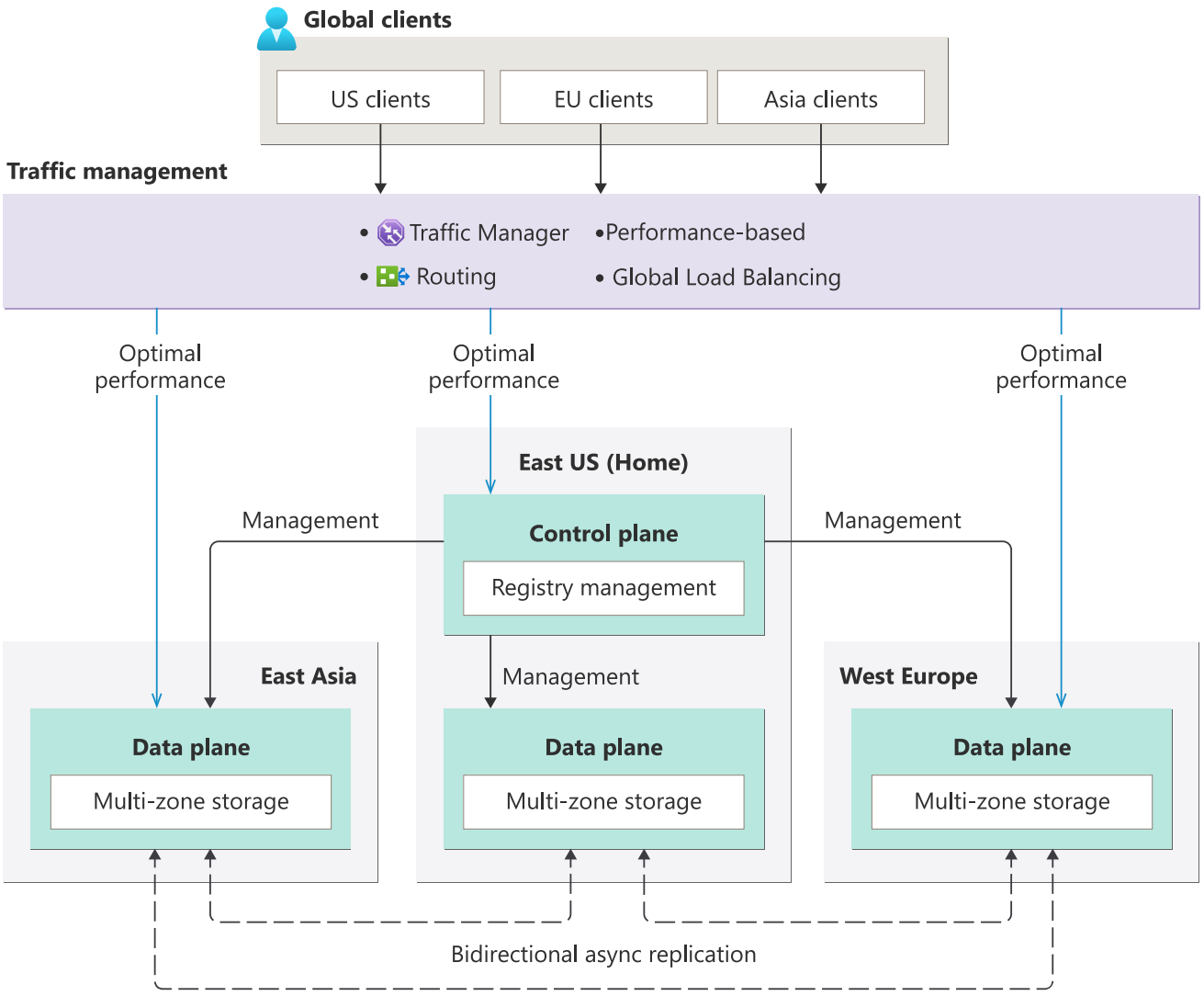
Configuración de la compatibilidad con varias regiones

La replicación geográfica se puede configurar durante la creación del registro o agregarse a los registros Premium existentes. La replicación geográfica se puede configurar a través de Azure Portal, la CLI de Azure, Azure PowerShell o las plantillas de Azure Resource Manager.

- **Crear un registro con replicación geográfica** Configure la replicación geográfica después de crear el registro especificando regiones adicionales.
- **Habilitar la replicación geográfica en un registro existente.** Para habilitar las funciones de replicación geográfica, actualice los registros existentes de nivel Básico o Estándar al nivel Premium. Puede cambiar las regiones de replicación en cualquier momento. Para obtener más información, consulte Configuración de replicación geográfica.
- **Deshabilitar la replicación geográfica.** Quite las réplicas regionales individuales a través de Azure Portal o las herramientas de línea de comandos. No se puede quitar el registro de la región principal.

Comportamiento cuando todas las regiones están en buen estado

En esta sección se describe qué esperar cuando un registro está configurado para la replicación geográfica y todas las regiones están operativas.



- **Enrutamiento del tráfico entre regiones:** Container Registry funciona en una configuración activa-activa donde cada punto de conexión regional puede atender todas las operaciones del plano de datos de forma independiente, incluidas las lecturas y escrituras. Las operaciones del plano de datos, como las operaciones de envío y extracción de contenedores, se enrutan automáticamente mediante Traffic Manager con criterios basados en el rendimiento para determinar el punto de conexión regional óptimo para el rendimiento.

- **Replicación de datos entre regiones:** La replicación geográfica sincroniza automáticamente las imágenes de contenedor y los artefactos en todas las regiones configuradas mediante la replicación asincrónica con coherencia final. El servicio utiliza almacenamiento direccionable por contenido para replicar de manera eficiente solo las capas de imagen únicas. Este enfoque minimiza el uso del ancho de banda y el tiempo de replicación. Las operaciones de lectura y escritura funcionan en todas las regiones con replicación geográfica. Los cambios realizados en cualquier región se replican en todas las demás regiones.

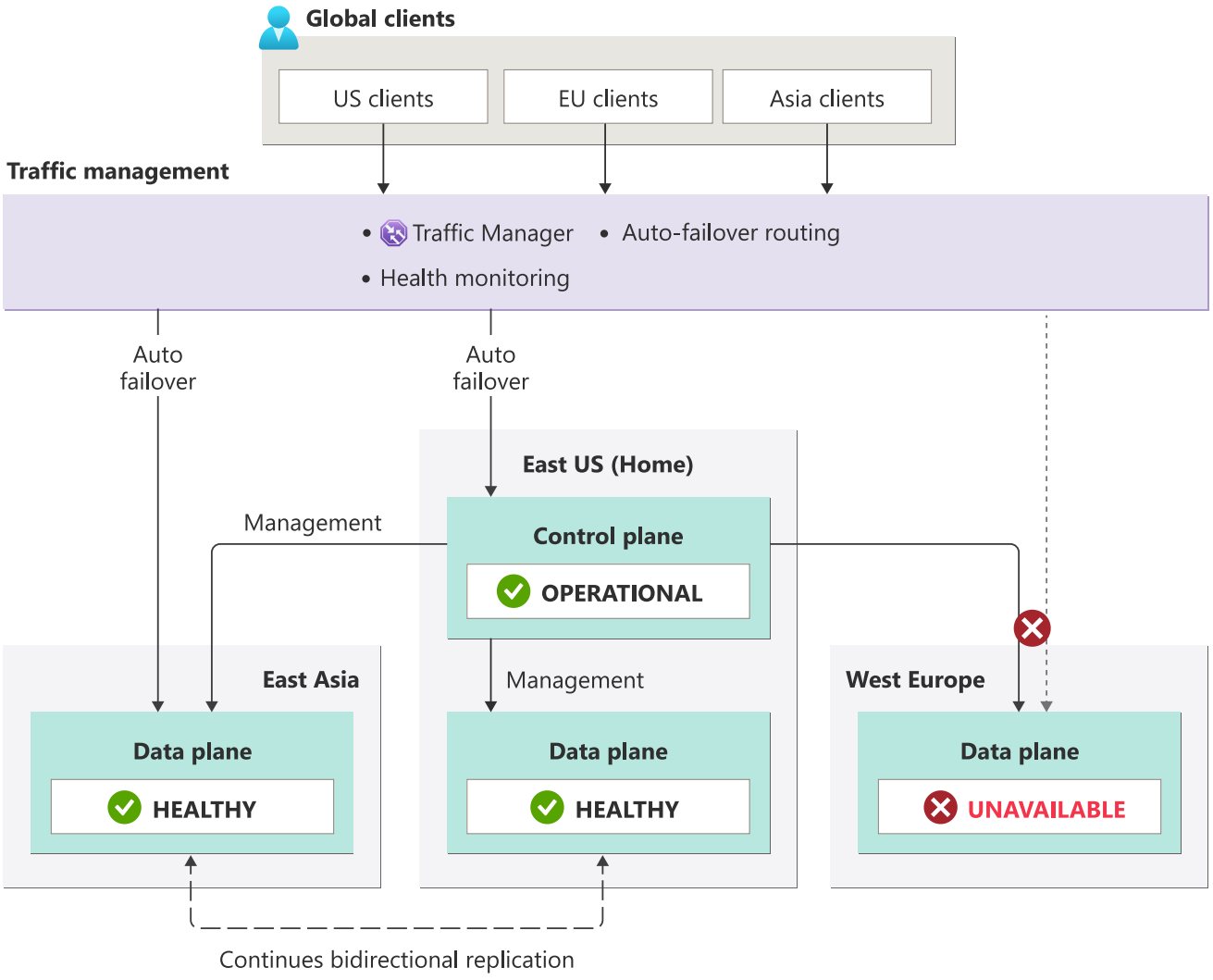
  La replicación se completa normalmente en cuestión de minutos después de los cambios. Sin embargo, no hay garantía sobre el tiempo de replicación de los datos. Las imágenes

de contenedores grandes o las actualizaciones de alta frecuencia pueden tardar más tiempo en replicarse en todas las regiones.

Comportamiento durante una falla de región

En esta sección se describe qué esperar cuando se configura un registro para la replicación geográfica y hay una interrupción en la región primaria.

Cuando una región deja de estar disponible, las operaciones de contenedor pueden seguir usando puntos de conexión regionales alternativos.



- **Detección y respuesta:** Container Registry supervisa el estado de cada réplica regional y es responsable de redirigir el tráfico a otra región.
- **Notificaciones:** Microsoft no le notifica automáticamente cuando una región está inactiva. Sin embargo, puede usar Azure Service Health para comprender el estado general del servicio, incluidos los errores de región, y puede configurar alertas de Service Health para notificarle problemas.

También puede supervisar las métricas de disponibilidad del Registro para cada punto de conexión regional en Azure Monitor.

- **Solicitudes activas:** Cualquier solicitud activa actualmente en tránsito hacia una región no disponible tendrá un error y deberá reintentarse para que pueda dirigirse a una región correcta.
- **Pérdida de datos esperada:** Es posible que las escrituras recientes realizadas en la región defectuosa no se repliquen en otras regiones. Este error significa que podrían perderse hasta que se recupere la región. Normalmente, se espera que la pérdida de datos sea inferior a 15 minutos, pero no se garantiza.
- **Tiempo de inactividad esperado:** En cuanto a las operaciones del plano de datos, se espera un breve periodo de inactividad mientras se completa la conmutación por error, lo que suele tardar entre 1 y 2 minutos.

  Si la región principal no está disponible, las operaciones del plano de control no estarán disponibles hasta que se recupere la región.
- **Reenrutamiento del tráfico:** Cuando una región deja de estar disponible, las operaciones de contenedor se enrutan automáticamente a otra réplica dentro de una región saludable. Los clientes no necesitan cambiar el punto de conexión en el que interactúan con el registro. Microsoft controla automáticamente el enrutamiento, la conmutación por error y la conmutación por recuperación.

Recuperación de regiones

Cuando se recupera una región, las operaciones del plano de datos se reanudan automáticamente para ese punto de conexión regional a través del enrutamiento de Traffic Manager. El servicio sincroniza cualquier cambio que se produzca durante la interrupción mediante la replicación asíncrona con coherencia eventual.

Prueba de fallos de región

No se puede simular el error de una de las regiones asociadas con el registro, pero se puede probar la capacidad de la aplicación para realizar una conmutación por error entre regiones. Puede simular una conmutación por error regional deshabilitando temporalmente las réplicas geográficas, lo que las quita del enrutamiento de Traffic Manager. A continuación, puede verificar que las operaciones de contenedores se transfieren correctamente a regiones alternativas en caso de error sin experimentar realmente una interrupción regional. Para obtener más información, consulte Deshabilitar temporalmente el enrutamiento a la replicación.

Al volver a activar la réplica, Traffic Manager reanuda el enrutamiento del tráfico a la réplica que se ha vuelto a habilitar. Además, los metadatos y las imágenes se sincronizan con coherencia eventual con la réplica reactivada para garantizar la coherencia de los datos en todas las regiones.

Copias de seguridad y restauración

Container Registry admite la exportación de imágenes de contenedor y artefactos desde el registro al almacenamiento externo o a registros alternativos. Use las funciones de importación y exportación de Container Registry o los comandos estándar de Docker para crear copias de imágenes de contenedores críticas para escenarios de recuperación ante desastres.

Para la mayoría de las soluciones, no debe confiar exclusivamente en copias de seguridad. En su lugar, utilice las otras capacidades descritas en esta guía para apoyar los requisitos de resiliencia. Sin embargo, las copias de seguridad protegen contra algunos riesgos que otros enfoques no. Para más información, consulte ¿Qué son la redundancia, la replicación y la copia de seguridad?.

Acuerdo de nivel de servicio

El contrato de nivel de servicio (SLA) para los servicios de Azure describe la disponibilidad esperada de cada servicio y las condiciones que la solución deberá cumplir para lograr esa expectativa de disponibilidad. Para obtener más información, consulte Acuerdos de Nivel de Servicio para servicios en línea.

Contenido relacionado

- Fiabilidad de Azure
- Niveles de servicio de Container Registry
- Procedimientos recomendados de Container Registry
- Supervisión de Container Registry
- Precios de Container Registry
- Importación de imágenes de contenedor en un registro de contenedor

---

Last updated on 22/01/2026

Confiabilidad en Azure Cosmos DB

Azure Cosmos DB para NoSQL es un servicio de base de datos multimodelo distribuido globalmente que admite modelos de datos de documentos con esquemas flexibles. Azure Cosmos DB ofrece características de confiabilidad completas, incluidos varios niveles de coherencia que permiten equilibrar el rendimiento y la disponibilidad, las implementaciones con redundancia de zona de disponibilidad que protegen frente a errores de zona de disponibilidad, replicación multiregión con conmutación por error administrada por el servicio o administrada por el cliente, y opciones de copia de seguridad continuas y periódicas para la protección de datos.

Cuando se usa Azure, la confiabilidad es una responsabilidad compartida. Microsoft proporciona una variedad de capacidades para apoyar la resiliencia y la recuperación. Es responsable de comprender cómo funcionan esas funcionalidades dentro de todos los servicios que usa y de seleccionar las funcionalidades que necesita para cumplir los objetivos empresariales y los objetivos de tiempo de actividad.

En este artículo se describe cómo hacer que Azure Cosmos DB sea resistente a varias posibles interrupciones y problemas, incluidos errores transitorios, interrupciones de zona de disponibilidad, interrupciones de regiones y mantenimiento del servicio. También se describe cómo usar copias de seguridad para recuperarse de otros tipos de problemas y se resalta información clave sobre el acuerdo de nivel de servicio (SLA) de Azure Cosmos DB.

Recomendaciones de implementación de producción

Azure Well-Architected Framework proporciona recomendaciones sobre confiabilidad, seguridad, costo, operaciones y rendimiento. Para comprender cómo estas áreas influyen entre sí y contribuir a una solución de Azure Cosmos DB confiable, consulte Procedimientos recomendados para Azure Cosmos DB.

Introducción a la arquitectura de confiabilidad

En esta sección se describen algunos de los aspectos importantes de cómo funciona el servicio que es más relevante desde una perspectiva de confiabilidad. En la sección se presenta la arquitectura lógica, que incluye algunos de los recursos y características que se implementan y usan. También se describe la arquitectura física, que proporciona detalles sobre cómo funciona el servicio en segundo plano.

Arquitectura lógica

El recurso principal que implemente es un Azure Cosmos DB *account*. Cada cuenta puede tener varias *bases de datos* con varios *contenedores*. Los contenedores sirven como unidades lógicas de distribución y escalabilidad. Puede crear contenedores como colecciones, tablas y gráficos, en función de la API que use para interactuar con Azure Cosmos DB. Para obtener más información sobre el modelo de recursos, consulte Databases, contenedores y elementos en Azure Cosmos DB. Cada contenedor usa particionamiento, que admite alto rendimiento y alta escalabilidad.

Configure el *rendimiento*, que representa la cantidad de recursos del sistema que puede utilizar para consultar y trabajar con sus datos. Puede aprovisionar manualmente el rendimiento, usar la escalabilidad automática para ajustar dinámicamente la capacidad en función de los requisitos de la carga de trabajo o usar el tipo de cuenta sin servidor para que se cobre por el uso real.

Una sola cuenta puede abarcar varias regiones de Azure, lo que aumenta su resiliencia ante interrupciones regionales. Puede configurar varias regiones para leer y, si usa el nivel Crítico para la empresa, puede usar varias regiones para escribir. Azure Cosmos DB replica automáticamente sus datos geográficamente. El comportamiento de la replicación geográfica se ve afectado por la configuración que se usa, como el nivel de coherencia, que indica cómo desea compensar el equilibrio entre la coherencia de los datos, la disponibilidad, la latencia y el rendimiento. Diferentes niveles de coherencia optimizan para diferentes preocupaciones, admiten diferentes garantías y proporcionan diferentes tipos de replicación entre regiones.

Arquitectura física

Azure Cosmos DB almacena varios *replicas* de los datos para la redundancia. El servicio mitiga automáticamente las interrupciones de las réplicas manteniendo el quórum entre las réplicas de cada región. Este enfoque garantiza la alta disponibilidad y protege contra la pérdida de datos durante errores de nodo individuales, sin necesidad de cambios o configuración de la aplicación.

Internamente, Azure Cosmos DB administra los datos a través de varias construcciones, incluidas las particiones *físicas*, *conjuntos de partición* y *conjuntos de réplicas*. Para obtener información más detallada sobre cómo funciona Azure Cosmos DB, consulte Distribución global de datos con Azure Cosmos DB: a fondo.

Resistencia a errores transitorios

Los errores transitorios son errores breves e intermitentes en los componentes. Se producen con frecuencia en un entorno distribuido como la nube y son una parte normal de las operaciones.

Los errores transitorios se corrigen después de un breve período de tiempo. Es importante que las aplicaciones puedan controlar errores transitorios, normalmente mediante el reintento de solicitudes afectadas.

Todas las aplicaciones hospedadas en la nube deben seguir las instrucciones de control de errores transitorios de Azure cuando se comunican con cualquier API, bases de datos y otros componentes hospedados en la nube. Para obtener más información, consulte Recomendaciones para controlar errores transitorios.

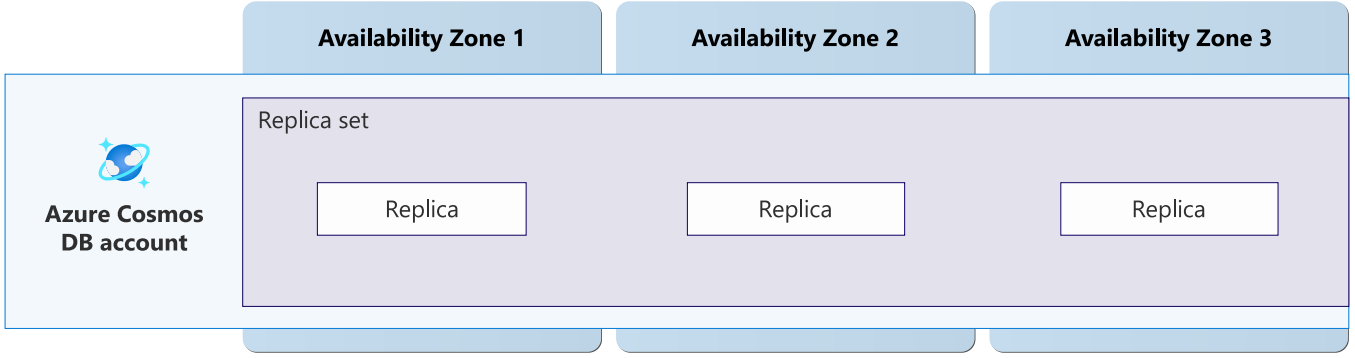
Use los SDK de Azure Cosmos DB. Los SDK implementan automáticamente la compatibilidad con una serie de consideraciones de resistencia, incluido el control de errores transitorios a través de reintentos automáticos, y respetan las respuestas de límite de velocidad enviadas por el servicio. Para obtener más información, consulte Design resilient applications with Azure Cosmos DB SDKs.

Al trabajar con una cuenta multirregional, el SDK también admite una estrategia de disponibilidad basada en un umbral, también denominada *hedging*, en la que se envían solicitudes de lectura en paralelo a varias regiones y se acepta la respuesta más rápida. Este enfoque puede mejorar el rendimiento de la aplicación cuando una región experimenta temporalmente una latencia mayor de lo habitual.

Resistencia a errores de zona de disponibilidad

Las zonas de disponibilidad son grupos físicamente independientes de centros de datos dentro de una región de Azure. Cuando una zona falla, los servicios pueden transferirse a una de las zonas restantes.

Azure Cosmos DB admite *redundancia de zona*. Al habilitar la redundancia de zona, Azure distribuye las réplicas de los datos en varias zonas de disponibilidad, lo que proporciona resistencia a problemas y interrupciones del centro de datos. Microsoft selecciona las zonas de disponibilidad que se van a usar.



Una cuenta de Azure Cosmos DB puede utilizar varias regiones (ubicaciones) para la distribución global, la escalabilidad y la conmutación por error. La redundancia de zona se configura por separado para cada región de la cuenta.

El uso de redundancia de zona en Azure Cosmos DB no tiene ningún impacto perceptible en el rendimiento o la latencia. No requiere ningún ajuste en el modo de coherencia seleccionado y no requiere ninguna modificación en el código de la aplicación.

Use la redundancia de zona en las regiones en las que se admite, especialmente para las cuentas de una sola región. Dado que las zonas de disponibilidad son físicamente independientes y proporcionan distintas fuentes de alimentación, redes y refrigeración, los acuerdos de nivel de servicio de disponibilidad para Azure Cosmos DB son más altos para las cuentas con redundancia de zona que las cuentas que no usan zonas de disponibilidad.

> **Sugerencia**
>
> Habilitar la redundancia de zona es una excelente manera de aumentar la resistencia de la base de datos de Azure Cosmos DB sin introducir complejidades de aplicación adicionales ni afectar al rendimiento. En función de la configuración de la cuenta, es posible que ni siquiera incurra en costos adicionales.

Si no habilita la redundancia de zona, la cuenta no es *zonal* en esa región. Las cuentas no zonales podrían localizar réplicas en una sola zona de disponibilidad, lo que provocaría un posible tiempo de inactividad si esa zona específica experimenta un problema.

Requirements

- **Region support:** Puede habilitar la redundancia de zona en las regiones de Azure que admiten zonas de disponibilidad. Para ver si la región admite zonas de disponibilidad, consulte la lista de regiones admitidas.

  La redundancia de zona no es una configuración para toda la cuenta. Azure Cosmos DB cuentas pueden abarcar varias regiones y cada región se puede configurar de forma independiente para usar zonas de disponibilidad. Las regiones que no admiten zonas de disponibilidad no le impiden habilitar la redundancia de zona en otras regiones dentro de la misma cuenta.

- **Cuentas sin servidor:** Solo puede configurar cuentas sin servidor con redundancia de zona al crearlas. No se pueden convertir cuentas sin servidor existentes sin zonas de

disponibilidad a una configuración de zona de disponibilidad. En el caso de las cargas de trabajo críticas, use el rendimiento aprovisionado.

Consideraciones

- **Varias interrupciones de zona simultáneas:** Una cuenta de una sola región con redundancia de zona puede mantener la disponibilidad de lectura y escritura cuando una interrupción afecta a una sola zona de disponibilidad. Sin embargo, si la interrupción afecta a varias zonas de disponibilidad o a toda la región, las cuentas de una sola región pierden el acceso de lectura y escritura hasta que se restaure el servicio. Considere desplegar una cuenta multirregional si necesita ser resiliente al fallo simultáneo de varias zonas.

- **Cuentas de varias regiones:** Si tiene una cuenta de varias regiones, opcionalmente puede habilitar la redundancia de zona en cualquiera o todas las regiones de cuenta que admiten zonas de disponibilidad. Se recomienda encarecidamente habilitar la redundancia de zona cuando la cuenta esté configurada para usar una sola región o si está configurada para usar una sola región de escritura con varias regiones de lectura.

Cost

Las regiones en las que se habilita la redundancia de zona se cobran a un precio superior. Sin embargo, se renuncia a los precios premium de las zonas de disponibilidad para las cuentas configuradas con escrituras de varias regiones y para las colecciones configuradas para usar el modo de rendimiento de escalabilidad automática. Para obtener más información, consulte Precios de Azure Cosmos DB.

Configurar soporte de zonas de disponibilidad

Para la mayoría de las cuentas, solo se habilita la redundancia de zona cuando se agrega una nueva región a una cuenta de Azure Cosmos DB. Para habilitar la compatibilidad con zonas de disponibilidad en una cuenta existente, agregue una región y habilite la redundancia de zona en ella. Puede seguir un proceso para agregar una región temporal para que pueda configurar la redundancia de zona en la región original. Para obtener pasos detallados, consulte Habilitar redundancia de zona en una cuenta de Azure Cosmos DB.

Para las cuentas sin servidor, debe habilitar la redundancia de zona al crear la cuenta.

Comportamiento cuando todas las zonas están en buen estado

En esta sección se describe qué esperar al configurar una cuenta de Azure Cosmos DB para la redundancia de zona y todas las zonas están operativas.

- **Operación entre zonas:** Azure Cosmos DB enruta automáticamente las solicitudes a las réplicas entre zonas de disponibilidad, por lo que cualquier réplica puede servir una solicitud.

- **Replicación de datos entre zonas:** Cuando un cliente realiza un cambio en los datos, ese cambio se aplica a varias réplicas en distintas zonas para lograr el cuórum. Este enfoque se conoce como *replicación sincrónica*. La replicación sincrónica garantiza un alto nivel de coherencia de datos, lo que reduce la probabilidad de pérdida de datos durante un error de zona. Las zonas de disponibilidad se encuentran relativamente cerca, lo que significa que hay un efecto mínimo en la latencia o el rendimiento.

Comportamiento durante un fallo de zona

En esta sección se describe qué esperar al configurar una cuenta de Azure Cosmos DB para la redundancia de zona y hay una interrupción en una de las zonas.

- **Detección y respuesta:** La plataforma Azure Cosmos DB es responsable de detectar un error en una zona de disponibilidad. No es necesario hacer nada para iniciar una conmutación por error de la zona.

- **Notificación**: Microsoft no le notifica automáticamente cuando una zona está inactiva. Sin embargo, puede usar Azure Resource Health para supervisar el estado de un recurso individual y puede configurar alertas de Resource Health para notificarle problemas. También puede usar Azure Service Health para comprender el estado general del servicio, incluidos los errores de zona, y puede configurar alertas de Service Health para notificarle problemas.

- **Active requests:** Cuando una zona de disponibilidad deja de estar disponible, Azure Cosmos DB finaliza las solicitudes en curso conectadas a las réplicas de la zona afectada y la aplicación debe reintentar esas solicitudes. Asegúrese de que la aplicación está preparada siguiendo las instrucciones de control de errores transitorios.

- **Pérdida de datos esperada:** No hay ninguna pérdida de datos esperada de un error de zona.

- **Tiempo de inactividad esperado:** Durante las interrupciones de zona, las conexiones pueden experimentar interrupciones breves que suelen durar unos segundos a medida que

se redistribuye el tráfico. Asegúrese de que las aplicaciones están preparadas siguiendo las instrucciones de control de errores transitorios.

- **Redistribution:** Azure Cosmos DB redirige automáticamente las solicitudes entrantes a réplicas saludables en otras zonas de disponibilidad. Cuando una zona de disponibilidad tiene una interrupción, la plataforma reasigna automáticamente el rendimiento aprovisionado a otras réplicas.

Recuperación de zona

Cuando se recupera la zona de disponibilidad, Azure Cosmos DB restaura automáticamente las réplicas en la zona de disponibilidad y vuelve a enrutar el tráfico entre réplicas como normal.

Prueba de fallos de zona

La conmutación por error y la recuperación de las zonas de disponibilidad de Azure Cosmos DB están totalmente administradas por Microsoft. No es necesario iniciar ni validar los procesos de fallo de las zonas de disponibilidad.

Resistencia a errores en toda la región

Al implementar una cuenta de Azure Cosmos DB en una sola región, una interrupción en toda la región que afecta a todos los nodos de Azure Cosmos DB normalmente no causa pérdida de datos, pero impide que la aplicación acceda a los datos. Azure Cosmos DB restaura el acceso a los datos después de que el servicio se recupere en la región afectada. La pérdida de datos solo se produce si la región experimenta un desastre irrecuperable.

Para prepararse para los casos excepcionales de interrupciones de regiones, puede configurar Azure Cosmos DB para admitir varios niveles de durabilidad y disponibilidad mediante uno de estos enfoques:

- Varias regiones de lectura con una sola región de escritura. Opcionalmente, puede habilitar la conmutación por error administrada por servicio o la conmutación automática por partición (PPAF).
- Regiones de escritura múltiples.

En la tabla siguiente se resumen las opciones de recuperación disponibles en función de la configuración de la cuenta y el tipo de interrupción. En las secciones posteriores de este artículo se proporcionan detalles exhaustivos de estas opciones y el comportamiento asociado.

Expandir tabla

Configuration	Tipo de interrupción	Impacto en la disponibilidad	Impacto de durabilidad	Qué hacer
Cuenta de una sola región	Interrupción regional	Se pierde el acceso de lectura y escritura hasta que se restaura el servicio.	No se pierden datos a menos que la región experimente un desastre irrecuperable.	Espere a que se restaure el servicio o solicite la restauración de la cuenta de copia de seguridad en otra región.
Región de escritura única, cuenta de varias regiones	Interrupción de la región de lectura	El SDK redirige a las regiones disponibles según la configuración de regiones preferidas. En el caso de las cuentas que utilizan consistencia fuerte con solo dos regiones o un retraso limitado que supere la ventana de retraso, también se pierde la disponibilidad de escritura a menos que se desconecte la región afectada.	No se produce pérdida de datos.	Asegúrese de que el rendimiento es suficiente en las regiones restantes. Para la consistencia fuerte o con caducidad limitada, considere la posibilidad de desconectar la región afectada.
Región de escritura única, cuenta de varias regiones	Interrupción de la región de escritura (con PPAF habilitado)	Conmutación por error automática a nivel de partición; no se requiere intervención manual.	Si la cuenta usa una coherencia fuerte, no se pierden datos. Si la cuenta no usa una coherencia fuerte, los datos no replicados podrían perderse en el improbable caso de que la región sufra una pérdida de datos permanente.	No es necesaria ninguna acción. PPAF administra automáticamente la conmutación por error.
Región de escritura única, cuenta de varias regiones	Interrupción de la región de escritura de datos (sin PPAF)	Se pierde la disponibilidad de escritura hasta que se complete una operación de desconexión de la	Si la cuenta usa una coherencia fuerte, no se pierden datos. Si la cuenta no usa una coherencia fuerte, los datos no	Realice una operación de desconexión de la región. Si la conmutación por error administrada por el servicio está habilitada, Azure Cosmos

Configuration	Tipo de interrupción	Impacto en la disponibilidad	Impacto de durabilidad	Qué hacer
		región o una conmutación por error administrada por el servicio. Las lecturas continúan desde regiones saludables.	replicados podrían perderse en el improbable caso de que la región sufra una pérdida de datos permanente.	DB inicia automáticamente la conmutación por error, pero esto puede tardar una hora o más. No cambie la región de escritura durante la interrupción.
Cuenta de región de escritura múltiple	Interrupción en cualquier región	Enrutamiento automático a regiones correctas a través de la configuración del SDK; no se requiere intervención manual.	Los datos actualizados recientemente en la región con errores podrían no estar disponibles en las regiones restantes. En el improbable caso de que la región sufre una pérdida permanente de datos, se podrían perder datos no replicados.	Asegúrese de que el rendimiento es suficiente en las regiones restantes. Después de la recuperación, Azure Cosmos DB recupera automáticamente los datos no replicados mediante el método de resolución de conflictos configurado.
Cualquier configuración de cuenta	Datos dañados o eliminación accidental	Sin impacto en la disponibilidad.	Posible pérdida de datos en función de cuándo se detectan daños o eliminaciones.	Restauración a un momento dado (copia de seguridad continua) o restauración a partir de una copia de seguridad periódica.

Nota:

Este artículo se centra en los aspectos de confiabilidad de las características de varias regiones de Azure Cosmos DB. Hay otras ventajas para varias regiones de lectura y escritura, como un mayor rendimiento y escalado para aplicaciones distribuidas globalmente. Debe evaluar toda la arquitectura de la solución y tener en cuenta todas las ventajas de usar estas funcionalidades.

SDK y resiliencia

Los SDK de Azure Cosmos DB son una parte importante de la estrategia de resistencia de la aplicación. Cuando tiene una cuenta de varias regiones, la configuración del SDK afecta a cómo

se enrutan las solicitudes entre regiones, incluidas las regiones preferidas a las que conectarse y las regiones a las que se debe excluir. Los SDK supervisan la disponibilidad de regiones y particiones, y pueden reconfigurarse dinámicamente para utilizar regiones y particiones saludables, como a través del disyuntor a nivel de partición.

Para obtener más información sobre cómo el SDK admite la alta disponibilidad, consulte la documentación de alta disponibilidad para el SDK que usa:

- SDK de Azure Cosmos DB .NET v3
- SDK v4 de Azure Cosmos DB Java
- SDK de Azure Cosmos DB para Python

Posible pérdida de datos durante fallos regionales

Al implementar una cuenta de Azure Cosmos DB en varias regiones, la durabilidad de los datos depende del nivel de coherencia que configure en la cuenta. En la tabla siguiente se detallan todos los niveles de coherencia, el objetivo de punto de recuperación (RPO) de una cuenta de Azure Cosmos DB que se implementa en al menos dos regiones. El RPO representa la pérdida potencial de datos durante una interrupción del servicio en una región.

Expandir tabla

Nivel de coherencia	RPO para la interrupción de la región
Sesión, prefijo coherente, eventual	Menos de 15 minutos
Uso vinculado	*K* y *T*
Alta	0

K = el número de versiones (es decir, actualizaciones) de un elemento.

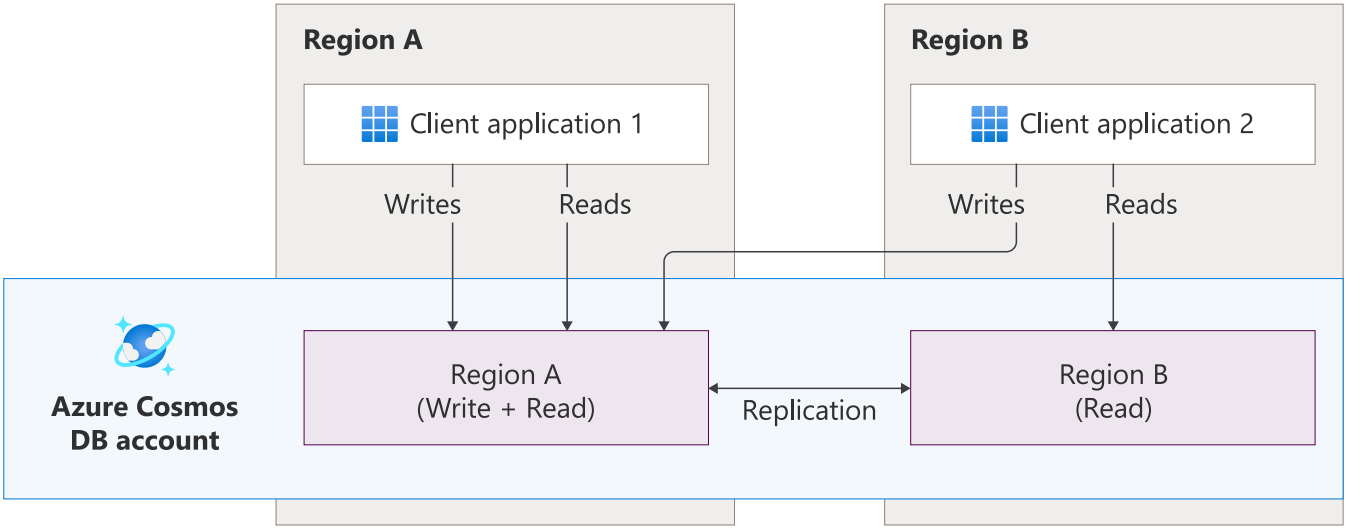
T = intervalo de tiempo desde la última actualización.

En el caso de cuentas de varias regiones, el valor mínimo de *K* y *T* es de 100 000 operaciones de escritura o de 300 segundos. Este valor define el RPO mínimo para los datos cuando se utiliza la obsolescencia limitada.

Para obtener más información sobre las diferencias entre los niveles de coherencia, consulte Niveles de coherencia en Azure Cosmos DB.

Varias regiones de lectura con una sola región de escritura

Si la solución requiere un tiempo de actividad continuo durante las interrupciones de la región, puede configurar Azure Cosmos DB para replicar los datos en varias regiones, con escrituras controladas por la región primaria. Opcionalmente, puede configurar las aplicaciones para conectarse a regiones de lectura específicas, lo que puede ayudar a mejorar su rendimiento. Si una región sufre una interrupción del servicio, la cuenta puede seguir funcionando desde las regiones que están en buen estado.



Conmutación por error entre regiones

Puede configurar el SDK de Azure Cosmos DB con una lista prioritaria de regiones de lectura. El SDK conecta la aplicación a la primera región disponible de la lista. Durante una interrupción de la región de lectura, el SDK detecta la interrupción de la región a través de códigos de respuesta de back-end, la marca como no disponible y enruta las operaciones futuras a la siguiente región disponible en la lista de preferencias. Asegúrese de que la lista de regiones preferidas se ha establecido correctamente y se alinea con los requisitos de negocio y latencia. Para obtener instrucciones detalladas, consulte Solución de problemas de disponibilidad del SDK de Azure Cosmos DB.

La conmutación por error es el proceso de hacer que una de las regiones de su cuenta no esté disponible, ya sea completamente o en parte. El efecto de una conmutación por error depende de si la región es una región de escritura o una región de lectura:

- Si una región de escritura deja de estar disponible, otra región se convierte en la región de escritura.
- Si una región de lectura deja de estar disponible, esa región no puede atender solicitudes de lectura y otras regiones se usan en su lugar para las operaciones de lectura.

Azure Cosmos DB proporciona varios tipos de conmutación por error:

- **Conmutación por error automática por partición (PPAF):** Internamente, Azure Cosmos DB distribuye los datos entre varias particiones físicas. Si se produce un problema con la infraestructura que admite una partición, es posible que otras particiones no se vean afectadas. PPAF permite a las cuentas de región de escritura única realizar automáticamente la conmutación por error de particiones individuales a una región secundaria, al tiempo que se mantienen las particiones en buen estado en la región principal. PPAF puede ayudar a minimizar el tiempo de inactividad y a permitir una recuperación más rápida durante un error de región parcial. Para obtener más información, consulte Cómo adoptar Per-Partition Automatic Failover (PPAF) para Azure Cosmos DB.

- **Conmutación por error forzada:** Puede desconectar una de las regiones de su cuenta. Esto también se conoce como *conmutación por error administrada por el cliente* o operación de *región fuera de línea*. Este es el enfoque recomendado para restaurar rápidamente la disponibilidad durante una interrupción. Usted es responsable de detectar la interrupción del servicio y de activar la conmutación por error. También puede utilizar las conmutaciones por error forzadas para simular escenarios de caída de una región con fines de prueba, como durante un simulacro de recuperación ante desastres.

  Si desconecta la región de escritura, la región de lectura con la siguiente prioridad más alta se convierte en la nueva región de escritura. Si deja sin conexión una región de lectura, las aplicaciones pueden conectarse a cualquier otra región de lectura de la cuenta.

  Una conmutación por error forzada de su región de escritura conlleva la posibilidad de pérdida de datos en el caso de las escrituras no replicadas.

  Tras una conmutación por error forzada, Microsoft debe volver a poner la región en línea. Para regiones saludables, este proceso está automatizado, pero puede tardar varios días. Si la región no vuelve a estar en línea en un día o dos, abra un caso de soporte técnico para solicitar asistencia.

- **Cambiar la región de escritura:** Cuando las regiones estén en buen estado, puede cambiar la región de escritura de la cuenta. Este cambio equivale, en la práctica, a una conmutación por error planificada de la región de escritura de su cuenta.

  Cambiar la región de escritura no produce ninguna pérdida de datos, ya que la replicación de datos se recupera antes de que se promueva la nueva región de escritura. Puede haber una breve interrupción, pero los clientes que usan lógica de reintento y otras técnicas de control de errores transitorios no suelen experimentar un impacto significativo.

Esta operación requiere que las regiones estén en buen estado, **por lo que no se puede usar durante una interrupción de la región.**

- **Conmutación por error administrada por el servicio:** Cuando la cuenta usa la conmutación por error administrada por el servicio, Microsoft es responsable de decidir cuándo realizar la conmutación por error entre regiones. Para habilitar la conmutación por error administrada por el servicio, especifique prioridades para cada región. Sin embargo, el proceso de declarar una interrupción del servicio y activar la conmutación por error administrada por el servicio puede llevar bastante tiempo, posiblemente una hora o más. Para una recuperación más rápida, realice una conmutación por error forzada en lugar de esperar a que se desencadene la conmutación por error administrada por el servicio.

  Si Microsoft activa una conmutación por error administrada por el servicio para la región de escritura de la cuenta, cualquier escritura no replicada podría perderse.

  Tras una conmutación por error administrada por el servicio, Microsoft debe volver a poner la región en línea. Microsoft trae automáticamente la región en línea, pero este proceso puede tardar varios días.

Requirements

Region support: Puede configurar cualquier región de Azure como región de lectura para la cuenta de Azure Cosmos DB.

Cost

Agregar una región de lectura adicional a una cuenta de Azure Cosmos DB aumenta los costos existentes para cada región. Para obtener más información, consulte Precios de Azure Cosmos DB.

Configuración de varias regiones de lectura

- **Agregue regiones de lectura a una cuenta:** Puede configurar varias regiones en su cuenta al crear la cuenta o en cualquier momento después de crearla. Para obtener más información, consulte Agregar o quitar regiones de la cuenta de base de datos.

- **Habilitación de la conmutación por error:** Algunos tipos de conmutación por error deben configurarse de antemano:

  - *Conmutación automática por error por partición*: Para obtener más información, consulte Cómo incorporar y adoptar la conmutación automática por error por partición (PPAF)

para Azure Cosmos DB.

- *Conmutación por error administrada por el servicio:* En primer lugar, habilite la conmutación por error administrada por el servicio. A continuación, establezca las prioridades de conmutación por error para cada región de su cuenta.

Planeamiento y administración de capacidad

Si la aplicación distribuye las solicitudes entre regiones y una región se queda sin conexión, las regiones restantes experimentan un volumen de solicitudes mayor. Utilice el escalado automático del rendimiento para adaptar dinámicamente la capacidad en función de la demanda. Si usa el rendimiento aprovisionado, planee la capacidad adecuada para controlar la pérdida de una región sin degradación del servicio y considere la posibilidad de sobreaprovisionar. Para más información, consulte Administración de la capacidad con sobreaprovisionamiento.

Comportamiento cuando todas las regiones están en buen estado

En esta sección se describe qué esperar al configurar una cuenta de Azure Cosmos DB con varias regiones de lectura y todas las regiones están operativas.

- **Operación entre regiones:** La aplicación configura la región que debe recibir operaciones de lectura. Puede configurar la aplicación con una lista prioritaria de regiones o excluir algunas regiones. Para obtener más información sobre cómo funciona la selección de regiones, consulte Diagnose y solucione la disponibilidad de los SDK de Azure Cosmos DB en entornos multirregionales.

  Todas las operaciones de escritura se dirigen a la región de escritura asignada a su cuenta.

- **Replicación de datos entre regiones:** Todas las operaciones de escritura se producen en la región primaria de la cuenta. Las operaciones de escritura se replican a las otras regiones de lectura según el nivel de coherencia configurado para la cuenta. Para obtener información sobre el retraso máximo de replicación, consulte Posible pérdida de datos durante las interrupciones de la región.

Comportamiento durante un error de región de lectura

En esta sección se describe qué esperar al configurar una cuenta de Azure Cosmos DB con varias regiones de lectura y se produce una interrupción en una de las regiones de lectura de la cuenta.

> **Importante**

Lo ideal es que las interrupciones de la región de lectura se controlen en el nivel de cliente configurando correctamente la lista de regiones preferidas en la configuración del SDK. Cuando se configura correctamente, el SDK detecta automáticamente la interrupción y vuelve a enrutar las operaciones de lectura a la siguiente región disponible sin necesidad de realizar ninguna conmutación por error del lado del servicio.

- **Detección y respuesta:** La responsabilidad de detectar la interrupción y responder depende del tipo de conmutación por error que usa su cuenta.
  - *PPAF:* PPAF normalmente no se aplica a interrupciones de la región de lectura. Sin embargo, en el caso de las cuentas con consistencia fuerte y solo dos regiones, la pérdida de la región de lectura reduce la cuenta a una única región, lo que impide mantener el quórum dinámico. En este escenario, PPAF puede activarse para conservar la disponibilidad cambiando las particiones afectadas a la región correcta.
  - *Conmutación por error forzada:* Usted es responsable de realizar una conmutación por error forzada. Para conocer los pasos detallados, consulte Realizar una conmutación por error forzada para su cuenta de Azure Cosmos DB.

    Si no realiza una conmutación por error, el comportamiento de su cuenta dependerá de su nivel de consistencia:

    - *Coherencia fuerte*: la coherencia fuerte requiere dos o más regiones para mantener el cuórum dinámico. Si hay menos de dos regiones disponibles y no realiza una conmutación por error, la cuenta perderá la disponibilidad de escritura hasta que se restablezca el servicio.
    - *Coherencia de obsolescencia limitada:* La coherencia de obsolescencia limitada se basa en el mantenimiento de un umbral de obsolescencia específico entre regiones. Si la duración de la interrupción de la región supera el umbral, el sistema no podrá mantener la consistencia entre las escrituras. Si no se realiza una conmutación por error, la cuenta perderá la disponibilidad de escritura hasta que se restablezca el servicio.
  - *Conmutación por error administrada por el servicio:* Si la conmutación por error administrada por el servicio está habilitada, Microsoft detectará la interrupción e iniciará una conmutación por error de la cuenta. Sin embargo, este proceso puede tardar mucho tiempo, potencialmente una hora o más. Para una recuperación más rápida, realice una

conmutación por error forzada en lugar de esperar a que se desencadene la conmutación por error administrada por el servicio.

- **Notificación**: Microsoft no le notifica automáticamente cuando una región está inactiva. Sin embargo:
  - Puede usar Azure Resource Health para supervisar el estado de un recurso individual y puede configurar alertas de Resource Health para notificarle problemas.
  - Puede usar Azure Service Health para comprender el estado general del servicio, incluidos los errores de región, y puede configurar alertas de Service Health para notificarle los problemas.
- **Solicitudes activas:** Es posible que cualquier solicitud activa se interrumpa y que el cliente tenga que volver a intentarla una vez finalizada la conmutación por error. Si los clientes controlan correctamente los errores transitorios mediante el reintento después de un breve período de tiempo, normalmente evitan un impacto significativo.
- **Pérdida de datos esperada:** Una interrupción en una región de lectura no causa pérdida de datos. Azure Cosmos DB sigue cumpliendo las garantías de coherencia de lectura.
- **Tiempo de inactividad esperado:** La cantidad de tiempo de inactividad que experimenta la cuenta depende del tipo de conmutación por error que usa la cuenta.
  - *PPAF:* Cuando PPAF está habilitado, el sistema detecta y recupera automáticamente el error, normalmente en un plazo de 3 minutos, sin intervención manual.
  - *Conmutación por error forzada:* El tiempo de inactividad depende de:
    - El tiempo que tardes en detectar la interrupción e iniciar la conmutación por error.
    - El tiempo que tarda la conmutación por error, que suele ser de unos segundos.

> ⚠ **Advertencia**
>
> No realice ninguna operación de configuración (plano de control) en la región afectada durante escenarios de interrupción, ya que dan como resultado una incoherencia de la cuenta y una recuperación retrasada. Algunos de los ejemplos de operaciones del plano de control que se deben evitar son:
>
> - Cambiar la región de escritura o modificar la prioridad de conmutación por error

- Actualice la cuenta a la configuración para escritura múltiple
- Actualización de los niveles de coherencia u otra configuración de la cuenta
- Actualización de configuraciones de punto de conexión privado o configuración de red
- Actualización del rendimiento de la cuenta u operaciones de escalado
- Cualquier otra operación que modifique la configuración de la cuenta o la configuración de región

  - *Conmutación por error administrada por el servicio:* Microsoft es responsable de iniciar la conmutación por error administrada por el servicio y el tiempo de inactividad que experimenta la cuenta se basa en el tiempo que tarda Microsoft declarar la interrupción e iniciar la conmutación por error. En algunas situaciones, puede tardar una hora o más. Si tu cuenta sufre una interrupción en las operaciones de escritura y necesitas restablecer rápidamente la disponibilidad de escritura, realiza una conmutación por error forzada.

- **Redistribución:** Para la conmutación por error forzada o la conmutación por error administrada por el servicio, la región afectada se desconecta y se marca como sin conexión.

  No es necesario realizar ningún cambio en el código de su aplicación para administrar la interrupción de la región de lectura. Los SDK Azure Cosmos DB redirigen las operaciones de lectura a la siguiente región disponible en la lista de regiones preferidas. Si ninguna de las regiones de la lista de regiones preferidas está disponible, las operaciones de lectura se revierten automáticamente a la región de escritura actual de la cuenta tal como está configurada en el servicio.

  > ⓘ **Nota:**
  >
  > Si usa puntos de conexión privados con una cuenta de Azure Cosmos DB, asegúrese de que el DNS privado se enruta correctamente después de la operación de región sin conexión. Para obtener instrucciones detalladas, consulte **Consideraciones sobre la conmutación por error para puntos de conexión privados**.

Comportamiento durante un fallo de región de escritura

En esta sección se describe qué esperar al configurar una cuenta de Azure Cosmos DB con varias regiones de lectura y se produce una interrupción en la región de escritura de la cuenta.

- **Detección y respuesta:** La responsabilidad de detectar la interrupción y responder depende del tipo de conmutación por error que usa su cuenta.
  - *PPAF:* Microsoft detecta automáticamente la interrupción e inicia una conmutación por error de algunas particiones, si procede. La aplicación no necesita realizar ninguna acción.
  - *Conmutación por error forzada:* Usted es responsable de realizar una conmutación por error forzada. Para conocer los pasos detallados, consulte Realizar una conmutación por error forzada para su cuenta de Azure Cosmos DB.

    Si no se realiza una conmutación por error, la cuenta perderá la disponibilidad de escritura hasta que se restablezca el servicio.

    Si hay una interrupción en la región de escritura de la cuenta, evite realizar una operación de *cambiar la región de escritura*. Los cambios en la región de escritura no se completan con éxito si se produce una interrupción en la región de origen o de destino. El motivo es que el procedimiento de cambio de región incluye una comprobación de coherencia que requiere conectividad entre las regiones.
  - *Conmutación por error gestionada por el servicio:* Microsoft detecta automáticamente la interrupción e inicia una conmutación por error de su cuenta. La aplicación no necesita realizar ninguna acción.
- **Notificación**: Microsoft no le notifica automáticamente cuando una región está inactiva. Sin embargo:
  - Puede usar Azure Resource Health para supervisar el estado de un recurso individual y puede configurar alertas de Resource Health para notificarle problemas.
  - Puede usar Azure Service Health para comprender el estado general del servicio, incluidos los errores de región, y puede configurar alertas de Service Health para notificarle los problemas.
- **Solicitudes activas:** Es posible que cualquier solicitud activa se interrumpa y que el cliente tenga que volver a intentarla una vez finalizada la conmutación por error. Si los clientes controlan correctamente los errores transitorios mediante el reintento después de un breve período de tiempo, normalmente evitan un impacto significativo.
- **Pérdida de datos esperada:** Si configura la cuenta con coherencia fuerte, no se produce ninguna pérdida de datos. De lo contrario, es posible que se pierdan las escrituras no

replicadas después de que se complete la conmutación por error. Para obtener información sobre la pérdida máxima de datos esperada durante una interrupción de la región, consulte Posibles pérdidas de datos durante las interrupciones de la región.

- **Tiempo de inactividad esperado:** La cantidad de tiempo de inactividad que experimenta la cuenta depende del tipo de conmutación por error que usa la cuenta.
  - *PPAF:* Cuando PPAF está habilitado, espere una breve interrupción, que normalmente es de aproximadamente 3 minutos.
  - *Conmutación por error forzada:* El tiempo de inactividad depende de:
    - El tiempo que tardes en detectar la interrupción e iniciar la conmutación por error.
    - El tiempo que tarda la conmutación por error, que suele ser de unos segundos.

    > ⚠ **Advertencia**
    >
    > No realice ninguna operación de plano de control en la región afectada durante los escenarios de interrupción, ya que provocan inconsistencias en la cuenta y retrasan la recuperación. Algunos de los ejemplos de operaciones del plano de control que se deben evitar son:
    > - Cambiar la región de escritura o modificar la prioridad de conmutación por error
    > - Actualice la cuenta a la configuración para escritura múltiple
    > - Actualización de los niveles de coherencia u otra configuración de la cuenta
    > - Actualización de configuraciones de punto de conexión privado o configuración de red
    > - Actualización del rendimiento de la cuenta u operaciones de escalado
    > - Cualquier otra operación que modifique la configuración de la cuenta o la configuración de región

  - *Conmutación por error administrada por el servicio:* Microsoft es responsable de iniciar la conmutación por error administrada por el servicio y el tiempo de inactividad que experimenta la cuenta se basa en el tiempo que tarda Microsoft declarar la interrupción e iniciar la conmutación por error. En algunas situaciones, puede tardar una hora o más. Para restaurar rápidamente la disponibilidad de escritura, realice una conmutación por error forzada.

- **Redistribución:** La redistribución del tráfico de escritura depende del tipo de conmutación por error que utilice su cuenta.

- *PPAF:* Azure Cosmos DB realiza automáticamente la conmutación por error de la partición defectuosa a una región en buen estado.
- *Conmutación por error forzada:* Al realizar una conmutación por error forzada, la región de escritura de la cuenta cambia a la región que especifique.

> ⓘ **Nota:**
>
> Si usa puntos de conexión privados con una cuenta de Azure Cosmos DB, asegúrese de que el DNS privado se enruta correctamente después de la operación de región sin conexión. Para obtener instrucciones detalladas, consulte **Consideraciones sobre la conmutación por error para puntos de conexión privados**.

- *Conmutación por error administrada por el servicio:* Azure Cosmos DB promueve automáticamente una de las regiones secundarias de la cuenta para que sea la nueva región de escritura principal. La conmutación por error se produce a otra región en el orden de prioridad de región que especifique.

Recuperación de regiones

Microsoft debe restaurar la operatividad de una región. Cuando una región se recupera después de una interrupción, Microsoft pone automáticamente la región en línea. Sin embargo, este proceso puede tardar varios días.

> ⓘ **Importante**
>
> Después de una conmutación por error forzada, Microsoft restablece automáticamente la conectividad de la región para las regiones saludables. Si la región no vuelve a estar en línea en un día o dos, abra un caso de soporte técnico para solicitar asistencia.

Una vez que la región está en línea, las acciones que realice son diferentes en función de si la interrupción se encontraba en una región de lectura o en una región de escritura.

- **Después de interrupciones de la región de lectura:** Cuando la región afectada vuelve a estar en línea, se sincroniza con la región de escritura actual y está disponible de nuevo para atender solicitudes de lectura después de haberse puesto al día por completo. Las siguientes lecturas se redirigen a la región recuperada sin necesidad de realizar cambios en el código de la aplicación. Tanto durante la conmutación por error como durante el

restablecimiento de una región con errores anteriores, Azure Cosmos DB sigue respetando las garantías de coherencia de lectura.

- **Después de interrupciones en la región de escritura:** Cuando la región afectada vuelve a estar en línea, la región se muestra como "en línea" en el portal de Azure y se vuelve disponible como una región de lectura. En este momento, es seguro cambiar la región de escritura a la región recuperada.

  > ⓘ **Importante**
  >
  > La región recuperada **no se promoverá automáticamente de nuevo como la región de escritura** una vez que se recupere. Es su responsabilidad **volver a cambiar a la región recuperada como región de escritura**, una vez que es seguro hacerlo.

  No hay *ninguna pérdida de datos o disponibilidad* antes, mientras, o después de cambiar la región de escritura. La aplicación sigue teniendo alta disponibilidad.

  Si alguna escritura no se replicó antes de que la región quedara fuera de línea, puede leer las escrituras no replicadas desde el canal de conflictos. La aplicación puede leer la fuente de conflictos, resolver los conflictos en función de la lógica específica de la aplicación y volver a escribir los datos actualizados en el contenedor según corresponda.

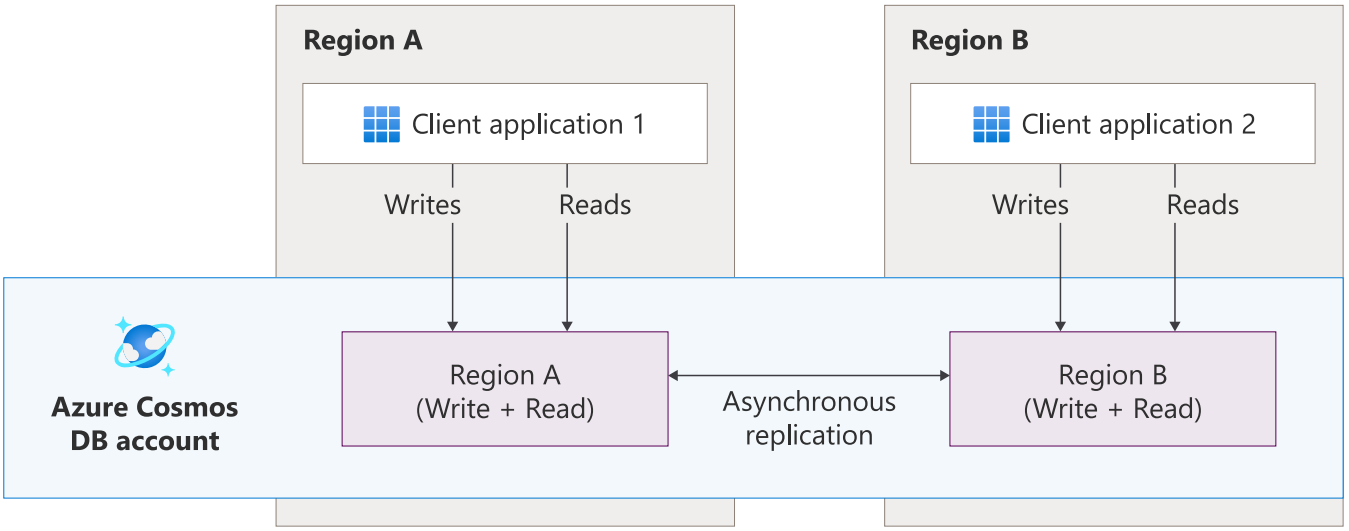
Prueba de fallos de región

Es posible que su aplicación no administre correctamente las conmutaciones por error de región, incluso si su cuenta de Azure Cosmos DB tiene alta disponibilidad. Para probar la alta disponibilidad de extremo a extremo de su aplicación como parte de sus pruebas de aplicaciones o simulacros de recuperación ante desastres (DR), deshabilite temporalmente la conmutación por error administrada por el servicio para la cuenta. Invoque la conmutación por error forzada mediante PowerShell, el CLI de Azure o el portal de Azure y, a continuación, supervise la aplicación. Una vez completada la prueba, puede volver a la región principal cuando esta vuelva a estar en línea automáticamente y, a continuación, restablecer la conmutación por error administrada por el servicio para la cuenta. Si la región no vuelve a estar en línea en un día o dos, abra un caso de soporte técnico para solicitar asistencia.

Si su cuenta usa PPAF, puede simular una conmutación por error de partición. Para obtener más información, consulte Probar la configuración de PPAF (simular error).

Varias regiones de escritura

Puede configurar Azure Cosmos DB para que acepte escrituras en varias regiones. Esta configuración puede proporcionar una resistencia muy alta a las interrupciones de la región. También es útil para reducir la latencia de escritura en aplicaciones distribuidas geográficamente.



Cuando se configura una cuenta de Azure Cosmos DB para varias regiones de escritura, no se admite la coherencia fuerte y pueden surgir conflictos de escritura. La región del hub actúa como árbitro en los conflictos de escritura. Para obtener más información sobre cómo resolver estos conflictos, consulte Tipos de conflictos y directivas de resolución al utilizar varias regiones de escritura.

Es importante tener en cuenta el diseño de la aplicación y cómo funciona con varias regiones de escritura. Revise los procedimientos recomendados para las escrituras multiregión.

Requirements

Region support: Puede configurar cualquier región de Azure como región de lectura o escritura para la cuenta de Azure Cosmos DB.

Cost

Agregar una región de escritura adicional a una cuenta de Azure Cosmos DB aumenta los costos existentes para cada región. Para obtener más información, consulte Precios de Azure Cosmos DB.

Configuración de varias regiones de escritura

Puede configurar varias regiones de escritura en su cuenta al crear la cuenta o en cualquier momento después de crearla. Para obtener más información, consulte Configuración de varias regiones de escritura.

Para usar eficazmente varias regiones de escritura, la aplicación también debe configurarse correctamente. Consulte Configuración de escrituras multiregión en aplicaciones que usan Azure Cosmos DB.

Planeamiento y administración de capacidad

Si la aplicación distribuye las solicitudes entre regiones y una región se queda sin conexión, las regiones restantes experimentan un volumen de solicitudes mayor. Utilice el escalado automático del rendimiento para adaptar dinámicamente la capacidad en función de la demanda. Si usa el rendimiento aprovisionado, planee la capacidad adecuada para controlar la pérdida de una región sin degradación del servicio y considere la posibilidad de sobreaprovisionar. Para más información, consulte Administración de la capacidad con sobreaprovisionamiento.

Comportamiento cuando todas las regiones están en buen estado

En esta sección se describe qué esperar al configurar una cuenta de Azure Cosmos DB con varias regiones de escritura y todas las regiones están operativas.

- **Operación entre regiones:** Cuando una cuenta está configurada con varias regiones de escritura, la aplicación configura la región que se debe usar para las operaciones de lectura y escritura. Puede configurar la aplicación con una lista prioritaria de regiones o excluir algunas regiones. Para obtener más información sobre cómo funciona la selección de regiones, consulte Diagnose y solucione la disponibilidad de los SDK de Azure Cosmos DB en entornos multirregionales. Para obtener información sobre cómo configurar la aplicación, consulte Configuración de escrituras multiregión en aplicaciones que usan Azure Cosmos DB.

- **Replicación de datos entre regiones:** Los datos se replican entre regiones de forma asincrónica. El retraso de replicación depende del nivel de coherencia de la cuenta. No se puede usar una coherencia fuerte para las escrituras de varias regiones. Para obtener más información, consulte Posibles pérdidas de datos durante interrupciones de región.

  Cuando una cuenta está configurada para varias regiones de escritura, las aplicaciones de diferentes regiones pueden realizar cambios que entren en conflicto entre sí. Azure Cosmos DB proporciona funcionalidades de resolución de conflictos. Para obtener más información, consulte Tipos de conflictos y directivas de resolución al usar varias regiones de escritura. Para obtener información sobre cómo configurar su propia directiva de resolución de conflictos, consulte Administrar directivas de resolución de conflictos en Azure Cosmos DB.

> ⓘ **Nota:**
>
> Actualizar el mismo identificador de documento con frecuencia o volver a crear el mismo identificador de documento con frecuencia después de que expire su TTL o se elimine, afecta negativamente al rendimiento de la replicación debido a un mayor número de conflictos generados en el sistema.

Comportamiento durante una falla de región

En esta sección se describe qué esperar al configurar una cuenta de Azure Cosmos DB con varias regiones de escritura y se produce una interrupción en una de las regiones de lectura o escritura de la cuenta.

- **Detección y respuesta:** La aplicación detecta la pérdida de la región. Azure Cosmos DB SDKs proporcionan funcionalidades automáticas de selección de regiones que enrutan las operaciones de lectura y escritura a regiones funcionales.
- **Notificación**: Microsoft no le notifica automáticamente cuando una región está inactiva. Sin embargo:
  - Puede usar Azure Resource Health para supervisar el estado de un recurso individual y puede configurar alertas de Resource Health para notificarle problemas.
  - Puede usar Azure Service Health para comprender el estado general del servicio, incluidos los errores de región, y puede configurar alertas de Service Health para notificarle los problemas.
- **Solicitudes activas:** Es posible que cualquier solicitud activa se interrumpa y que el cliente tenga que volver a intentarla una vez finalizada la conmutación por error. Si los clientes controlan correctamente los errores transitorios mediante el reintento después de un breve período de tiempo, normalmente evitan un impacto significativo.
- **Pérdida de datos esperada:** Los datos actualizados recientemente podrían dejar de estar disponibles en otras regiones. Para obtener información sobre la pérdida máxima de datos esperada durante una interrupción de la región, consulte Posibles pérdidas de datos durante las interrupciones de la región. En el improbable caso de que la región afectada sufre una pérdida de datos permanente, es posible que pierda datos no replicados.
- **Tiempo de inactividad esperado:** No hay tiempo de inactividad esperado en configuraciones de escritura múltiple, siempre que los SDK estén configurados

correctamente con `ApplicationRegions` o `PreferredRegions`.

> **Sugerencia**
>
> Para obtener los mejores resultados, las aplicaciones distribuidas globalmente deben estar delante de un servicio de equilibrio de carga global, como Azure Front Door o Azure Traffic Manager. Estos servicios pueden detectar la degradación regional y enrutar automáticamente el tráfico a instancias de aplicación en una región correcta.

- **Redistribution:** Los SDK de Azure Cosmos DB detectan automáticamente que la región es defectuosa y redireccionan las operaciones de lectura y escritura a la siguiente región disponible en la lista de regiones preferidas. No se requieren cambios en el código de la aplicación.

> **Sugerencia**
>
> Si la aplicación está gestionada por Azure Front Door o Traffic Manager, esos servicios también detectan la degradación regional y enrutan el tráfico a una región que esté funcionando correctamente.

Recuperación de regiones

Cuando la región afectada vuelve a estar en línea, la región se muestra como "en línea" en el portal de Azure y vuelve a estar disponible.

Cualquier dato de escritura que no se haya replicado cuando se produjo el error de la región queda disponible a través del canal de conflictos. Las aplicaciones pueden leer la fuente de conflictos, resolver los conflictos de acuerdo con la lógica específica de la aplicación y escribir los datos actualizados de nuevo en el contenedor de Azure Cosmos DB según corresponda.

Prueba de fallos de región

Para probar escenarios de conmutación por error de escritura multiregión, puede desconectar una región de escritura mediante una conmutación por error forzada. Este proceso simula una interrupción de la región y puede observar cómo responde la aplicación.

Copias de seguridad y restauración

Para la mayoría de las soluciones, no debe confiar exclusivamente en copias de seguridad. En su lugar, utilice las otras capacidades descritas en esta guía para apoyar los requisitos de resiliencia. Sin embargo, las copias de seguridad protegen contra algunos riesgos que otros enfoques no. Para más información, consulte ¿Qué son la redundancia, la replicación y la copia de seguridad?.

La pérdida de datos puede producirse debido a eliminaciones accidentales u otros problemas en la aplicación que provocan daños en los datos. Cuando se usa una cuenta de una sola región, la pérdida de datos también puede producirse debido a un desastre irrecuperable en la región de Azure Cosmos DB. Para ayudarle a protegerse contra la pérdida de datos, Azure Cosmos DB proporciona un conjunto de funcionalidades de copia de seguridad y restauración. Puede configurar copias de seguridad y retención en función de los requisitos de capacidad de recuperación y los requisitos de costo. Para más información, consulte Copia de seguridad en línea y restauración de datos a petición en Azure Cosmos DB.

Resistencia al mantenimiento del servicio

Azure Cosmos DB administra de forma transparente todos los detalles de los nodos de proceso individuales y realiza automáticamente la aplicación de revisiones y otros tipos de mantenimiento planeado. Los SLA de Azure Cosmos DB para la disponibilidad y la latencia se aplican a través de todas las operaciones de mantenimiento automáticas que realiza el sistema.

Acuerdo de nivel de servicio

El acuerdo de nivel de servicio (SLA) para Azure servicios describe la disponibilidad esperada de cada servicio y las condiciones que la solución debe cumplir para lograr esa expectativa de disponibilidad. Para obtener más información, consulte Acuerdos de Nivel de Servicio para servicios en línea.

Azure Cosmos DB proporciona acuerdos de nivel de servicio para una variedad de configuraciones y características de servicio, incluida la disponibilidad, la latencia, el rendimiento y la coherencia.

Los Acuerdos de Nivel de Servicio de disponibilidad son diferentes en función de si usa cualquiera de las siguientes funcionalidades de producto:

- Rendimiento aprovisionado
- Cuenta regional única con soporte para zona de disponibilidad (redundancia de zona)
- Cuentas que usan varias regiones de lectura
- Cuentas que usan varias regiones de escritura (nivel Crítico para la empresa)

Contenido relacionado

- confiabilidad Azure
- Información general de Azure Cosmos DB
- Niveles de coherencia en Azure Cosmos DB
- Distribución de datos global con Azure Cosmos DB
- Diagnose y solucione los problemas de disponibilidad de los SDK de Azure Cosmos DB en entornos multirregionales

ⓘ**Nota:** El autor ha creado este artículo con ayuda de inteligencia artificial. Más información

---

Last updated on 08/05/2026

Confiabilidad en Azure Database for MySQL

Azure Database for MySQL es un servicio de base de datos totalmente administrado que proporciona control y flexibilidad pormenorizados sobre las funciones de administración de bases de datos y las opciones de configuración. El servicio proporciona funcionalidades de alta disponibilidad (HA) y recuperación ante desastres (DR) en función de sus requisitos.

Al usar Azure, relibilidad es una responsabilidad compartida. Microsoft proporciona una variedad de capacidades para admitir resiliencia y recuperación. Es responsable de comprender cómo funcionan esas funcionalidades dentro de todos los servicios que usa y de seleccionar las funcionalidades que necesita para cumplir los objetivos empresariales y los objetivos de tiempo de actividad.

En este artículo se describe cómo hacer que Azure Database for MySQL sea resistente a diversas interrupciones y problemas potenciales, incluidos los errores transitorios, las interrupciones de zonas de disponibilidad, las interrupciones regionales y el mantenimiento del servicio. También se describe cómo puede usar copias de seguridad para recuperarse de otros tipos de problemas y se resalta cierta información clave sobre el acuerdo de nivel de servicio (SLA) de Azure Database for MySQL.

Recomendaciones de implementación de producción

Azure Well-Architected Framework proporciona recomendaciones sobre confiabilidad, seguridad, costo, operaciones y rendimiento. Para comprender cómo estas áreas influyen entre sí y contribuyen a una solución de Azure Database for MySQL confiable, consulte Procedimientos recomendados de arquitectura para Azure Database for MySQL.

Introducción a la arquitectura de confiabilidad

En esta sección se describen algunos de los aspectos importantes de cómo funciona el servicio que es más relevante desde una perspectiva de confiabilidad. En la sección se presenta la arquitectura lógica, que incluye algunos de los recursos y características que se implementan y usan. También se describe la arquitectura física, que proporciona detalles sobre cómo funciona el servicio en segundo plano.

Arquitectura lógica

Al trabajar con Azure Database for MySQL, se implementa un *server*, que representa los recursos de proceso y almacenamiento necesarios para admitir el servidor de bases de datos. Puede implementar una o varias *bases de datos* en el servidor.

Puede implementar servidores en los niveles de proceso ampliables, de uso general y *optimizados* para memoria. Cada nivel de proceso está optimizado para diferentes tipos de cargas de trabajo.

Para obtener más información sobre la arquitectura general del servicio y los modelos de implementación, consulte Azure Database for MySQL información general.

Arquitectura física

- **Separación entre proceso de cómputo y almacenamiento:** Azure Database for MySQL usa una arquitectura de separación entre proceso de cómputo y almacenamiento para ofrecer alta disponibilidad. El motor de base de datos se ejecuta en una máquina virtual (VM). Los archivos de datos se almacenan en Azure Storage, que mantiene sincrónicamente tres copias de los datos para protegerse frente a errores de hardware de almacenamiento. Según la configuración de alta disponibilidad del servidor, los archivos de datos se pueden almacenar en el almacenamiento con redundancia de zona (ZRS) o en el almacenamiento con redundancia local (LRS).

- **Alta disponibilidad:** Opcionalmente, puede habilitar una *configuración de alta disponibilidad* en el servidor. Al habilitar la configuración de alta disponibilidad, el servicio aprovisiona y mantiene un servidor de réplica en espera activa. Los cambios de datos en el servidor principal se replican sincrónicamente en el servidor de réplica en espera para garantizar una pérdida de datos cero durante un error del servidor principal.

  La arquitectura separa la capa de proceso de la capa de almacenamiento, lo que permite al servicio controlar los distintos tipos de errores de forma adecuada. Para lograr una mayor resistencia, puede distribuir los servidores entre zonas de disponibilidad.

  Un servidor de réplica en espera se implementa en la misma configuración de máquina virtual que el servidor principal, incluidos los núcleos virtuales, el almacenamiento y la configuración de red.

  Puede cambiar entre servidores mediante la realización de una *conmutación por error*. Use *conmutaciones por error no planeadas* cuando se produzca un error en el servidor principal y use *conmutaciones por error planeadas* cuando necesite minimizar el tiempo de inactividad de la aplicación durante una conmutación por error.

Para obtener más información, consulte HA en Azure Database for MySQL.

- **Backups:** Azure Database for MySQL crea automáticamente copias de seguridad del servidor. Para obtener más información, consulte Copia de seguridad y restauración.

Resistencia a errores transitorios

Los errores transitorios son errores breves e intermitentes en los componentes. Se producen con frecuencia en un entorno distribuido como la nube y son una parte normal de las operaciones. Los errores transitorios se corrigen después de un breve período de tiempo. Es importante que las aplicaciones puedan controlar errores transitorios, normalmente mediante el reintento de solicitudes afectadas.

Todas las aplicaciones hospedadas en la nube deben seguir las instrucciones de control de errores transitorios Azure cuando se comunican con cualquier API, bases de datos y otros componentes hospedados en la nube. Para obtener más información, consulte Recomendaciones para controlar errores transitorios.

Las aplicaciones deben controlar los errores de conectividad transitorios que pueden producirse durante el mantenimiento, las operaciones de escalado o las interrupciones de red. Siga estas recomendaciones:

- ✓ Cuando la aplicación encuentre errores transitorios, vuelva a intentar la operación mediante retroceso exponencial. Aumente el retraso entre reintentos y limite el número de intentos. Si la operación sigue fallando después del número máximo de reintentos, trátela como un fallo.
- ✓ Cuando sea posible, use bibliotecas cliente que gestionen automáticamente los reintentos.
- ✓ Los errores transitorios que se producen durante las operaciones de escritura requieren una consideración más cuidadosa. Considere la posibilidad de hacer que las operaciones de escritura sean idempotentes para poder ejecutarlas varias veces.

Resistencia a errores de zona de disponibilidad

Availability zones son grupos de centros de datos físicamente independientes dentro de una región de Azure. Cuando una zona falla, los servicios pueden transferirse a una de las zonas restantes.

Seleccione el tipo de soporte para zonas de disponibilidad mediante la configuración de alta disponibilidad. Al habilitar la alta disponibilidad, Azure Database for MySQL implementa un *servidor de réplica en espera* junto con el servidor principal. Este modelo de alta disponibilidad

ayuda a garantizar que los datos confirmados nunca se pierdan durante los errores. Sea cual sea el modelo de implementación de alta disponibilidad que elija, el servicio confirma los datos de forma sincrónica en ambos servidores de réplica, el principal y el en espera. Si se interrumpe el servidor principal, el sistema cambia automáticamente al servidor réplica en espera.

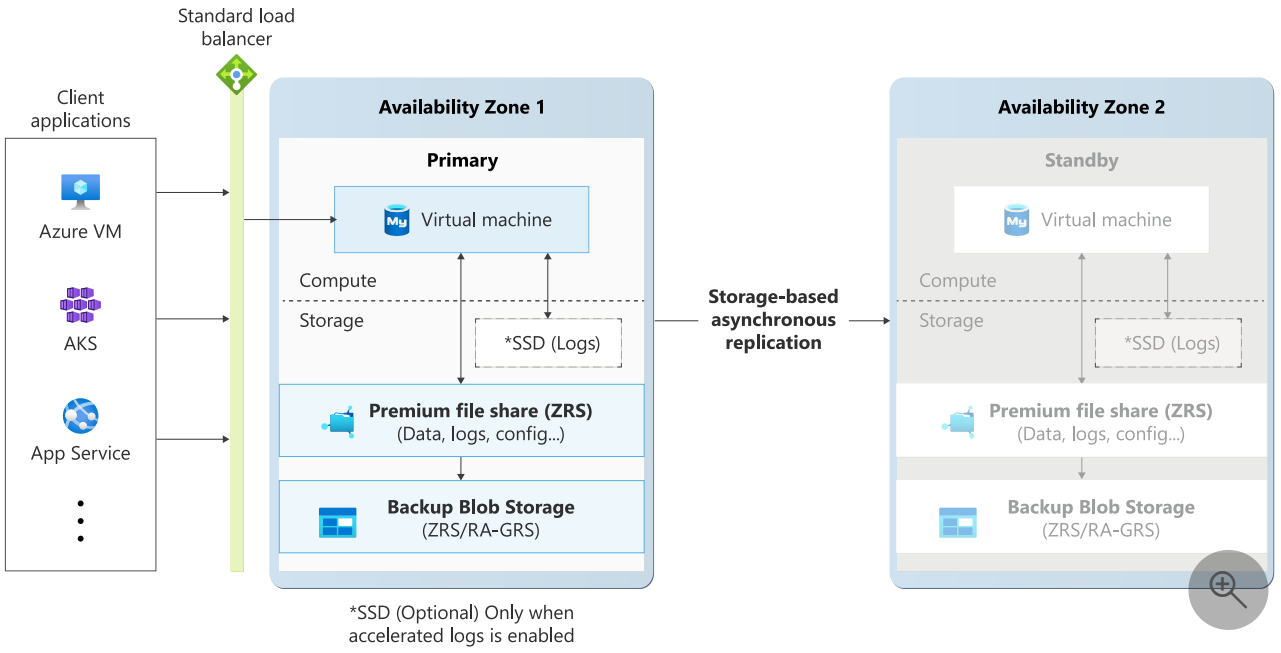
El servicio almacena datos en Azure Files Premium Storage. En función de la configuración de alta disponibilidad del servidor, usa ZRS o LRS, que almacena tres copias de datos dentro o entre zonas de disponibilidad.

Azure Database for MySQL admite dos tipos de configuración de zona de disponibilidad cuando se usa alta disponibilidad:

- **Alta disponibilidad con redundancia de zona:** La redundancia de zona proporciona el nivel más alto de resistencia de zona mediante la implementación de un servidor principal en una zona de disponibilidad y un servidor de réplica en espera en una zona de disponibilidad diferente. El servidor de réplicas en espera usa una configuración de red, almacenamiento y proceso similar al servidor principal. Una configuración con redundancia de zona proporciona aislamiento físico de toda la pila entre los servidores principal y en espera.

  Seleccione las zonas de disponibilidad para los servidores principales y en espera.

  Use implementaciones con redundancia de zona para servidores de producción.

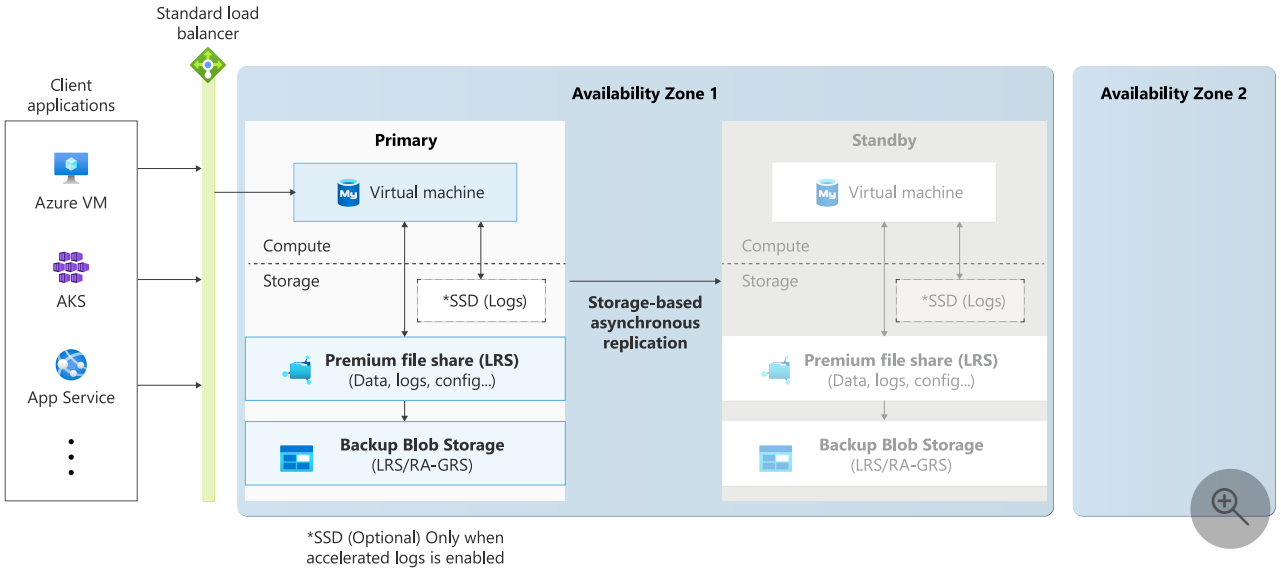


  Las operaciones de escritura pueden experimentar un pequeño aumento de la latencia de confirmación porque el servicio replica de forma sincrónica los datos en el servidor en espera. De media, cabe esperar un aumento de la latencia del 5 % al 10 % en las

operaciones de escritura y confirmación de la aplicación. El impacto varía según la carga de trabajo, la SKU seleccionada y la región.

- **Alta disponibilidad con redundancia local:** Los servidores principales y en espera usan la misma zona de disponibilidad. Si se produce una interrupción en el servidor principal, pero la zona sigue operativa, el servidor conmuta automáticamente al servidor en espera.

  Una implementación con redundancia local proporciona alta disponibilidad dentro de una sola zona de disponibilidad. Protege contra errores de nivel de nodo y también ayuda a reducir el tiempo de inactividad de la aplicación durante los eventos de tiempo de inactividad planeados y no planeados. Sin embargo, no protege contra una interrupción en esa zona. En las regiones que tienen zonas de disponibilidad, este tipo de configuración también se denomina a veces *zonal* o *de una sola zona*.



  Use HA con redundancia local solo en los siguientes escenarios:

  - Cuando se tienen aplicaciones inusualmente sensibles a la latencia, se valida la necesidad de minimizar la latencia entre la réplica principal y la secundaria, y se planifica la resiliencia entre zonas utilizando otros enfoques arquitectónicos.

  - Al implementar en una región que no admite zonas de disponibilidad. En este escenario, la región funciona como una sola zona, por lo que la HA con redundancia local es la única opción de alta disponibilidad.

  Los servidores están en la misma zona, lo que puede reducir la latencia de escritura a las aplicaciones que se implementan dentro de esa zona.

Si configura su servidor sin alta disponibilidad, funcionará en un solo servidor. Si se produce un error en ese servidor o su zona, el servidor no está disponible.

Requisitos

- **Compatibilidad con regiones:** Azure Database for MySQL admite diferentes configuraciones de zona de disponibilidad en función de la región de Azure. Para obtener una lista completa de las regiones, incluidos los tipos de compatibilidad con zonas de disponibilidad y las consideraciones específicas de cada región, consulte Regiones de Azure.
- **Nivel de servicio:** HA requiere los niveles General Purpose o Memory Optimized. El nivel de ráfaga no admite HA (alta disponibilidad), ni con redundancia de zona ni con redundancia local.

Costo

Cuando habilita la HA, crea y paga el servidor en espera a la misma tarifa que el servidor principal. La configuración de la zona de disponibilidad no afecta al costo. La replicación de datos no conlleva ningún cargo dentro o entre zonas de disponibilidad. Dependiendo del volumen de almacenamiento de copia de seguridad, también se le puede facturar por el almacenamiento de copia de seguridad. Para obtener información detallada sobre los precios, consulte los precios de Azure Database for MySQL.

Consideraciones

- **Claves principales:** Use claves principales en todas las tablas porque este enfoque reduce el tiempo de replicación y conmutación por error.
- **Limitaciones y problemas conocidos:** Revise la lista de limitaciones y problemas conocidos.

Configurar soporte de zonas de disponibilidad

Para configurar la compatibilidad con zonas de disponibilidad para un servidor, configure las opciones de alta disponibilidad.

> **Nota:**
>
> Cuando selecciona qué zonas de disponibilidad usar, en realidad está seleccionando la *zona de disponibilidad lógica*. Si implementa otros componentes de carga de trabajo en una suscripción de Azure diferente, pueden usar un número de zona de disponibilidad lógica diferente para acceder a la misma zona de disponibilidad física. Para obtener más información, consulte zonas de disponibilidad física y lógica.

- **Cree un servidor con redundancia de zona.** Para obtener información sobre cómo crear un servidor con alta disponibilidad y redundancia de zona habilitada, consulte los siguientes artículos:
  - Azure portal: administre alta disponibilidad con redundancia de zona en Azure Database for MySQL mediante el portal de Azure.
  - CLI de Azure: Administrar la alta disponibilidad con redundancia de zona en Azure Database for MySQL mediante CLI de Azure.
- **Cree un servidor con redundancia local.** Para crear un servidor con alta disponibilidad con redundancia local en una sola zona de disponibilidad, debe usar el CLI de Azure u otro método de implementación mediante programación. Para obtener las instrucciones de CLI de Azure, consulte Habilitación de alta disponibilidad durante la creación del servidor.
- **Cambie la configuración de zona de disponibilidad para los servidores existentes.** Si tiene un servidor existente, el enfoque que sigue para habilitar la compatibilidad con la zona de disponibilidad depende de la configuración inicial del servidor.
  - Para cambiar un servidor existente a alta disponibilidad con redundancia de zona, debe migrar a un nuevo servidor. Para obtener más información, consulte Migración de un servidor existente a un servidor con redundancia de zona.
  - Para cambiar un servidor existente a alta disponibilidad con redundancia local:
    1. Desactive HA, si está habilitada.
    2. Habilitar la HA con redundancia local. Debe usar el CLI de Azure u otro método de implementación mediante programación. Para obtener instrucciones sobre la CLI de Azure, consulte Administrar la HA con redundancia de zona en Azure Database for MySQL mediante la CLI de Azure.
- **Desactive HA.** Al deshabilitar la alta disponibilidad, se elimina el servidor réplica en espera, por lo que el servidor no es resiliente frente a interrupciones a nivel de zona. Sin embargo, si las copias de seguridad con redundancia geográfica están habilitadas, todavía puede recuperar el servidor en otra región mediante esas copias de seguridad. Para obtener más información, consulte Deshabilitar HA.

Comportamiento cuando todas las zonas están en buen estado

En esta sección se describe qué esperar al configurar servidores con compatibilidad con alta disponibilidad y zona de disponibilidad, y todas las zonas de disponibilidad están operativas.

- **Operación entre zonas:** Las aplicaciones cliente mySQL se conectan al servidor principal mediante el nombre de dominio completo (FQDN) del servidor de bases de datos. Evite usar la dirección IP del servidor principal porque la dirección IP puede cambiar, incluso durante las conmutaciones por error.

  Azure Database for MySQL usa una configuración activa-pasiva en la que el servidor principal controla todas las conexiones de base de datos y las consultas de la zona de disponibilidad principal. El servidor de réplicas en espera no atiende el tráfico de cliente durante las operaciones normales.

- **Replicación de datos entre zonas:** Las escrituras se confirman en el servidor principal y se escriben sincrónicamente en los registros del servidor en espera mediante ZRS. El servidor principal no espera a que el servidor en espera aplique los registros, pero dado que los registros están en ZRS, están disponibles incluso si se produce un error de réplica o zona.

  Los efectos de la replicación son diferentes en función de la configuración de zona de disponibilidad que use el servidor:

  - *Con redundancia de zona:* Dado que los servidores están en zonas independientes, este enfoque garantiza una pérdida de datos cero durante un error de zona. Esta situación también se conoce como lograr un objetivo de punto de recuperación (RPO) de cero en caso de fallos de zona.

    Sin embargo, la replicación entre zonas podría introducir una pequeña cantidad de latencia adicional. En promedio, puede esperar 5% a 10% mayor latencia para las escrituras y confirmaciones de la aplicación, pero el impacto varía según la carga de trabajo, la SKU seleccionada y la región.

  - *Redundancia local:* No se replica ningún tráfico entre zonas.

  > ⓘ **Nota:**
  >
  > El sistema replica todos los cambios en tiempo real en el servidor de réplicas en espera, incluidos los errores de usuario no deseados, como una eliminación accidental de una tabla o actualizaciones de datos incorrectas. Debido a la replicación inmediata, no puede usar la réplica en espera para la recuperación. Para recuperarse de errores de

usuario, debe realizar una restauración en un punto específico en el tiempo desde un respaldo. Para obtener más información, consulte Copia de seguridad y restauración.

Comportamiento durante un fallo de zona

En esta sección se describe qué cabe esperar cuando se configuran servidores con compatibilidad con HA y con zonas de disponibilidad, y se produce una interrupción en una zona de disponibilidad.

- **Detección y respuesta:** Azure comprueba periódicamente el estado de los servidores principal y en espera. Después de varios pings, si la supervisión de estado detecta que no se puede acceder a un servidor principal, el servicio inicia una conmutación automática al servidor en espera. El algoritmo de supervisión de estado usa varios puntos de datos para evitar situaciones de falsos positivos.

  Si se produce un error de zona, el comportamiento difiere en función de la configuración de zona de disponibilidad que use el servidor:

  - *Zone-redundant:* Azure Database for MySQL detecta automáticamente errores de zona de disponibilidad supervisando continuamente varios puntos de conexión de servidor. Para obtener más información, consulte Cómo funciona la detección automática de conmutación por error en servidores habilitados para alta disponibilidad.

    Para ver los posibles estados de HA, consulte Monitor HA. Cuando se produce un error en una zona, Azure inicia una conmutación por error no planificada al servidor en espera sin necesidad de que usted tome medidas.

  - *Con redundancia local:* Los servidores principales y en espera no están disponibles si la zona de disponibilidad que hospeda un servidor con redundancia local deja de estar disponible. En este escenario, el servicio no proporciona conmutación automática por error. Es responsable de detectar la interrupción de zona y realizar acciones de recuperación, como restaurar copias de seguridad con redundancia de zona en un servidor independiente en otra zona de disponibilidad o región.

- **Notificación:** Microsoft no le notifica automáticamente cuando una zona está inactiva. Sin embargo, puede usar Azure Resource Health para supervisar el estado de un recurso individual y puede configurar Resource Health alertas para notificarle problemas. También puede usar Azure Service Health para comprender el estado general del servicio, incluidos los errores de zona, y puede configurar alertas de Service Health para notificarle problemas.

Azure Database for MySQL genera un evento de Azure Resource Health cuando ocurre una conmutación por error no planificada.

- **Solicitudes activas:** Cuando una zona de disponibilidad deja de estar disponible, es posible que finalicen las solicitudes en curso a los servidores de la zona afectada. Las aplicaciones deben reintentar estas solicitudes. Si los clientes controlan correctamente los errores transitorios mediante el reintento después de un breve período de tiempo, normalmente evitan un impacto significativo.

- **Pérdida de datos esperada:** La cantidad de pérdida de datos depende de la configuración de la zona de disponibilidad del servidor.

  - *Redundante por zona:* Se anticipa una pérdida de datos cero durante el fallo de zona debido a la replicación sincrónica entre los servidores principal y en espera en distintas zonas.

  - *Con redundancia local:* Los datos de los servidores de la zona afectada no están disponibles hasta que se recupere la zona.

- **Tiempo de inactividad esperado:** La cantidad de tiempo de inactividad depende de la configuración de zona de disponibilidad que usa el servidor.

  - *Con redundancia de zona:* La conmutación por error se completa normalmente en un plazo de 60 a 120 segundos. Si los clientes controlan correctamente los errores transitorios mediante el reintento después de un breve período de tiempo, normalmente evitan un impacto significativo.

  - *Con redundancia local:* Los servidores de una zona afectada no están disponibles hasta que se recupere la zona de disponibilidad.

- **Redistribución:** El comportamiento de reenrutamiento del tráfico depende de la configuración de zona de disponibilidad que usa el servidor.

  - *Con redundancia entre zonas:* Después de la conmutación por error, el servidor en espera se convierte en el nuevo servidor principal y comienza a aceptar nuevas conexiones. Azure establece automáticamente un servidor en espera en la zona principal original después de recuperarse. Para obtener más información, consulte Conmutación por error no planeada.

  - *Con redundancia local:* Cuando una zona no está disponible, el servidor no está disponible. Si tiene un servidor independiente que ha creado previamente en otra zona

de disponibilidad o región, es responsable de redirigir el tráfico a ese servidor.

Recuperación de zona

El comportamiento de recuperación de zona depende de la configuración de zona de disponibilidad que usa el servidor.

- *Con redundancia de zona:* Cuando se recupera la zona de disponibilidad, Azure Database for MySQL vuelve a generar automáticamente el servidor en espera en la zona recuperada y lo sincroniza con el servidor principal actual. A continuación, la zona recuperada actúa como ubicación en espera. El servicio no mueve automáticamente el rol principal a la zona original para evitar interrupciones innecesarias. Si desea devolver el primario a la zona original, puede iniciar manualmente una conmutación por error planificada.

- *Con redundancia local:* Una vez que la zona está en buen estado, los servidores de la zona están disponibles de nuevo. Es responsable de los procedimientos de recuperación de zona y la sincronización de datos que requieren las cargas de trabajo.

Prueba de fallos de zona

Las opciones para probar errores de zona dependen de la configuración de zona de disponibilidad que usa la instancia.

- *Con redundancia de zona:* Puede probar la resistencia de su aplicación ante la conmutación por error iniciando una conmutación por error planeada. Utilice una conmutación por error planeada para simular una interrupción imprevista mientras la carga de trabajo se ejecuta y observe el tiempo de inactividad de su aplicación. Ejecutar simulaciones en entornos que no son de producción o en un momento silencioso. Para obtener más información, consulte Conmutación por error programada.

- *Con redundancia local:* No se puede simular una interrupción de zona completa, pero puede simular que el servidor no está disponible de forma similar a lo que ocurre durante una interrupción de zona. Para obtener más información, consulte los artículos siguientes:

  - Portal de Azure: Detener e iniciar una instancia de servidor flexible de Azure Database for MySQL.

  - CLI de Azure: reinicie, detenga e inicie una instancia de servidor flexible Azure Database for MySQL.

Resistencia a errores en toda la región

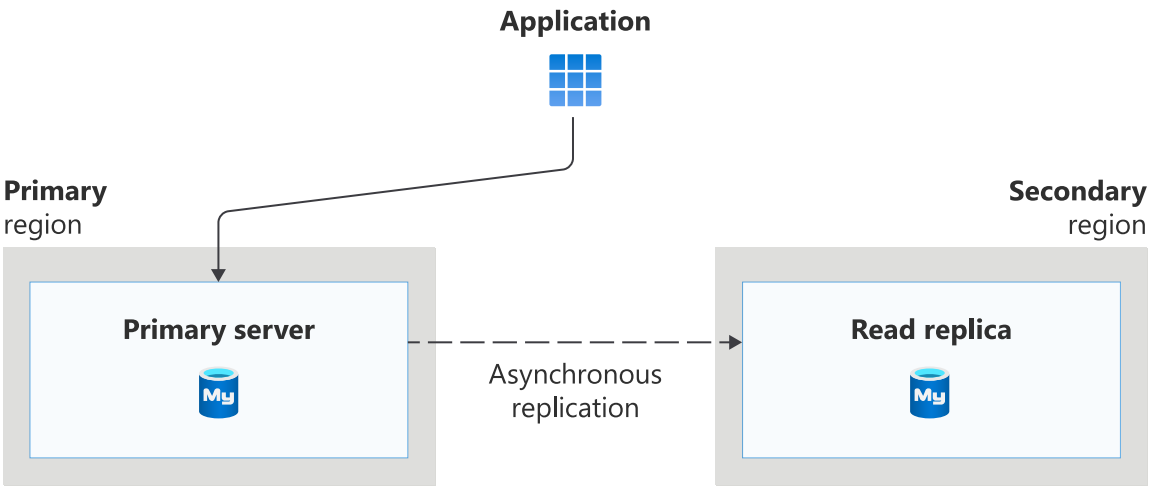
Azure Database for MySQL admite *réplicas de lectura entre regiones*, que puede usar para mantener una copia sincronizada de la base de datos en otra región para una recuperación más rápida.

También puede usar copias de seguridad con redundancia geográfica, en regiones admitidas, para proporcionar recuperación entre regiones. Sin embargo, las copias de seguridad suelen implicar más tiempo de inactividad y pérdida de datos que la replicación. Para obtener más información, consulte Copia de seguridad y restauración.

Réplicas de lectura entre regiones

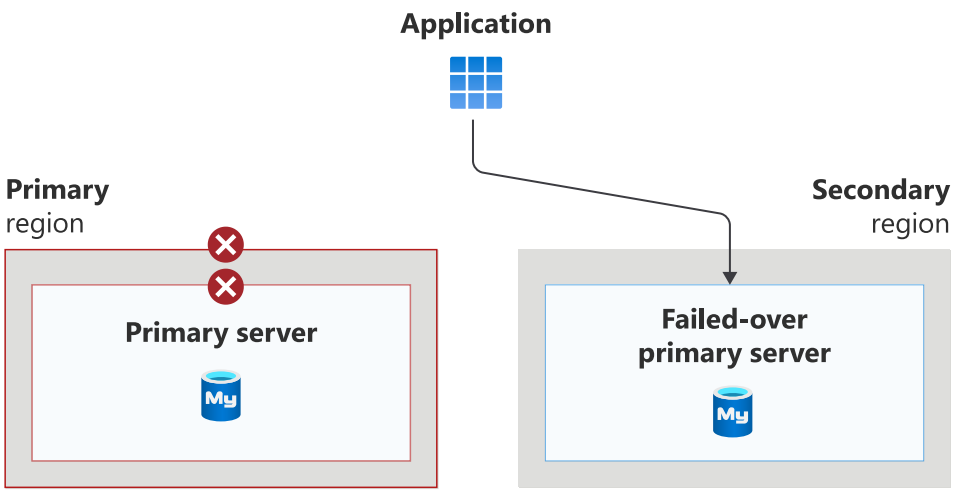
Implemente réplicas de lectura para proteger las bases de datos frente a errores de nivel de región. Cada réplica de lectura es un servidor Azure Database for MySQL independiente. Cuando colocas una réplica de lectura en una segunda región de Azure, tu servidor de bases de datos puede ofrecer resiliencia ante un problema que afecte a toda la región. Puede implementar hasta 10 réplicas de lectura, que opcionalmente pueden estar en diferentes regiones de Azure.

La tecnología de replicación física de MySQL actualiza las réplicas de lectura de forma asincrónica a partir del servidor de origen de la región principal, lo que significa que las réplicas pueden ir con retraso respecto al origen. Las réplicas de lectura entre regiones pueden servir opcionalmente cargas de trabajo de solo lectura para reducir la latencia de las aplicaciones distribuidas globalmente o descargar el tráfico de lectura desde el servidor de origen. Para obtener más información sobre las características y consideraciones relacionadas con las réplicas de lectura, consulte Réplicas de lectura.



Si la región primaria falla, puede realizar una conmutación por error manual para convertir la réplica secundaria en el servidor principal. Para realizar manualmente una conmutación por error, se detiene el proceso de replicación, lo que convierte la réplica de lectura en un servidor de

lectura/escritura. Debido a la replicación asincrónica, la conmutación por error puede provocar la pérdida de datos. La aplicación debe conectarse al nuevo servidor principal y es responsable de la reconfiguración de la aplicación.



> ⓘ **Nota:**
>
> En esta sección se resume alguna de la información importante sobre cómo las réplicas de lectura pueden admitir resiliencia ante fallos a nivel regional. También puede utilizar réplicas de lectura para mejorar el rendimiento y dar soporte a bases de usuarios geográficamente distribuidas y a gran escala.

Requisitos

- **Region support:** Puede crear réplicas de lectura entre regiones en cualquier región que admita Azure Database for MySQL. No se limita a las regiones emparejadas de Azure.
- **Niveles de cómputo:** Los niveles de cómputo de uso general y optimizados para memoria admiten réplicas de lectura. El nivel Ampliable no admite réplicas de lectura.

Consideraciones

- **Diferencias de configuración:** Al crear una réplica, hereda varias opciones de configuración del servidor de origen, incluida la generación de procesos, los núcleos virtuales y el almacenamiento. Puede personalizar estos valores en la réplica de lectura después de crearla, pero es mejor usar valores iguales o superiores para asegurarse de que la réplica pueda mantenerse al día con los cambios en la fuente.
- **Monitorización del retraso de replicación:** El proceso de replicación asíncrona implica un retraso de replicación, que puede variar en función de varios factores. Cuando el retraso de

replicación es muy alto, es posible que el servidor experimente problemas. Es importante supervisar el retraso de replicación para que pueda mitigar los problemas antes de que se escalen. Para más información, consulte Supervisión de la replicación.

- **HA:** Las réplicas de lectura no pueden tener HA habilitada y, cuando se realiza una conmutación por error para convertirlas en el servidor principal, tampoco tienen HA. Usted es responsable de configurar la alta disponibilidad tras la conmutación por error a una réplica.

Costo

Las réplicas de lectura incurren en costos de proceso y almacenamiento, así como cargos por transferencia de datos entre regiones para la replicación. Para obtener información detallada sobre los precios, consulte precios de Azure Database for MySQL y precios de ancho de banda.

Configuración de la compatibilidad con varias regiones

- **Cree una réplica de lectura:** Para obtener información sobre cómo crear una réplica de lectura, consulte los artículos siguientes:
  - Azure portal: Cómo crear y administrar réplicas de lectura en Servidor flexible de Azure Database for MySQL mediante el portal de Azure
  - CLI de Azure: Cómo crear y administrar réplicas de lectura en el servidor flexible de Azure Database for MySQL mediante CLI de Azure

  Puede configurar réplicas después de crear el servidor de origen solo si el servidor de origen está en ejecución y es accesible.

- **Detener la replicación:** Para obtener información sobre cómo detener la replicación, consulte Detener la replicación en un servidor de réplicas.

- **Eliminar una réplica de lectura:** Para obtener información sobre cómo eliminar una réplica de lectura, consulte Eliminación de un servidor de réplicas.

Comportamiento cuando todas las regiones están en buen estado

En esta sección se describe qué esperar al configurar el servidor con una réplica de lectura en otra región y todas las regiones están operativas:

- **Enrutamiento del tráfico entre regiones:** Durante las operaciones normales, la aplicación debe dirigir el tráfico de lectura y escritura al servidor de origen en la región primaria. Opcionalmente, puede dirigir las solicitudes de lectura a la réplica de lectura.

- **Replicación de datos entre regiones:** Las réplicas de lectura entre regiones usan la replicación asincrónica para minimizar el impacto en el rendimiento del servidor de origen. La cantidad de retraso de replicación depende de varios factores, incluida la carga de escritura y la latencia entre el servidor de origen y las réplicas. El retraso de replicación suele ser al menos varios minutos, pero puede ser mucho más largo. Para obtener más información, consulte Monitor replication y para obtener instrucciones detalladas, consulte Monitor replication in the Azure portal.

Comportamiento durante una falla de región

En esta sección se describe qué cabe esperar cuando se configura un servidor con compatibilidad con réplicas de lectura entre regiones y se produce una caída en la región primaria.

- **Detección y respuesta:** Eres responsable de detectar una caída en la región principal y activar manualmente una conmutación por error. Esta acción puede provocar la pérdida de datos no replicados.

  > ⓘ **Importante**
  >
  > Usted es responsable de desencadenar la conmutación por error. Azure no conmuta por error para leer las réplicas automáticamente, incluso si se produce un error en una región.

  La conmutación por error requiere que realice los pasos siguientes:

  1. Detenga la replicación. Este procedimiento es irreversible y el servidor no se puede volver a convertir en una réplica. El proceso da como resultado la pérdida de datos. Para obtener más información sobre las implicaciones de esta acción, vea Detener la replicación.

  2. Vuelva a configurar la aplicación para usar el nuevo servidor principal.

  Para obtener más información, consulte Conmutación por error.

- **Notification:** Microsoft no le notifica automáticamente cuando una región está inactiva. Sin embargo, puede usar Azure Service Health para comprender el estado general del servicio,

incluidos los errores de región, y puede configurar alertas de Service Health para notificarle problemas.

- **Solicitudes activas:** Todas las conexiones activas a la región de origen se quitan si el servidor de origen no está disponible. Las aplicaciones deben reintentar la realización de conexiones al nuevo servidor principal una vez completado el proceso de conmutación por error.

- **Pérdida de datos esperada:** Durante una interrupción en la región, debe realizar una conmutación por error para detener la replicación. Este proceso produce una pérdida permanente de datos no replicados.

  La cantidad de pérdida de datos depende del retraso de replicación en el momento de la interrupción. El retraso de replicación suele ser al menos varios minutos, pero puede ser mucho más largo. Para más información, consulte Supervisión de la replicación.

- **Tiempo de inactividad esperado:** La detención de la replicación se completa normalmente en dos minutos después de desencadenar la acción. Es responsable de volver a configurar las aplicaciones para conectarse al nuevo servidor principal. El tiempo que se tarda en realizar la reconfiguración también contribuye al tiempo de inactividad general.

- **Reenrutamiento del tráfico:** Es responsable de volver a configurar las aplicaciones para conectarse al nuevo servidor principal.

  > ⓘ **Nota:**
  >
  > Después de conmutar por error una réplica de lectura para convertirse en el servidor principal, el servidor no tiene habilitada la alta disponibilidad. Debe activar HA manualmente o incluirla en su automatización.

Recuperación de regiones

Cuando la región se recupere, usted será responsable de las actividades de retorno tras la recuperación para reanudar las operaciones en la región primaria. Microsoft no mueve automáticamente el servidor principal. Puede crear una nueva réplica de lectura en la región primaria y, a continuación, realizar otro proceso de conmutación por error para restaurar operaciones en la región primaria. Tenga en cuenta uno de los enfoques siguientes, en función de si la aplicación puede tolerar tiempo de inactividad o pérdida de datos:

- Ponga la aplicación fuera de servicio y espere a que la replicación se complete con todos los cambios. Este enfoque requiere tiempo de inactividad de la aplicación que sea aproximadamente el mismo que el retraso de replicación.
- Realice la conmutación por error y acepte la pérdida de datos no replicados.

Recuerde que también es responsable de volver a configurar las aplicaciones para conectarse al nuevo servidor principal según sea necesario.

Prueba de fallos de región

Pruebe periódicamente los procedimientos de conmutación por error de las réplicas de lectura para garantizar que sus procesos sean válidos y que las capacidades cumplan su objetivo de tiempo de recuperación (RTO) y sus requisitos de RPO.

Puede realizar la conmutación de una réplica de lectura para que se convierta en el servidor principal en cualquier momento, incluso cuando todas las regiones están en buen estado. Recomendamos realizar estas pruebas en un entorno de no producción, porque hacerlas puede provocar la pérdida de datos y requiere una conmutación por recuperación manual.

Como parte de la estrategia de recuperación ante desastres, ejecute periódicamente simulacros de recuperación completa. Estos simulacros incluyen la validación de datos, las pruebas de funcionalidad de la aplicación y los procedimientos documentados de reversión.

Copias de seguridad y restauración

Azure Database for MySQL realiza automáticamente una copia de seguridad de los datos, por lo que puede restaurarlos a cualquier momento dado dentro del período de retención de copia de seguridad. Esta protección le ayuda a evitar daños accidentales y la eliminación de datos. Microsoft administra completamente las copias de seguridad sin interrumpir la disponibilidad del servidor. Las copias de seguridad incluyen copias de seguridad completas y copias de seguridad del registro de transacciones.

- **Almacenamiento de copia de seguridad:** Si configura el servidor con alta disponibilidad con redundancia entre zonas, el sistema almacena las copias de seguridad en ZRS. En el caso de los servidores configurados sin alta disponibilidad o con alta disponibilidad con redundancia local, el sistema almacena copias de seguridad en LRS.

  En las regiones de Azure que tienen una región emparejada, puede configurar el almacenamiento con redundancia geográfica (GRS) para las copias de seguridad cuando cree el servidor. Este enfoque replica las copias de seguridad en la región emparejada de

Azure para una protección adicional frente a errores de región. El sistema replica las copias de seguridad de forma asincrónica.

El período de retención de copia de seguridad predeterminado es de 7 días, pero puede ampliar la retención hasta 35 días. Todas las copias de seguridad están cifradas.

- **Restaurar:** La restauración a un momento dado (PITR) le permite restaurar la base de datos en cualquier momento dentro del período de retención de copia de seguridad. El proceso de restauración crea un nuevo servidor de bases de datos con un nuevo nombre de servidor proporcionado por el usuario. Puede usar el nuevo servidor as-is o copiar datos de él.

  Al restaurar una copia de seguridad con redundancia geográfica, se crea un nuevo servidor en la región emparejada. En algunas regiones, puede usar Universal Geo-Restore para restaurar una copia de seguridad con redundancia geográfica en una región que no sea la emparejada con la primaria.

  Use esta funcionalidad para recuperarse de modificaciones accidentales de datos, errores de aplicación o escenarios de prueba.

Para la mayoría de las soluciones, no debe confiar exclusivamente en copias de seguridad. En su lugar, utilice las otras capacidades descritas en esta guía para apoyar los requisitos de resiliencia. Sin embargo, las copias de seguridad protegen contra algunos riesgos que otros enfoques no. Para más información, consulte ¿Qué son la redundancia, la replicación y la copia de seguridad?.

Para obtener más información, vea Backup and restore in Azure Database for MySQL.

Resistencia al mantenimiento del servicio

Azure Database for MySQL controla automáticamente las tareas de mantenimiento críticas, incluida la aplicación de revisiones del hardware subyacente, el sistema operativo y el motor de base de datos. El servicio incluye actualizaciones de seguridad, actualizaciones de software y actualizaciones de versiones secundarias como parte del mantenimiento planeado. Para obtener más información, consulte Scheduled maintenance in Azure Database for MySQL.

Para asegurarse de que el servidor permanece disponible durante las ventanas de mantenimiento, siga estas recomendaciones:

- ✓ **Evite las operaciones de administración durante los períodos de mantenimiento.** No realice operaciones de administración del servidor mientras el mantenimiento está en curso porque estas operaciones pueden afectar a la confiabilidad del servidor.

- **Utilice el mantenimiento con tiempo de inactividad casi nulo.** Si el servidor tiene habilitada la alta disponibilidad y cumple otros criterios de idoneidad, las operaciones de mantenimiento normalmente se completan en un plazo de 10 a 30 segundos. Si habilitas HA, las operaciones de mantenimiento suelen usar actualizaciones progresivas para minimizar el tiempo de inactividad. Las actividades de mantenimiento periódicas, como las actualizaciones de versiones secundarias, se producen primero en la réplica en espera. Para reducir el tiempo de inactividad, el servidor en espera se promueve a principal para que las cargas de trabajo puedan seguir ejecutándose mientras se aplican las tareas de mantenimiento en el nodo restante. Esta secuencia se aplica independientemente de si el servidor usa HA con redundancia de zona o con redundancia local. Para obtener más información, consulte Mantenimiento casi sin tiempo de inactividad.

- **Configurar ventanas de mantenimiento personalizadas.** Puede configurar la programación de mantenimiento para que se administre por el sistema o definir una ventana de mantenimiento personalizada para minimizar el impacto en las operaciones empresariales. También puede reprogramar las operaciones de mantenimiento planeado. Programe el mantenimiento durante períodos de baja actividad para minimizar el impacto empresarial. Para obtener más información, consulte Administrar la configuración de mantenimiento programado para Azure Database for MySQL.

- **Implemente la lógica de reintento.** Asegúrese de que las aplicaciones pueden controlar breves interrupciones de conectividad que pueden producirse durante los reinicios de mantenimiento. Para que las aplicaciones sean resistentes a estos tipos de problemas, consulte Resistencia a errores transitorios.

- **Habilite el mantenimiento de Virtual Canary en servidores de desarrollo y pruebas.** El mantenimiento de Virtual Canary proporciona acceso anticipado a las actualizaciones. Al habilitarlo en servidores de desarrollo y pruebas, puede comprobar que las próximas actualizaciones no afectan a la carga de trabajo antes de que lleguen a los servidores de producción. Para obtener más información, consulte Mantenimiento de Canario virtual.

Acuerdo de nivel de servicio

El acuerdo de nivel de servicio (SLA) para Azure servicios describe la disponibilidad esperada de cada servicio y las condiciones que la solución debe cumplir para lograr esa expectativa de disponibilidad. Para obtener más información, vea SLAs for servicios en línea.

Azure Database for MySQL proporciona acuerdos de nivel de servicio de disponibilidad diferentes en función de la configuración del servidor:

- Servidores configurados con alta disponibilidad con redundancia entre zonas.
- Servidores configurados con alta disponibilidad y redundancia local.
- Servidores configurados sin alta disponibilidad.

Contenido relacionado

- confiabilidad Azure
- mejores prácticas de arquitectura para Azure Database for MySQL
- Información general de la continuidad empresarial con Azure Database for MySQL

ⓘ**Nota:** El autor ha creado este artículo con ayuda de inteligencia artificial. Más información

---

Last updated on 18/06/2026

Confiabilidad en Azure Database for PostgreSQL

Azure Database for PostgreSQL es un servicio de base de datos totalmente administrado que proporciona control y flexibilidad pormenorizados sobre las funciones de administración de bases de datos y las opciones de configuración. El servicio proporciona funcionalidades de alta disponibilidad y recuperación ante desastres en función de sus requisitos.

Cuando se usa Azure, la confiabilidad es una responsabilidad compartida. Microsoft proporciona una variedad de funcionalidades para admitir resistencia y recuperación. Es responsable de comprender cómo funcionan esas funcionalidades dentro de todos los servicios que usa y de seleccionar las funcionalidades que necesita para cumplir los objetivos empresariales y los objetivos de tiempo de actividad.

En este artículo se describe cómo hacer que Azure Database for PostgreSQL sea resistente a diversas interrupciones y problemas potenciales, incluidos fallos transitorios, interrupciones de zonas de disponibilidad, interrupciones regionales y tareas de mantenimiento del servicio. También se describe cómo puede usar copias de seguridad para recuperarse de otros tipos de problemas y se resalta información clave sobre el acuerdo de nivel de servicio (SLA) de Azure Database for PostgreSQL.

Recomendaciones de implementación de producción

Para obtener información sobre cómo implementar Azure Database for PostgreSQL para admitir los requisitos de confiabilidad de la solución y cómo afecta la confiabilidad a otros aspectos de la arquitectura, consulte Procedimientos recomendados de arquitectura para Azure Database for PostgreSQL en Azure Well-Architected Framework.

Introducción a la arquitectura de confiabilidad

En esta sección se describen algunos de los aspectos importantes de cómo funciona el servicio que es más relevante desde una perspectiva de confiabilidad. En la sección se presenta la arquitectura lógica, que incluye algunos de los recursos y características que se implementan y usan. También se describe la arquitectura física, que proporciona detalles sobre cómo funciona el servicio en segundo plano.

Arquitectura lógica

Al trabajar con Azure Database for PostgreSQL, se implementa un *servidor*, que representa los recursos de proceso y almacenamiento necesarios para admitir las *bases de datos* que se implementan en el servidor.

Puede implementar servidores en varios *niveles de computación*: de ráfaga, de propósito general y optimizado para memoria. Cada nivel está optimizado para diferentes tipos de cargas de trabajo. En algunas regiones de Azure, puede implementar servidores con Azure Confidential Computing.

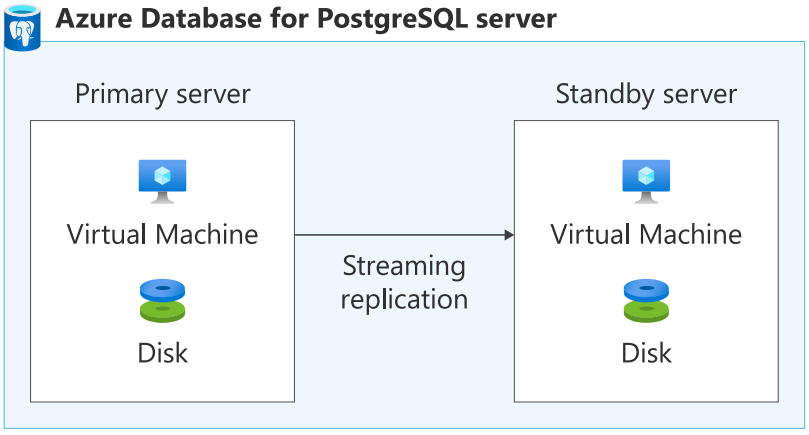
Para obtener más información sobre la arquitectura general del servicio y los modelos de implementación, consulte Azure Database for PostgreSQL información general.

Arquitectura física

- **Separación de proceso y almacenamiento:** Azure Database for PostgreSQL usa una arquitectura de separación de proceso y almacenamiento para admitir la alta disponibilidad. El motor de base de datos se ejecuta en una máquina virtual Linux, mientras que Azure Storage contiene los archivos de datos y mantiene tres copias sincrónicas con redundancia local de los archivos de base de datos para garantizar la durabilidad de los datos.

- **Alta disponibilidad:** Puede habilitar una configuración de alta disponibilidad en el servidor. Al habilitar la configuración de alta disponibilidad, el servicio aprovisiona y mantiene un servidor en espera activo. El servidor principal replica de forma sincrónica los cambios de datos en el servidor en espera para garantizar una pérdida de datos cero durante un error del servidor principal.

  La arquitectura separa la capa de proceso de la capa de almacenamiento, por lo que el servicio puede controlar los distintos tipos de errores de forma adecuada. Para lograr una mayor resistencia, puede distribuir los servidores entre zonas de disponibilidad.



Un servidor en espera se implementa en la misma configuración de máquina virtual que el servidor principal, incluidos los núcleos virtuales, el almacenamiento y la configuración de red.

Puede cambiar entre servidores mediante la realización de una *conmutación por error*. Existen dos tipos de conmutación por error: *las conmutaciones por error forzadas*, que se usan cuando se produce un error en el servidor principal y *las conmutaciones por error planeadas*, que se usan durante algunas operaciones de mantenimiento y en otros escenarios en los que es necesario minimizar el tiempo de inactividad de la aplicación durante una conmutación por error.

Al realizar operaciones como detener, iniciar y reiniciar, se producen en servidores de bases de datos principales y en espera al mismo tiempo. Los eventos planeados, como el escalado de proceso y el escalado de almacenamiento, se producen primero en espera y, a continuación, en el servidor principal. Actualmente, el servidor no conmuta por error para estas operaciones planeadas.

Para más información, consulte Alta disponibilidad en Azure Database for PostgreSQL.

- **Copias de seguridad:** Azure Database for PostgreSQL crea automáticamente copias de seguridad del servidor. Para obtener más información, consulte Copia de seguridad y restauración.

Resistencia a errores transitorios

Los errores transitorios son errores breves e intermitentes en los componentes. Se producen con frecuencia en un entorno distribuido como la nube y son una parte normal de las operaciones. Los errores transitorios se corrigen después de un breve período de tiempo. Es importante que las aplicaciones puedan controlar errores transitorios, normalmente mediante el reintento de solicitudes afectadas.

Todas las aplicaciones hospedadas en la nube deben seguir las instrucciones de control de errores transitorios de Azure cuando se comunican con cualquier API, bases de datos y otros componentes hospedados en la nube. Para obtener más información, consulte Recomendaciones para controlar errores transitorios.

Las aplicaciones deben controlar los errores de conectividad transitorios que pueden producirse durante el mantenimiento, las operaciones de escalado o las interrupciones de red. Siga estas recomendaciones:

- ✓ Cuando la aplicación encuentre errores transitorios, vuelva a intentar la operación mediante retroceso exponencial. Aumente el retraso entre reintentos y limite el número de intentos. Si la operación sigue produciendo un error después de los reintentos máximos, considérela un fracaso.
- ✓ Siempre que sea posible, use bibliotecas cliente (también denominadas controladores) que controlen automáticamente los reintentos.
- ✓ Los errores transitorios que se producen durante las operaciones de escritura requieren una consideración más cuidadosa. Considere la posibilidad de hacer que las operaciones de escritura sean idempotentes, por lo que se pueden ejecutar de forma segura varias veces.

Para más información, consulte Control de errores de conectividad transitorios en Azure Database for PostgreSQL.

Resistencia a errores de zona de disponibilidad

Las zonas de disponibilidad son grupos físicamente independientes de centros de datos dentro de una región de Azure. Cuando una zona falla, los servicios pueden transferirse a una de las zonas restantes.

Seleccione el tipo de compatibilidad con zonas de disponibilidad mediante la configuración de alta disponibilidad. Al habilitar la alta disponibilidad, el servicio implementa un servidor en espera junto con el servidor principal. Este modelo de alta disponibilidad ayuda a garantizar que los datos confirmados nunca se pierdan durante los errores. Sea cual sea el modelo de implementación de alta disponibilidad que utilice su servidor, confirma de forma síncrona los datos tanto en el servidor principal como en el servidor en espera. Si se produce una interrupción en el servidor principal, se conmuta automáticamente al servidor de respaldo.

Cada zona de disponibilidad almacena archivos de datos y registros de escritura anticipada (WAL) en discos administrados Premium con almacenamiento con redundancia local (LRS) que almacena automáticamente tres copias de datos dentro de cada zona.

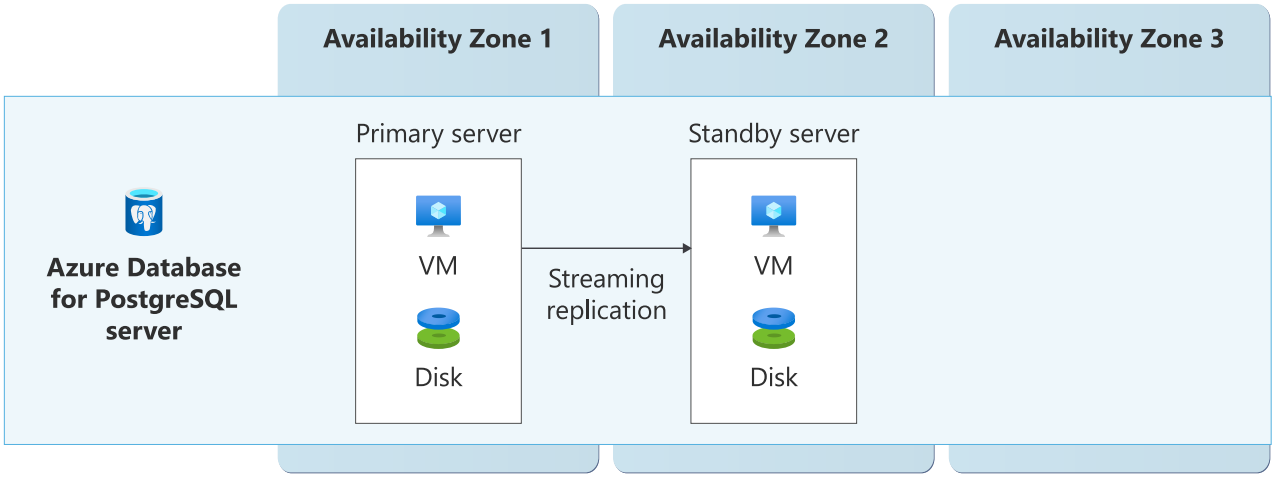
Azure Database for PostgreSQL admite dos tipos de configuración de zona de disponibilidad cuando se usa alta disponibilidad:

- **Alta disponibilidad con redundancia de zona:** La redundancia de zona proporciona el nivel más alto de resistencia de zona mediante la implementación de un servidor principal en una zona de disponibilidad y un servidor en espera en una zona de disponibilidad diferente. El servidor en espera usa la configuración de proceso, almacenamiento y red similar a la del

servidor principal. Una configuración con redundancia de zona proporciona aislamiento físico de toda la pila entre los servidores principal y en espera.

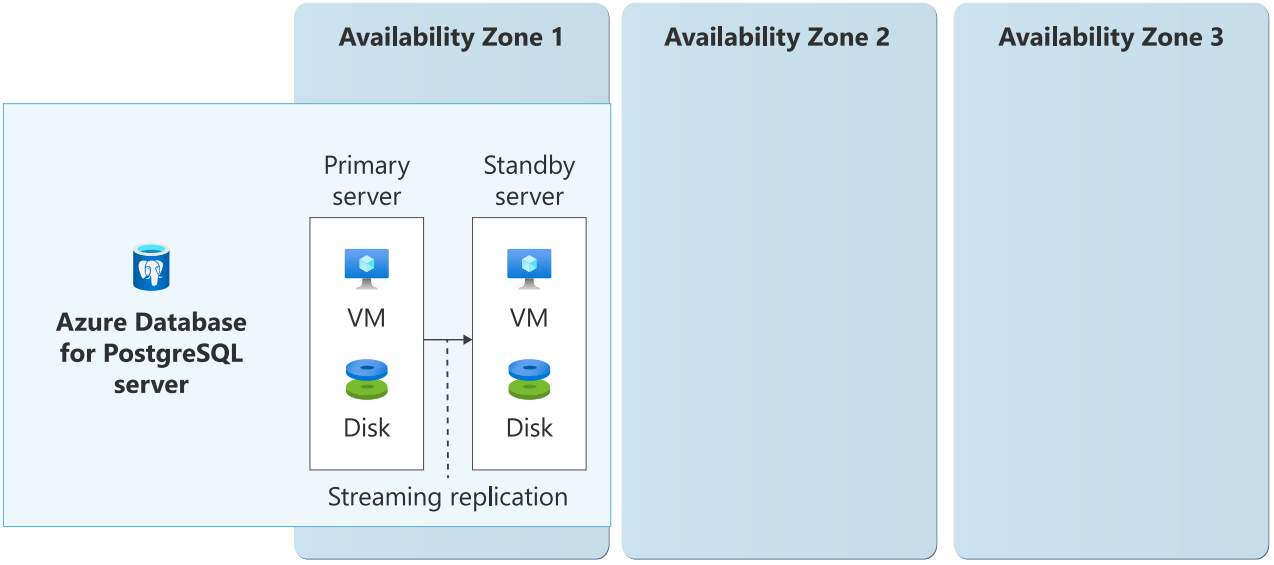
Puede seleccionar las zonas de disponibilidad para los servidores principales y en espera o permitir que Microsoft los elija.

Se recomiendan implementaciones con redundancia de zona para servidores de producción.



Las operaciones de escritura pueden experimentar un pequeño aumento de la latencia de confirmación porque el servicio replica de forma sincrónica los datos en el servidor en espera. El impacto varía según la carga de trabajo, la SKU seleccionada y la región.

- **Alta disponibilidad zonal (misma zona):** Los servidores principales y en espera usan la misma zona de disponibilidad. Si se produce una interrupción en el servidor principal, pero la zona sigue operativa, el servidor conmuta automáticamente al servidor en espera. Una implementación zonal proporciona alta disponibilidad dentro de una sola zona de disponibilidad. Protege contra errores de nivel de nodo y también ayuda a reducir el tiempo de inactividad de la aplicación durante los eventos de tiempo de inactividad planeados y no planeados. Sin embargo, no protege contra una interrupción en esa zona.



La alta disponibilidad zonal (misma zona) solo está disponible en las situaciones siguientes:

- La región no admite zonas de disponibilidad. La región funciona eficazmente como una sola zona, por lo que la única configuración de alta disponibilidad que puede seleccionar es la misma zona.
- Si una región no tiene capacidad suficiente para una implementación con redundancia de zona, el servicio puede colocar inicialmente ambos servidores en la misma zona de disponibilidad y, a continuación, migrarlos automáticamente a zonas independientes cuando la capacidad esté disponible. Esta opción está disponible cuando se usa Azure Portal o la CLI de Azure para implementar un servidor. Para obtener más información, consulte Configuración de opciones críticas para la empresa (alta disponibilidad).

Colocar los servidores en la misma zona puede reducir la latencia de escritura en las aplicaciones que implemente dentro de la misma zona.

Cuando los servidores están en la misma zona, se puede reducir la latencia de escritura en las aplicaciones que implemente dentro de la misma zona.

Si configura el servidor sin alta disponibilidad, se ejecuta en un solo servidor. Si ese servidor o su zona deja de funcionar, el servidor no está disponible. Para obtener más información, consulte Configuraciones sin zonas de disponibilidad.

Requisitos

- **Compatibilidad regional:** Azure Database for PostgreSQL admite configuraciones de zona de disponibilidad de manera diferente según las regiones de Azure. Para obtener una lista completa de las regiones, los tipos de compatibilidad con las zonas de disponibilidad y las consideraciones específicas de cada región, consulte Regiones de Azure.

- **Nivel de proceso:** En la tabla siguiente se muestra la compatibilidad con el nivel de proceso para cada tipo de compatibilidad con zonas de disponibilidad:

Expandir tabla

Nivel de computación	Con redundancia de zona	Zonal (misma zona)
Ampliable	No soportado	No soportado
General Purpose	Soportado	Soportado
Memoria optimizada	Soportado	Soportado

- **Nivel de servicio:** Ambos tipos de alta disponibilidad requieren niveles de uso general o Optimizado para memoria.

Consideraciones

Capacidad de región: Si una región no tiene capacidad suficiente para una implementación con redundancia de zona, el servicio puede colocar inicialmente ambos servidores en la misma zona de disponibilidad y migrarlos automáticamente a zonas independientes cuando la capacidad esté disponible. Esta opción está disponible cuando se usa Azure Portal o la CLI de Azure para implementar un servidor. Para obtener más información, consulte Configuración de opciones críticas para la empresa (alta disponibilidad).

Costo

Al habilitar la alta disponibilidad, se crea un servidor en espera y se factura a la misma velocidad que el servidor principal. La configuración de la zona de disponibilidad no afecta al costo. No hay cargos por la replicación de datos dentro o entre zonas de disponibilidad. Dependiendo del volumen de almacenamiento de copia de seguridad, también se le puede facturar por el almacenamiento de copia de seguridad. Para obtener información detallada sobre los precios, consulte Precios de Azure Database for PostgreSQL.

Configurar soporte de zonas de disponibilidad

Para configurar la compatibilidad con zonas de disponibilidad para un servidor, configure las opciones de alta disponibilidad.

- **Cree un servidor con redundancia de zona:** Para obtener información sobre cómo crear un servidor con alta disponibilidad y redundancia de zona habilitada, consulte Inicio rápido: Creación de un servidor de Azure Database for PostgreSQL.

- **Cambie la configuración de zona de disponibilidad para los servidores existentes:** Cambie la configuración de zona de disponibilidad para los servidores existentes cambiando la configuración de alta disponibilidad. Para ver los pasos detallados, consulte Habilitación de la alta disponibilidad para los servidores existentes.

  No se puede cambiar la zona usada para el servidor principal o en espera. Necesita volver a crear el servidor.

  > **Sugerencia**
  >
  > Se recomienda esperar hasta que la actividad del servidor sea baja antes de cambiar la configuración de alta disponibilidad.

- **Deshabilitar alta disponibilidad:** Al deshabilitar la alta disponibilidad, se quita el servidor en espera, por lo que el servidor no es resistente a interrupciones en su zona de disponibilidad. Para obtener más información, consulte Deshabilitación de la alta disponibilidad.

Comportamiento cuando todas las zonas están en buen estado

En esta sección se describe lo que cabe esperar al configurar servidores con alta disponibilidad y compatibilidad con zonas de disponibilidad, cuando todas las zonas de disponibilidad están operativas.

- **Operación entre zonas:** Las aplicaciones cliente de PostgreSQL se conectan al servidor principal mediante el nombre del servidor de base de datos. Azure Database for PostgreSQL usa una configuración activa-pasiva en la que el servidor principal de la zona de disponibilidad principal controla todas las conexiones y consultas de base de datos. El servidor en espera no atiende el tráfico de cliente durante las operaciones normales.

- **Replicación de datos entre zonas:** El servidor principal replica de forma sincrónica los cambios en el servidor en espera. Las transacciones no se consideran completas hasta que los servidores principal y en espera confirmen la escritura.

  Cuando una aplicación escribe y confirma los datos, PostgreSQL registra primero el cambio en el WAL en el servidor primario. El servidor principal transmite estos registros al servidor en espera mediante el protocolo de streaming de PostgreSQL. Después de que el servidor en espera almacene duraderamente el WAL, el servidor principal confirma la escritura. La aplicación confirma su transacción solo después de esta confirmación. Este proceso de confirmación no espera a que los registros se apliquen al servidor en espera.

Los efectos de la replicación son diferentes en función de la configuración de zona de disponibilidad que use el servidor:

- *Con redundancia de zona:* Dado que los servidores están en zonas independientes, este enfoque garantiza una pérdida de datos cero durante un error de zona. Esta situación también se denomina a veces lograr un objetivo de punto de recuperación (RPO) de cero para los errores de zona.

  Sin embargo, la replicación entre zonas podría introducir una pequeña cantidad de latencia adicional. El impacto de la latencia depende de la aplicación. Para la mayoría de las aplicaciones, la latencia adicional es insignificante.

- *Zonal*: dado que ambos servidores están en la misma zona, no se replica ningún tráfico entre zonas.

> ⓘ **Nota:**
>
> El sistema replica los datos de registro en tiempo real en el servidor en espera. Los errores de usuario en el servidor principal, como una eliminación accidental de una tabla o actualizaciones de datos incorrectas, se replican en el servidor en espera. No puede usar el modo de espera para recuperarse de estos tipos de errores y debe realizar una restauración a un momento dado desde la copia de seguridad. Para obtener más información, consulte **Copia de seguridad y restauración**.

Comportamiento durante un fallo de zona

En esta sección se describe qué puede esperar al configurar servidores con alta disponibilidad y compatibilidad con zonas de disponibilidad, si se produce una interrupción en una zona de disponibilidad.

- **Detección y respuesta:** Azure comprueba periódicamente el estado de los servidores principal y en espera. Después de varios pings, si la supervisión de estado detecta que no se puede acceder a un servidor principal, el servicio inicia una conmutación automática al servidor en espera. El algoritmo de supervisión de estado usa varios puntos de datos para evitar situaciones de falsos positivos.

  Si se produce un error en una zona de disponibilidad, el comportamiento es diferente en función de la configuración de la zona de disponibilidad que usa el servidor:

- *Con redundancia de zona:* Azure Database for PostgreSQL detecta automáticamente fallos en la zona de disponibilidad. Para ver los posibles tipos de estado de alta disponibilidad, consulte Supervisión del estado de alta disponibilidad (HA). Cuando se produce un error en una zona, Azure inicia una conmutación por error forzada al servidor en espera sin que sea necesario realizar ninguna acción.

- *Zonal:* Si la zona de disponibilidad que hospeda un servidor zonal deja de estar disponible, los servidores principal y en espera no están disponibles. En este escenario, el servicio no proporciona conmutación automática por error. Es responsable de detectar la interrupción de zona y realizar acciones de recuperación, como restaurar copias de seguridad con redundancia de zona en un servidor independiente en otra zona de disponibilidad o región.

- **Notificación:** La supervisión del estado de mantenimiento de alta disponibilidad en Azure Database for PostgreSQL proporciona información general continua sobre el estado y la preparación de instancias habilitadas para alta disponibilidad. La característica de supervisión se basa en Azure Resource Health y puede detectar y alertar sobre cualquier problema que pueda afectar a la preparación de la conmutación por error de la base de datos o a la disponibilidad general. Evalúe métricas clave como el estado de conexión, el estado de conmutación por error y el estado de replicación de datos para habilitar la solución de problemas proactiva y ayudar a mantener el tiempo de actividad y el rendimiento de la base de datos.

  Para obtener una guía detallada sobre la configuración y la interpretación de los estados de salud de alta disponibilidad, consulte Supervisión de los estados de salud de alta disponibilidad (HA).

- **Solicitudes activas:** Cuando una zona de disponibilidad deja de estar disponible, es posible que se finalicen las solicitudes en curso a los servidores de la zona afectada. Las aplicaciones deben reintentar estas solicitudes. Si los clientes controlan correctamente los errores transitorios mediante el reintento después de un breve período de tiempo, normalmente evitan un impacto significativo.

- **Pérdida de datos esperada:** La cantidad de pérdida de datos depende de la configuración de zona de disponibilidad que usa el servidor.

  - *Redundante por zona:* Se anticipa una pérdida de datos cero durante el fallo de zona debido a la replicación sincrónica entre los servidores principal y en espera en distintas zonas.

- *Zonal:* Los datos de los servidores de la zona afectada no están disponibles hasta que se recupere la zona.

- **Tiempo de inactividad esperado:** La cantidad de tiempo de inactividad depende de la configuración de zona de disponibilidad que usa el servidor.
  - *Con redundancia de zona:* La conmutación por error se completa normalmente en un plazo de 60 a 120 segundos. Si los clientes controlan correctamente los errores transitorios mediante el reintento después de un breve período de tiempo, normalmente evitan un impacto significativo.
  - *Zonal:* Los servidores de una zona afectada no están disponibles hasta que se recupere la zona de disponibilidad.

- **Redistribución:** El comportamiento de reenrutamiento del tráfico depende de la configuración de zona de disponibilidad que usa el servidor.
  - *Con redundancia de zona:* Después de la conmutación por error, el servidor en espera anterior se convierte en el nuevo servidor principal y comienza a aceptar nuevas conexiones. Azure establece automáticamente un nuevo servidor en espera en la zona principal original después de recuperarse. Para más información, consulte Conmutación por error forzada.
  - *Zonal:* Cuando una zona no está disponible, el servidor no está disponible. Si tiene un servidor independiente que creó con antelación en otra región o zona de disponibilidad, es responsable de redirigir el tráfico a ese servidor.

Recuperación de zona

El comportamiento de recuperación de zona depende de la configuración de zona de disponibilidad que usa el servidor.

- *Con redundancia de zona:* Cuando se recupera la zona de disponibilidad, Azure Database for PostgreSQL vuelve a generar automáticamente el servidor de reserva en la zona recuperada y lo sincroniza con el servidor principal actual. A continuación, la zona recuperada actúa como ubicación en espera. Para evitar interrupciones innecesarias, el servicio no traslada automáticamente de nuevo el rol principal a la zona original. Puede iniciar manualmente una conmutación por error planeada si desea devolver la principal a la zona original.

- *Zonal:* Una vez que la zona está en buen estado, los servidores de la zona están disponibles de nuevo. Es responsable de cualquier procedimiento de recuperación de zona y

sincronización de datos que requieran las cargas de trabajo.

Prueba de fallos de zona

Las opciones para probar errores de zona dependen de la configuración de zona de disponibilidad que usa la instancia.

- *Con redundancia de zona:* Puede probar la resistencia de la aplicación para la conmutación por error iniciando una *conmutación por error forzada*. Una conmutación por error forzada le permite simular un evento no planificado durante la ejecución de su carga de trabajo, permitiéndole observar el tiempo de inactividad de su aplicación. Se recomienda ejecutar simulaciones en un entorno que no sea de producción o en un momento silencioso. Para obtener más información, consulte Iniciar una conmutación por error forzada.

- *Zonal:* Aunque no se puede simular una interrupción de zona completa, puede simular que el servidor no está disponible de una manera similar a una interrupción de zona. Para obtener más información, vea Detener el proceso de un servidor.

Resistencia a errores en toda la región

Azure Database for PostgreSQL admite *réplicas de lectura entre regiones*, que puede usar para mantener una copia sincronizada de la base de datos en otra región para una recuperación más rápida.

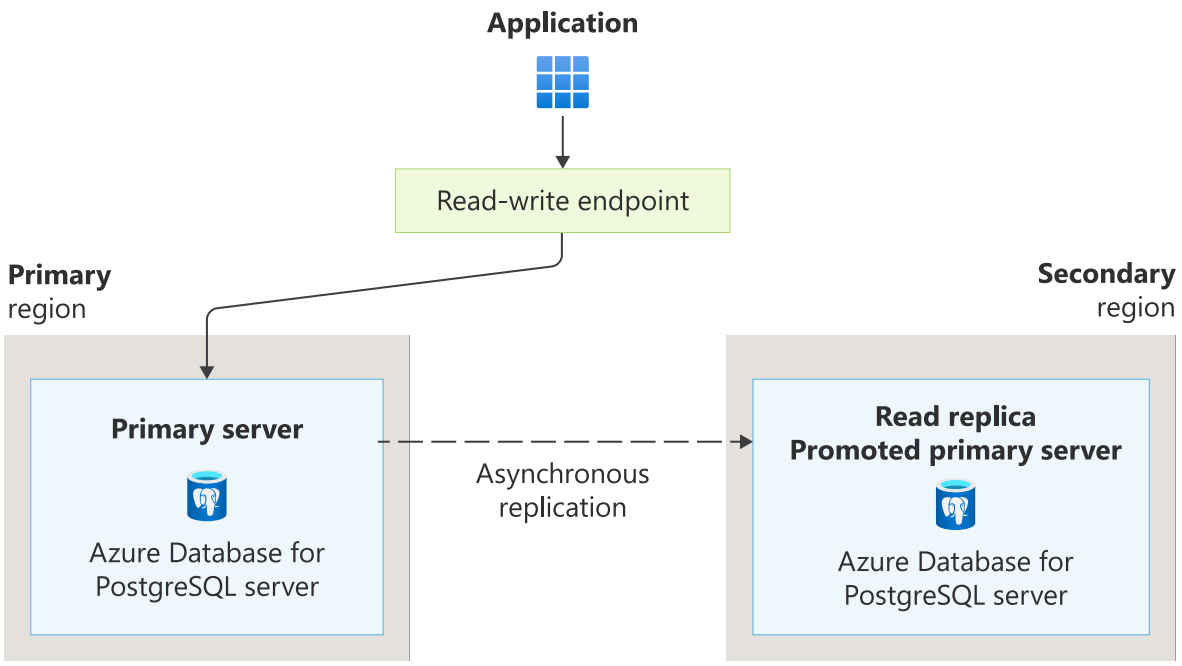
También puede usar copias de seguridad con redundancia geográfica, en regiones admitidas, para proporcionar recuperación entre regiones. Sin embargo, las copias de seguridad suelen implicar más tiempo de inactividad y pérdida de datos que la replicación. Para obtener más información, consulte Copia de seguridad y restauración.

Réplicas de lectura entre regiones

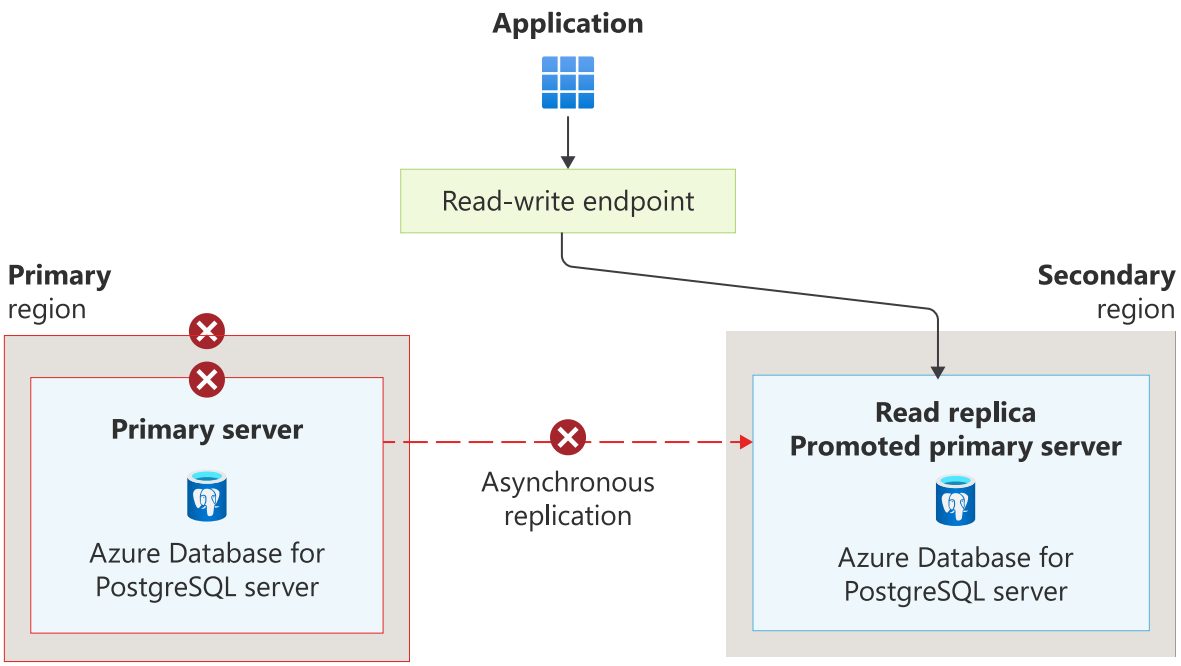
Puede implementar réplicas de lectura para proteger las bases de datos frente a errores de nivel de región. Cada réplica de lectura es un servidor independiente de Azure Database for PostgreSQL. Cuando coloca una réplica de lectura en una segunda región de Azure, el servidor de bases de datos puede proporcionar resiliencia ante un problema que afecte a toda la región. Puede implementar hasta cinco réplicas de lectura, las cuales opcionalmente pueden estar en distintas regiones de Azure. La tecnología de replicación física de PostgreSQL actualiza las réplicas de lectura de forma asincrónica y pueden retardar la principal. Las réplicas de lectura entre regiones pueden servir opcionalmente cargas de trabajo de solo lectura para reducir la latencia de las aplicaciones distribuidas globalmente o descargar el tráfico de lectura desde el

servidor principal. Para obtener más información sobre las características y consideraciones de réplica de lectura, consulte Réplicas de lectura.

Los puntos de conexión virtuales proporcionan puntos de conexión de lectura-escritura y de solo lectura, y redirigen automáticamente el tráfico cuando se promueve una réplica, lo que ayuda a minimizar el tiempo de inactividad durante las conmutaciones por error. Se recomienda encarecidamente usar puntos de conexión virtuales con réplicas de lectura entre regiones para mejorar la resistencia de la aplicación. Para más información, consulte Puntos de conexión virtuales para réplicas de lectura en Azure Database for PostgreSQL.



Si se produce un error en la región primaria, puede desencadenar una *promoción* para que la réplica secundaria se convierta en la principal. Es posible que distintos tipos de conmutación por error sean adecuados en función de cómo se usen las réplicas de lectura. Cuando se usan réplicas de lectura para proporcionar resistencia a errores de región, normalmente se usa el enfoque *de promoción al servidor principal* , que actualiza el punto de conexión virtual. Durante una interrupción de la región, debe realizar una *promoción forzada*, lo que puede dar lugar a una pérdida de datos para los datos no replicados. En escenarios planificados en los que la región primaria está en buen estado, puede optar por realizar una promoción planificada para evitar la pérdida de datos. Para más información, consulte Promoción de réplicas de lectura en Azure Database for PostgreSQL.



> **Nota:**
>
> En esta sección se resume información importante sobre cómo las réplicas de lectura pueden contribuir a la resiliencia frente a fallos que afectan a toda una región. También puede usar réplicas de lectura para mejorar el rendimiento y admitir bases de usuarios distribuidas geográficamente a gran escala. Para obtener más información, consulte Réplicas de lectura.

Requisitos

- **Compatibilidad con regiones:** Puede crear réplicas de lectura entre regiones en cualquier región que admita Azure Database for PostgreSQL. No está limitado a las regiones emparejadas de Azure.
- **Niveles de cómputo:** Los niveles de cómputo de uso general y optimizados para memoria admiten réplicas de lectura. El nivel Ampliable no admite réplicas de lectura.

Consideraciones

- **Diferencias de configuración:** Es posible que las réplicas de lectura no hereden todas las opciones de configuración del servidor principal. Planee configurar las opciones necesarias después de la conmutación por error. El servidor principal y las réplicas deben ser *simétricos*, lo que significa que deben tener los mismos niveles, tamaños de almacenamiento y valores para algunas configuraciones. Durante fallos en la región, se puede renunciar al requisito de tener un servidor simétrico para promociones forzadas,

pero se recomienda mantener una configuración simétrica siempre que sea posible para evitar problemas inesperados. Para obtener más información, consulte Administración de configuración.

- **Supervisión del retraso de replicación:** El proceso de replicación asíncrona conlleva un retraso de replicación, que puede variar en función de muchos factores. Cuando el retraso de replicación es alto, el servidor podría experimentar problemas. Es importante supervisar el retraso de replicación para que pueda mitigar los problemas antes de que se escalen. Para más información, consulte Supervisión de la replicación.

- **Alta disponibilidad:** Las réplicas de lectura no pueden tener alta disponibilidad habilitada y, cuando se promueven, tampoco tienen alta disponibilidad. Eres responsable de configurar la alta disponibilidad después de promover una réplica.

Para ver otros factores sobre el proceso de promoción que se debe tener en cuenta, consulte Consideraciones.

Costo

Las réplicas de lectura generan costes de computación y almacenamiento, además de cargos por transferencia de datos entre regiones para la replicación. Para obtener información detallada sobre los precios, consulte Precios de Azure Database for PostgreSQL y Precios de ancho de banda.

Configuración de la compatibilidad con varias regiones

- **Cree una réplica de lectura:** Para obtener información sobre cómo crear una réplica de lectura, consulte Creación de una réplica de lectura. Puede configurar réplicas después de crear el servidor principal, siempre que el servidor principal se ejecute y sea accesible.

  Para crear un punto de conexión virtual, consulte Creación de puntos de conexión virtuales.

- **Eliminar una réplica de lectura:** Para obtener información sobre cómo eliminar una réplica de lectura, consulte Eliminación de una réplica de lectura.

Comportamiento cuando todas las regiones están en buen estado

En esta sección se describe qué esperar cuando el servidor está configurado con una réplica de lectura en otra región y un punto de conexión virtual, y todas las regiones están operativas:

- **Enrutamiento del tráfico entre regiones:** En las operaciones normales, el punto de conexión virtual dirige el tráfico del punto de conexión de lectura y escritura al servidor principal de la región primaria. Si también utiliza el punto de conexión virtual de solo lectura, dirige el tráfico a la réplica que haya configurado.

- **Replicación de datos entre regiones:** Las réplicas de lectura entre regiones usan la replicación asincrónica para minimizar el impacto en el rendimiento del servidor principal. La cantidad de retraso de replicación depende de muchos factores, incluida la carga de escritura y la latencia entre el servidor principal y las réplicas. El retraso de replicación suele ser al menos varios minutos, pero puede ser más largo. Para más información, consulte Supervisión de la replicación.

Comportamiento durante una falla de región

En esta sección se describe qué esperar cuando el servidor está configurado con una réplica de lectura en otra región y un punto de conexión virtual, y hay una interrupción en la región primaria:

- **Detección y respuesta:** Usted es responsable de detectar una interrupción en la región primaria y promover manualmente una réplica de lectura como el nuevo servidor principal. Durante una interrupción regional, debe realizar una promoción forzada, lo que resulta en la pérdida de cualquier dato no replicado.

  > **Importante**
  >
  > Usted es responsable de activar la promoción. Azure no promueve las réplicas de lectura automáticamente, incluso si se produce una falla en una región.

  Para obtener pasos detallados para comenzar una promoción, consulte Convertir réplica de lectura en principal.

- **Notificación:** Microsoft no le notifica automáticamente cuando una región está inactiva. Sin embargo, puede usar Azure Service Health para comprender el estado general del servicio, incluidos los errores de región, y puede configurar alertas de Service Health para notificarle problemas.

- **Peticiones activas:** El proceso de promoción interrumpe todas las conexiones activas con la región principal. Una vez completado el proceso de promoción, las aplicaciones deben reintentar la realización de conexiones a la réplica promocionada.

- **Pérdida de datos esperada:** Durante una interrupción de la región, debe realizar una promoción obligatoria, lo que da lugar a la pérdida permanente de los datos no replicados.

  La cantidad de pérdida de datos depende del retraso de replicación en el momento de la interrupción. El retraso de replicación suele ser al menos varios minutos, pero puede ser más largo. Para más información, consulte Supervisión de la replicación.

- **Tiempo de inactividad esperado:** La promoción forzada se completa normalmente en un plazo de 1 a 3 minutos después de desencadenarse. Es posible que las aplicaciones también necesiten volver a conectarse al punto de conexión correcto. Los puntos de conexión virtuales se actualizan como parte del proceso de promoción forzada. Las aplicaciones deben respetar el período de vida (TTL) de los registros DNS del punto de conexión para asegurarse de que se vuelven a conectar rápidamente a la réplica correcta una vez completada la promoción.

- **Reenrutamiento del tráfico:** El punto de conexión virtual del servidor redirige automáticamente el tráfico de la aplicación a la nueva réplica principal.

  > ⓘ **Nota:**
  >
  > Después de promover una réplica de lectura para que sea el servidor principal, no tiene habilitada la configuración de alta disponibilidad. Debe habilitar la configuración de alta disponibilidad manualmente o agregarla a sus propios procesos de automatización.

Recuperación de regiones

Cuando se usan puntos de conexión virtuales, una vez recuperada la región primaria, el servidor principal antiguo se configura automáticamente como una réplica de lectura. Puede realizar otra promoción para devolver las operaciones principales a la región primaria preferida.

Prueba de fallos de región

Pruebe periódicamente los procedimientos de promoción de réplica de lectura para asegurarse de que los procesos son válidos y que las funcionalidades cumplen los requisitos del objetivo de tiempo de recuperación (RTO) y del objetivo de punto de recuperación (RPO).

Puede promover una réplica de lectura para convertirse en el servidor principal en cualquier momento, incluso cuando todas las regiones estén en buen estado. Para las pruebas:

- Puede realizar pruebas de promoción forzadas. Se recomienda realizar estas pruebas en un entorno que no sea de producción, ya que puede provocar la pérdida de datos. Las pruebas de promoción forzadas ayudan a simular el comportamiento que se observa durante una interrupción de una región.
- Para el mantenimiento planeado o los escenarios de prueba en los que desea evitar la pérdida de datos, use una promoción planeada en su lugar. Sin embargo, la promoción planeada sigue un proceso diferente al de la promoción durante una interrupción de la región, por lo que es posible que no refleje el comportamiento durante una interrupción de región verdadera.

Para obtener instrucciones paso a paso, consulte Conmutar réplica de lectura a principal.

Como parte de la estrategia de recuperación ante desastres, ejecute periódicamente los simulacros de recuperación completos. Estos simulacros deben incluir la validación de datos, las pruebas de funcionalidad de la aplicación y los procedimientos de reversión documentados.

Copias de seguridad y restauración

Azure Database for PostgreSQL realiza automáticamente una copia de seguridad de los datos. Estas copias de seguridad proporcionan funcionalidades de recuperación a un momento dado y le ayudan a protegerse frente a daños accidentales y la eliminación de datos. Microsoft administra completamente las copias de seguridad. No interrumpen la disponibilidad del servidor e incluyen copias de seguridad completas y copias de seguridad del registro de transacciones.

- **Almacenamiento de copia de seguridad:** Si implementa el servidor en una región con zonas de disponibilidad, el servicio almacena copias de seguridad en el almacenamiento con redundancia de zona (ZRS), independientemente de la configuración de alta disponibilidad del servidor. En el caso de los servidores implementados en regiones sin zonas de disponibilidad, el servicio almacena copias de seguridad en el almacenamiento con redundancia local (LRS).

  En las regiones de Azure con pares, puede configurar el almacenamiento de copia de seguridad con redundancia geográfica al crear el servidor para replicar las copias de seguridad en la región emparejada de Azure y obtener protección adicional frente a fallos regionales. El servicio replica las copias de seguridad de forma asincrónica.

  El período de retención de copia de seguridad predeterminado es de siete días, pero puede ampliar la retención hasta 35 días. También puede usar Azure Backup para el

almacenamiento a largo plazo de copias de seguridad manuales durante hasta 10 años. Todas las copias de seguridad están cifradas.

- **Restaurar:** La recuperación a un momento dado le permite restaurar la base de datos a cualquier momento dentro del período de retención de copia de seguridad. El proceso de restauración crea un nuevo servidor de bases de datos con un nuevo nombre de servidor proporcionado por el usuario. Puede usar el nuevo servidor as-is o copiar datos de él.

  Al restaurar una copia de seguridad con redundancia geográfica, se crea un nuevo servidor en la región emparejada.

  Esta funcionalidad es útil para recuperarse de modificaciones accidentales de datos, errores de aplicación o escenarios de prueba.

Para la mayoría de las soluciones, no debe confiar exclusivamente en copias de seguridad. En su lugar, utilice las otras capacidades descritas en esta guía para apoyar los requisitos de resiliencia. Sin embargo, las copias de seguridad protegen contra algunos riesgos que otros enfoques no. Para más información, consulte ¿Qué son la redundancia, la replicación y la copia de seguridad?.

Para más información, consulte Copia de seguridad y restauración en Azure Database for PostgreSQL.

Resistencia al mantenimiento del servicio

Azure Database for PostgreSQL controla automáticamente las tareas de mantenimiento críticas, incluida la aplicación de revisiones al hardware subyacente, el sistema operativo y el motor de base de datos. El servicio incluye actualizaciones de seguridad, actualizaciones de software y actualizaciones de versiones secundarias como parte del mantenimiento planeado.

Para asegurarse de que el servidor permanece disponible durante las ventanas de mantenimiento, siga estas recomendaciones:

- ✓ **Habilitar alta disponibilidad:** Durante el mantenimiento, es posible que el servidor tenga que reiniciarse como parte del proceso de actualización. Si habilita alta disponibilidad, las operaciones de mantenimiento suelen usar actualizaciones graduales para minimizar el tiempo de inactividad. Las actividades de mantenimiento periódicas, como las actualizaciones de versiones secundarias, se producen primero en la réplica en espera. Para reducir el tiempo de inactividad, el modo de espera se promueve a principal para que las cargas de trabajo puedan continuar en el nodo promocionado mientras se aplican las tareas

de mantenimiento al otro nodo. Esta secuenciación se aplica tanto si el servidor utiliza alta disponibilidad con redundancia de zona como si utiliza alta disponibilidad zonal.

En el caso de los servidores sin alta disponibilidad habilitada, espere un breve tiempo de inactividad durante las operaciones de mantenimiento. Con alta disponibilidad habilitada, las operaciones de mantenimiento normalmente se completan con un tiempo de inactividad mínimo o sin tiempo de inactividad.

- ✓ **Configurar ventanas de mantenimiento personalizadas:** Puede configurar la programación de mantenimiento para administrar el sistema o definir una ventana de mantenimiento personalizada para minimizar el impacto en las operaciones empresariales. Programe el mantenimiento durante períodos de baja actividad para minimizar el impacto empresarial. Para obtener más información, consulte Programación del mantenimiento.
- ✓ **Implemente la lógica de reintento:** Asegúrese de que las aplicaciones pueden controlar breves interrupciones de conectividad que pueden producirse durante los reinicios de mantenimiento. Para que las aplicaciones sean resistentes a estos tipos de problemas, consulte Guía sobre resistencia a errores transitorios .

Acuerdo de nivel de servicio

El acuerdo de nivel de servicio (SLA) para Azure servicios describe la disponibilidad esperada de cada servicio y las condiciones que la solución debe cumplir para lograr esa expectativa de disponibilidad. Para obtener más información, consulte Acuerdos de Nivel de Servicio para servicios en línea .

Azure Database for PostgreSQL proporciona diferentes acuerdos de nivel de servicio de disponibilidad, en función de la configuración del servidor:

- Los servidores configurados con redundancia de zona de alta disponibilidad ofrecen un acuerdo de nivel de servicio (SLA) de tiempo de actividad del 99,99%.
- Los servidores configurados con alta disponibilidad zonal ofrecen un SLA con un tiempo de actividad del 99,95%.
- Los servidores configurados sin alta disponibilidad ofrecen un SLA del 99,9%.

Contenido relacionado

- confiabilidad Azure
- Procedimientos recomendados de arquitectura para Azure Database for PostgreSQL
- Introducción a la continuidad empresarial con Azure Database for PostgreSQL

- Recuperación ante desastres geográfica en Azure Database for PostgreSQL
- Scripts de Terraform para implementar principios de resiliencia y recuperación

ⓘ**Nota:** El autor ha creado este artículo con ayuda de inteligencia artificial. Más información

---

Last updated on 19/06/2026

Confiabilidad en Azure Elastic SAN

Azure Elastic SAN es un servicio de red de área de almacenamiento nativo en la nube (SAN) que proporciona una solución de almacenamiento escalable, rentable, de alto rendimiento y completa para una variedad de opciones de proceso. Elastic SAN permite crear y administrar volúmenes, que son discos virtuales que se pueden conectar a las máquinas virtuales, contenedores u otros servicios de Azure a través del protocolo iSCSI.

Cuando se usa Azure, la confiabilidad es una responsabilidad compartida. Microsoft proporciona una variedad de capacidades para apoyar la resiliencia y la recuperación. Es responsable de comprender cómo funcionan esas funcionalidades dentro de todos los servicios que usa y de seleccionar las funcionalidades que necesita para cumplir los objetivos empresariales y los objetivos de tiempo de actividad.

En este artículo se describe cómo hacer que Azure Elastic SAN sea resiliente a una variedad de posibles interrupciones y problemas, incluidos fallos transitorios, fallos de zona de disponibilidad y fallos a nivel regional. También describe las opciones de copia de seguridad y recuperación, y resalta la información clave sobre el acuerdo de nivel de servicio (SLA) de Azure Elastic SAN.

Recomendaciones de implementación de producción para la confiabilidad

En el caso de las cargas de trabajo de producción, se recomienda que usted:

- ✓ **Uso del almacenamiento con redundancia de zona (ZRS):** ZRS distribuye copias de los datos en tres zonas de disponibilidad.
- ✓ **Configure el acceso a la red para reducir la intervención manual durante los errores de zona:** Los puntos de conexión privados permiten la conmutación automática por error de zona sin intervención manual, mientras que los puntos de conexión de servicio pueden requerir intervención manual.
- ✓ **Para cargas de trabajo que requieren recuperación ante desastres entre regiones:** Cree instantáneas de los volúmenes con regularidad y expórtelas a instantáneas de disco administrado para la protección de datos. Copie las instantáneas en una región secundaria que esté geográficamente distante de la región primaria.

Introducción a la arquitectura de confiabilidad

Elastic SAN tiene una jerarquía de recursos de tres niveles:

- **ELASTIC SAN**: el recurso de nivel superior donde se configura la redundancia, se asigna la capacidad de almacenamiento y se establecen los límites de rendimiento. El número de unidades base asignadas a elastic SAN determina el total de IOPS y el rendimiento disponibles en toda la SAN.
- **Grupos de volúmenes**: construcciones de administración usadas para administrar volúmenes a escala. La configuración de acceso a la red, como puntos de conexión privados o puntos de conexión de servicio, se configura en el nivel de grupo de volúmenes y se hereda por todos los volúmenes del grupo.
- **Volúmenes**: Volúmenes de almacenamiento individuales divididos de la capacidad total del SAN. Los volúmenes están conectados a recursos de proceso a través del protocolo iSCSI.

Para más información sobre los recursos que implemente, consulte Recursos elásticos de SAN.

Internamente, Elastic SAN almacena los datos dentro de *los clústeres de almacenamiento*. Al configurar la SAN elástica para usar el almacenamiento con redundancia local (LRS), los datos se replican tres veces dentro de un único clúster de almacenamiento en un centro de datos. Puede configurar el almacenamiento con redundancia de zona (ZRS) para almacenar copias de los datos en tres zonas de disponibilidad. Para obtener más información, consulte Resilience to availability zone failures (Resistencia a errores de zona de disponibilidad).

Resistencia a errores transitorios

Los errores transitorios son errores breves e intermitentes en los componentes. Se producen con frecuencia en un entorno distribuido como la nube y son una parte normal de las operaciones. Los errores transitorios se corrigen después de un breve período de tiempo. Es importante que las aplicaciones puedan controlar errores transitorios, normalmente mediante el reintento de solicitudes afectadas.

Todas las aplicaciones hospedadas en la nube deben seguir las instrucciones de control de errores transitorios de Azure cuando se comunican con cualquier API, bases de datos y otros componentes hospedados en la nube. Para obtener más información, consulte Recomendaciones para controlar errores transitorios.

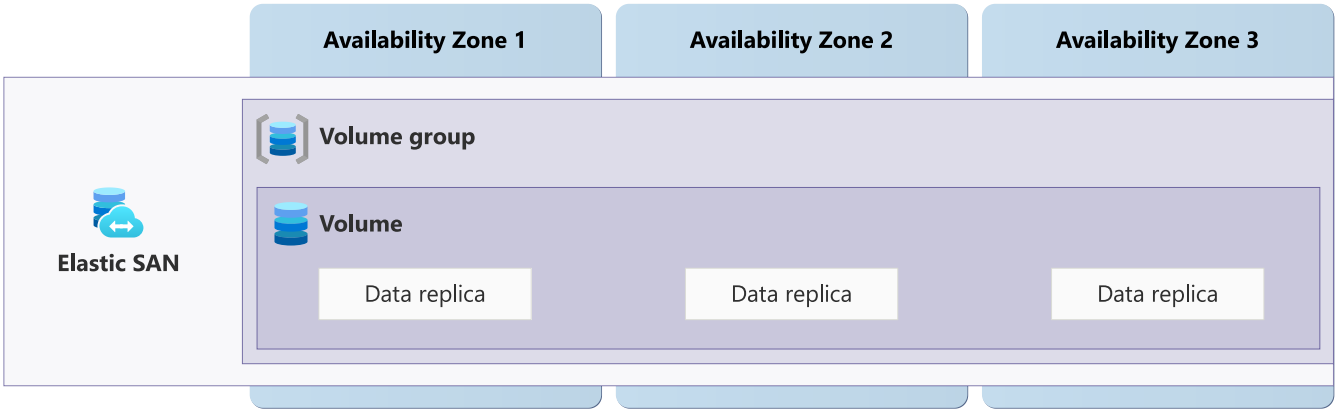
Si se interrumpe la conexión iSCSI a un volumen de Elastic SAN, el iniciador iSCSI en el cliente intenta volver a conectarse automáticamente. Es posible que experimente una breve pausa en las operaciones de E/S durante la reconexión. Configure el iniciador iSCSI con los valores de reintento y tiempo de espera adecuados para controlar las interrupciones transitorias. Para

obtener recomendaciones detalladas sobre la configuración de iSCSI y MPIO, consulte Optimización del rendimiento de elastic SAN.

Resistencia a errores de zona de disponibilidad

Las zonas de disponibilidad son grupos físicamente independientes de centros de datos dentro de una región de Azure. Cuando una zona falla, los servicios pueden transferirse a una de las zonas restantes.

Azure Elastic SAN se pueden configurar para usar el almacenamiento con redundancia de zona (ZRS), lo que significa que los datos se replican de forma sincrónica en tres zonas de disponibilidad de la región. La redundancia de zona le ayuda a lograr resistencia y confiabilidad para las cargas de trabajo de producción.



El enfoque de conectividad afecta a la capacidad de la carga de trabajo para conmutar de manera transparente durante una falla de zona. Se recomienda usar puntos de conexión privados para conectarse a los volúmenes. Los puntos de conexión privados admiten la conmutación automática por error. Si usa puntos de conexión de servicio, es posible que la conmutación por error requiera una intervención manual. Para obtener más información sobre los enfoques de conectividad, consulte Más información sobre las configuraciones de red para Elastic SAN.

El uso del almacenamiento con redundancia local (LRS) deja la SAN elástica desprotegida frente a errores de zona de disponibilidad. LRS hace que el SAN *elástico no es zonal* almacenando datos en una sola zona de disponibilidad en lugar de distribuirlos entre zonas con ZRS.

Requisitos

Compatibilidad con regiones: Los recursos de SAN elásticos con redundancia de zona se pueden implementar en un subconjunto de regiones. Para obtener una lista de las regiones, consulte Objetivos de escalado de Elastic SAN.

Cost

Al crear una SAN elástica con ZRS, el costo es mayor que LRS. Para obtener más información sobre los precios, consulte precios Azure Elastic SAN.

Configurar soporte de zonas de disponibilidad

- **Cree una nueva SAN elástica con ZRS:** Al crear una SAN elástica y seleccionar ZRS como opción de redundancia, la SAN elástica es automáticamente redundante por zona. No se puede cambiar la opción de redundancia después de crear la SAN elástica. Para obtener más información sobre cómo crear un nuevo recurso de Elastic SAN, consulte Implementación de una SAN elástica.

- **Habilite la redundancia de zona en una SAN elástica de LRS existente:** No se puede convertir una SAN elástica de LRS a ZRS directamente. Para migrar, realice una instantánea de los volúmenes de Elastic SAN, expórtelos a instantáneas de disco administrado, implemente una nueva SAN elástica en ZRS y, a continuación, cree volúmenes en el nuevo SAN elástico mediante esas instantáneas de disco. Para obtener más información, consulte Snapshot Azure Elastic SAN volúmenes.

Comportamiento cuando todas las zonas están en buen estado

En esta sección se describe qué esperar al configurar una SAN elástica para la redundancia de zona y todas las zonas están operativas.

- **Operación entre zonas:** Cuando se conecta a un volumen de Elastic SAN, la conexión iSCSI se enruta a un clúster en una de las zonas de disponibilidad. La plataforma enruta automáticamente el tráfico entre zonas.

- **Replicación de datos entre zonas:** Cuando un cliente escribe datos en un volumen de Elastic SAN, esos datos se escriben sincrónicamente en clústeres dentro de tres zonas de disponibilidad antes de que se confirme la operación de escritura. La replicación sincrónica garantiza un alto nivel de coherencia de datos y garantiza que no se pierdan datos durante un error de zona.

  La implementación de una Elastic SAN de ZRS proporciona más confiabilidad que una Elastic SAN de LRS, pero agrega más latencia de escritura. Haga pruebas comparativas de su Elastic SAN y simule la carga de trabajo de la aplicación para comparar la latencia entre LRS y ZRS, así podrá ver si afecta su carga de trabajo.

Comportamiento durante un fallo de zona

En esta sección se describe qué se puede anticipar al configurar una SAN elástica para la redundancia de zona, cuando ocurre una interrupción en una de las zonas.

- **Detección y respuesta:** La plataforma elastic SAN detecta los errores en una zona de disponibilidad. No es necesario hacer nada para iniciar una conmutación por error de zona para las SAN Elastic de ZRS.

- **Notificación:** Microsoft no le notifica automáticamente cuando una zona está inactiva. Sin embargo, puede usar Azure Service Health para comprender el estado general del servicio, incluidos los errores de zona, y puede configurar alertas de Service Health para notificarle problemas.

- **Solicitudes activas:** Cuando una zona de disponibilidad deja de estar disponible, el servicio puede finalizar las operaciones de E/S en curso que se conectan a las réplicas de la zona afectada y debe reintentar esas operaciones. Si usa puntos de conexión privados, el servicio conmuta automáticamente. Si usa puntos de conexión de servicio, es posible que deba reiniciar el iniciador iSCSI para cambiar a una zona sana.

- **Pérdida de datos esperada:** Elastic SAN evita la pérdida de datos durante un error de zona mediante la replicación sincrónica de datos entre tres zonas de disponibilidad.

- **Tiempo de inactividad esperado:** Cuando se usan puntos de conexión privados, la conmutación por error de zona se produce automáticamente. Es posible que experimente una degradación del rendimiento y la disponibilidad durante unos minutos después de una conmutación por error mientras la SAN se reequilibra por sí misma.

  Cuando usas puntos de conexión de servicio, Elastic SAN no cambia automáticamente a una zona saludable. Es posible que tenga que reiniciar el iniciador iSCSI para iniciar una conmutación por error a una zona correcta diferente.

- **Reenrutamiento del tráfico:** Cuando una zona no está disponible, la plataforma Elastic SAN detecta la pérdida de la zona y enruta el tráfico a las zonas correctas restantes.

  Para garantizar la conmutación automática por error de zona sin intervención manual, use puntos de conexión privados para conectarse a la SAN elástica. Al usar puntos de conexión de servicio en lugar de puntos de conexión privados, es posible que tenga que reiniciar el iniciador iSCSI manualmente para iniciar una conmutación por error a una zona en buen estado.

Recuperación de zona

Cuando se recupera la zona de disponibilidad, la plataforma Elastic SAN restaura automáticamente las operaciones normales y reanuda la replicación en tres zonas. No tiene que realizar ninguna acción.

Prueba de fallos de zona

La plataforma Azure Elastic SAN administra el enrutamiento del tráfico, la conmutación por error y la recuperación de zona para los recursos con redundancia de zona. Dado que esta característica está totalmente administrada, no es necesario validar los procesos de error de zona de disponibilidad.

Resistencia a errores en toda la región

Azure Elastic SAN es un servicio de una sola región. Si la región deja de estar disponible, el recurso de Elastic SAN tampoco está disponible. No hay replicación entre regiones ni conmutación por error integrada de una región a otra para Elastic SAN. Es responsable de diseñar su propia solución de recuperación ante desastres de varias regiones si la carga de trabajo requiere resistencia de nivel de región.

Soluciones personalizadas de varias regiones para la resistencia

Es responsable de implementar la recuperación ante desastres de varias regiones para los datos de Elastic SAN. El enfoque recomendado es usar instantáneas de volumen:

1. **Cree instantáneas periódicamente.** Utiliza instantáneas de volumen para capturar copias puntuales de tus volúmenes de Elastic SAN.

   El objetivo de punto de recuperación (RPO) depende de la frecuencia con la que cree y copie instantáneas en la región secundaria. Cuanto más frecuentemente cree instantáneas y cópielas, menor será la pérdida de datos potencial durante un desastre.

2. **Exportar instantáneas a instantáneas de disco administrado.**Exporte las instantáneas de volumen a instantáneas de disco administrado, que se pueden copiar en otras regiones.

3. **Copie instantáneas en una región secundaria.**Copie la instantánea incremental en una nueva región que esté geográficamente distante de la región primaria. Esto reduce el riesgo de que varias regiones se vean afectadas por un único desastre.

4. **Restaurar desde capturas.** En un escenario de recuperación ante desastres, cree nuevos volúmenes en la SAN elástica secundaria a partir de las instantáneas de disco administrado copiadas.

El objetivo de tiempo de recuperación (RTO) depende del tamaño de los datos, el tiempo necesario para copiar instantáneas entre regiones y el tiempo necesario para implementar y configurar una nueva SAN elástica en la región secundaria. Para reducir el tiempo de recuperación, considere la posibilidad de implementar una SAN elástica secundaria en la región de recuperación antes de que se produzca un desastre. Esto también ayuda a evitar restricciones de capacidad durante una interrupción.

Copias de seguridad y restauración

Para la mayoría de las soluciones, no debe confiar exclusivamente en copias de seguridad. En su lugar, utilice las otras capacidades descritas en esta guía para apoyar los requisitos de resiliencia. Sin embargo, las copias de seguridad protegen contra algunos riesgos que otros enfoques no. Para más información, consulte ¿Qué son la redundancia, la replicación y la copia de seguridad?.

Elastic SAN admite dos tipos de copia de seguridad:

- **Azure Elastic SAN backup:** Esta funcionalidad de Azure Backup ofrece una solución totalmente administrada para programar copias de seguridad, establecer escalas de tiempo de expiración para puntos de recuperación y datos de recuperación en un nuevo volumen. Ayuda a protegerse contra la pérdida de datos frente a eliminaciones accidentales, ransomware y actualizaciones de aplicaciones. Para obtener más información, consulte Azure Elastic SAN backup (versión preliminar).

  > **Importante**
  >
  > Como esta solución está cubierta por los términos complementarios de Microsoft para las versiones preliminares de Azure, úsela para pruebas y no para uso de producción.

- **Instantáneas:** Como alternativa, puede crear y administrar instantáneas en función de los requisitos de protección de datos.

  Azure Elastic SAN admite instantáneas de volumen para la protección de datos. Las instantáneas son copias incrementales en un momento específico de sus volúmenes que consumen espacio de la capacidad total de su Elastic SAN. Para proteger los datos, cree instantáneas periódicamente. La frecuencia depende de la cantidad de datos que puede

permitirse perder (su RPO). Puede crear instantáneas manualmente o crear su propia automatización para crearlas según una programación.

Las instantáneas se almacenan en la misma SAN elástica que tus volúmenes y usan la misma configuración de redundancia. Para protegerse frente a errores en toda la región, exporte las instantáneas a instantáneas de disco administrado y cópielas en otra región. Para obtener más información, consulte Exportación de instantáneas de volumen y Copiar una instantánea incremental a una nueva región.

Puede crear un nuevo volumen SAN elástico a partir de una instantánea o de una instantánea de un disco administrado. Para obtener más información, consulte Creación de un volumen a partir de una instantánea.

Resistencia al mantenimiento del servicio

Microsoft aplica periódicamente actualizaciones de servicio y realiza otro mantenimiento. La plataforma Azure controla estas actividades automáticamente, lo que garantiza que el mantenimiento sea transparente y sin problemas. No se espera ningún tiempo de inactividad durante los eventos de mantenimiento a menos que se le haya informado a través del mantenimiento planeado de Azure Service Health.

Acuerdo de nivel de servicio

El acuerdo de nivel de servicio (SLA) para Azure servicios describe la disponibilidad esperada de cada servicio y las condiciones que la solución debe cumplir para lograr esa expectativa de disponibilidad. Para obtener más información, consulte Acuerdos de Nivel de Servicio para servicios en línea.

Contenido relacionado

- ¿Qué es Azure Elastic SAN?
- Plan para implementar una instancia de Elastic SAN
- Confiabilidad en Azure

Last updated on 28/04/2026

Confiabilidad de Azure Functions

Azure Functions es un servicio de proceso controlado por eventos que ejecuta pequeños bloques de código o *funciones* sin aprovisionamiento o administración de infraestructuras explícitas. Las funciones pueden responder a eventos como solicitudes HTTP, temporizadores, mensajes de cola y cambios en otros servicios de Azure. Esta funcionalidad hace que Functions sea adecuado para el procesamiento de datos, la integración del sistema y las tareas en segundo plano.

Cuando se usa Azure, la confiabilidad es una responsabilidad compartida. Microsoft proporciona una variedad de funcionalidades para admitir resistencia y recuperación. Es responsable de comprender cómo funcionan esas funcionalidades dentro de todos los servicios que usa y de seleccionar las funcionalidades que necesita para cumplir los objetivos empresariales y los objetivos de tiempo de actividad.

En este artículo se describe cómo hacer que Functions sea resistente a varias posibles interrupciones y problemas, incluidos errores transitorios, errores de zona de disponibilidad y errores en toda la región. También resalta información clave sobre el acuerdo de nivel de servicio (SLA) de Functions.

Recomendaciones de implementación de producción

Azure Well-Architected Framework proporciona recomendaciones sobre confiabilidad, seguridad, costo, operaciones y rendimiento. Para obtener información sobre cómo estas áreas influyen entre sí y contribuyen a una solución confiable de Functions, consulte Procedimientos recomendados de arquitectura para Functions.

Introducción a la arquitectura de confiabilidad

Al implementar Functions, es importante familiarizarse con estos conceptos:

- **Planes de hospedaje:** Los planes representan el entorno de hospedaje de las aplicaciones de funciones. El plan determina los recursos de proceso disponibles, el modelo de precios y el comportamiento de escalado.
- **Cuentas de almacenamiento:** Al crear una aplicación de funciones, debe especificar una cuenta de almacenamiento de host. La cuenta de almacenamiento gestiona aspectos de las operaciones internas de la aplicación de funciones, incluidos el almacenamiento de código

de funciones, el registro y la gestión de concurrencia (como en el caso de las concesiones de blobs para tipos específicos de desencadenadores).

También puede usar una cuenta de almacenamiento para la implementación. Esta cuenta de almacenamiento puede ser la misma que la cuenta de almacenamiento de host o una cuenta de almacenamiento diferente.

> ⓘ **Importante**
>
> Las cuentas de almacenamiento son una parte fundamental de la arquitectura de confiabilidad de Functions. Configúrelos para satisfacer los requisitos de resistencia de la aplicación de funciones.

- **Desencadenadores y enlaces**: Los desencadenadores y enlaces permiten que la función responda a eventos, reciba datos de otros servicios y escriba datos en otros servicios.
- **Durable Functions**: Durable Functions es una característica de Functions. Proporciona funciones con estado, como orquestaciones de larga duración y entidades con estado.

  Cuando se usan funciones duraderas, se configura un proveedor de almacenamiento que almacena el estado. Evalúe las características de confiabilidad del almacén de estado que elija y configúrela para satisfacer los requisitos de resistencia.

Resistencia a errores transitorios

Los errores transitorios son errores breves e intermitentes en los componentes. Se producen con frecuencia en un entorno distribuido como la nube y son una parte normal de las operaciones. Los errores transitorios se corrigen después de un breve período de tiempo. Es importante que las aplicaciones puedan controlar errores transitorios, normalmente mediante el reintento de solicitudes afectadas.

Todas las aplicaciones hospedadas en la nube deben seguir las instrucciones de control de errores transitorios de Azure cuando se comunican con cualquier API, bases de datos y otros componentes hospedados en la nube. Para obtener más información, consulte Recomendaciones para controlar errores transitorios.

Tenga en cuenta las siguientes recomendaciones para controlar errores transitorios en las aplicaciones de funciones:

- **Desencadenadores y enlaces:** La plataforma Functions incluye el control de errores transitorios integrado para desencadenadores y enlaces. Cuando se produce un error transitorio mientras se desencadena un desencadenador compatible o se leen o escriben datos de enlace admitidos, la plataforma puede reintentar automáticamente la operación. Este comportamiento de reintento integrado garantiza que los problemas de conectividad temporal o las interrupciones del servicio no impidan que la función se ejecute. Para obtener más información, consulte Control de errores y reintentos de Functions.

  Esta protección solo cubre errores transitorios. La plataforma no reintenta los errores persistentes, como un cadena de conexión mal configurado o un recurso eliminado.

  La plataforma trata los errores persistentes y los errores transitorios repetidos como errores. Configure el registro para capturar información sobre los errores de ejecución de funciones. Para obtener más información, consulte Configuración de la supervisión de Funciones.

- **Tu código de función:** En tu código de función, eres responsable de controlar errores transitorios cuando tu función llame a servicios externos. Implemente la lógica de reintento, los tiempos de espera y los patrones de cortacircuito según corresponda para las llamadas de servicio externas realizadas en su código de función. Diseñe las funciones para que sean idempotentes siempre que sea posible, de modo que los reintentos no creen efectos secundarios repetidos.

- **Clientes:** Las aplicaciones cliente que se conectan a funciones de forma sincrónica, como a través de una conexión HTTP, deben ser resistentes a errores transitorios.

Resistencia a errores de zona de disponibilidad

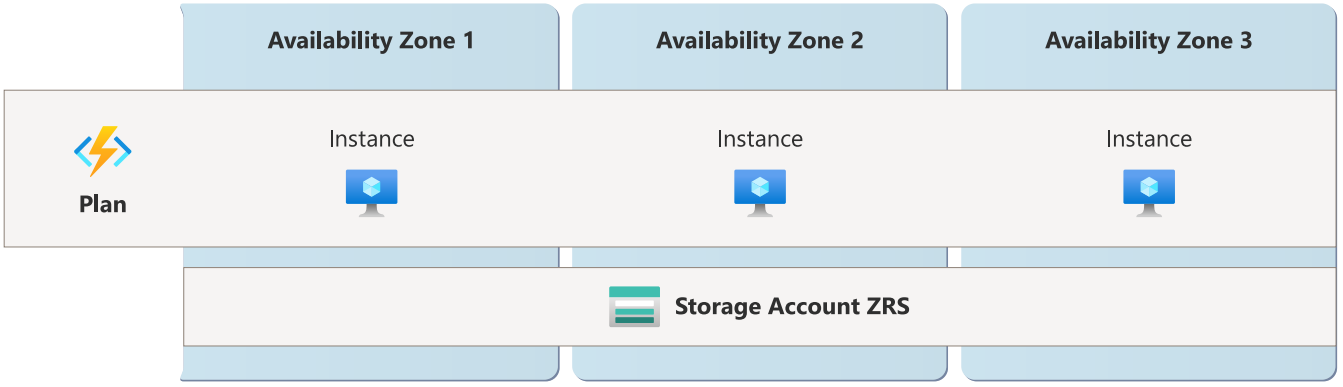
Las zonas de disponibilidad son grupos físicamente independientes de centros de datos dentro de una región de Azure. Cuando una zona falla, los servicios pueden transferirse a una de las zonas restantes.

Los planes Flex Consumption admiten implementaciones con redundancia de zona.

Cuando se habilita la redundancia de zona, la plataforma distribuye automáticamente las instancias del plan en todas las zonas de disponibilidad de la región seleccionada. Si alguna zona de disponibilidad de la región tiene un problema, las funciones continúan ejecutándose en instancias de zonas saludables.

Debe habilitar el almacenamiento con redundancia de zona (ZRS) en la cuenta de almacenamiento del host, para garantizar que también sea resistente a las interrupciones de

zona.



El plan es *no zonal* o *regional* cuando no se habilita la redundancia de zona. La plataforma puede ubicar instancias de plan en cualquier zona de disponibilidad de la región o en la misma zona. Las instancias del plan no son tolerantes a fallos en las zonas de disponibilidad. Es posible que su plan experimente tiempo de inactividad durante una interrupción en cualquier zona de la región.

Requisitos

- **Compatibilidad con regiones:** Puede implementar planes de consumo flexible con redundancia de zona en un conjunto específico de regiones. Puede recuperar la lista actual de regiones admitidas mediante la CLI de Azure. Para más información, consulte Visualización de regiones que admiten zonas de disponibilidad.

- **Cuenta de almacenamiento de host:** Debe configurar la cuenta de almacenamiento de host predeterminada de la aplicación de funciones para usar ZRS. Si usa una cuenta de almacenamiento de host que no está configurada para ZRS, es posible que la aplicación se comporte inesperadamente durante una interrupción de zona.

- **Cuenta de almacenamiento de contenedor de implementación:** Si usa una cuenta de almacenamiento independiente para el contenedor de implementación de la aplicación, actualícela también para que sea redundante de zona.

Consideraciones

Comportamientos no en tiempo de ejecución: La redundancia de zona solo garantiza un tiempo de actividad continuado para las aplicaciones implementadas. Una interrupción de zona de disponibilidad puede afectar a aspectos de Functions, aunque la aplicación siga sirviendo tráfico. Estos comportamientos incluyen el escalado del plan, la creación de aplicaciones, la configuración de la aplicación y la publicación de aplicaciones.

Distribución de instancias a través de zonas

Al configurar las aplicaciones del plan de consumo Flex como redundantes por zona, la plataforma distribuye automáticamente las instancias del plan en varias zonas de la región seleccionada, aplicando reglas diferentes para instancias siempre activas en comparación con las instancias bajo demanda.

- **Las instancias siempre disponibles** se distribuyen entre al menos dos zonas mediante la distribución circular.

  Para garantizar la resistencia de zona, la plataforma mantiene automáticamente al menos dos instancias listas para siempre para cada función o grupo de escalado por función, independientemente de la configuración siempre lista para la aplicación. Las instancias que crea la plataforma son administradas por la plataforma, facturadas como instancias listas para siempre y no cambian las opciones de configuración siempre preparadas.

- **Las instancias a petición** se generan a partir de volúmenes fuente de eventos cuando la aplicación se escala más allá del número de instancias siempre listas. Las instancias a petición se distribuyen entre zonas de disponibilidad de la mejor manera posible. La plataforma prioriza un escalado horizontal más rápido en lugar de una distribución uniforme entre zonas. La plataforma intenta equilibrar la distribución a lo largo del tiempo.

Costo

No se incurre en ningún costo adicional al habilitar la redundancia de zona. Los precios de un plan con redundancia de zona son los mismos que los de un plan de zona única. Sin embargo, habilitar la redundancia de zona afecta al número mínimo de instancias del plan.

Cuando se habilitan zonas de disponibilidad en una aplicación con una configuración de instancia del tipo siempre lista, y hay menos de dos instancias para cada función o grupo de escalado por función, la plataforma crea automáticamente dos instancias del tipo siempre lista para cada función o grupo de escalado por función. Estas nuevas instancias también se facturan como instancias siempre listas.

Para más información sobre los precios completos, consulte Precios de Functions.

Configurar soporte de zonas de disponibilidad

- **Cree un nuevo plan de Funciones con redundancia de zona.** Puede habilitar la redundancia de zona al crear un nuevo plan. Para obtener más información, consulte Creación de una

aplicación de funciones con redundancia de zona.

- **Habilite la redundancia de zona en un plan existente.** Puede activar o desactivar las zonas de disponibilidad para los planes de Elastic Premium existentes. Los planes elastic Premium tienen un comportamiento de capacidad específico que difiere de los planes dedicados (App Service) y requiere pasos de configuración adicionales. Para ver los pasos detallados, consulte Habilitación de la redundancia de zona en un plan existente.

Planeamiento y administración de capacidad

Las aplicaciones de funciones con redundancia de zona siguen ejecutándose incluso cuando las zonas de la región experimentan una interrupción.

Durante una interrupción en la zona, Functions detecta instancias perdidas e intenta automáticamente buscar o crear instancias de reemplazo en las zonas saludables. La plataforma realiza este proceso en base al mejor esfuerzo y no garantiza el éxito. Si la carga de trabajo requiere un número específico de instancias para mantener el nivel de servicio esperado, considere *la posibilidad de sobreaprovisionar* el número de instancias siempre listas. El sobreaprovisionamiento permite a la solución tolerar la pérdida de capacidad y seguir funcionando sin reducir el rendimiento. Para más información, consulte Administración de la capacidad mediante el sobreaprovisionamiento.

Comportamiento cuando todas las zonas están en buen estado

En esta sección se describe qué esperar cuando un plan es con redundancia de zona, la cuenta de almacenamiento de host usa ZRS y todas las zonas de disponibilidad están operativas.

- **Operación entre zonas:** Al configurar la redundancia de zona en Functions, las solicitudes se distribuyen automáticamente entre las instancias de cada zona de disponibilidad. Una solicitud puede ir a cualquier instancia de cualquier zona de disponibilidad.

- **Replicación de datos entre zonas:** Functions es un servicio de proceso sin estado, por lo que no hay datos para replicar entre zonas. La plataforma replica la configuración entre zonas automáticamente.

  Si la cuenta de almacenamiento de host usa ZRS, Azure Storage replica sincrónicamente sus datos en varias zonas de disponibilidad.

  Para funciones duraderas, revise el proveedor de almacenamiento para obtener información sobre cómo replica datos entre zonas.

Comportamiento durante un fallo de zona

En esta sección se describe qué esperar cuando un plan tiene redundancia de zona, la cuenta de almacenamiento del host usa ZRS y hay una interrupción en una zona de disponibilidad.

- **Detección y respuesta:** La plataforma Functions es responsable de detectar un error en una zona de disponibilidad. No es necesario iniciar una conmutación por error de zona.
- **Notificación**: Microsoft no le notifica automáticamente cuando una zona está inactiva. Sin embargo, puede usar Azure Resource Health para supervisar el estado de un recurso individual y puede configurar alertas de Resource Health para notificarle problemas. También puede usar Azure Service Health para comprender el estado general del servicio, incluidos los errores de zona, y puede configurar alertas de Service Health para notificarle problemas.
- **Solicitudes activas:** Cuando una zona de disponibilidad no está disponible, las solicitudes en curso que se conectan a una instancia de la zona de disponibilidad errónea finalizan y se deben reintentar. Asegúrese de que las aplicaciones están preparadas siguiendo las instrucciones de control de errores transitorios.
- **Pérdida de datos esperada:** No se espera que los errores de zona causen pérdida de datos porque Functions es un servicio sin estado.

  Si la cuenta de almacenamiento de host utiliza ZRS, el almacenamiento garantiza que no se pierdan datos debido a un fallo de zona.

  Para las funciones duraderas, compruebe su proveedor de almacenamiento para saber si la pérdida de datos sea posible durante un fallo de zona.
- **Tiempo de inactividad esperado:** Durante las interrupciones de zona, las conexiones pueden experimentar interrupciones breves que suelen durar unos segundos a medida que se redistribuye el tráfico. Asegúrese de que las aplicaciones están preparadas siguiendo las instrucciones de control de errores transitorios.
- **Reenrutamiento del tráfico:** Functions detecta las instancias perdidas de esa zona e intenta buscar nuevas instancias de reemplazo. Una vez que Functions encuentra reemplazos, distribuye el tráfico entre las nuevas instancias según sea necesario.

  > ⓘ **Importante**

> Azure no garantiza que las solicitudes de más instancias se realicen correctamente en un escenario de zona descendente. La plataforma intenta reponer instancias perdidas de forma óptima. Si necesita capacidad garantizada durante un error de zona de disponibilidad, cree y configure los planes para tener en cuenta la pérdida de zona mediante el sobreaprovisionamiento de la capacidad.

- **Comportamientos no en tiempo de ejecución:** Las aplicaciones de un plan de aplicación de funciones con redundancia de zona siguen ejecutándose y atendiendo el tráfico incluso si una zona de disponibilidad experimenta una interrupción. Sin embargo, los comportamientos que no son de ejecución pueden verse afectados durante una interrupción de la zona de disponibilidad. Estos comportamientos incluyen el escalado de aplicaciones de funciones, la creación de aplicaciones, la configuración de la aplicación y la publicación de aplicaciones.

Recuperación de zona

Cuando se recupera la zona de disponibilidad, Functions restaura automáticamente las instancias de la zona de disponibilidad, quita las instancias temporales creadas en las otras zonas de disponibilidad y vuelve a enrutar el tráfico entre las instancias como normal.

Prueba de fallos de zona

La plataforma Functions administra el enrutamiento del tráfico, la conmutación por error y la recuperación de zona para los recursos con redundancia de zona. No es necesario iniciar nada. Azure administra completamente esta característica, por lo que no es necesario validar los procesos de error de zona de disponibilidad.

Resistencia a errores en toda la región

Functions es un servicio de una sola región. Si la región deja de estar disponible, el recurso de Functions tampoco está disponible.

Soluciones personalizadas de varias regiones para la resistencia

Para evitar interrupciones en su servicio durante caídas regionales, puede implementar de forma redundante las mismas funciones en aplicaciones de funciones en varias regiones.

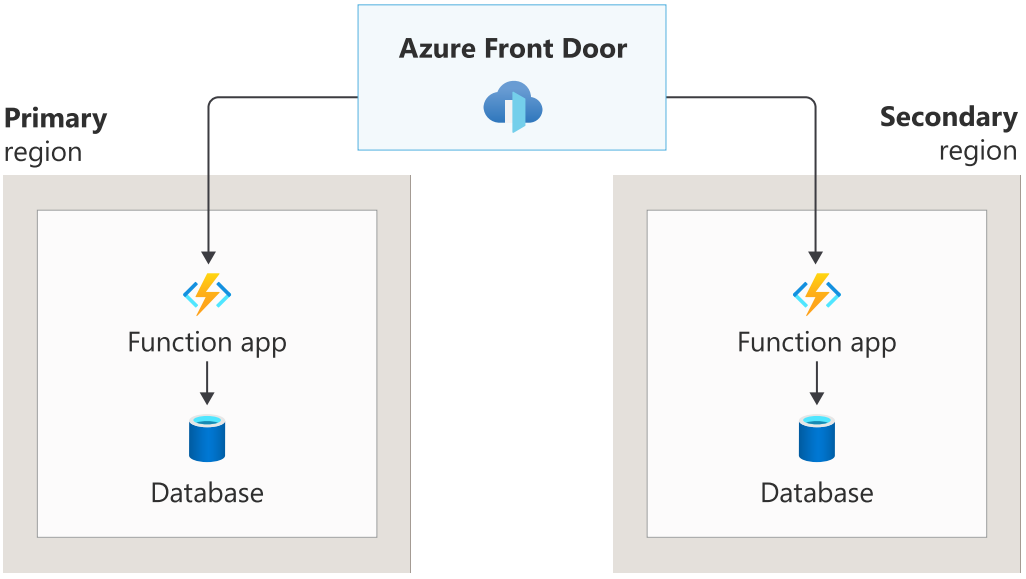
Usted es responsable de:

- Implementación de aplicaciones de funciones en varias regiones.
- Administración de la distribución del tráfico entre regiones.
- Implementación de mecanismos de conmutación por error.
- Garantizar la coherencia de los datos entre regiones (si procede).
- Supervisión y administración de implementaciones entre regiones.

Al ejecutar el mismo código de función en varias regiones, tenga en cuenta los patrones *activo-activo* y *activo-pasivo* . En las secciones siguientes se presentan estos patrones, pero no se proporcionan instrucciones detalladas ni pasos de configuración.

Patrón activo-activo para funciones de desencadenador HTTP

En un patrón activo-activo, las funciones en las dos regiones ejecutan activamente y procesan eventos, ya sea de forma duplicada o de manera alternada. Utilice un patrón activo-activo en combinación con Azure Front Door para sus funciones críticas desencadenadas por HTTP, que pueden enrutar y distribuir solicitudes HTTP entre funciones que se están ejecutando en varias regiones. Azure Front Door también puede comprobar periódicamente el estado de cada punto de conexión. Si un servicio de una región deja de responder a las comprobaciones de estado, Azure Front Door lo quita de la rotación y solo reenvía el tráfico a los servicios saludables restantes.



Patrón activo-pasivo para funciones de desencadenador que no son HTTP

En el caso de funciones dirigidas por eventos que no se activan mediante HTTP (como activadores de Azure Service Bus y Azure Event Hubs), use un patrón activo-pasivo. En un patrón

activo-pasivo, las instancias de función se ejecutan en la región que recibe eventos, mientras que las instancias de la región secundaria permanecen inactivas. Este patrón garantiza que solo una función procese cada mensaje, lo que ayuda a mantener la coherencia de los datos. Permite conmutar por error a la región secundaria durante un desastre, como una interrupción de la región.

Considere la conmutación por error de la app de funciones junto con los comportamientos de conmutación por error de otros servicios que use, como:

- Service Bus, replicación geográfica y recuperación ante desastres
- Replicación geográfica de Event Hubs y recuperación ante desastres geográfica

Considere una topología de ejemplo que usa un activador de Event Hubs, donde el espacio de nombres de Event Hubs está configurado para la recuperación ante desastres geográficos. En este caso, el patrón activo-pasivo requiere los siguientes componentes:

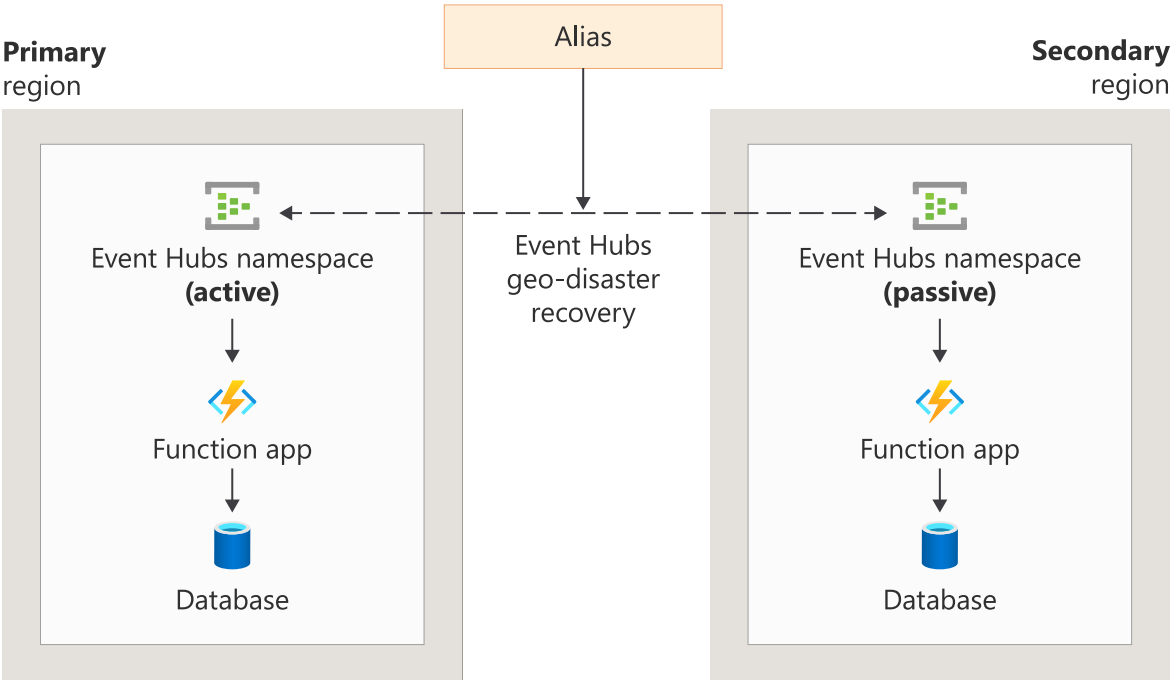
- Espacios de nombres de Event Hubs implementados tanto en una región primaria como en una secundaria.

- Se habilitó la recuperación ante desastres geográfica para vincular los centros de eventos principales y secundarios. Esta configuración también crea un *alias* que puede usar para conectarse al espacio de nombres de Event Hubs y cambiar el alias de la base de datos principal a la secundaria sin cambios en la información de conexión.

- Aplicaciones de funciones desplegadas tanto en la región primaria como en la secundaria. La aplicación de la región secundaria permanece inactiva porque no recibe mensajes.

- Los desencadenadores de cada aplicación de funciones utilizan la cadena de conexión *direct* (noalias) para su espacio de nombres correspondiente de Event Hubs.

- Los publicadores del espacio de nombres de Event Hubs publican al alias de la cadena de conexión.



Antes de la conmutación por error, los publicadores que envían eventos al alias compartido dirigen el tráfico hacia el centro de eventos principal. La aplicación de funciones principal escucha exclusivamente al hub de eventos principal. La aplicación de funciones secundaria permanece pasiva e inactiva.

Cuando se inicia el failover, los publicadores que envían eventos al alias compartido son dirigidos al centro de eventos secundario. La aplicación de funciones secundaria se activa y se desencadena automáticamente. El centro de eventos puede controlar todo el proceso de conmutación por error y cada aplicación de funciones solo se ejecuta cuando su centro de eventos correspondiente está activo.

Durable Functions

Para la recuperación ante desastres en varias regiones para funciones duraderas, consulte Disaster recovery and geo-distribution in Azure durable functions.

Resistencia al mantenimiento del servicio

Functions realiza actualizaciones de servicio periódicas y otras tareas de mantenimiento.

- **Resistencia de errores transitorios:** Durante el mantenimiento del servicio, las instancias que ejecutan la aplicación de funciones pueden reiniciarse o experimentar interrupciones temporales. Asegúrese de que las aplicaciones cliente que interactúan con la aplicación de funciones son resistentes a errores transitorios.

- **Habilite la redundancia de zona:** Al habilitar la redundancia de zona en el plan, también se mejora la resistencia durante las actualizaciones de la plataforma. Implementar varias instancias en tu plan y habilitar la redundancia de zonas proporciona una capa adicional de resiliencia si una instancia o una zona se vuelven defectuosas durante un proceso de actualización.

Resistencia a las implementaciones de aplicaciones

Las implementaciones de aplicaciones presentan el riesgo de problemas en un entorno de producción. Prepárese para revertir una actualización si provoca problemas. Controlar cómo se implementan las actualizaciones para reducir la interrupción de los reinicios de la aplicación.

Los planes de consumo flexible admiten estrategias de actualización del sitio, que proporcionan varias maneras de implementar las actualizaciones de la aplicación. Estas estrategias incluyen actualizaciones graduales para implementaciones sin tiempo de inactividad.

Acuerdo de nivel de servicio

El acuerdo de nivel de servicio (SLA) para Azure servicios describe la disponibilidad esperada de cada servicio y las condiciones que la solución debe cumplir para lograr esa expectativa de disponibilidad. Para obtener más información, consulte Acuerdos de Nivel de Servicio para servicios en línea.

Functions proporciona acuerdos de nivel de servicio de disponibilidad distintos para el plan de consumo y para otros tipos de plan.

Contenido relacionado

- Recuperación ante desastres y distribución geográfica en funciones duraderas de Azure
- Creación de una instancia de Azure Front Door
- Consideraciones sobre la conmutación por error de Event Hubs
- Estrategias de arquitectura para usar regiones y zonas de disponibilidad
- Confiabilidad en Azure

Last updated on 28/04/2026

Confiabilidad en Azure HDInsight

En este artículo se describe la compatibilidad con la confiabilidad en Azure HDInsighty se tratan zonas de disponibilidad y recuperación entre regiones y continuidad empresarial. Para obtener información general más detallada sobre la confiabilidad de Azure, consulte Confiabilidad de Azure.

Soporte para zonas de disponibilidad

Las zonas de disponibilidad son grupos físicamente independientes de centros de datos dentro de una región de Azure. Cuando una zona falla, los servicios pueden transferirse a una de las zonas restantes.

Azure HDInsight admite una configuración de implementación zonal. Los nodos de clúster de Azure HDInsight se colocan en una sola zona que se selecciona en la región seleccionada. Un clúster de HDInsight zonal está aislado de las interrupciones que se producen en otras zonas. Sin embargo, si una interrupción afecta a la zona específica elegida para el clúster de HDInsight, el clúster no estará disponible. Este modelo de implementación proporciona conectividad de red de baja latencia y económica dentro del clúster. La replicación de este modelo de implementación en varias zonas de disponibilidad puede proporcionar un mayor nivel de disponibilidad para protegerse frente a errores de hardware.

> **ⓘ Importante**
>
> En el caso de las implementaciones en las que los usuarios no especifican una zona específica, los tipos de nodo no son resistentes a la zona y pueden experimentar tiempo de inactividad durante una interrupción en cualquier zona de esa región.

Prerequisites

- Las zonas de disponibilidad solo se admiten para los clústeres creados después del 15 de junio de 2023. La configuración de la zona de disponibilidad no se puede actualizar después de crear el clúster. Tampoco puede actualizar un clúster de zona sin disponibilidad para usar zonas de disponibilidad.
- Los clústeres se deben crear en una red virtual personalizada.

- Debe traer su propia base de datos SQL para la base de datos de Ambari y el metastore externo, como metastore de Hive, para que pueda configurar estas bases de datos en la misma zona de disponibilidad.

- Los clústeres de HDInsight deben crearse con la opción de zona de disponibilidad en una de las siguientes regiones:
  - Australia East
  - Sur de Brasil
  - Canada Central
  - Central US
  - East US
  - Este de EE. UU. 2
  - France Central
  - Centro-oeste de Alemania
  - Japan East
  - Korea Central
  - Norte de Europa
  - Centro de Qatar
  - Sudeste asiático
  - Centro-sur de EE. UU.
  - UK South
  - Gobierno de EE. UU. - Virginia
  - West Europe
  - Oeste de EE. UU. 2

Creación de un clúster de HDInsight con zona de disponibilidad

Puede usar la plantilla de Azure Resource Manager (ARM) para iniciar un clúster de HDInsight en una zona de disponibilidad especificada.

En la sección de recursos, debe agregar una sección de "zonas" y proporcionar en qué zona de disponibilidad desea que se implemente este clúster.

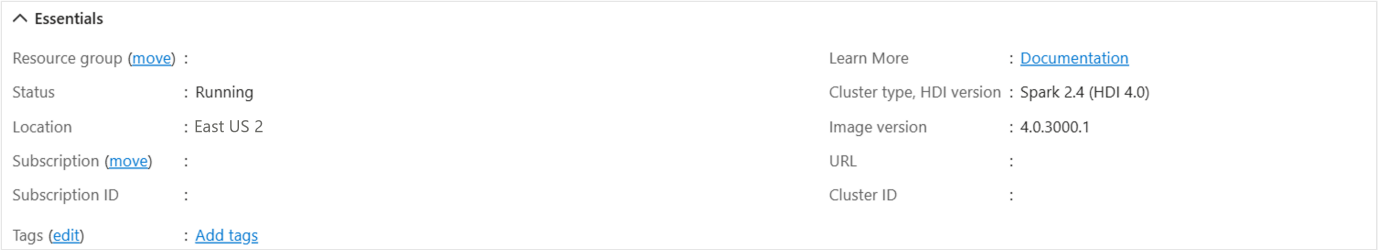
JSON

```
"resources": [
    {
        "type": "Microsoft.HDInsight/clusters",
        "apiVersion": "2021-06-01",
```

```
            "name": "[parameters('cluster name')]",
            "location": "East US 2",
            "zones": [
                "1"
            ],
        }
    ]
```

Comprobación de nodos dentro de una zona de disponibilidad entre zonas

Cuando el clúster de HDInsight esté listo, puede comprobar la ubicación para ver en qué zona de disponibilidad se implementan.



Obtener la respuesta de la API:

JSON

```
[
        {
            "location": "East US 2",
            "zones": [
                "1"
            ],
        }
]
```

Escalado vertical del clúster

Puede escalar verticalmente un clúster de HDInsight con más nodos de trabajo. Los nodos de trabajo recién agregados se colocarán en la misma zona de disponibilidad de este clúster.

Migración de zonas de disponibilidad

Actualmente, los clústeres de Azure HDInsight no admiten la migración local de instancias de clúster existentes a compatibilidad con zonas de disponibilidad. Sin embargo, puede elegir volver a crear el clústery elegir una región o zona de disponibilidad diferente durante la creación del

clúster. Un clúster en espera secundario en una región diferente y una zona de disponibilidad diferente se puede usar en escenarios de recuperación ante desastres.

Experiencia de reducción de zona

Cuando una zona de disponibilidad deja de funcionar:

- No se puede usar ssh en este clúster.
- No se puede eliminar, escalar verticalmente o reducir verticalmente este clúster.
- No se pueden enviar trabajos ni ver el historial de trabajos.
- Todavía puede enviar una nueva solicitud de creación de clústeres en otra región.

Recuperación ante desastres entre regiones y continuidad empresarial

La recuperación ante desastres (DR) hace referencia a las prácticas que las organizaciones usan para recuperarse de eventos de alto impacto, como desastres naturales o implementaciones con errores que producen tiempo de inactividad y pérdida de datos. Independientemente de la causa, el mejor remedio para un desastre es un plan de recuperación ante desastres bien definido y probado y un diseño de aplicaciones que apoye activamente la recuperación ante desastres. Antes de empezar a crear el plan de recuperación ante desastres, consulte Recomendaciones para diseñar una estrategia de recuperación ante desastres.

Para la recuperación ante desastres, Microsoft usa el modelo de responsabilidad compartida. En este modelo, Microsoft garantiza que la infraestructura de línea base y los servicios de plataforma estén disponibles. Sin embargo, muchos servicios de Azure no replican automáticamente datos ni se revierten de una región con errores para realizar la replicación cruzada en otra región habilitada. En esos servicios, es responsable de configurar un plan de recuperación ante desastres válido para su carga de trabajo. La mayoría de los servicios que se ejecutan en ofertas de plataforma como servicio (PaaS) de Azure proporcionan características e instrucciones para admitir la DR. Puede usar características específicas del servicio para admitir la recuperación rápida con el fin de contribuir al desarrollo del plan de recuperación ante desastres.

Los clústeres de Azure HDInsight dependen de muchos servicios de Azure, como almacenamiento, bases de datos, Active Directory, Active Directory Domain Services, redes y Key Vault. Una aplicación de análisis bien diseñada, de alta disponibilidad y tolerante a errores debe diseñarse con suficiente redundancia para resistir las interrupciones regionales o locales en uno o varios de estos servicios. En esta sección se proporciona información general sobre los

procedimientos recomendados, la disponibilidad de una sola región y varias regiones, y las opciones de optimización para el planeamiento de la continuidad empresarial.

Recuperación ante desastres en la geografía de varias regiones

La mejora de la continuidad empresarial mediante la recuperación ante desastres de alta disponibilidad entre regiones requiere diseños arquitectónicos de mayor complejidad y un costo más alto. En las siguientes tablas se detallan algunas áreas técnicas que pueden aumentar el costo total de propiedad.

Optimizaciones de costos

Expandir tabla

Área	Causa de la escalación de costo	Estrategias de optimización
Almacenamiento de datos	Duplicación de tablas o datos principales en una región secundaria.	Replique solo los datos mantenidos.
Salida de datos	Las transferencias de datos entre regiones salientes tienen un precio. Revise las directrices de precios de ancho de banda.	Replique solo los datos mantenidos para reducir la superficie de salida de la región.
Cálculo del clúster	Clústeres de HDInsight adicionales en la región secundaria.	Use scripts automatizados para implementar un proceso secundario después del error principal. Use el escalado automático para mantener el tamaño del clúster secundario al mínimo. Use SKU de máquinas virtuales más baratas. Cree los clústeres secundarios en las regiones donde se pueden descontar las SKU de máquinas virtuales.
Autenticación	Los escenarios multiusuario de la región secundaria incurren en configuraciones adicionales de Microsoft Entra Domain Services	Evite las configuraciones multiusuario en la región secundaria.

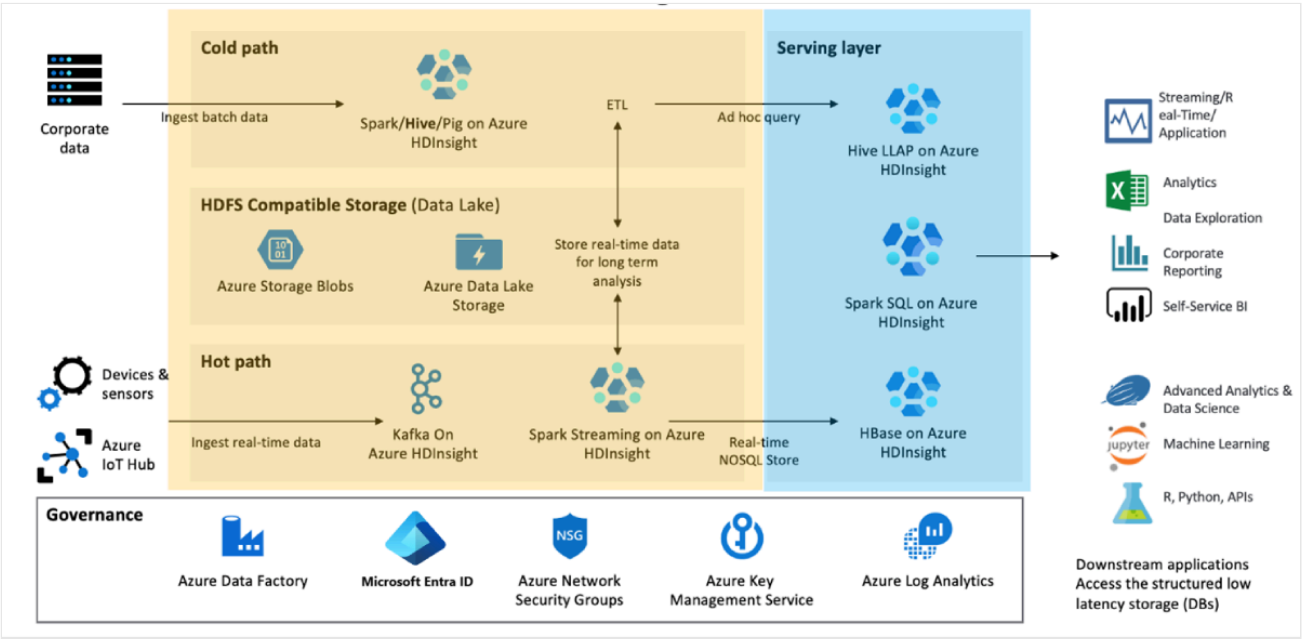
Optimizaciones de complejidad

Expandir tabla

Área	Causa de la escalación de complejidad	Estrategias de optimización
Patrones de lectura y escritura	Requerir que los elementos primarios y secundarios estén habilitados para lectura y escritura.	Diseñe el elemento secundario para que sea de solo lectura.
RTO y RPO cero	Requerir pérdida de datos cero (RPO = 0) y tiempo de inactividad cero (RTO = 0).	Diseñe RPO y RTO de maneras para reducir el número de componentes que necesitan conmutar por error. Para obtener más información sobre RTO y RPO, consulte Azure HDInsight arquitecturas de continuidad empresarial.
Funciones empresariales	Requerir toda la funcionalidad empresarial de la primaria en la secundaria.	Evalúe si puede realizar la ejecución con un subconjunto crítico mínimo básico de la funcionalidad empresarial en la secundaria.
Connectivity	Requerir que todos los sistemas ascendentes y descendentes de la conectividad primaria se conecten también a la secundaria.	Limite la conectividad secundaria a un subconjunto crítico mínimo básico.

Cuando el plan de recuperación ante desastres de varias regiones, tenga en cuenta las siguientes recomendaciones:

- Determine la funcionalidad de negocio mínima que necesitará si se produce un desastre y por qué. Por ejemplo, evalúe si necesita funcionalidades de conmutación por error para la capa de transformación de datos (que se muestra en amarillo) *y* la capa de servicio de datos (que se muestra en azul) o si solo necesita conmutación por error para la capa de servicio de datos.



- Segmente los clústeres en función de la carga de trabajo, el ciclo de vida de desarrollo y los departamentos. El hecho de tener más clústeres reduce las posibilidades de que un único error grande afecte a varios procesos empresariales diferentes.
- Haga que sus regiones secundarias sean de solo lectura. Las regiones de conmutación por error con capacidades de lectura y escritura pueden conducir a arquitecturas complejas.
- Los clústeres transitorios son más fáciles de administrar cuando se produce un desastre. Diseñe sus cargas de trabajo de forma que los clústeres se puedan reciclar y no se mantenga ningún estado en ellos.
- A menudo, las cargas de trabajo se dejan sin terminar si se produce un desastre y es necesario reiniciar en la nueva región. Diseñe las cargas de trabajo para que sean idempotentes por naturaleza.
- Use la automatización durante las implementaciones de clústeres y asegúrese de que los valores de configuración del clúster se incluyen en el script en la medida de lo posible para garantizar una implementación rápida y totalmente automatizada si se produce un desastre.

Detección, notificación y administración de interrupciones

- Use las herramientas de supervisión de Azure en HDInsight para detectar un comportamiento anómalo en el clúster y establecer las notificaciones de alerta correspondientes. Puede implementar las soluciones de administración específicas del clúster de HDInsight preconfiguradas que recopilan las métricas de rendimiento importantes del tipo de clúster específico. Para más información, consulte Supervisión de Azure para HDInsight.
- Suscríbase a las alertas de estado de Azure para recibir notificaciones sobre problemas de servicio, mantenimiento planeado, avisos de estado y seguridad para una suscripción, un servicio o una región. Las notificaciones de estado que incluyen la causa del problema y la hora de llegada estimada resuelta le ayudan a ejecutar mejor la conmutación por error y la conmutación por recuperación. Para obtener más información, consulte Documentación de Azure Service Health.

Recuperación ante desastres en una sola región geográfica

Cada componente de un sistema HDInsight básico tiene sus propios mecanismos de tolerancia a errores de una sola región. Tenga en cuenta que no siempre toma un evento catastrófico para afectar a la funcionalidad empresarial. Los incidentes de servicio en uno o varios de los servicios

siguientes en una sola región también pueden provocar la pérdida de la funcionalidad empresarial esperada.

- **Proceso (máquinas virtuales): clúster de Azure HDInsight**. HDInsight ofrece un Acuerdo de Nivel de Servicio de disponibilidad del 99,9 %. Para proporcionar alta disponibilidad en una sola implementación, HDInsight viene acompañado de muchos servicios que están en modo de alta disponibilidad de manera predeterminada. Los mecanismos de tolerancia a errores de HDInsight los proporcionan los servicios de alta disponibilidad de los ecosistemas de Microsoft y Apache OSS.

  Los siguientes componentes de infraestructura están diseñados para ser de alta disponibilidad:
  - Nodos principales activos y en espera
  - Varios nodos de puerta de enlace
  - Tres nodos de Quorum de Zookeeper
  - Nodos de trabajo distribuidos por dominios de actualización y de error

  Los siguientes servicios también están diseñados para ser de alta disponibilidad:
  - Servidor Apache Ambari
  - Servidores de escala de tiempo de la aplicación para YARN
  - Servidor de Historial de trabajos para MapReduce de Hadoop
  - Apache Livy
  - HDFS
  - Administrador de Recursos YARN
  - HBase Master

  Para más información, consulte servicios de alta disponibilidad admitidos con Azure HDInsight.

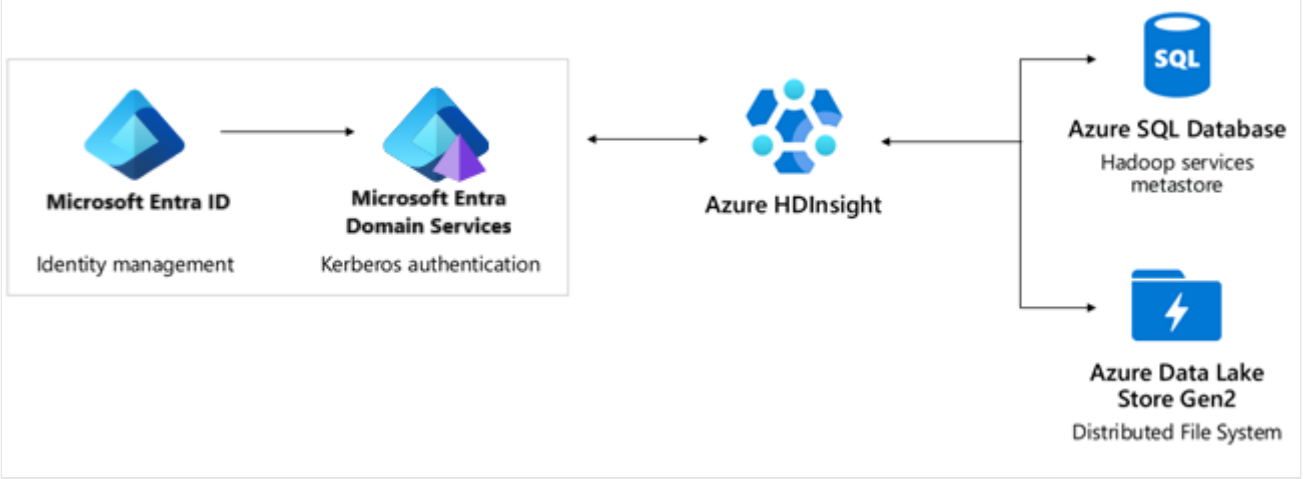
- **Metastores: Azure SQL Database**. HDInsight usa Azure SQL Database como metastore, lo que proporciona un Acuerdo de Nivel de Servicio del 99,99 %. Se conservan tres réplicas de datos dentro de un centro de datos con replicación sincrónica. Si se produce una pérdida de réplica, se proporciona una réplica alternativa sin problemas. Replicación geográfica activa es compatible de forma integrada con un máximo de cuatro centros de datos. Cuando hay una conmutación por error, ya sea manual o del centro de datos, la primera réplica de la jerarquía se convierte automáticamente en compatible con lectura y escritura. Para más información, consulte Continuidad empresarial de Azure SQL Database.

- **Storage: Azure Data Lake Gen2 o Blob Storage**. HDInsight recomienda Azure Data Lake Storage Gen2 como capa de almacenamiento subyacente. Azure Storage, incluido Azure

Data Lake Storage Gen2, proporciona un Acuerdo de Nivel de Servicio del 99,9 %. HDInsight usa el servicio LRS en el que se conservan tres réplicas de datos dentro de un centro de datos y la replicación es sincrónica. Si se produce una pérdida de réplica, se proporciona una réplica sin problemas.

- **Autenticación: Microsoft Entra ID, Microsoft Entra Domain Services, Enterprise Security Package**.
  - Microsoft Entra ID proporciona un Acuerdo de Nivel de Servicio del 99,9 %. Active Directory es un servicio global con varios niveles de redundancia interna y capacidad de recuperación automática. Para obtener más información, consulte cómo Microsoft en la mejora continua de la confiabilidad del Microsoft Entra ID.
  - Microsoft Entra Domain Services proporciona un Acuerdo de Nivel de Servicio del 99,9 %. Microsoft Entra Domain Services es un servicio de alta disponibilidad hospedado en centros de datos distribuidos globalmente. Los conjuntos de réplicas son una característica en vista previa de Microsoft Entra Domain Services que habilita la recuperación geográfica ante desastres si una región de Azure se queda sin conexión. Para obtener más información, consulte Conceptos y características de los conjuntos de réplicas de Microsoft Entra Domain Services para obtener más información.
  - Azure DNS proporciona un Acuerdo de Nivel de Servicio del 100 %. HDInsight usa Azure DNS en varios lugares para la resolución de nombres de dominio.

- **Servicios opcionales**, como Azure Key Vault y Azure Data Factory.



Contenido relacionado

- Arquitecturas de continuidad empresarial de Azure HDInsight
- Caso práctico de arquitectura de solución de alta disponibilidad de Azure HDInsight
- ¿Qué son Apache Hive y HiveQL en Azure HDInsight?
- confiabilidad Azure

Last updated on 01/05/2026

Confiabilidad en Azure Kubernetes Service (AKS)

Azure Kubernetes Service (AKS) es un servicio de orquestación de contenedores administrado que simplifica la implementación, la administración y las operaciones de Kubernetes.

Cuando se usa Azure, la confiabilidad es una responsabilidad compartida. Microsoft proporciona una variedad de funcionalidades para admitir resistencia y recuperación. Es responsable de comprender cómo funcionan esas funcionalidades dentro de todos los servicios que usa y de seleccionar las funcionalidades que necesita para cumplir los objetivos empresariales y los objetivos de tiempo de actividad.

En este artículo se describe cómo hacer que Azure Kubernetes Service (AKS) sea resistente a una variedad de posibles interrupciones y problemas, incluidos errores transitorios, interrupciones de zona de disponibilidad y interrupciones en regiones. También se describe cómo puede usar copias de seguridad para recuperarse de otros tipos de problemas y se resalta cierta información clave sobre el acuerdo de nivel de servicio (SLA) de Azure Kubernetes Service (AKS).

Recomendaciones de implementación de producción

Para obtener recomendaciones sobre cómo implementar cargas de trabajo de producción confiables en AKS, consulte los siguientes artículos:

- Procedimientos recomendados de implementación y confiabilidad de clústeres para AKS
- Descripción general de la alta disponibilidad (HA) y la recuperación ante desastres (DR) para AKS
- Consideraciones de resistencia de zona para AKS

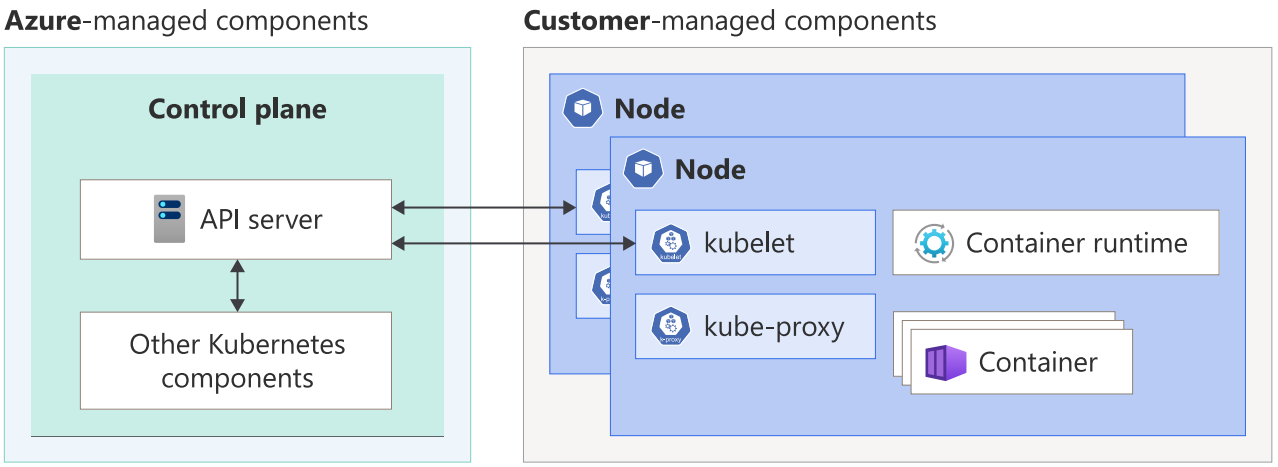
Introducción a la arquitectura de confiabilidad

Al crear un clúster de AKS, la plataforma Azure crea y configura automáticamente:

- Un plano de control que tiene el servidor de API, etcd, el programador y otros pods necesarios para administrar la carga de trabajo.

- Un grupo de nodos del sistema a tu suscripción que hospeda los complementos y pods adicionales que se ejecutan en el espacio de nombres *kube-system*.

Una vez completada la configuración del grupo de nodos inicial, puede agregar o eliminar grupos de nodos para sus propias cargas de trabajo de usuario. AKS no administra grupos de nodos para la confiabilidad y debe asegurarse de que las cargas de trabajo son resistentes a los errores de infraestructura.



La resistencia es una responsabilidad compartida entre usted y Microsoft. Como servicio de cálculo, AKS administra algunos aspectos de la fiabilidad del clúster, pero usted es responsable de gestionar otros aspectos.

- **Microsoft administra** el plano de control y otros componentes administrados de AKS.
- **Es su responsabilidad**:
  - *Defina cómo deben configurarse los componentes, incluidos los grupos de nodos y los equilibradores de carga que se conectan a los servicios, para satisfacer los requisitos de confiabilidad.* Después de definir los componentes, Microsoft los implementa y administra en su nombre.
  - *Administre los componentes fuera del clúster de AKS, incluido el almacenamiento y las bases de datos.* Compruebe que estos componentes cumplen los requisitos de confiabilidad. Al implementar las cargas de trabajo, asegúrese de que otros componentes de Azure también están configurados para la resistencia siguiendo los procedimientos recomendados para esos servicios.

Resistencia a errores transitorios

Los errores transitorios son errores breves e intermitentes en los componentes. Se producen con frecuencia en un entorno distribuido como la nube y son una parte normal de las operaciones. Los errores transitorios se corrigen después de un breve período de tiempo. Es importante que las aplicaciones puedan controlar errores transitorios, normalmente mediante el reintento de solicitudes afectadas.

Todas las aplicaciones hospedadas en la nube deben seguir las instrucciones de control de errores transitorios de Azure cuando se comunican con cualquier API, bases de datos y otros componentes hospedados en la nube. Para obtener más información, consulte Recomendaciones para controlar errores transitorios.

Cuando utilizas AKS, pueden ocurrir fallos transitorios por varias razones, como fallos de aplicaciones, operaciones de escalado y balanceo de pods, actualización de nodos y fallos temporales en la infraestructura, como problemas de hardware o de redes.

Es imposible eliminar todos los errores transitorios, por lo que los clientes que acceden a las aplicaciones hospedadas en AKS deben estar preparados para reintentar las solicitudes con errores y seguir otras recomendaciones de control de errores transitorios. Puede minimizar la probabilidad de errores transitorios y evitar o mitigar el tiempo de inactividad que podrían provocar siguiendo los procedimientos recomendados de Kubernetes y Azure en la implementación.

- ✓ Establece los presupuestos de interrupciones de pod (PDB) en el pod YAML para especificar cuántos pods necesitas tener en un estado `Ready` en un momento dado. Al establecer archivos PDB, AKS garantiza una disponibilidad mínima de réplicas cuando ejecuta operaciones para acordonar y purgar los nodos. Si no se puede satisfacer una PDB durante procesos como actualizaciones, el pod continúa funcionando y es posible que se produzca un error en la operación. Para obtener más información, consulte PDB.
- ✓ Use `maxUnavailable` para definir el número máximo de réplicas que pueden dejar de estar disponibles en un momento dado. Por ejemplo, cuando se realiza un reinicio gradual, AKS garantiza que no más del número `maxUnavailable` de pods se reinician en un momento dado. Para obtener más información, consulte maxUnavailable.
- ✓ Siga los procedimientos recomendados de implementación. Las réplicas de pod también pueden fallar debido a problemas de la aplicación. Para más información, consulte Procedimientos recomendados de nivel de implementación para la confiabilidad del clúster de AKS.

> ⓘ **Nota:**
>
> Si quiere que AKS valide las implementaciones para cumplir los procedimientos recomendados y proporcione notificaciones de bloqueo o advertencia, puede usar medidas de seguridad de implementación. Las medidas de seguridad de implementación son una oferta administrada que ayuda a aplicar los procedimientos recomendados del producto antes de que el código se implemente en el clúster.

Resistencia a errores de zona de disponibilidad

Las zonas de disponibilidad son grupos físicamente independientes de centros de datos dentro de una región de Azure. Cuando una zona falla, los servicios pueden transferirse a una de las zonas restantes.

Al implementar un clúster de AKS en una región que admita zonas de disponibilidad, los distintos componentes requieren distintos tipos de configuración.

El plano de control de AKS es resiliente a zonas por defecto. Si se produce un error en una zona, el plano de control no requiere ninguna configuración ni administración para lograr resistencia. Sin embargo, la resistencia del plano de control no es suficiente para que el clúster permanezca operativo cuando se produce un error en una zona. Para el grupo de nodos del sistema y los grupos de nodos de usuario que implemente, debe habilitar la compatibilidad con zonas de disponibilidad para ayudar a garantizar que las cargas de trabajo son resistentes a los errores de zona de disponibilidad.

Requisitos

Compatibilidad con regiones: Puede implementar clústeres de AKS resistentes a zonas en cualquier región que admita zonas de disponibilidad.

Consideraciones

Para mejorar la confiabilidad y la resistencia de las cargas de trabajo de producción de AKS en una región, debe configurar AKS para la redundancia de zona mediante la realización de las siguientes configuraciones:

- **Implemente varias réplicas.** Kubernetes distribuye los pods entre nodos en función de las etiquetas de nodo. Para distribuir la carga de trabajo entre zonas, debe implementar varias réplicas del pod. Por ejemplo, si configura el grupo de nodos con tres zonas pero solo implementa una única réplica de su pod, su implementación no es resiliente a las zonas.

- **Habilite el escalado automático.** Los grupos de nodos de Kubernetes proporcionan opciones de escalado manuales y automáticas. Con el escalado manual, puedes agregar o eliminar nodos según sea necesario y los pods pendientes esperan hasta que escales verticalmente un grupo de nodos. El escalado administrado por AKS usa el escalador automático del clúster o el aprovisionamiento automático de nodos (NAP). AKS aumenta o reduce el grupo de nodos según los requisitos del pod y teniendo en cuenta la cuota y capacidad de la SKU de tu suscripción. Este método ayuda a garantizar que los pods se programen en los nodos disponibles en las zonas de disponibilidad.

- **Establece las restricciones de la topología del pod.** Usa restricciones de propagación de topología de pods para controlar cómo se distribuyen los pods entre distintos nodos o zonas. Las restricciones le ayudan a lograr la alta disponibilidad, la resistencia y el uso eficaz de los recursos. Si prefieres propagar pods estrictamente entre zonas, puedes establecer restricciones para forzar un pod en un estado pendiente a fin de mantener el equilibrio de pods entre zonas. Para más información, consulta Restricciones de propagación de topología de pod.

- **Configurar redes con resiliencia zonal.** Si los pods sirven tráfico externo, configure la arquitectura de red del clúster mediante servicios como Azure Application Gateway, Azure Load Balancer o Azure Front Door.

- **Asegúrese de que las dependencias son resistentes a la zona.** La mayoría de las aplicaciones de AKS usan otros servicios para el almacenamiento, la seguridad o las redes. Asegúrese de revisar las recomendaciones de resistencia de zona para esos servicios.

Cost

No hay ningún cargo adicional para habilitar la compatibilidad con la zona de disponibilidad en AKS. Paga por las máquinas virtuales (VM) y otros recursos que implemente en las zonas de disponibilidad.

Configurar soporte de zonas de disponibilidad

- **Cree un nuevo clúster de AKS que tenga compatibilidad con la zona de disponibilidad:** Para configurar la compatibilidad con zonas de disponibilidad, consulte Creación de un clúster de Azure Kubernetes Service (AKS) que use zonas de disponibilidad.
- **Migración:** No se puede habilitar la compatibilidad con zonas de disponibilidad después de crear un clúster. En su lugar, debe crear un nuevo clúster que tenga habilitada la compatibilidad con la zona de disponibilidad y eliminar el clúster existente.
- **Deshabilite la compatibilidad con zonas de disponibilidad:** No se puede deshabilitar la compatibilidad con zonas de disponibilidad después de crear un clúster. En su lugar, debe crear un clúster nuevo que tenga compatibilidad con la zona de disponibilidad deshabilitada y eliminar el clúster existente.

Comportamiento cuando todas las zonas están en buen estado

En esta sección se describe qué esperar cuando los clústeres de AKS están configurados para la compatibilidad con zonas de disponibilidad y todas las zonas de disponibilidad están operativas.

- **Enrutamiento de tráfico entre zonas:** Al implementar un clúster de AKS que usa zonas de disponibilidad, es importante asegurarse de que los componentes de red también son resistentes a la zona. Según los equilibradores de carga y otros componentes de red que use, es posible que tenga que configurar explícitamente los componentes para enrutar el tráfico a los nodos correctos de las zonas correctas y responder a interrupciones de zona. Para más información, consulte Consideraciones de resistencia de zona para AKS.

- **Replicación de datos entre zonas:** Si ejecuta una carga de trabajo sin estado, debe usar servicios administrados de Azure, como bases de datos de Azure, Azure Managed Redis o Azure Storage para almacenar los datos de la aplicación. Puede usar estos servicios para ayudar a garantizar que el tráfico se pueda mover entre nodos y zonas sin riesgo de pérdida de datos ni que afecte a la experiencia del usuario. Puedes usar implementaciones, servicios y sondeos de estado de Kubernetes para administrar pods sin estado y garantizar la distribución uniforme entre zonas.

  Si necesita almacenar el estado dentro del clúster mediante discos de Azure, use el almacenamiento con redundancia de zona de Azure para asegurarse de que los datos se replican en varias zonas de disponibilidad. Para obtener más información, consulte Elección del tipo de disco adecuado en función de las necesidades de la aplicación.

Comportamiento durante un fallo de zona

En esta sección se describe qué esperar cuando se produce una interrupción de zona de disponibilidad mientras los clústeres de AKS están configurados con compatibilidad con zonas de disponibilidad.

- **Detección y respuesta**: cuando se produce una interrupción de zona, el plano de control conmuta automáticamente por error. Si los grupos de nodos usan zonas de disponibilidad y siguen losprocedimientos recomendados de resistencia de zona, puedes esperar que AKS abra nodos y réplicas en las zonas que están operativas. AKS realiza esta tarea automáticamente cuando se usan soluciones administradas como cluster autoscaler o NAP. Sin el escalado automático, los nodos y las réplicas permanecen en estado *Pendiente* y esperan a que la intervención manual escale verticalmente el grupo de nodos.

  AKS también intenta reequilibrar los pods en las zonas correctas. Si decides escalar manualmente el grupo de nodos en un escenario de reducción de zona, los pods pueden dejarse en el estado *Pendiente* cuando no haya nodos disponibles en las zonas correctas. El escalado horizontal en las zonas restantes también está sujeto a la disponibilidad de la cuota y la capacidad de la SKU de máquina virtual que uses.

- **Notificación**: Microsoft no le notifica automáticamente cuando una zona está inactiva. Sin embargo, puede usar Azure Resource Health para supervisar el estado de un recurso

individual y puede configurar alertas de Resource Health para notificarle problemas. También puede usar Azure Service Health para comprender el estado general del servicio, incluidos los errores de zona, y puede configurar alertas de Service Health para notificarle problemas.

También puede usar las métricas de estado del nodo o pod para supervisar el estado de los nodos y pods.

- **Solicitudes activas:** Las solicitudes activas pueden experimentar interrupciones. Algunas solicitudes pueden fallar, y la latencia podría aumentar mientras la carga de trabajo se transfiere a otra zona.
- **Pérdida de datos esperada:** Si almacena el estado dentro del clúster mediante discos de Azure y usa almacenamiento con redundancia de zona, no se espera que un error de zona cause ninguna pérdida de datos.
- **Tiempo de inactividad esperado:** Si configura correctamente la resistencia de zona para el clúster y los pods, no se espera que un error de zona cause tiempo de inactividad para la carga de trabajo de AKS.
- **Redireccionamiento del tráfico:** los equilibradores de carga redirigen las nuevas solicitudes entrantes a pods que se ejecutan en nodos saludables.

Para más información, consulte Consideraciones de resistencia de zona para AKS.

Recuperación de zona

Cuando se recupera la zona de disponibilidad, el comportamiento de la conmutación por recuperación depende del componente:

- **Plano de control:** AKS restaura automáticamente las operaciones del plano de control en todas las zonas de disponibilidad. No se requiere intervención manual.
- **Grupos de nodos y nodos**: inmediatamente después de la conmutación por recuperación, los nodos permanecen en las zonas previamente correctas y no se restauran en la zona recuperada. Pero la próxima vez que se realice una operación de escalado de nodos, como al escalar horizontalmente el grupo de nodos, el grupo de nodos puede crear nodos en la zona recuperada.
- **Pods**: inmediatamente después de la conmutación por recuperación, los pods siguen ejecutándose en los nodos en los que se están ejecutando. Cuando se crean nuevos pods o se vuelven a crear los pods existentes, son aptos para usar nodos en la zona recuperada.

- **Almacenamiento:** cualquier almacenamiento conectado a pods se recupera en función de cómo funciona el almacenamiento con redundancia de zona.

Prueba de fallos de zona

Puede probar la resistencia a los errores de zona de disponibilidad mediante los métodos siguientes:

- Aislar y drenar nodos en una única zona de disponibilidad
- Simulación de un error de zona de disponibilidad mediante Azure Chaos Studio

Resistencia a errores en toda la región

Los clústeres de AKS son recursos de una sola región. Si la región no está disponible, el clúster de AKS tampoco está disponible.

Soluciones personalizadas de varias regiones para la resistencia

Si necesita implementar la carga de trabajo de Kubernetes en varias regiones de Azure, tiene dos opciones para administrar la orquestación de estos clústeres.

- Use Fleet Manager de Azure Kubernetes si desea una experiencia administrada más sencilla. Con Fleet Manager de Azure Kubernetes, puede hacer lo siguiente:
  - Administre un conjunto de clústeres de AKS como una sola unidad y esos clústeres se pueden distribuir entre varias regiones de Azure.
  - Automatice aspectos específicos de la administración de clústeres, como las actualizaciones de imágenes de clúster y nodo.
  - Utilice las capacidades de distribución de tráfico para repartir el tráfico entre los clústeres y realizar una conmutación automática si una región no está disponible.
- Organice la conmutación por error mediante un modelo de implementación activo-activo o activo-pasivo manual si la carga de trabajo requiere un control más matizado sobre los distintos componentes de las conmutaciones por error interregionales. Para obtener más información, consulte Información general sobre alta disponibilidad y recuperación ante desastres para AKS.

Copias de seguridad y restauración

Azure Backup tiene una extensión que puede usar para realizar copias de seguridad de los recursos del clúster de AKS y volúmenes persistentes que se conectan al clúster. El almacén de Backup se comunica con el clúster de AKS a través de la extensión para realizar operaciones de copia de seguridad y restauración.

Si el clúster de AKS está en una región emparejada, puede configurar las copias de seguridad que se almacenarán en el almacenamiento con redundancia geográfica. Puede restaurar copias de seguridad georedundantes en la región emparejada.

Para obtener más información, consulte los artículos siguientes:

- ¿Qué es la copia de seguridad de Azure Kubernetes Service?
- Copia de seguridad de AKS mediante Azure Backup

Para la mayoría de las soluciones, no debe confiar exclusivamente en copias de seguridad. En su lugar, utilice las otras capacidades descritas en esta guía para apoyar los requisitos de resiliencia. Sin embargo, las copias de seguridad protegen contra algunos riesgos que otros enfoques no. Para más información, consulte ¿Qué son la redundancia, la replicación y la copia de seguridad?.

Intente usar clústeres sin estado que minimicen la necesidad de copia de seguridad. Almacene datos en bases de datos y sistemas de almacenamiento externos en lugar de en el clúster.

Resistencia al mantenimiento del servicio

AKS realiza el mantenimiento en el clúster, incluidas las actualizaciones del clúster y las imágenes de nodo. Para asegurarse de que Kubernetes mantiene el número mínimo de instancias de pod necesarias para atender el tráfico de producción incluso durante las actualizaciones, debe configurar los pods para que usen presupuestos de interrupciones del pod.

Para reducir las interrupciones del servicio durante períodos de tiempo críticos, AKS proporciona controles para que pueda especificar los tiempos de mantenimiento planeados. Para más información, consulte Uso del mantenimiento planeado para programar y controlar las actualizaciones del clúster de Azure Kubernetes Service.

Acuerdo de nivel de servicio

El contrato de nivel de servicio (SLA) para los servicios de Azure describe la disponibilidad esperada de cada servicio y las condiciones que la solución deberá cumplir para lograr esa expectativa de disponibilidad. Para obtener más información, consulte Acuerdos de Nivel de Servicio para servicios en línea.

AKS proporciona tres planes de tarifa para la administración de clústeres: **Gratis**, **Estándar** y **Premium**. El nivel Gratis permite usar AKS para probar las cargas de trabajo. Los niveles Estándar y Premium están diseñados para cargas de trabajo de producción. Al implementar un clúster de AKS con zonas de disponibilidad habilitadas, aumenta el porcentaje de tiempo de actividad definido en el Acuerdo de Nivel de Servicio. Sin embargo, el Acuerdo de Nivel de Servicio solo se aplica si implementa un clúster en el plan de tarifa Estándar o Premium.

Contenido relacionado

- Confiabilidad en Azure

Last updated on 22/01/2026

Confiabilidad en Azure Load Balancer

Azure Load Balancer es un servicio de equilibrio de carga de nivel 4 para el tráfico del Protocolo de control de transmisión (TCP) y el protocolo de datagramas de usuario (UDP) que distribuye las solicitudes entrantes entre las instancias correctas de los servicios. Load Balancer proporciona alta disponibilidad y rendimiento de red de latencia ultra baja.

Cuando se usa Azure, la confiabilidad es una responsabilidad compartida. Microsoft proporciona una variedad de funcionalidades para admitir resistencia y recuperación. Es responsable de comprender cómo funcionan esas funcionalidades dentro de todos los servicios que usa y de seleccionar las funcionalidades que necesita para cumplir los objetivos empresariales y los objetivos de tiempo de actividad.

En este artículo se describe cómo hacer que Load Balancer sea resistente a una variedad de posibles interrupciones y problemas, incluidos errores transitorios, interrupciones de zona de disponibilidad y interrupciones de regiones. También resalta información clave sobre el acuerdo de nivel de servicio (SLA) de Load Balancer.

> **ⓘ Importante**
>
> La confiabilidad de la solución general depende de la configuración de las instancias de back-end (servidores), como máquinas virtuales (VM) de Azure o conjuntos de escalado de máquinas virtuales a los que el equilibrador de carga enruta el tráfico.
>
> En este artículo no se tratan las instancias de back-end, pero sus configuraciones de disponibilidad afectan directamente a la resistencia de la aplicación. Revise las **guías de confiabilidad de los servicios de Azure de la solución** para obtener información sobre cómo cada servicio admite sus requisitos de confiabilidad. Al configurar las instancias de back-end para alta disponibilidad y redundancia de zona, puede lograr una confiabilidad completa para la aplicación.

Recomendaciones de implementación de producción

Azure Well-Architected Framework proporciona recomendaciones sobre confiabilidad, seguridad, costo, operaciones y rendimiento. Para obtener información sobre cómo estas áreas influyen entre sí y contribuyen a una solución confiable de Load Balancer, consulte Procedimientos recomendados de arquitectura para Load Balancer.

Introducción a la arquitectura de confiabilidad

Un equilibrador de carga puede ser público o interno. Se puede acceder a un equilibrador de carga público desde Internet a través de un recurso de dirección IP pública. Un equilibrador de carga interno solo es accesible desde dentro de la red virtual y otras redes que se conectan a la red virtual.

Cada equilibrador de carga consta de varios componentes:

- *Configuraciones ip de front-end* que reciben tráfico. Un equilibrador de carga público recibe tráfico en una dirección IP pública. Un equilibrador de carga interno recibe tráfico en una dirección IP dentro de la red virtual.
- *Grupos de back-end* que contienen una colección de *instancias de back-end* que pueden recibir tráfico, como máquinas virtuales individuales que ejecutan la aplicación.
- *Reglas del equilibrador de carga* que definen cómo el equilibrador de carga distribuye el tráfico desde un front-end a un grupo de instancias en el back-end.
- *Sondas de estado* que supervisan la disponibilidad de las instancias de servidor del back-end.

Para obtener más información, consulte Componentes de Load Balancer.

En el caso de las soluciones implementadas globalmente, puede implementar un *equilibrador de carga global*, que es un tipo único de equilibrador de carga público para enrutar el tráfico entre diferentes implementaciones regionales de la solución. Un equilibrador de carga global proporciona una única dirección IP de anycast y enruta el tráfico al equilibrador de carga regional más cercano basándose en la proximidad del cliente y el estado de salud regional. Para obtener más información, consulte Resilience to region-wide failures (Resistencia a errores en toda la región).

Resistencia a errores transitorios

Los errores transitorios son errores breves e intermitentes en los componentes. Se producen con frecuencia en un entorno distribuido como la nube y son una parte normal de las operaciones. Los errores transitorios se corrigen después de un breve período de tiempo. Es importante que las aplicaciones puedan controlar errores transitorios, normalmente mediante el reintento de solicitudes afectadas.

Todas las aplicaciones hospedadas en la nube deben seguir las instrucciones de control de errores transitorios de Azure cuando se comunican con cualquier API, bases de datos y otros componentes hospedados en la nube. Para obtener más información, consulte Recomendaciones para controlar errores transitorios.

Al usar Load Balancer, tenga en cuenta los procedimientos recomendados siguientes para minimizar el riesgo de errores transitorios que afectan a la aplicación:

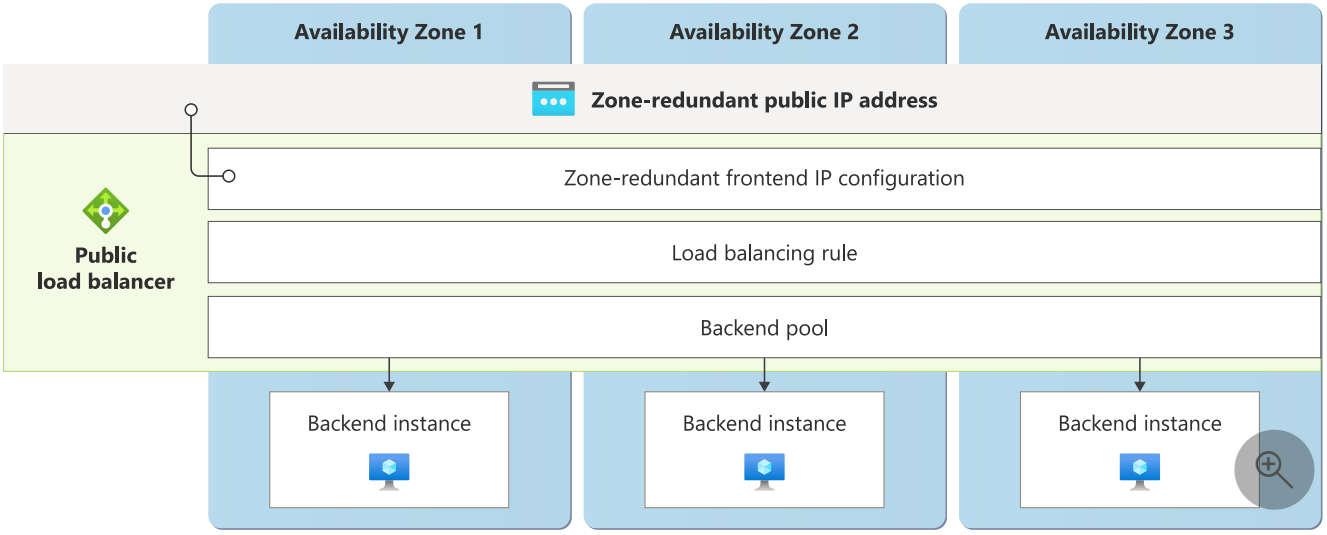
- **Implementación de la lógica de reintento.** Los clientes deben implementar mecanismos de reintento adecuados para errores de conexión transitorios, que incluyen estrategias de retroceso exponenciales.
- **Configure sondeos de estado con tolerancia.** Configure los sondeos de estado para equilibrar la detección rápida de errores con la necesidad de evitar falsos positivos durante problemas transitorios.
- **Supervisar la asignación de puertos SNAT.** Para las conexiones salientes, supervise la asignación de puertos de traducción de direcciones de red de origen (SNAT) y configure reglas de salida para evitar errores de conexión transitorios que se producen debido al agotamiento de puertos.

Resistencia a errores de zona de disponibilidad

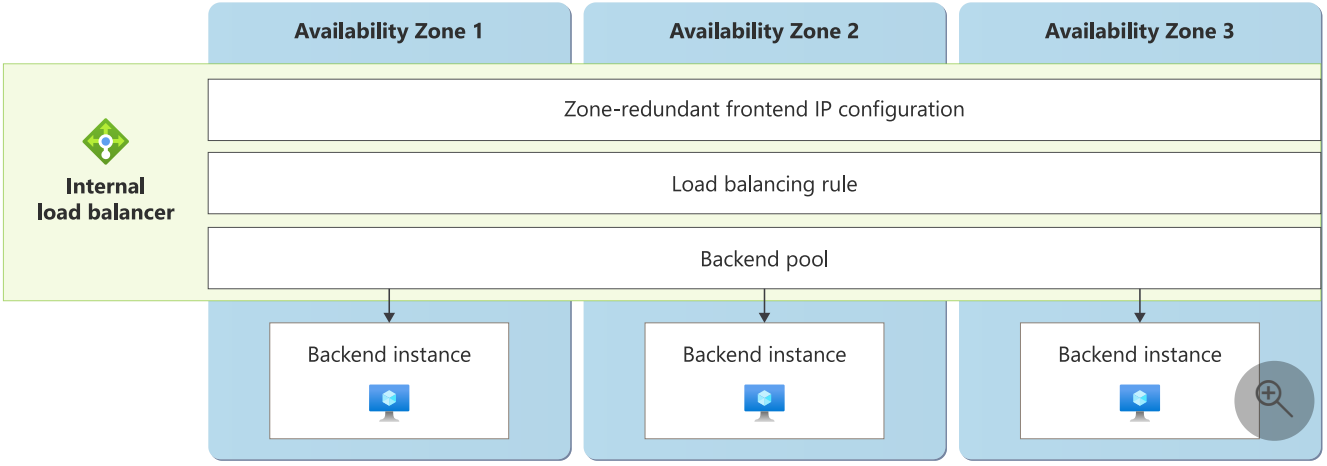
Las zonas de disponibilidad son grupos físicamente independientes de centros de datos dentro de una región de Azure. Cuando una zona falla, los servicios pueden transferirse a una de las zonas restantes.

Puede implementar el Load Balancer como *redundante por zona* configurando cada configuración de IP de front-end que usted cree. Una configuración de IP de front-end con redundancia de zona usa una infraestructura independiente en varias zonas para atender el tráfico simultáneamente. Esta configuración garantiza que los errores de zona no afecten a la capacidad del equilibrador de carga para recibir y distribuir el tráfico.

En el diagrama siguiente se muestra un equilibrador de carga público con redundancia de zona, que se configura mediante la creación de una dirección IP pública con redundancia de zona.



En el diagrama siguiente se muestra un equilibrador de carga interno que usa una configuración similar con redundancia de zona.



> **Nota:**
>
> Puede implementar equilibradores de carga zonales, pero se recomienda usar equilibradores de carga con redundancia de zona para todas las cargas de trabajo, incluidas las cargas de trabajo que implemente en una sola zona.

En regiones sin zonas de disponibilidad, Azure crea todos los equilibradores de carga en una configuración *regional* o no *zonal* mediante una configuración de front-end sin zona configurada. Si la región experimenta una interrupción, los equilibradores de carga nozonales pueden experimentar tiempo de inactividad. Las direcciones IP públicas creadas en una región que se actualice más adelante para tener zonas de disponibilidad serán redundantes a nivel de zona una vez que la región esté en estado de disponibilidad general con varias zonas de disponibilidad.

Instancias de back-end y zonas de disponibilidad

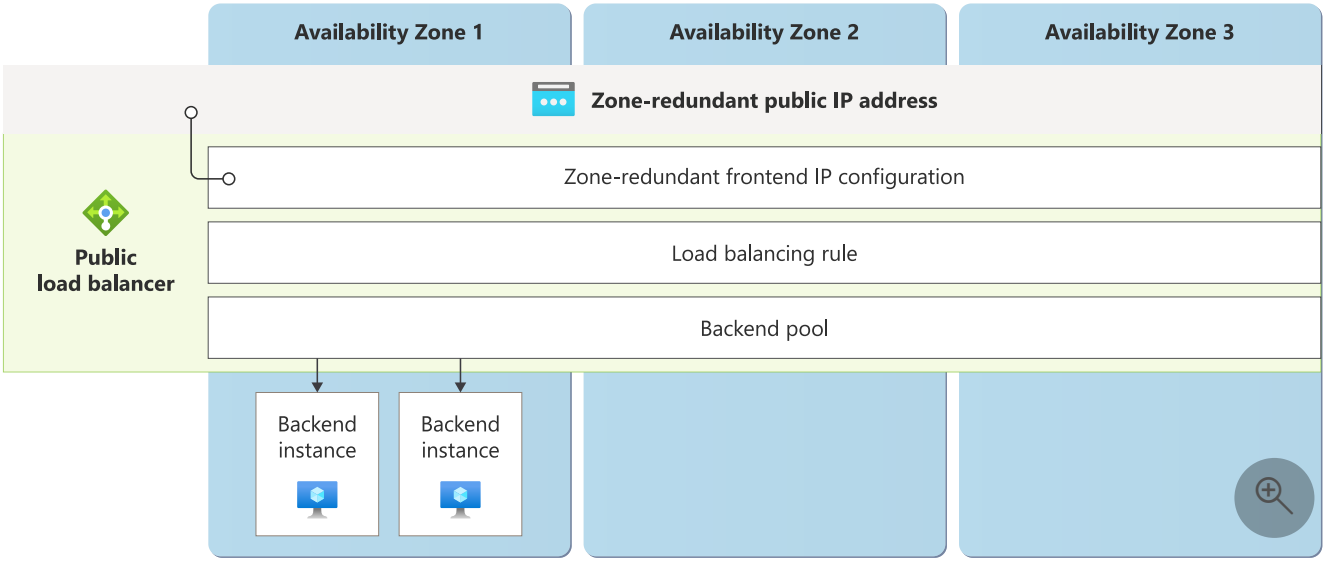
La configuración de zona de disponibilidad de las instancias de back-end funciona independientemente de la configuración de IP de front-end del equilibrador de carga.

Para distribuir las instancias de back-end entre zonas, configure el servicio pertinente según sus características de confiabilidad y sus requisitos de arquitectura.

> **Nota:**
>
> Para lograr resiliencia, distribuya las instancias de back-end entre varias zonas de disponibilidad. Si todas las instancias de back-end se ejecutan en una sola zona, una interrupción en esa zona hace que la aplicación no esté disponible, incluso si usa un equilibrador de carga con redundancia de zona.

Por ejemplo, cuando se usan máquinas virtuales, un enfoque de diseño común para cargas de trabajo de producción consiste en colocar varias máquinas virtuales zonales en las zonas 1, 2 y 3 para lograr resistencia de zona. Para realizar el equilibrio de carga, puede crear un equilibrador de carga redundante por zonas y configurar esas máquinas virtuales como instancias de back-end para el equilibrador. Los sondeos de estado del balanceador de carga eliminan automáticamente las máquinas virtuales no saludables de la rotación independientemente de su ubicación de zona.

Si implementa las máquinas virtuales en la misma zona de disponibilidad, puede implementar una configuración ip de front-end con redundancia de zona en el equilibrador de carga, como se muestra en el diagrama siguiente.



Requisitos

Compatibilidad con regiones: Puede implementar equilibradores de carga con redundancia de zona en cualquier región que admita zonas de disponibilidad.

Cost

La configuración de la zona de disponibilidad no cambia la forma en que Azure factura un equilibrador de carga. Azure basa la facturación en el número de reglas que configure y los datos que procesa, independientemente de la configuración de zona. Para más información, consulte los precios de Azure Load Balancer.

Configurar soporte de zonas de disponibilidad

Al utilizar un Load Balancer, se activa la compatibilidad de zona de disponibilidad en la configuración del IP de front-end.

- **Cree un nuevo equilibrador de carga con soporte para zonas de disponibilidad.**
  - En el caso de *los equilibradores de carga públicos*, la configuración de IP de front-end adopta automáticamente la configuración de zona de disponibilidad del recurso de dirección IP pública que se asocia a él. Para que la zona de configuración de IP de front-end sea redundante, cree o elija una dirección IP pública con redundancia de zona. Azure configura las direcciones IP públicas como redundantes de zona de forma predeterminada. Para más información, consulte Creación de un equilibrador de carga público para equilibrar la carga de máquinas virtuales mediante Azure Portal.
  - En el caso de *los equilibradores de carga internos*, al configurar la dirección IP de front-end del equilibrador de carga, se establece el tipo de compatibilidad de zona de disponibilidad en la configuración ip de front-end. Para más información, consulte Creación de un equilibrador de carga interno para equilibrar la carga de las máquinas virtuales mediante Azure Portal.
- **Cambie la configuración de zona de disponibilidad de un equilibrador de carga existente.** Para cambiar la configuración de zona de disponibilidad de un equilibrador de carga existente, reemplace la configuración de IP de front-end. Puede usar este enfoque para pasar de una configuración zonal a una configuración ip de front-end con redundancia de zona:

  1. Cree una nueva configuración de IP de front-end que tenga la configuración de zona de disponibilidad deseada.

     En el *caso de los equilibradores de carga públicos*, cree una nueva dirección IP pública que use la configuración de zona de disponibilidad deseada. A continuación, vuelva a configurar el equilibrador de carga para agregar una configuración ip de front-end que haga referencia a esa dirección IP pública.

En el *caso de los equilibradores de carga internos*, vuelva a configurar el equilibrador de carga para agregar una nueva configuración ip de front-end con la configuración de zona de disponibilidad deseada. Este paso asigna una nueva dirección IP privada desde la subred.

2. Vuelva a configurar las reglas de equilibrio de carga para usar la nueva configuración de IP de front-end.

   > **Importante**
   >
   > Esta operación requiere que vuelva a configurar los clientes para enviar tráfico a la nueva dirección IP de front-end. En función de los clientes, el proceso puede requerir tiempo de inactividad.

3. Quite la configuración anterior de IP de front-end.

Comportamiento cuando todas las zonas están en buen estado

En esta sección se describe qué esperar cuando un equilibrador de carga usa una configuración ip de front-end con redundancia de zona y todas las zonas de disponibilidad funcionan normalmente.

- **Enrutamiento de tráfico entre zonas:** El equilibrio de carga puede producirse en cualquier zona de disponibilidad. El equilibrador de carga envía tráfico a instancias de back-end saludables que especificáis en el grupo de back-end, sin tener en cuenta la región de disponibilidad que contiene la instancia de back-end.

- **Replicación de datos entre zonas:** Load Balancer funciona como un servicio de paso a través de red que no almacena ni replica los datos de la aplicación. Incluso si configura la persistencia de sesión en el equilibrador de carga, el equilibrador de carga no almacena ningún estado. La persistencia de la sesión ajusta el proceso de hash para dirigir las solicitudes a la misma instancia de back-end, aunque el equilibrador de carga no garantiza dicha persistencia. Cuando cambia el grupo de back-end, el equilibrador de carga vuelve a calcular la distribución de las solicitudes de cliente y no almacena ni sincroniza el estado.

  El servicio mantiene su estado de configuración a través de la replicación sincrónica entre zonas, lo que garantiza la coherencia inmediata de las reglas de equilibrio de carga, las configuraciones de sondeo de estado y la pertenencia al grupo de back-end en todas las zonas.

Comportamiento durante un fallo de zona

En esta sección se describe qué esperar cuando un balanceador de carga usa una configuración IP frontal con redundancia de zona y ocurre una falla en la zona de disponibilidad.

- **Detección y respuesta:** La plataforma Azure es responsable de detectar un error en una zona de disponibilidad y responder. No es necesario hacer nada para iniciar una conmutación por error de la zona.

- **Notificación**: Microsoft no le notifica automáticamente cuando una zona está inactiva. Sin embargo, puede usar Azure Resource Health para supervisar el estado de un recurso individual y puede configurar alertas de Resource Health para notificarle problemas. También puede usar Azure Service Health para comprender el estado general del servicio, incluidos los errores de zona, y puede configurar alertas de Service Health para notificarle problemas.

- **Solicitudes activas:** Cuando se produce un error en una zona, los flujos TCP y UDP existentes dentro de la zona se restablecen automáticamente y los clientes deben reintentar esos flujos. Los clientes deben implementar un control de errores transitorio suficiente, incluidos los reintentos automatizados.

- **Pérdida de datos esperada:** Load Balancer es un servicio de red sin estado, por lo que no almacena los datos de la aplicación y no se produce ninguna pérdida de datos en la capa del equilibrador de carga.

- **Tiempo de inactividad esperado:** No se espera ningún tiempo de inactividad para los equilibradores de carga con redundancia zonal ya que los equilibradores de carga siguen funcionando desde zonas saludables.

  Si el error afecta a los servicios de proceso de la zona, es posible que las máquinas virtuales u otros recursos de la zona afectada no estén disponibles. Los sondeos de estado del equilibrador de carga detectan estos errores y enrutan el tráfico a instancias alternativas en otra zona, según el algoritmo de equilibrio de carga y el estado de salud de las instancias de back-end.

- **Reenrutamiento del tráfico:** El equilibrador de carga sigue funcionando desde las zonas correctas. El Balanceador de Carga mantiene la misma dirección IP de front-end durante fallos de zona. Este comportamiento significa que no es necesario aplicar actualizaciones del sistema de nombres de dominio (DNS) ni volver a configurar clientes. La plataforma Azure establece nuevas conexiones de cliente a través de zonas correctas restantes.

Recuperación de zona

Cuando se recupera una zona de disponibilidad, Load Balancer reanuda automáticamente las operaciones normales. El front-end con redundancia de zona comienza automáticamente a atender el tráfico desde la zona recuperada junto con otras zonas operativas. Los sondeos de estado de la zona recuperada reanudan la evaluación de las instancias de back-end.

Si el error de zona también afecta a los servicios de proceso de esa zona, Load Balancer agrega automáticamente instancias de back-end al grupo de instancias de back-end saludable después de que las instancias se recuperen y pasen las comprobaciones de estado. La distribución del tráfico se rebalancea en todas las zonas disponibles en función del algoritmo de equilibrio de carga y del estado de salud de las instancias de fondo.

Prueba de fallos de zona

La plataforma Azure administra el enrutamiento del tráfico, la respuesta de zona inactiva y la recuperación, por lo que no es necesario iniciar ni validar los procesos de error de zona de disponibilidad.

Puede usar Azure Chaos Studio para simular el error de una máquina virtual en una sola zona. Chaos Studio proporciona errores integrados para las máquinas virtuales, incluido un error que apaga una máquina virtual. Puede usar estas funcionalidades para simular los fallos de zona y probar los procesos de failover.

Resistencia a errores en toda la región

Los equilibradores de carga públicos e internos se implementan en una sola región de Azure. Si la región deja de estar disponible, los equilibradores de carga de esa región también no estarán disponibles. Load Balancer proporciona compatibilidad nativa con varias regiones a través de un equilibrador de carga global, que admite el equilibrio de carga entre regiones de Azure. Además, puede desplegar otros servicios de balanceo de carga para enrutar y realizar conmutación por error entre las regiones de Azure.

Equilibradores de carga globales

Los equilibradores de carga globales proporcionan una única dirección IP anycast estática que automáticamente enruta el tráfico a la implementación regional óptima, basada en la proximidad del cliente y la salud regional. Los equilibradores de carga globales mejoran la confiabilidad y el rendimiento de la aplicación.

Con los equilibradores de carga globales, se implementan varios equilibradores de carga públicos en distintas regiones y un equilibrador de carga global actúa como front-end global. Si los servidores back-end de una región tienen un problema, el tráfico cambia

automáticamente a regiones saludables sin cambios de DNS porque la dirección IP anycast permanece constante y enruta el tráfico a otra región.

Para más información, consulte Equilibrador de carga global.

Soluciones personalizadas de varias regiones para la resistencia

Azure proporciona una gama de servicios de equilibrio de carga que se ajustan a diferentes requisitos. Elija un equilibrador de carga que cumpla los requisitos de resistencia y se adapte al tipo de aplicación. Para obtener más información, consulte Opciones de equilibrio de carga.

Acuerdo de nivel de servicio

El contrato de nivel de servicio (SLA) para los servicios de Azure describe la disponibilidad esperada de cada servicio y las condiciones que la solución deberá cumplir para lograr esa expectativa de disponibilidad. Para obtener más información, consulte Acuerdos de Nivel de Servicio para servicios en línea.

El Acuerdo de Nivel de Servicio de Load Balancer se aplica cuando tiene al menos dos máquinas virtuales correctas configuradas como instancias de back-end. El Acuerdo de Nivel de Servicio excluye el tiempo de inactividad debido al agotamiento de puertos SNAT o a los grupos de seguridad de red (NSG) que bloquean el tráfico.

Contenido relacionado

- Documentación de Load Balancer
- Introducción a la confiabilidad de Azure

ⓘ **Nota:** El autor ha creado este artículo con ayuda de inteligencia artificial. Más información

---

Last updated on 20/03/2026

Confiabilidad en Azure Service Bus

Azure Service Bus es un servicio de agente de mensajes empresariales totalmente administrado que proporciona funcionalidades de mensajería asincrónica confiables para desacoplar aplicaciones y servicios. Service Bus admite colas para la comunicación de punto a punto y temas con suscripciones para patrones de mensajería de publicación y suscripción. El servicio proporciona características de confiabilidad integradas, incluida la durabilidad del mensaje, garantías de entrega al menos una vez y colas de mensajes fallidos para controlar el procesamiento de mensajes erróneos.

Cuando se usa Azure, la confiabilidad es una responsabilidad compartida. Microsoft proporciona una variedad de funcionalidades para admitir resistencia y recuperación. Es responsable de comprender cómo funcionan esas funcionalidades dentro de todos los servicios que usa y de seleccionar las funcionalidades que necesita para cumplir los objetivos empresariales y los objetivos de tiempo de actividad.

En este artículo se describe cómo hacer que Service Bus sea resistente a diferentes posibles interrupciones y problemas, incluidos errores transitorios, interrupciones de zona de disponibilidad y interrupciones de región. También resalta cierta información clave sobre el acuerdo de nivel de servicio (SLA) de Service Bus.

Recomendaciones de implementación de producción

Azure Well-Architected Framework proporciona recomendaciones sobre confiabilidad, rendimiento, seguridad, costo y operaciones. Para obtener información sobre cómo estas áreas influyen entre sí y contribuyen a una solución confiable de Azure Service Bus, consulte Procedimientos recomendados de arquitectura para Service Bus.

Introducción a la arquitectura de confiabilidad

En esta sección se describen algunos de los aspectos importantes de cómo funciona el servicio que es más relevante desde una perspectiva de confiabilidad. En la sección se presenta la arquitectura lógica, que incluye algunos de los recursos y características que se implementan y usan. También se describe la arquitectura física, que proporciona detalles sobre cómo funciona el servicio en segundo plano.

Arquitectura lógica

Un *espacio de nombres* actúa como contenedor de administración para Service Bus y se puede configurar para usar el nivel Básico, Estándar o Premium. Para configurar el servicio a nivel de espacio de nombres, asigne capacidad, configure la seguridad de red y habilite la replicación geográfica y la recuperación ante desastres geográficos.

Dentro de un espacio de nombres, se implementan *colas* y *temas*, que son entidades de mensajería que tienen una semántica diferente. Para más información, consulte Colas, temas y suscripciones de Service Bus.

Puede configurar opcionalmente *particiones* en el espacio de nombres, las cuales distribuyen colas y tópicos en varios agentes de mensajes y almacenes de mensajería. Un espacio de nombres puede usar varias particiones para controlar el procesamiento paralelo y el escalado horizontal. Service Bus solo garantiza la ordenación dentro de una sola partición. La creación de particiones desempeña un papel clave en el diseño de confiabilidad de la aplicación. Al diseñar la aplicación, realice un equilibrio entre maximizar la disponibilidad y la coherencia. Para el nivel Premium, habilite la creación de particiones en el espacio de nombres. Para los espacios de nombres de nivel Básico y Estándar, configure particiones en la entidad y opcionalmente cuando envíe mensajes.

Puede usar Service Bus y sus enfoques de diseño asincrónicos para aumentar la disponibilidad de las aplicaciones. Para más información, consulte Patrones de mensajería asincrónica y alta disponibilidad.

Arquitectura física

Service Bus proporciona los recursos de proceso y almacenamiento subyacentes. Para cada espacio de nombres, varios *agentes de mensajes* procesan los mensajes y varios *almacenes de mensajería* almacenan los mensajes. Cada almacén de mensajería tiene tres copias: una copia principal y dos copias secundarias. Service Bus mantiene las tres copias sincronizadas para las operaciones de datos y administración. Si se produce un error en la copia principal, Service Bus promueve una copia secundaria a principal sin tiempo de inactividad para los clientes.

En el caso de los espacios de nombres que usan el nivel Básico o Estándar, Service Bus proporciona redundancia a través de una infraestructura multiinquilino compartida que replica automáticamente los mensajes entre zonas de disponibilidad donde estén disponibles. El servicio mantiene varios almacenes de mensajería y mantiene todas las copias sincronizadas para las operaciones de administración y datos.

En el caso de los espacios de nombres de nivel Premium, Service Bus proporciona unidades de mensajería dedicadas (RU) que cada una tiene recursos dedicados de CPU y memoria. Los espacios de nombres de nivel Premium se pueden escalar automáticamente en función de las

demandas de carga de trabajo. Para obtener más información, consulte Actualización automática de las unidades de mensajería de un espacio de nombres de Service Bus.

La infraestructura de Service Bus abarca varias máquinas físicas y bastidores que se distribuyen entre dominios de error, lo que reduce el riesgo de fallos catastróficos que afectan al espacio de nombres. En las regiones que tienen zonas de disponibilidad, la infraestructura se extiende entre centros de datos físicos independientes. El servicio implementa mecanismos transparentes de detección de errores y conmutación por error para que siga funcionando dentro de los niveles de servicio garantizados y normalmente sin interrupciones perceptibles cuando se produzcan estos errores.

Resistencia a errores transitorios

Los errores transitorios son errores breves e intermitentes en los componentes. Se producen con frecuencia en un entorno distribuido como la nube y son una parte normal de las operaciones. Los errores transitorios se corrigen después de un breve período de tiempo. Es importante que las aplicaciones puedan controlar errores transitorios, normalmente mediante el reintento de solicitudes afectadas.

Todas las aplicaciones hospedadas en la nube deben seguir las instrucciones de control de errores transitorios de Azure cuando se comunican con cualquier API, bases de datos y otros componentes hospedados en la nube. Para obtener más información, consulte Recomendaciones para controlar errores transitorios.

El SDK de Service Bus incluye lógica de reintento automático con retroceso exponencial para las operaciones que producen errores debido a condiciones transitorias, como tiempos de espera de red o falta de disponibilidad temporal del servicio. Cuando las aplicaciones experimentan desconectaciones transitorias de Service Bus, el SDK intenta volver a conectarse automáticamente mediante la directiva de reintento configurada.

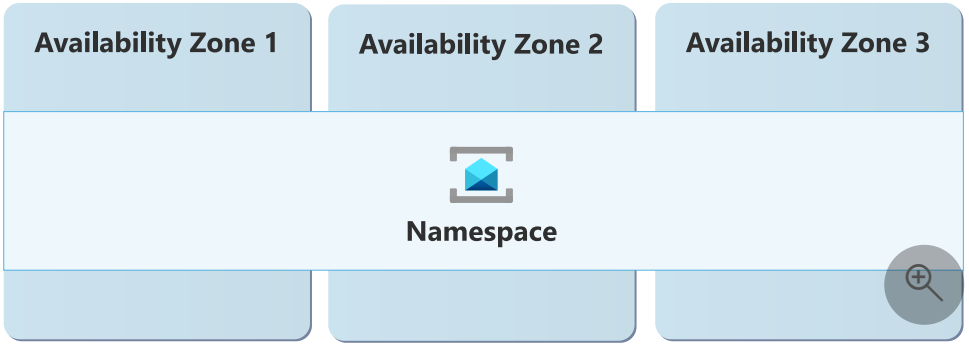
Para optimizar el control de errores transitorios en las aplicaciones, use el SDK de Service Bus más reciente, que incluye las características de administración de conexiones y lógica de reintento más actuales. Para más información, consulte Biblioteca cliente de Service Bus para .NET.

Resistencia a errores de zona de disponibilidad

Las zonas de disponibilidad son grupos físicamente independientes de centros de datos dentro de una región de Azure. Cuando una zona falla, los servicios pueden transferirse a una de las zonas restantes.

Service Bus admite implementaciones con redundancia de zona en todos los niveles de servicio. Al crear un espacio de nombres de Service Bus en una región admitida, la redundancia de zona se habilita automáticamente sin costo adicional. El modelo de implementación con redundancia de zona se aplica a todas las características de Service Bus, incluidas las particiones y las sesiones.

Service Bus replica de forma transparente los datos de configuración, metadatos y mensajes en varias zonas de disponibilidad de la región. La redundancia de zona proporciona una conmutación automática por error sin su intervención. Todos los componentes de Service Bus, incluidos los procesos, las redes y el almacenamiento, se replican entre zonas. Service Bus tiene suficientes reservas de capacidad para controlar instantáneamente la pérdida completa de una zona. Sigue funcionando sin pérdida de datos ni interrupción en las aplicaciones de mensajería, incluso si una zona de disponibilidad completa deja de estar disponible.



Requisitos

- **Compatibilidad con regiones:** Puede implementar espacios de nombres de Service Bus con redundancia de zona en regiones de Azure que admiten zonas de disponibilidad. Service Bus habilita automáticamente la compatibilidad con la zona de disponibilidad al crear un espacio de nombres en una región admitida, sin que se requiera ninguna configuración adicional.

- **Niveles:** Todos los niveles de Service Bus (Básico, Estándar y Premium) admiten zonas de disponibilidad sin requisitos adicionales.

Consideraciones

Los espacios de nombres de Service Bus incluyen una propiedad `zoneRedundant`. Anteriormente, esta propiedad era necesaria para habilitar zonas de disponibilidad, pero este requisito ha cambiado y la `zoneRedundant` propiedad está en desuso. Esta propiedad puede aparecer como `false` incluso cuando está habilitada la redundancia de zona. Todos los espacios de nombres en regiones que tienen zonas de disponibilidad son redundantes a nivel de zona.

Cost

La redundancia de zona en Service Bus no agrega ningún costo adicional.

Configurar soporte de zonas de disponibilidad

Los espacios de nombres de Service Bus admiten automáticamente la redundancia de zona al implementarlos en regiones admitidas. No es necesario realizar ninguna otra configuración.

Comportamiento cuando todas las zonas están en buen estado

En esta sección se describe qué esperar cuando los espacios de nombres de Service Bus están configurados para la redundancia de zona y todas las zonas de disponibilidad están operativas.

- **Enrutamiento de tráfico entre zonas:** Service Bus usa un modelo activo-activo en el que los mensajes se distribuyen entre varias zonas de disponibilidad. Las conexiones de cliente se equilibran automáticamente entre zonas y el servicio enruta las operaciones a la infraestructura de mensajería disponible independientemente de la zona.

- **Replicación de datos entre zonas:** Service Bus usa la replicación sincrónica entre zonas de disponibilidad, incluidos los metadatos y los datos de mensajes. Varias copias del almacén de mensajería deben confirmar las operaciones de escritura antes de que el servicio las considere completadas, lo que garantiza la coherencia de los datos entre zonas durante las operaciones normales.

Comportamiento durante un fallo de zona

En esta sección se describe qué esperar cuando los espacios de nombres de Service Bus están configurados para redundancia zonal y ocurre una interrupción en una zona de disponibilidad.

- **Detección y respuesta:** Microsoft detecta automáticamente errores de zona e inicia la conmutación de emergencia a zonas saludables. No se requiere ninguna acción del cliente para la conmutación por error de nivel de zona.

- **Notificación:** Microsoft no le notifica automáticamente cuando una zona está inactiva. Sin embargo, puede usar Azure Service Health para comprender el estado general del servicio, incluidos los errores de zona, y puede configurar alertas de Service Health para notificarle problemas.

- **Solicitudes activas:** Durante un error de zona, Service Bus podría quitar las solicitudes activas. Si los clientes controlan correctamente los errores transitorios mediante el

reintento después de un breve período de tiempo, normalmente evitan un impacto significativo.

- **Pérdida de datos esperada:** No se produce ninguna pérdida de datos durante un error de zona porque Service Bus replica de forma sincrónica mensajes entre zonas antes de la confirmación.
- **Tiempo de inactividad esperado:** Un error de zona puede provocar unos segundos de tiempo de inactividad. Si los clientes controlan correctamente los errores transitorios mediante el reintento después de un breve período de tiempo, normalmente evitan un impacto significativo.
- **Reenrutamiento del tráfico:** Service Bus detecta la pérdida de la zona y redirige automáticamente las nuevas solicitudes a otra réplica en una de las zonas de disponibilidad correctas.

  El SDK de Service Bus normalmente controla la administración de conexiones y la lógica de reintento de forma transparente.

Recuperación de zona

Cuando se recupera una zona de disponibilidad, Service Bus reintegra automáticamente la zona en la topología activa del servicio. A continuación, la zona recuperada acepta nuevas conexiones y procesa mensajes junto con las demás zonas. Los datos que se replican en zonas supervivientes durante la interrupción permanecen intactos y la replicación sincrónica normal se reanuda en todas las zonas. No es necesario tomar medidas para la recuperación o la reintegración de zonas.

Prueba de fallos de zona

Service Bus administra el enrutamiento del tráfico, la conmutación por error y la recuperación de zona para los fallos de zona, por lo que no es necesario validar los procesos de fallo de zona de disponibilidad ni proporcionar datos adicionales.

Resistencia a errores en toda la región

Service Bus ofrece dos tipos de compatibilidad con varias regiones, y ambos requieren espacios de nombres de nivel Premium.

- La replicación geográfica proporciona replicación activa-pasiva de metadatos y datos de mensajes entre una región primaria y una región secundaria. Use Geo-Replication para la

mayoría de las aplicaciones que deben permanecer resistentes a las interrupciones de regiones y tener baja tolerancia a la pérdida de datos de mensajes.

- Metadata Geo-Disaster Recovery proporciona la replicación activa-pasiva de la configuración y los metadatos entre una región primaria y secundaria, pero no replica los datos del mensaje. Considere usar Recuperación ante Desastres Geográficos para aplicaciones que controlan su propia replicación de datos o no necesitan replicación de datos.

Tanto la replicación geográfica como la recuperación ante desastres geográfica de metadatos requieren que inicie manualmente la conmutación por error o la promoción de una región secundaria para que se convierta en la nueva región primaria. Microsoft no inicia automáticamente la conmutación por error ni la promoción, incluso si la región principal deja de funcionar.

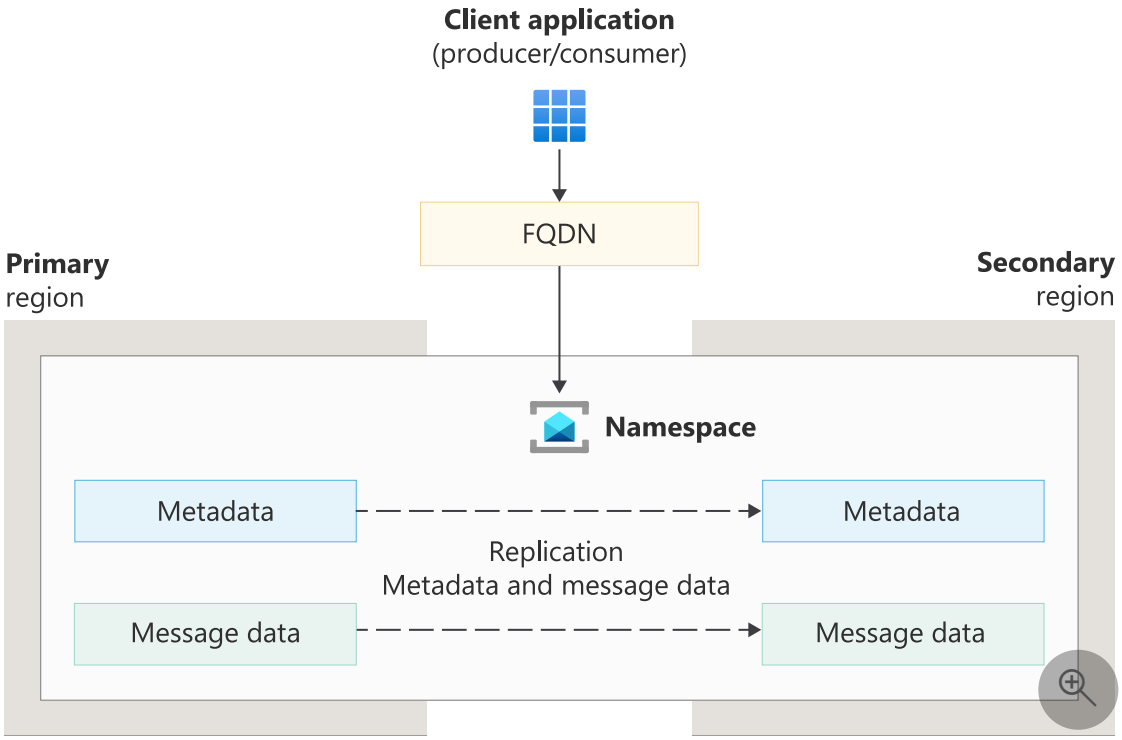
Los espacios de nombres de los niveles Básico y Estándar no incluyen características nativas de varias regiones, pero puede implementar patrones de replicación de nivel de aplicación mediante varios espacios de nombres entre regiones. Para obtener más información, consulte Soluciones personalizadas de varias regiones para lograr resistencia.

Geo-Replication

El nivel Premium admite la replicación geográfica. Esta característica replica metadatos, como entidades, configuración y propiedades, para el espacio de nombres. También replica datos, como los mensajes de las colas y los temas, junto con las propiedades y el estado del mensaje. Configure el enfoque de replicación para la configuración y los datos del espacio de nombres. Esta característica garantiza que los mensajes permanecen disponibles en otra región y le permite cambiar a la región secundaria cuando sea necesario.

Use Geo-Replication para escenarios que requieran resistencia durante interrupciones de regiones y que tengan una tolerancia baja a la pérdida de datos de mensajes.

El espacio de nombres se extiende esencialmente entre regiones. Una región actúa como principal y las demás regiones sirven como secundaria. La suscripción de Azure muestra un único espacio de nombres.



En cualquier momento, puede *promover* la región secundaria a una región primaria. Al promover la región secundaria, Service Bus redirige el nombre de dominio completamente calificado (FQDN) del espacio de nombres a la región secundaria seleccionada y convierte la región primaria anterior en una región secundaria. Decide si desea realizar una *promoción planeada*, lo que significa que espera a que se complete la replicación de datos o una *promoción forzada*, lo que podría provocar la pérdida de datos.

> ⓘ **Nota:**
>
> Service Bus Geo-Replication usa el término *promoción* porque representa mejor el proceso de promoción de una región secundaria a una región primaria (y, posteriormente, degradación de una región primaria a una región secundaria). El término *conmutación por error* también se emplea para describir este proceso general.

En esta sección se resumen aspectos importantes de la replicación geográfica. Revise la documentación completa para obtener información sobre cómo funciona. Para más información, consulte Replicación geográfica de Service Bus.

Requisitos

- **Compatibilidad con regiones:** Puede elegir cualquier región de Azure que admita Service Bus como región primaria o secundaria. No es necesario usar regiones emparejadas de Azure, por lo que seleccione regiones secundarias en función de los requisitos de latencia, cumplimiento o residencia de datos.

- **Nivel:** Para habilitar la replicación geográfica, el espacio de nombres debe usar el nivel Premium.
- **Recuperación ante Desastres Geográficos de Metadatos:** No se puede configurar un espacio de nombres para usar tanto la Geo-Replicación como la Recuperación ante Desastres Geográficos.

Consideraciones

- **Limitaciones de características:** Al habilitar la replicación geográfica, se aplican algunas restricciones. Para más información, consulte Replicación geográfica de Service Bus.
- **Puntos de conexión privados:** Si usa puntos de conexión privados para conectarse al espacio de nombres, también debe configurar las redes en las regiones primarias y secundarias. Para obtener más información, consulte Puntos de conexión privados.

Cost

Para obtener información sobre cómo funcionan los precios para la replicación geográfica, consulte Precios.

Configuración de la compatibilidad con varias regiones

- **Habilite Geo-Replication en un nuevo espacio de nombres.** Para habilitar Geo-Replication en un espacio de nombres durante la creación, consulte Set up Geo-Replication.
- **Migración de metadatos de Recuperación ante desastres geográfica a replicación geográfica.**Cambie del uso de metadatos de Recuperación ante desastres geográfica a replicación geográfica.
- **Cambie el enfoque de replicación.** Para cambiar entre modos de replicación sincrónicos y asincrónicos, consulte Cambiar el modo de replicación.
- **Desactive la replicación geográfica.** Para desactivar Geo-Replication para una región secundaria, consulte Eliminar región secundaria.

Comportamiento cuando todas las regiones están en buen estado

En esta sección se describe qué esperar cuando se configura un espacio de nombres de Service Bus para Geo-Replication y la región primaria está operativa.

- **Enrutamiento del tráfico entre regiones:** Las aplicaciones cliente se conectan a través del FQDN de su espacio de nombres, y su tráfico se dirige a la región primaria.

  Solo la región primaria procesa activamente los mensajes de los clientes durante las operaciones normales. La región secundaria recibe mensajes replicados, pero por lo demás permanece en modo de espera.

- **Replicación de datos entre regiones:** El comportamiento de replicación de datos entre la región principal y secundaria depende de si el emparejamiento de replicación usa la replicación sincrónica o asincrónica.

  - *Síncrono:* Los mensajes se replican en la región secundaria antes de que se complete la operación de escritura.

    Este modo proporciona la mayor garantía de que los datos del mensaje son seguros porque se deben confirmar en la región primaria y secundaria. Pero la replicación sincrónica aumenta considerablemente la latencia de escritura para los mensajes entrantes. También requiere que la región secundaria esté disponible para aceptar la operación de escritura, por lo que una interrupción en la región secundaria hace que se produzca un error en la operación de escritura.

  - *Asincrónico:* El servicio escribe mensajes en la región primaria y, a continuación, completa la operación de escritura. Poco tiempo después, replica los mensajes en la región secundaria.

    Este modo proporciona un mayor rendimiento de escritura que la replicación sincrónica porque no hay ninguna latencia de replicación entre regiones durante las operaciones de escritura. También puede tolerar la pérdida de la región secundaria, a la vez que permite operaciones de escritura en la región primaria. Pero si se produce una interrupción en la región primaria, es posible que los datos que no se repliquen en la región secundaria no estén disponibles o se pierdan.

    Al configurar la replicación asincrónica, se establece el tiempo de retraso máximo aceptable para la replicación. En cualquier momento, puede comprobar el retraso de replicación actual mediante métricas de Azure Monitor.

    Si el retraso de replicación asincrónica aumenta más allá del máximo que especifique, la región primaria comienza a limitar las peticiones entrantes para que la replicación sea capaz de alcanzar el nivel esperado. Para evitar esta situación, seleccione regiones secundarias que no estén demasiado distantes geográficamente y asegúrese de que la capacidad sea suficiente para el rendimiento.

    Algunos tipos de metadatos se replican de forma sincrónica incluso cuando se selecciona el modo de replicación asincrónica.

Para obtener más información, consulte Modos de replicación.

Comportamiento durante una interrupción regional

En esta sección se describe qué esperar cuando se configura un espacio de nombres de Service Bus para la replicación geográfica y hay una interrupción en la región principal o secundaria.

- **Detección y respuesta:** Usted es responsable de decidir cuándo promover la región secundaria del espacio de nombres para convertirse en la nueva región primaria. Microsoft no toma esta decisión ni inicia el proceso por ti, incluso si hay una interrupción de la región. Para conocer los criterios que se deben tener en cuenta al decidir si se debe realizar una conmutación por error, consulte Escenarios recomendados para desencadenar la promoción.

  Para obtener más información sobre cómo promover una región secundaria a la nueva principal, consulte Flujo de promoción.

  Al promover una región secundaria, elija entre una *promoción planeada* o una *promoción forzada*. Una promoción planeada espera a que la región secundaria se ponga al día antes de aceptar tráfico nuevo. Este enfoque evita la pérdida de datos, pero produce tiempo de inactividad.

  Durante una interrupción en la región primaria, generalmente se necesita realizar una promoción forzada. Si la región primaria está disponible y desencadena una promoción por otro motivo, puede elegir una promoción planeada en su lugar.

- **Notificación:** Microsoft no le notifica automáticamente cuando una región está inactiva. Sin embargo, puede usar Azure Service Health para comprender el estado general del servicio, incluidos los errores de región, y puede configurar alertas de Service Health para notificarle problemas.

- **Solicitudes activas:** El comportamiento depende de si la interrupción de la región se produce en la región primaria o en la región secundaria:

  - *Interrupción de la región primaria:* Si la región primaria no está disponible, se finalizan todas las solicitudes activas. Las aplicaciones cliente deben reintentar las operaciones una vez completada la promoción.

  - *Interrupción de la región secundaria:* Una interrupción en la región secundaria puede provocar problemas para las solicitudes activas en las situaciones siguientes:

    - Si usa el modo de replicación sincrónica, la región primaria no puede completar las operaciones de escritura si alguna región secundaria no está disponible.

- Si utiliza el modo de replicación asincrónica, su espacio de nombres limita el flujo y no acepta nuevos mensajes una vez que el retraso de replicación alcanza el máximo que haya configurado.

  Para seguir usando el espacio de nombres en la región primaria, quite el espacio de nombres secundario de la configuración de Geo-Replication.

- **Pérdida de datos esperada:** La cantidad de pérdida de datos depende de si realiza una promoción planeada o forzada y si el modo de replicación es sincrónico o asincrónico:

  - *Promoción planeada:* No se espera ninguna pérdida de datos. Pero durante una interrupción de la región, es posible que no sea posible una promoción planeada porque requiere que todas las regiones primarias y secundarias estén disponibles.

  - *Promoción forzada, replicación sincrónica:* No se espera ninguna pérdida de datos.

  - *Promoción forzada, replicación asincrónica:* Es posible que experimente alguna pérdida de datos para los mensajes recientes que no se replican en la región secundaria y para los cambios de estado que no se replican. La cantidad depende del retraso de replicación. Para comprobar el retraso de replicación actual, use las métricas de Azure Monitor.

  Si realiza una promoción forzada, no puede recuperar los datos perdidos, incluso después de que la región primaria esté disponible.

- **Tiempo de inactividad esperado:** La cantidad de tiempo de inactividad esperado depende de si realiza una promoción planeada o forzada:

  - *Promoción planeada:* El primer paso de una promoción planeada replica los datos en la región secundaria. Ese proceso suele completarse rápidamente, pero en algunas situaciones, puede tardar hasta la longitud del retraso de replicación. Una vez completada la replicación, el proceso de promoción suele tardar entre 5 y 10 minutos. A veces, los servidores del sistema de nombres de dominio (DNS) pueden tardar más tiempo en actualizar las entradas y replicar completamente sus registros en los clientes.

    La región primaria no admite operaciones de escritura durante el proceso completo de promoción.

    Es posible que esta opción no sea posible durante una interrupción de la región porque requiere que todas las regiones primarias y secundarias estén disponibles.

  - *Promoción forzada:* Durante una promoción forzada, Service Bus no espera a que la replicación de datos se complete e inicie la promoción inmediatamente. El proceso de

promoción suele tardar entre 5 y 10 minutos. A veces, las entradas DNS pueden tardar más tiempo en replicarse y actualizarse por completo en todos los clientes.

La región primaria no admite operaciones de escritura durante el proceso completo de promoción.

- **Reenrutamiento del tráfico:** Una vez completada la promoción, el FQDN del espacio de nombres apunta a la nueva región primaria. Pero este redireccionamiento depende de la rapidez con la que se actualizan los registros DNS de los clientes, incluido si sus servidores DNS respetan el período de vida (TTL) de los registros DNS del espacio de nombres.

Recuperación de regiones

Una vez que la región primaria original se haya recuperado, si desea devolver el espacio de nombres a esa región, siga el mismo proceso de promoción regional.

Si ha realizado una promoción forzada durante la interrupción de la región, no puede recuperar los datos perdidos, incluso después de que la región primaria esté disponible.

Prueba de fallos de región

Para probar la replicación geográfica, promueva temporalmente la región secundaria a principal y compruebe que las aplicaciones cliente pueden cambiar entre regiones con una interrupción mínima.

Supervise la duración de la promoción y compruebe que los runbooks y la automatización funcionan correctamente. Después de realizar las pruebas, puede revertir a la configuración original.

Comprenda el posible tiempo de inactividad y pérdida de datos que puede experimentar durante y después del proceso de promoción. Pruebe la Geo-Replication en un entorno no de producción que reproduzca la configuración de su espacio de nombres de producción.

Recuperación de metadatos ante desastres geográficos

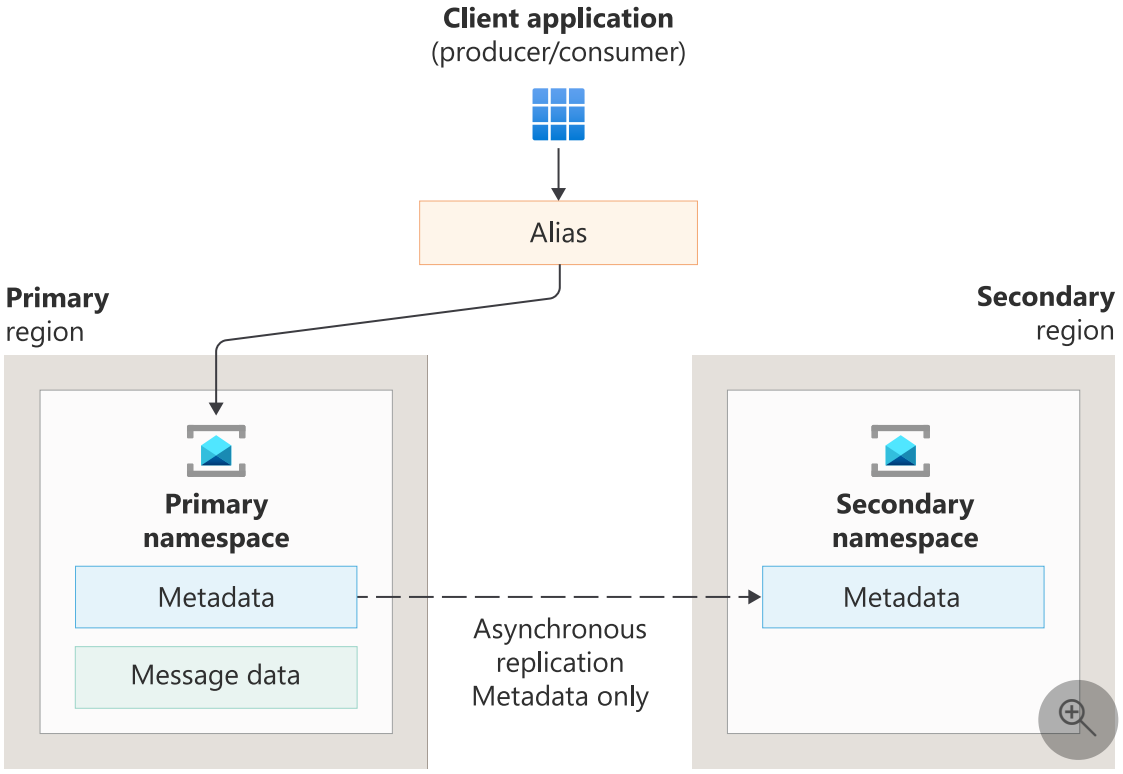
El nivel Premium admite metadatos Geo-Disaster Recovery. Esta característica mejora la recuperación de escenarios de desastres, incluida la pérdida grave de una región. Geo-Disaster Recovery solo replica la configuración y los metadatos del espacio de nombres, y no replica los datos del mensaje. Para admitir la recuperación ante desastres, esta característica garantiza que un espacio de nombres de otra región esté preconfigurado y listo para aceptar inmediatamente mensajes de los clientes. La recuperación ante desastres geográfica sirve

como una solución de recuperación unidireccional y no admite la conmutación por recuperación en la región primaria anterior.

La recuperación de metadatos Geo-Disaster funciona mejor para las aplicaciones que no necesitan mantener todos los mensajes y pueden tolerar cierta pérdida de datos durante un escenario de desastre. La recuperación de metadatos Geo-Disaster Recovery también puede ser adecuada para aplicaciones que replican los datos por sí mismas o que no necesitan replicación de datos en absoluto. Por ejemplo, cuando los mensajes contienen imágenes grandes que posteriormente se convierten en miniaturas, la pérdida de algunos mensajes de una región con errores puede ser aceptable si puede reanudar rápidamente el procesamiento de mensajes nuevos en otra región y reconstruir los mensajes perdidos más adelante.

> **Importante**
>
> Recuperación ante Desastres Geográficos permite la continuidad de las operaciones que tienen la misma configuración, pero **no duplica los datos de los mensajes**. Si debe replicar datos de mensajes, considere la posibilidad de usar la replicación geográfica.

Al configurar metadatos Geo-Disaster Recovery, se crea un *alias* al que se conectan las aplicaciones cliente. El alias es un FQDN que dirige todo el tráfico al espacio de nombres principal de forma predeterminada.



Si se produce un error en la región primaria u otro tipo de desastre, puede iniciar manualmente una conmutación por error unidireccional desde la región primaria a la región

secundaria en cualquier momento. Puede realizar una *conmutación por error segura*, que espera a que se complete la replicación antes de cambiar a la secundaria. Es posible que esta opción no esté disponible durante una interrupción de la región. Después de que se inicie una conmutación por error, se completa casi al instante. Durante el proceso de conmutación por error, el alias de recuperación ante desastres geográfica se redirige al espacio de nombres secundario.

En esta sección se resumen los aspectos importantes de Geo-Disaster Recovery. Revise la documentación completa para obtener información sobre cómo funciona. Para obtener más información, consulte Service Bus Geo-Disaster Recovery.

Requisitos

- **Compatibilidad con regiones:** Puede elegir cualquier región de Azure que admita Service Bus como espacio de nombres principal o secundario. No es necesario usar regiones emparejadas de Azure, por lo que seleccione regiones secundarias en función de los requisitos de latencia, cumplimiento o residencia de datos.
- **Nivel:** Para habilitar los metadatos Geo-Disaster Recovery, ambos espacios de nombres deben usar el nivel Premium.
- **Particionamiento:** No es posible asociar un espacio de nombres particionado con un espacio de nombres no particionado.
- **Recuperación ante Desastres Geográficos de Metadatos:** No se puede configurar un espacio de nombres para usar tanto la Geo-Replicación como la Recuperación ante Desastres Geográficos.

Consideraciones

- **Limitaciones de características:** Al habilitar Geo-Disaster Recuperación, se aplican algunas restricciones. Para obtener más información, vea Puntos importantes que se deben tener en cuenta y Consideraciones.
- **Asignaciones de roles:** las asignaciones de control de acceso basado en roles (RBAC) de Microsoft Entra a entidades en el espacio de nombres principal no se replican en el espacio de nombres secundario. Cree asignaciones de roles manualmente en el espacio de nombres secundario para proteger el acceso a esas entidades.
- **Diseño de aplicaciones:** La Recuperación ante desastres geográficos requiere consideraciones específicas al diseñar sus aplicaciones cliente. Para más información, consulte Consideraciones.

- **Puntos de conexión privados:** Si usa puntos de conexión privados para conectarse al espacio de nombres, configure las redes en la región primaria y secundaria. Para obtener más información, consulte Puntos de conexión privados.
- **Espacios de nombres migrados de Estándar a Premium:** Si tu espacio de nombres está en el nivel Estándar y lo migras al nivel Premium, necesitas gestionar el alias de una manera diferente. Para obtener más información, consulte Service Bus de Standard a Premium.

Cost

Al habilitar la recuperación ante desastres geográfica de metadatos, usted paga por los espacios de nombres principal y secundario.

Configuración de la compatibilidad con varias regiones

- **Cree un emparejamiento de recuperación de desastres geográfico de metadatos.** Para configurar la recuperación ante desastres entre los espacios de nombres principal y secundario, consulte Configuración y proceso de conmutación por error.
- **Desactive los metadatos de Recuperación de Geo-Desastres.** Para interrumpir el emparejamiento entre espacios de nombres, consulte Configuración y flujo de conmutación por error.

Planeamiento y administración de capacidad

Cuando planee implementaciones de varias regiones, asegúrese de que ambas regiones tienen capacidad suficiente para controlar la carga completa si se produce un error en una región. La región secundaria permanece pasiva durante las operaciones normales, pero debe controlar inmediatamente el tráfico después de la conmutación por error. Planee cómo escalar la capacidad del espacio de nombres secundario para que pueda recibir tráfico de producción sin demora. Si puede tolerar un tiempo de inactividad adicional durante el proceso de conmutación por error, puede escalar la capacidad del espacio de nombres secundario durante o después de la conmutación por error. Para reducir el tiempo de inactividad, aprovisione la capacidad con antelación en el espacio de nombres secundario para que permanezca lista para recibir la carga de producción.

Comportamiento cuando todas las regiones están en buen estado

En esta sección se describe qué esperar cuando se configura un espacio de nombres de Service Bus para Geo-Disaster Recovery y la región primaria está operativa.

- **Enrutamiento de tráfico entre regiones:** Las aplicaciones de cliente se conectan a través del alias de Recuperación ante Desastres Geo para el espacio de nombres, y su tráfico se dirige al espacio de nombres primario en la región primaria.

  Solo el espacio de nombres principal procesa activamente los mensajes de los clientes durante las operaciones normales. El espacio de nombres secundario permanece en modo de espera y se produce un error en las solicitudes de acceso a los datos.

- **Replicación de datos entre regiones:** Solo los metadatos de configuración se replican entre los espacios de nombres. La replicación de la configuración se produce de forma continua y asincrónica.

  Todos los datos del mensaje permanecen solo en el espacio de nombres principal y no se replican en el espacio de nombres secundario.

Comportamiento durante una interrupción regional

En esta sección se describe qué esperar cuando se configura un espacio de nombres de Service Bus para Geo-Disaster Recovery y se produce una interrupción en la región primaria.

- **Detección y respuesta:** usted es responsable de monitorear la salud de la región e iniciar la conmutación por error manualmente. Microsoft no inicia la conmutación por error ni promueve automáticamente una región secundaria, incluso si la región primaria está inactiva.

  Para obtener más información sobre cómo iniciar la conmutación por error, consulte el Flujo de conmutación por error.

  Al iniciar una conmutación por error, elija si desea realizar una *conmutación por error segura* o una *conmutación por error estándar*, como una conmutación por error forzada o manual. Una conmutación por error segura espera a que la replicación se complete en la región secundaria antes de que se inicie la conmutación por error. Este enfoque reduce la pérdida de metadatos, pero presenta tiempo de inactividad. La conmutación por error segura requiere que los espacios de nombres estén en la misma suscripción de Azure.

  Durante una interrupción en la región primaria, normalmente debe realizar una conmutación por error forzada. Si la región primaria está disponible y desencadena una conmutación por error por otro motivo, puede elegir una conmutación por error planeada.

  La conmutación por error es una operación unidireccional, por lo que es necesario reestablecer el emparejamiento de Geo-Disaster Recovery más adelante. Para obtener más información, consulte Recuperación de regiones.

- **Notificación:** Microsoft no le notifica automáticamente cuando una región está inactiva. Sin embargo, puede usar Azure Service Health para comprender el estado general del servicio, incluidos los errores de región, y puede configurar alertas de Service Health para notificarle problemas.

- **Solicitudes activas:** Las solicitudes activas terminan cuando se inicia la conmutación por error. Las aplicaciones cliente deben reintentar las operaciones una vez completada la conmutación por error.

- **Pérdida de datos esperada:**

  - *Metadatos:* Normalmente, la configuración y los metadatos se replican en el espacio de nombres secundario. La replicación de metadatos se produce de forma asincrónica, por lo que es posible que los cambios recientes no se repliquen, especialmente los cambios complejos. Compruebe la configuración del espacio de nombres secundario antes de que los clientes accedan a él.

  - *Datos del mensaje:* Los datos del mensaje no se replican entre regiones. Si la región primaria deja de funcionar, los mensajes del espacio de nombres principal no estarán disponibles.

    Los mensajes no se pierden permanentemente a menos que un desastre catastrófico cause una pérdida total de la región primaria. Si la región se recupera, puede recuperar mensajes del espacio de nombres principal más adelante.

- **Tiempo de inactividad esperado:** Normalmente, la conmutación ocurre en un plazo de 5 a 10 minutos. Los clientes pueden tardar más tiempo en replicar y actualizar completamente las entradas DNS.

- **Reenrutamiento del tráfico:** Los clientes que usan el alias de recuperación de Geo-Disaster para conectarse al espacio de nombres se redirigen automáticamente al espacio de nombres secundario después de la conmutación por error. Pero este redireccionamiento depende de los servidores DNS que respetan el TTL de los registros DNS del espacio de nombres y los clientes que reciben esos registros DNS actualizados.

Recuperación de regiones

Una vez que la región primaria original se recupere, debe reestablecer manualmente el emparejamiento y, opcionalmente, realizar una conmutación por recuperación. Cree un nuevo emparejamiento de recuperación de Geo-Disaster con la región recuperada como secundaria y luego realice otra conmutación por error si desea regresar a la región original. Este proceso implica la posible pérdida de datos de los mensajes enviados al espacio de nombres principal temporal.

Si el desastre provoca la pérdida de todas las zonas de la región primaria, los datos podrían ser irrecuperables. En otros escenarios, los datos del mensaje que permanecen en el espacio de nombres principal antes de la conmutación por error son recuperables. Puede obtener mensajes históricos del espacio de nombres principal anterior después de restaurar el acceso. Usted es responsable de configurar las aplicaciones para recibir y procesar esos mensajes. Microsoft no los restaura automáticamente en tu región secundaria.

Prueba de fallos de región

Para probar los procesos de respuesta y recuperación ante desastres, realice una conmutación por error planeada durante una ventana de mantenimiento. Inicie la conmutación por error desde el espacio de nombres principal al espacio de nombres secundario y compruebe que las aplicaciones pueden conectarse y procesar mensajes desde el nuevo espacio de nombres principal.

Supervise la duración de la conmutación por error y compruebe que los runbooks y la automatización funcionan correctamente. Después de realizar las pruebas, puede revertir a la configuración original.

Comprenda el posible tiempo de inactividad y la pérdida de datos que puede experimentar durante y después del proceso de conmutación por error. Pruebe la recuperación ante desastres geográfica de metadatos en un entorno de preproducción que refleje la configuración de su espacio de nombres de producción.

Soluciones personalizadas de varias regiones para la resistencia

La Geo-Replicación y la Recuperación ante Desastres mediante Metadata proporcionan resiliencia frente a interrupciones regionales y otros problemas, y son adecuadas para la mayoría de las cargas de trabajo. Es posible que estas funcionalidades no se adapten a sus necesidades en las situaciones siguientes:

- Tiene requisitos para la replicación personalizada o para mantener varias regiones activas simultáneamente.
- Actualmente usas un nivel de Service Bus que no admite estas características.

Hay varios patrones de diseño que proporcionan distintos tipos de compatibilidad con varias regiones en Service Bus. Muchos de estos patrones requieren que implemente varios espacios de nombres y configure la aplicación para usarlos correctamente. Para obtener más información, consulte los siguientes recursos:

- Proteger aplicaciones de Service Bus contra interrupciones y desastres
- Replicación de mensajes y federación entre regiones

Resistencia al mantenimiento del servicio

Service Bus realiza un mantenimiento normal. Durante el mantenimiento planeado, los espacios de nombres se mueven a un nodo redundante que contiene las actualizaciones más recientes. Durante el traslado, el SDK del cliente se desconecta y, a continuación, se reconecta automáticamente al espacio de nombres. Las actualizaciones normalmente se completan en un plazo de 30 segundos. Es importante que las aplicaciones estén preparadas para las desconexiones de red transitorias que pueden producirse durante períodos de mantenimiento.

Para más información, consulte Eventos de mantenimiento de Azure para Service Bus.

Copias de seguridad y restauración

Service Bus no está diseñado para el almacenamiento de datos a largo plazo. Los datos normalmente se almacenan en un tema o cola durante un breve período de tiempo. Después, se procesa o se conserva en otro sistema de almacenamiento de datos y, a continuación, se elimina. Debido a este diseño, Service Bus mantiene automáticamente réplicas de datos de mensajes, pero no proporciona ninguna funcionalidad de copia de seguridad y restauración para esos datos.

En escenarios que requieren retención de mensajes a largo plazo, considere la posibilidad de implementar el archivado de nivel de aplicación en Azure Storage u otros servicios de almacenamiento duraderos.

Acuerdo de nivel de servicio

El contrato de nivel de servicio (SLA) para los servicios de Azure describe la disponibilidad esperada de cada servicio y las condiciones que la solución deberá cumplir para lograr esa expectativa de disponibilidad. Para obtener más información, consulte Acuerdos de Nivel de Servicio para servicios en línea.

Service Bus proporciona un SLA para todos los espacios de nombres. El SLA de disponibilidad es mayor cuando el espacio de nombres cumple los siguientes criterios:

- Usa el nivel Premium.
- Se encuentra en una región que tiene zonas de disponibilidad.
- Usa particionamiento.

Contenido relacionado

- Introducción a Service Bus
- Recuperación ante desastres geográficos de Service Bus
- Replicación geográfica de Service Bus

Last updated on 27/01/2026

Confiabilidad en Azure SQL Database

Azure SQL Database es un motor de base de datos de plataforma como servicio (PaaS) totalmente administrado que se encarga de la mayoría de las funciones de administración de bases de datos, como actualizar, aplicar revisiones, crear copias de seguridad y supervisar sin intervención del usuario.

Cuando se usa Azure, la confiabilidad es una responsabilidad compartida. Microsoft proporciona una variedad de funcionalidades para admitir resistencia y recuperación. Es responsable de comprender cómo funcionan esas funcionalidades dentro de todos los servicios que usa y de seleccionar las funcionalidades que necesita para cumplir los objetivos empresariales y los objetivos de tiempo de actividad.

En este artículo se describe cómo hacer que Azure SQL Database sea resistente a una variedad de posibles interrupciones y problemas, incluidos errores transitorios, interrupciones de zona de disponibilidad y interrupciones de regiones. También se describe cómo puede usar copias de seguridad para recuperarse de otros tipos de problemas, cómo controlar el mantenimiento del servicio y resaltar cierta información clave sobre el acuerdo de nivel de servicio (SLA) de Azure SQL Database.

Recomendaciones de implementación de producción

Para obtener información sobre cómo implementar Azure SQL Database para admitir los requisitos de confiabilidad de la solución y cómo afecta la confiabilidad a otros aspectos de la arquitectura, consulte Procedimientos recomendados de arquitectura para Azure SQL Database en Azure Well-Architected Framework.

Introducción a la arquitectura de confiabilidad

SQL Database se ejecuta en el motor de base de datos de SQL Server estable más reciente del sistema operativo Windows, incluidas todas las revisiones aplicables.

SQL Database logra redundancia almacenando tres copias de los datos en un único centro de datos de la región primaria de forma predeterminada. Este enfoque protege los datos si se produce un error localizado, como un error de red a pequeña escala o un error de energía, y también durante los siguientes eventos:

- Operaciones de administración iniciadas por el cliente que dan lugar a un breve tiempo de inactividad.
- Operaciones de mantenimiento de servicio
- Problemas y interrupciones del centro de datos, donde el centro de datos tiene los siguientes componentes:
  - Bastidores, donde se ejecutan las máquinas que impulsan el servicio
  - Máquinas físicas que hospedan la máquina virtual (VM) que ejecuta el motor de SQL Database
- Otros problemas con el motor de SQL Database
- Otras posibles interrupciones localizadas no planeadas

SQL Database usa Azure Service Fabric para administrar la replicación de la base de datos.

La redundancia se implementa de maneras diferentes para distintos niveles de servicio de SQL Database. Para obtener más información, consulte Disponibilidad mediante redundancia: SQL Database.

Resistencia a errores transitorios

Los errores transitorios son errores breves e intermitentes en los componentes. Se producen con frecuencia en un entorno distribuido como la nube y son una parte normal de las operaciones. Los errores transitorios se corrigen después de un breve período de tiempo. Es importante que las aplicaciones puedan controlar errores transitorios, normalmente mediante el reintento de solicitudes afectadas.

Todas las aplicaciones hospedadas en la nube deben seguir las instrucciones de control de errores transitorios de Azure cuando se comunican con cualquier API, bases de datos y otros componentes hospedados en la nube. Para obtener más información, consulte Recomendaciones para controlar errores transitorios.

SQL Database controla automáticamente las tareas de mantenimiento críticas, como la aplicación de revisiones, las copias de seguridad, Windows y las actualizaciones del motor de SQL Database. También controla automáticamente eventos no planeados, como errores de red, software o hardware subyacentes. SQL Database está diseñado para recuperarse rápidamente de errores críticos, lo que garantiza que los datos estén siempre disponibles. La mayoría de los usuarios no observan que las actualizaciones se realizan continuamente.

Cuando una base de datos se revisa o transfiere, el tiempo de inactividad no es disruptivo si emplea lógica de reintento en su aplicación.

Puede probar la resistencia de la aplicación a errores transitorios siguiendo las instrucciones de Prueba de resistencia de errores de la aplicación.

Resistencia a errores de zona de disponibilidad

Las zonas de disponibilidad son grupos físicamente independientes de centros de datos dentro de una región de Azure. Cuando una zona falla, los servicios pueden transferirse a una de las zonas restantes.

Para ver información sobre la compatibilidad de zona de disponibilidad con otros niveles de servicio, asegúrese de seleccionar el nivel de servicio adecuado al principio de esta página.

Requisitos

Los niveles de servicio Básico y Estándar no admiten la redundancia de zona.

La redundancia de zona está disponible para las bases de datos en los niveles de servicio Crítico para negocios, Propósito general e Hiperescala del modelo de compra basado en vCore, y únicamente el nivel de servicio Premium del modelo de compra basado en DTU.

Resistencia a errores en toda la región

En esta sección se proporciona información general sobre dos características relacionadas pero independientes que se pueden usar para la replicación geográfica en varias regiones de SQL Database:

- La replicación geográfica activa replica una base de datos única en una base de datos secundaria sincronizada.
- Los grupos de conmutación por error se basan en la replicación geográfica activa y permiten realizar una conmutación por error de un grupo de bases de datos.

Replicación geográfica activa

La replicación geográfica activa crea una base de datos secundaria legible sincronizada continuamente (que a veces se conoce como réplica geográfica *secundaria* o *geográfica*) en cualquier región para una base de datos principal única. La replicación geográfica activa puede crear bases de datos secundarias en la misma región, pero esta configuración no proporciona

protección contra una interrupción de la región. Al usar la replicación geográfica activa para lograr redundancia geográfica, busque la base de datos secundaria en una región diferente a la base de datos principal.

Requisitos

Al usar la replicación geográfica activa, tenga en cuenta los siguientes requisitos:

- **Compatibilidad con regiones:** la replicación geográfica activa se puede habilitar en todas las regiones de Azure y no requiere que use pares de regiones de Azure.

  > **Sugerencia**
  >
  > SQL Database sigue una práctica de implementación segura en la que Azure se esfuerza por no implementar actualizaciones en regiones emparejadas al mismo tiempo. Si configura la replicación geográfica activa para usar regiones no emparejadas, establezca diferentes ventanas de mantenimiento para los servidores de cada región. Este enfoque ayuda a reducir la posibilidad de que ambas regiones experimenten problemas de conectividad causados por el mantenimiento al mismo tiempo.

- **Configuración:** Tanto el principal como el secundario geográfico deben tener el mismo nivel de servicio y deben tener el mismo nivel de proceso, tamaño de proceso y redundancia de almacenamiento de copia de seguridad.

- **Firewall:** Tanto el servidor principal como el secundario geográfico deben tener las mismas reglas de firewall de direcciones IP.

- **Suscripciones de Azure:** La replicación geográfica activa es compatible con bases de datos en distintas suscripciones de Azure.

Consideraciones

- La replicación geográfica activa está diseñada para proporcionar tolerancia a fallos de una única base de datos. Si necesita conmutar por error varias bases de datos, considere la posibilidad de usar en su lugar grupos de conmutación por error.

- Dado que la replicación geográfica es asincrónica, la pérdida de datos es posible cuando se produce una conmutación por error. Si necesita eliminar la pérdida de datos de la replicación asincrónica durante los fallos, configure la aplicación para bloquear el

subproceso de llamada hasta que se transmita y endurezca la última transacción confirmada en el registro de transacciones de la base de datos secundaria. Este enfoque requiere desarrollo personalizado y reduce el rendimiento de la aplicación. Para obtener más información, consulte Evitar la pérdida de datos críticos.

- Para más información sobre las limitaciones y consideraciones, consulte Replicación geográfica activa.

Cost

Las bases de datos secundarias se facturan como bases de datos independientes.

Si no usa una base de datos secundaria para cargas de trabajo de lectura o escritura, considere si puede designarla como réplica en espera para reducir los costos.

Configuración de la compatibilidad con varias regiones

- **Habilite la replicación geográfica activa:** Para más información sobre cómo habilitar la replicación geográfica activa en Azure Portal, consulte Configuración de la replicación geográfica activa para SQL Database o Replicación geográfica activa.

  Después de habilitar la replicación geográfica activa, un paso de propagación inicial puede tardar algún tiempo.

- **Deshabilitar la replicación geográfica activa:** Para obtener más información sobre cómo deshabilitar la replicación geográfica activa en una base de datos, consulte Eliminación de la base de datos secundaria.

Comportamiento cuando todas las regiones están en buen estado

En esta sección se describe qué esperar cuando una base de datos está configurada para usar la replicación geográfica activa y todas las regiones están operativas.

- **Enrutamiento del tráfico entre regiones:** La aplicación debe configurarse para enviar solicitudes de lectura y escritura a la base de datos principal. Opcionalmente, puede enviar solicitudes de solo lectura a una base de datos secundaria, lo que ayuda a reducir el impacto de las cargas de trabajo de solo lectura en la base de datos principal.

- **Replicación de datos entre regiones:** La replicación entre las bases de datos principal y secundaria se produce de forma asincrónica, lo que significa que puede haber un retraso

entre el momento en que se aplica un cambio a la base de datos principal y cuando se replica en la base de datos secundaria.

Al realizar una conmutación por error, decide cómo gestionar el riesgo de pérdida de datos.

Comportamiento durante una falla de región

En esta sección se describe qué esperar cuando una base de datos está configurada para usar la replicación geográfica activa y hay una interrupción en la región primaria:

- **Detección y respuesta:** Usted es responsable tanto de detectar una interrupción en una base de datos o región como de activar la conmutación por error.
- **Notificación:** Microsoft no le notifica automáticamente cuando una región está inactiva. Sin embargo, puede usar Azure Service Health para comprender el estado general del servicio, incluidos los errores de región, y puede configurar alertas de Service Health para notificarle problemas.
- **Solicitudes activas:** Las solicitudes activas durante la conmutación por error se finalizan y se deben reintentar.
- **Pérdida de datos esperada:** Si la base de datos principal está disponible, opcionalmente puede realizar una conmutación por error sin pérdida de datos. El proceso de conmutación por error sincroniza los datos entre las bases de datos principal y secundaria antes de cambiar de roles.

  Si la base de datos principal no está disponible, es posible que tenga que realizar una *conmutación por error forzada*, lo que da lugar a la pérdida de datos. Puede calcular la cantidad de pérdida de datos mediante la supervisión del retraso de replicación. Para obtener más información, consulte Supervisión de SQL Database con métricas y alertas.
- **Tiempo de inactividad esperado:** Normalmente, hay hasta 60 segundos de tiempo de inactividad durante una conmutación por error. Asegúrese de que la aplicación controla los errores transitorios para que pueda recuperarse de breves períodos de tiempo de inactividad. Además, la rapidez con la que reconfigura la aplicación para conectarse a la nueva base de datos principal afecta al tiempo de inactividad que experimenta.
- **Reenrutamiento del tráfico:** Es responsable de volver a configurar la aplicación para enviar solicitudes a la nueva base de datos principal. Si necesita tener conmutación por error transparente, considere la posibilidad de usar grupos de conmutación por error.

Recuperación de regiones

Una vez recuperada la región primaria, puede realizar manualmente una conmutación por restablecimiento en la región primaria realizando otra conmutación por error.

Prueba de fallos de región

Puede simular una interrupción de región desencadenando una conmutación por error manual en cualquier momento. Puede desencadenar una conmutación por error (sin pérdida de datos) o una conmutación por error forzada.

Grupos de conmutación por error

Los grupos de conmutación por error se basan en la replicación geográfica activa. Con los grupos de conmutación por error, puede realizar las siguientes operaciones:

- Replique un conjunto de bases de datos desde un único servidor lógico en cualquier combinación de regiones de Azure.
- Realice la conmutación por error en las bases de datos como grupo.
- Use puntos de conexión que dirijan automáticamente las conexiones a la principal.

Requisitos

- **Compatibilidad regional:** Los grupos de conmutación de emergencia se pueden crear en todas las regiones de Azure y no es necesario utilizar pares de regiones de Azure.

  > **Sugerencia**
  >
  > Si configura un grupo de conmutación por error con regiones no emparejadas, considere la posibilidad de establecer diferentes ventanas de mantenimiento para los servidores de cada región. Este enfoque ayuda a reducir la posibilidad de que ambas regiones experimenten problemas de conectividad causados por el mantenimiento al mismo tiempo.

- **Configuración:** Las bases de datos secundarias de un grupo de conmutación por error deben tener el mismo nivel de servicio, nivel de proceso, tamaño de proceso, reglas de firewall de direcciones IP y redundancia de almacenamiento de copia de seguridad que la base de datos principal.

Consideraciones

- **Redundancia de zona:** Las bases de datos secundarias no tienen habilitada la redundancia de zona de forma predeterminada, pero puede habilitarla más adelante.
- **Posible pérdida de datos:** Dado que la replicación geográfica es asincrónica, es posible que se pierdan datos cuando se produce una conmutación por error. Si necesita eliminar la pérdida de datos de la replicación asincrónica durante las conmutaciones por error, puede configurar la aplicación para bloquear el subproceso de llamada hasta que la última transacción confirmada se ha transmitido y protegido en el registro de transacciones de la base de datos secundaria. Este enfoque requiere desarrollo personalizado y reduce el rendimiento de la aplicación. Para obtener más información, consulte Evitar la pérdida de datos críticos.
- Para obtener más información sobre las limitaciones y consideraciones, consulte Grupos de conmutación por error.

Cost

Las bases de datos secundarias se facturan como bases de datos independientes.

Si no usa una base de datos secundaria para cargas de trabajo de lectura o escritura, considere si puede designarla como réplica en espera para reducir los costos.

Configuración de la compatibilidad con varias regiones

- **Habilite los grupos de conmutación por error:** Usted configura un grupo de conmutación por error en un servidor lógico. Puede agregar todas las bases de datos del servidor lógico al grupo de conmutación por error, o bien puede seleccionar un subconjunto de bases de datos que se van a agregar.

  Al crear un grupo de conmutación por error, seleccione la directiva de conmutación por error, que especifica quién es responsable de detectar una interrupción y realizar una conmutación por error. Puede configurar la conmutación por error administrada por el cliente, que se recomienda, o bien la conmutación por error administrada por Microsoft.

  > **Importante**
  >
  > Es probable que se produzca una conmutación por error automatizada iniciada por Microsoft después de un retraso significativo y se lleva a cabo de la mejor manera posible. La conmutación por error de las bases de datos puede producirse en un

momento diferente a cualquier conmutación por error de otros servicios de Azure. Para obtener más información, consulte [Configuración de un grupo de conmutación por error para SQL Database](#).

Después de configurar el grupo de conmutación por error, el paso de inicialización puede tardar algún tiempo.

- **Deshabilitar grupos de conmutación por error:** Puede quitar una base de datos individual de un grupo de conmutación por error, quitar un grupo de conmutación por error completo o mover una base de datos a otro grupo de conmutación por error.

Comportamiento cuando todas las regiones están en buen estado

En esta sección se describe qué esperar cuando una base de datos está configurada dentro de un grupo de conmutación por error y todas las regiones están operativas.

- **Enrutamiento del tráfico entre regiones:** En el caso de las cargas de trabajo de lectura y escritura y de solo lectura, los grupos de conmutación por error proporcionan dos puntos de conexión de escucha donde puede conectar sus aplicaciones. Utilice los puntos de conexión del escucha del grupo de conmutación para minimizar el tiempo de inactividad durante las conmutaciones. Durante las operaciones normales, se produce el siguiente comportamiento de enrutamiento:
  - El punto de conexión del escuchador de lectura y escritura enruta todas las solicitudes a las bases de datos primarias.
  - El punto de conexión del agente de escucha de solo lectura enruta todas las solicitudes a una base de datos secundaria legible.
- **Replicación de datos entre regiones:** La replicación geográfica entre las bases de datos principal y secundaria se produce de forma asincrónica. Esta latencia significa que puede haber un retraso entre un cambio que se aplica a la base de datos principal y cuando se replica en la base de datos secundaria.

  Al realizar una conmutación por error, decide cómo gestionar el riesgo de pérdida de datos.

Comportamiento durante una falla de región

En esta sección se describe qué esperar cuando una base de datos está configurada dentro de un grupo de conmutación por error y hay una interrupción en la región primaria:

- **La detección y la respuesta** dependen de la política de conmutación por error que use.
  - *Conmutación por error iniciada por el cliente:* Es responsable de detectar la interrupción de una base de datos o región y de desencadenar la conmutación por error.
  - *Conmutación por error iniciada por Microsoft:* Microsoft desencadena la conmutación por error para todos los grupos de conmutación por error de la región afectada. Microsoft espera realizar este tipo de conmutación por error solo en situaciones excepcionales. No confíe en la conmutación por error administrada por Microsoft para la mayoría de las soluciones. Para obtener más información, consulte Directiva de conmutación por error: administrada por Microsoft.
- **Notificación:** Microsoft no le notifica automáticamente cuando una región está inactiva. Sin embargo, puede usar Azure Service Health para comprender el estado general del servicio, incluidos los errores de región, y puede configurar alertas de Service Health para notificarle problemas.
- **Solicitudes activas:** Las solicitudes activas durante la conmutación por error se finalizan y se deben reintentar.
- **Pérdida de datos esperada:** La pérdida de datos depende de la directiva de conmutación por error que use.
  - *Conmutación por error iniciada por el cliente:* Si la base de datos principal está disponible, opcionalmente puede realizar una conmutación por error sin pérdida de datos. El proceso de conmutación por error sincroniza los datos entre las bases de datos principal y secundaria antes de cambiar de roles.

    Si la base de datos principal no está disponible, es posible que tenga que realizar una *conmutación por error forzada*, lo que da lugar a la pérdida de datos. Puede calcular la cantidad de pérdida de datos mediante la supervisión del retraso de replicación. Para obtener más información, consulte Supervisión de SQL Database con métricas y alertas.
  - *Conmutación por error iniciada por Microsoft:* Una conmutación por error administrada por Microsoft solo se desencadena en situaciones excepcionales. Las conmutaciones por error administradas por Microsoft son conmutaciones por error forzadas, lo que significa que se espera una pérdida de datos. No puede controlar la cantidad de pérdida de datos que puede experimentar.
- **Tiempo de inactividad esperado:** El tiempo de inactividad depende de la directiva de conmutación por error que use.

- *Conmutación por error iniciada por el cliente:* Normalmente, hay hasta 60 segundos de tiempo de inactividad durante una conmutación por error. Asegúrese de que la aplicación controla los errores transitorios para que pueda recuperarse de breves períodos de tiempo de inactividad.
- *Conmutación por error iniciada por Microsoft:* Puede especificar un *período de gracia* que determine cuánto tiempo debe esperar Microsoft antes de iniciar la conmutación por error. El período de gracia mínimo es una hora. Sin embargo, es probable que el tiempo de respuesta de Microsoft sea de varias horas al menos.

- **Reenrutamiento del tráfico:** durante la conmutación por error, SQL Database actualiza los puntos de conexión de escucha de lectura y escritura y de solo lectura para dirigir el tráfico a la nueva base de datos primaria y secundaria según sea necesario.

  Si la nueva base de datos secundaria (anteriormente la base de datos principal) no está disponible, el punto de conexión del agente de escucha de solo lectura no se puede conectar hasta que esté disponible la nueva base de datos secundaria.

Recuperación de regiones

Usted es responsable de restablecer la región primaria si es necesario. Puede realizar manualmente una conmutación por recuperación en la región primaria mediante la realización de una conmutación por error administrada por el cliente.

Incluso si Microsoft inició la conmutación por error original, usted sigue siendo responsable de la conmutación por recuperación a la región anterior, si elige hacerlo.

Prueba de fallos de región

Puede simular una interrupción de región desencadenando una conmutación por error manual en cualquier momento. Puede desencadenar una conmutación por error (sin pérdida de datos) o una conmutación por error forzada.

Resistencia a la eliminación accidental del servidor lógico

Además de la resistencia frente a errores de infraestructura, Azure SQL Database proporciona protección contra la eliminación accidental o no deseada de los servidores lógicos.

Azure SQL Database admite la retención de eliminación temporal para servidores lógicos (versión preliminar). Al habilitar esta función, un servidor lógico eliminado no se elimina permanentemente de inmediato. En su lugar, se conserva en un estado eliminado temporalmente durante un período de retención configurable, durante el cual puede detectarlo y restaurarlo.

La retención del servidor lógico ayuda a mitigar los riesgos de:

- Eliminación accidental a través del portal de Azure, las API o los scripts
- Errores de automatización o implementación
- Errores operativos durante las pruebas o las actividades de limpieza

Azure conserva los metadatos cuando se elimina temporalmente un servidor lógico. Esta retención le permite restaurar el servidor a su estado anterior y recuperar sus bases de datos, siempre que sigan dentro del período de retención de copia de seguridad.

La retención de eliminación temporal complementa las funcionalidades de copia de seguridad y restauración existentes, como la restauración a un momento dado y la retención a largo plazo, proporcionando una ruta de recuperación de nivel de servidor lógico. Esta ruta de recuperación puede reducir el tiempo de recuperación y la complejidad operativa en comparación con la recreación de recursos y la restauración de bases de datos individuales.

> ⓘ **Nota:**
>
> La capacidad de configurar un período de retención para eliminación suave y restaurar un servidor lógico eliminado está actualmente en **versión preliminar**. Cualquier servidor lógico con más de dos años tiene automáticamente un período de retención de eliminación suave de siete días. Los servidores lógicos menores de dos años tienen deshabilitada la retención de eliminación suave de forma predeterminada.

Copias de seguridad y restauración

Realice copias de seguridad de las bases de datos para protegerse frente a diversos riesgos, incluida la pérdida de datos. Las copias de seguridad se pueden restaurar para recuperarse de pérdida accidental de datos, daños u otros problemas. Las copias de seguridad difieren de la redundancia de zona, la replicación geográfica activa o los grupos de conmutación por error, y tienen diferentes propósitos. Para obtener más información, consulte Redundancia, replicación y copia de seguridad.

SQL Database proporciona copias de seguridad automáticas de las bases de datos. Para obtener más información sobre la frecuencia de copia de seguridad, que puede afectar a la cantidad de pérdida de datos si necesita restaurar desde una copia de seguridad, consulte Copias de seguridad automatizadas en SQL Database.

Almacenamiento de copia de seguridad

Puede elegir almacenar las copias de seguridad automatizadas en LRS o ZRS. Si usa una región emparejada, puede optar por replicar las copias de seguridad automatizadas en la región emparejada mediante el almacenamiento con redundancia geográfica. Esta funcionalidad permite la restauración geográfica de las copias de seguridad en la región emparejada. Para obtener más información, consulte Copias de seguridad automatizadas en SQL Database.

Si usa una región no emparejada o si necesita replicar copias de seguridad en una región distinta de la región emparejada, considere la posibilidad de exportar la base de datos y almacenar el archivo exportado en una cuenta de almacenamiento que use la replicación de objetos de blob para replicar en una cuenta de almacenamiento de otra región. Para obtener más información, consulte Exportación de una base de datos.

Para una protección más sólida, puede habilitar la inmutabilidad de copia de seguridad para las copias de seguridad de retención a largo plazo (LTR). La inmutabilidad de copia de seguridad garantiza que las copias de seguridad de LTR para una base de datos se almacenan en un estado **Write Once, Read Many (WORM),** por lo que no se pueden modificar y no se pueden borrar durante un período de retención definido por el usuario. La inmutabilidad de copia de seguridad protege contra la eliminación o modificación accidental o malintencionada, incluso por parte de los administradores con privilegios. Para obtener más información, consulte Inmutabilidad de las copias de seguridad de LTR en Azure SQL Database.

Resistencia al mantenimiento del servicio

Cuando SQL Database realiza el mantenimiento en tus bases de datos y en los grupos elásticos, puede conmutar automáticamente las bases de datos para usar una réplica secundaria. Las aplicaciones cliente pueden observar breves interrupciones de conectividad cuando se produce una conmutación por error. Las aplicaciones cliente deben seguir las instrucciones de resistencia a errores transitorios para minimizar los efectos.

SQL Database permite especificar una ventana de mantenimiento que se usa normalmente para las actualizaciones de servicio y otras operaciones de mantenimiento. La configuración de una ventana de mantenimiento puede ayudarle a minimizar los efectos secundarios, como las

conmutaciones por error automáticas, durante el horario comercial. También puede recibir notificaciones anticipadas del mantenimiento planeado.

La plataforma mantiene automáticamente las puertas de enlace usadas para procesar las conexiones a SQL Database. Las actualizaciones o las operaciones de mantenimiento también pueden provocar interrupciones breves de conectividad que los clientes pueden reintentar.

Para obtener más información, vea Ventana de mantenimiento en SQL Database.

Acuerdo de nivel de servicio

El contrato de nivel de servicio (SLA) para los servicios de Azure describe la disponibilidad esperada de cada servicio y las condiciones que la solución deberá cumplir para lograr esa expectativa de disponibilidad. Para obtener más información, consulte Acuerdos de Nivel de Servicio para servicios en línea.

El acuerdo de nivel de servicio (SLA) para SQL Database describe la disponibilidad esperada del servicio y el punto de recuperación esperado y el tiempo de recuperación para la replicación geográfica activa.

Al implementar una base de datos con redundancia de zona o un grupo elástico y usar un nivel de servicio compatible, el SLA de tiempo de actividad es mayor.

Contenido relacionado

- Introducción a la continuidad empresarial con SQL Database
- Lista de comprobación de alta disponibilidad y recuperación ante desastres: SQL Database
- Confiabilidad en Azure

ⓘ**Nota:** El autor ha creado este artículo con ayuda de inteligencia artificial. Más información

---

Last updated on 27/05/2026

Confiabilidad en Azure SQL Managed Instance

Azure SQL Managed Instance es un motor de base de datos de plataforma como servicio (PaaS) totalmente administrado. Proporciona casi 100% compatibilidad de características con SQL Server. Azure SQL Managed Instance controla la mayoría de las funciones de administración de bases de datos, como la actualización, la aplicación de revisiones, las copias de seguridad y la supervisión sin intervención del usuario. Se ejecuta en la versión estable más reciente del motor de base de datos de SQL Server y un sistema operativo revisado con alta disponibilidad integrada.

Cuando se usa Azure, la confiabilidad es una responsabilidad compartida. Microsoft proporciona una variedad de funcionalidades para admitir resistencia y recuperación. Es responsable de comprender cómo funcionan esas funcionalidades dentro de todos los servicios que usa y de seleccionar las funcionalidades que necesita para cumplir los objetivos empresariales y los objetivos de tiempo de actividad.

En este artículo se describe cómo hacer que Azure SQL Managed Instance sea resistente a una variedad de posibles interrupciones y problemas, incluidos errores transitorios, interrupciones de zona de disponibilidad y interrupciones de regiones. También se describe cómo puede usar copias de seguridad para recuperarse de otros tipos de problemas, cómo controlar el mantenimiento del servicio y resaltar cierta información clave sobre el contrato de nivel de servicio (SLA) de Azure SQL Managed Instance.

Recomendaciones de implementación de producción para la confiabilidad

Para la mayoría de las implementaciones de producción de SQL Managed Instance, tenga en cuenta las siguientes recomendaciones:

- Siga las instrucciones proporcionadas en la lista de comprobación de alta disponibilidad y recuperación ante desastres (DR).
- Habilitación de la redundancia de zona.
- Configure copias de seguridad automatizadas y use almacenamiento con redundancia de zona (ZRS) o almacenamiento con redundancia geográfica (GRS) para las copias de seguridad.
- Planee probar periódicamente las copias de seguridad y el proceso de restauración.

Introducción a la arquitectura de confiabilidad

Las instancias administradas de SQL de uso general se ejecutan en un único nodo que administra Azure Service Fabric . Cada vez que se actualiza el motor de base de datos o el sistema operativo, o se detecta un error, SQL Managed Instance funciona con Service Fabric para mover el proceso del motor de base de datos sin estado a otro nodo de proceso sin estado que tenga suficiente capacidad libre. Los archivos de base de datos se almacenan en Azure Blob Storage, que tiene características de redundancia integradas. Los archivos de datos y registro se desasocian del nodo de proceso original y se adjuntan al proceso del motor de base de datos recién inicializado.

Redundancia

De forma predeterminada, SQL Managed Instance logra redundancia mediante la propagación de nodos de proceso y datos a lo largo de un único centro de datos de la región primaria. Este enfoque protege los datos durante los siguientes tiempos de inactividad esperados e inesperados:

- Operaciones de administración iniciadas por el cliente que dan lugar a un breve tiempo de inactividad
- Operaciones de mantenimiento de servicio
- Errores de red o energía a pequeña escala
- Problemas y interrupciones del centro de datos que implican los siguientes componentes:
  - *Bastidor* en el que se ejecutan las máquinas que impulsan el servicio
  - La *máquina física* que hospeda la máquina virtual (VM) que ejecuta el motor de base de datos SQL.
  - Máquina *virtual* que ejecuta el motor de base de datos de SQL
- Problemas con el motor de base de datos sql
- Posibles interrupciones localizadas no planeadas

Para obtener más información sobre cómo SQL Managed Instance proporciona redundancia, consulte Disponibilidad a través de redundancia de zona y local.

Resistencia a errores transitorios

Los errores transitorios son errores breves e intermitentes en los componentes. Se producen con frecuencia en un entorno distribuido como la nube y son una parte normal de las operaciones. Los errores transitorios se corrigen después de un breve período de tiempo. Es importante que las aplicaciones puedan controlar errores transitorios, normalmente mediante el reintento de solicitudes afectadas.

Todas las aplicaciones hospedadas en la nube deben seguir las instrucciones de control de errores transitorios de Azure cuando se comunican con cualquier API, bases de datos y otros componentes hospedados en la nube. Para obtener más información, consulte Recomendaciones para controlar errores transitorios.

Sql Managed Instance controla automáticamente las tareas de mantenimiento críticas, como la aplicación de revisiones, las copias de seguridad y las actualizaciones del motor de base de datos de SQL y Windows. También controla eventos no planeados, como errores de red, software o hardware subyacentes. Sql Managed Instance puede recuperarse rápidamente incluso en las circunstancias más críticas, lo que garantiza que los datos estén siempre disponibles. La mayoría de los usuarios no observan que las actualizaciones se realizan continuamente.

Cuando se revisa una instancia o se produce una conmutación por error, el tiempo de inactividad tiene un efecto mínimo si emplea un mecanismo de reintento en la aplicación. Puede probar la resistencia de la aplicación a errores transitorios.

Resistencia a errores de zona de disponibilidad

> ⓘ **Nota:**
>
> La redundancia de zona no está disponible actualmente para el nivel de servicio de próxima generación de uso general.

Las zonas de disponibilidad son grupos físicamente independientes de centros de datos dentro de una región de Azure. Cuando una zona falla, los servicios pueden transferirse a una de las zonas restantes.

Al habilitar una configuración con redundancia de zona, puede asegurarse de que la instancia administrada de SQL sea resistente a un gran conjunto de errores, incluidas interrupciones catastróficas del centro de datos, sin cambios en la lógica de la aplicación.

SQL Managed Instance logra la redundancia de zona al colocar nodos de cómputo sin estado en diferentes zonas de disponibilidad. Se basa en ZRS con estado que está asociado al nodo que actualmente contiene el proceso activo del motor de base de datos SQL. Si se produce una

interrupción, el proceso del motor de base de datos SQL se activa en uno de los nodos de cómputo sin estado, que luego accede a los datos en el almacenamiento con estado.

Requisitos

- **Compatibilidad con regiones:** La redundancia de zona para SQL Managed Instance se admite en regiones seleccionadas. Para obtener más información, consulte Regiones admitidas.

- **Redundancia de almacenamiento de copia de seguridad:** Para habilitar la redundancia de zona para la instancia administrada de SQL, establezca la redundancia del almacenamiento de copia de seguridad en *ZRS* o *almacenamiento con redundancia de zona geográfica* (GZRS).

Cost

Al habilitar la redundancia de zona, hay un costo adicional para la instancia administrada de SQL y para las copias de seguridad con redundancia de zona. Para obtener más información, consulte el apartado Precios.

Configurar soporte de zonas de disponibilidad

En esta sección se explica cómo configurar la compatibilidad con zonas de disponibilidad para las instancias administradas de SQL:

- **Habilite la redundancia de zona:** Para obtener información sobre cómo configurar la redundancia de zona en instancias nuevas y existentes, consulte Configuración de la redundancia de zona.

  Todas las operaciones de escalado de SQL Managed Instance, incluida la habilitación de la redundancia de zona, son operaciones en línea y requieren un tiempo de inactividad mínimo o nulo. Para obtener más información, consulte Duración de las operaciones de administración.

- **Deshabilitar la redundancia de zona:** puede deshabilitar la redundancia de zona siguiendo los mismos pasos para habilitar la redundancia de zona. Este proceso es una operación en línea similar a una actualización normal del objetivo de nivel de servicio. Al final del proceso, la instancia se migra de la infraestructura con redundancia de zona a la infraestructura de una sola zona.

Comportamiento cuando todas las zonas están en buen estado

En esta sección se describe qué esperar cuando la instancia administrada de SQL está configurada para que sea redundante de zona y todas las zonas de disponibilidad estén operativas:

- **Enrutamiento de tráfico entre zonas:** Durante las operaciones normales, las solicitudes se enrutan al nodo que ejecuta la capa de proceso de SQL Managed Instance.

- **Replicación de datos entre zonas:** Los archivos de base de datos se almacenan en Azure Storage mediante ZRS, que se adjunta al nodo que contiene actualmente el proceso activo del motor de base de datos de SQL.

  Las operaciones de escritura son sincrónicas y no se consideran completadas hasta que los datos se replican correctamente en todas las zonas de disponibilidad. Esta replicación sincrónica garantiza una fuerte coherencia y cero pérdida de datos durante los errores de zona. Sin embargo, podría dar lugar a una latencia de escritura ligeramente mayor en comparación con el almacenamiento con redundancia local.

Comportamiento durante un fallo de zona

En esta sección se describe qué esperar cuando la instancia administrada de SQL está configurada para que sea redundante de zona y una o varias zonas de disponibilidad no están disponibles:

- **Detección y respuesta:** Instancia administrada de SQL es responsable de detectar y responder a un error en una zona de disponibilidad. No es necesario hacer nada para iniciar una conmutación por error de la zona.

- **Notificación**: Microsoft no le notifica automáticamente cuando una zona está inactiva. Sin embargo, puede usar Azure Resource Health para supervisar el estado de un recurso individual y puede configurar alertas de Resource Health para notificarle problemas. También puede usar Azure Service Health para comprender el estado general del servicio, incluidos los errores de zona, y puede configurar alertas de Service Health para notificarle problemas.

- **Solicitudes activas:** Cuando una zona de disponibilidad no está disponible, las solicitudes que se procesan en la zona de disponibilidad defectuosa se finalizan y se deben reintentar. Para que las aplicaciones sean resistentes a estos tipos de problemas, consulte Guía sobre resistencia a errores transitorios .

- **Reenrutamiento del tráfico:** Sql Managed Instance funciona con Service Fabric para mover el motor de base de datos a un nodo de proceso sin estado adecuado que se encuentra en una zona de disponibilidad diferente y tiene suficiente capacidad libre. Una vez completada la conmutación por error, las nuevas conexiones se redirigen automáticamente al nuevo nodo de proceso principal.

  Una carga de trabajo pesada podría experimentar cierta degradación del rendimiento durante la transición de un nodo de proceso al otro porque el nuevo proceso del motor de base de datos comienza con una caché inactiva.

- **Tiempo de inactividad esperado:** puede haber una pequeña cantidad de tiempo de inactividad durante una conmutación por error de zona de disponibilidad. El tiempo de inactividad suele ser inferior a 30 segundos, que la aplicación debe tolerar si sigue las instrucciones de resistencia a errores transitorios .

- **Pérdida de datos esperada:** no se espera ninguna pérdida de datos para las transacciones confirmadas durante una conmutación por error de zona de disponibilidad. Es necesario reintentar las transacciones en curso.

Recuperación de zona

Cuando se recupera la zona de disponibilidad, SQL Managed Instance funciona con Service Fabric para restaurar operaciones en la zona recuperada. No se requiere intervención del cliente.

Prueba de fallos de zona

La plataforma instancia administrada de SQL administra el enrutamiento del tráfico, la conmutación por error y la conmutación por recuperación para las instancias con redundancia de zona. Dado que esta característica está totalmente administrada, no es necesario iniciar ni validar los procesos de error de zona de disponibilidad. Sin embargo, puede validar la gestión de fallos de su aplicación.

Resistencia a errores en toda la región

Una instancia administrada de SQL individual se implementa dentro de una sola región. Sin embargo, puede implementar una instancia administrada de SQL secundaria en una región de Azure independiente y configurar un *grupo de conmutación por error*.

Grupos de conmutación por error

Los grupos de conmutación por error replican automáticamente de forma geográfica sus datos y pueden conmutar por error automáticamente o manualmente si se produce un error regional, en función de la directiva de conmutación por error.

En esta sección se resume la información clave sobre los grupos de conmutación por error, pero es importante revisar la introducción a los grupos de conmutación por error y los procedimientos recomendados para obtener más información sobre cómo funcionan y cómo configurarlos.

Directivas de conmutación por error

Al crear un grupo de conmutación por error, seleccione la directiva de conmutación por error, que especifica quién es responsable de detectar una interrupción y realizar una conmutación por error. Puede configurar dos tipos de políticas de conmutación por error:

- **Conmutación por error administrada por el cliente (recomendada):** Cuando se usa una directiva de conmutación por error administrada por el cliente, puede decidir si se debe realizar una *conmutación* por error, que no incurre en pérdida de datos o una *conmutación por error forzada*, lo que podría suponer una pérdida de datos. La conmutación por error forzada se utiliza como método de recuperación durante las interrupciones del servicio cuando no se puede acceder a la instancia principal.
- **Conmutación por error administrada por Microsoft:** La conmutación por error administrada por Microsoft solo se usa en situaciones excepcionales para desencadenar una conmutación por error forzada.

> **Importante**
>
> Utilice las opciones de conmutación por error administradas por el cliente para desarrollar, probar e implementar tus planes de recuperación ante desastres. **No confíe en la conmutación por error administrada por Microsoft**, que solo se puede utilizar en circunstancias extremas. Es probable que se haya iniciado una conmutación por error administrada por Microsoft para toda una región. No se puede iniciar para grupos de conmutación por error individuales, instancias administradas de SQL, suscripciones o clientes. La conmutación por error puede producirse en momentos diferentes para distintos servicios de Azure. Recomendamos que utilice la conmutación por error administrada por el cliente.

Soporte para regiones

Puede seleccionar cualquier región de Azure para las instancias administradas de SQL en el grupo de conmutación por error. Debido a la alta latencia de las redes de área extensa, la replicación geográfica usa un mecanismo de replicación asincrónico. Para reducir los retrasos de red, seleccione regiones que tengan conexiones de baja latencia. Para más información sobre la latencia entre regiones de Azure, consulte Estadísticas de latencia de ida y vuelta de red de Azure.

Cost

Al crear varias instancias administradas de SQL en regiones diferentes, se le factura cada instancia administrada de SQL.

Sin embargo, si la instancia secundaria no tiene ninguna carga de trabajo de lectura o aplicaciones conectadas a ella, puede ahorrar en costes de licencias mediante la designación de la réplica como una instancia en espera. Para más información, consulte Configuración de una réplica en espera sin licencia para SQL Managed Instance.

Para más información sobre los precios de SQL Managed Instance, consulte Información de precios del servicio.

Configuración de la compatibilidad con varias regiones

Para obtener información sobre cómo configurar un grupo de conmutación por error, consulte Configuración de un grupo de conmutación por error para SQL Managed Instance.

Planeamiento y administración de capacidad

Durante una conmutación por error, el tráfico se redirige a una instancia administrada de SQL secundaria. Es importante que la instancia administrada de SQL secundaria esté lista para recibir tráfico. Cree una instancia administrada de SQL secundaria con el mismo nivel de servicio, generación de hardware y tamaño de proceso que la instancia principal.

Para más información sobre el escalado de instancias administradas de SQL en un grupo de conmutación por error, consulte Escalado de instancias.

Comportamiento cuando todas las regiones están en buen estado

En esta sección se describe qué esperar cuando las instancias administradas de SQL están configuradas para usar grupos de conmutación por error de varias regiones y todas las regiones están operativas:

- **Enrutamiento del tráfico entre regiones:** durante las operaciones normales, las solicitudes de lectura y escritura van a la instancia principal única de la región primaria.

  Los grupos de conmutación por error también proporcionan un punto de conexión de agente de escucha de solo lectura independiente. Durante las operaciones normales, este punto de conexión se conecta a la instancia secundaria para enrutar el tráfico de solo lectura especificado en la cadena de conexión.

  Para obtener más información sobre cómo los grupos de conmutación por error envían tráfico a cada instancia y cómo puede dirigir el tráfico a un punto de conexión del oyente de solo lectura, consulte Introducción a los grupos de conmutación por error y procedimientos recomendados.

- **Replicación de datos entre regiones:** de manera predeterminada, los datos se replican de forma asincrónica desde la instancia principal a la instancia administrada de SQL secundaria.

  Dado que la replicación geográfica es asincrónica, si realiza una conmutación por error forzada, es posible experimentar la pérdida de datos. Puede supervisar el retraso de replicación para comprender la posible pérdida de datos durante una conmutación por error forzada. Para más información, consulte lista de comprobación de recuperación ante desastres.

  Si necesita eliminar la pérdida de datos de la replicación asincrónica durante las conmutaciones por error, configure la aplicación para bloquear el subproceso de llamada hasta que confirme que la última transacción confirmada se ha transmitido y protegido en el registro de transacciones de la base de datos secundaria. Este enfoque requiere desarrollo personalizado y reduce el rendimiento de la aplicación. Para obtener más información, consulte Evitar la pérdida de datos críticos.

Comportamiento durante una falla de región

En esta sección se describe qué esperar cuando las instancias administradas de SQL están configuradas para usar grupos de conmutación por error de varias regiones y hay una interrupción en la región primaria:

- **Detección y respuesta:** la responsabilidad de la detección y respuesta depende de la directiva de conmutación por error utilizada por el grupo de conmutación por error.

  - *Directiva de conmutación por error administrada por el cliente:* es responsable de detectar el error en una región y desencadenar una conmutación por error o una conmutación por error forzada a la instancia secundaria del grupo de conmutación por error.

Si realiza una conmutación por error, SQL Managed Instance espera a que los datos se sincronicen con la instancia secundaria antes de realizar el procedimiento de conmutación por error.

Si realiza una conmutación por error forzada, SQL Managed Instance cambia inmediatamente la instancia secundaria al rol principal sin esperar a que los cambios recientes se propaguen desde la principal. Este tipo de conmutación por error puede incurrir en pérdida de datos.

- *Directiva de conmutación por error administrada por Microsoft:* las conmutaciones por error administradas por Microsoft se realizan en circunstancias excepcionales. Cuando Microsoft desencadena una conmutación por error, el grupo de conmutación por error realiza automáticamente una conmutación por error forzada a la instancia secundaria del grupo de conmutación por error. Sin embargo, se recomienda usar una directiva de conmutación por error administrada por el cliente para cargas de trabajo de producción para que pueda controlar cuándo se produce la conmutación por error.

- **Notificación**: Microsoft no le notifica automáticamente cuando una zona está inactiva. Sin embargo, puede usar Azure Resource Health para supervisar el estado de un recurso individual y puede configurar alertas de Resource Health para notificarle problemas. También puede usar Azure Service Health para comprender el estado general del servicio, incluidos los errores de zona, y puede configurar alertas de Service Health para notificarle problemas.

- **Solicitudes activas:** Cuando se produce una conmutación por error, se finalizan las solicitudes que se están procesando y se deben reintentar. Para que las aplicaciones sean resistentes a estos tipos de problemas, consulte Resistencia a errores transitorios.

- **Pérdida de datos esperada:** La cantidad de pérdida de datos depende de cómo configure la aplicación. Para más información, consulte Introducción a los grupos de conmutación por error y procedimientos recomendados.

- **Tiempo de inactividad esperado:** Puede haber una breve interrupción durante una conmutación por error de un grupo. El tiempo de inactividad suele ser inferior a 60 segundos.

- **Reenrutamiento del tráfico:** Una vez que el grupo de conmutación por error completa el proceso de conmutación por error, el tráfico de lectura y escritura se enruta automáticamente a la nueva instancia principal. Si las aplicaciones usan los puntos de conexión del grupo de conmutación por error en sus cadenas de conexión, no necesitan modificar sus cadenas de conexión después de la conmutación por error.

Recuperación de regiones

Los grupos de conmutación por error no conmutan automáticamente por recuperación a la región primaria cuando se restauran y, por tanto, es su responsabilidad iniciar una conmutación por recuperación.

Prueba de fallos de región

Puede probar la conmutación por error de un grupo de conmutación por error.

Probar un grupo de conmutación por error es solo una parte de realizar un simulacro de recuperación ante desastres. Para más información, consulte Realización de simulacros de DR.

Copias de seguridad y restauración

Realice copias de seguridad de las bases de datos para protegerse frente a diversos riesgos, incluida la pérdida de datos. Las copias de seguridad se pueden restaurar para recuperarse de pérdida accidental de datos, daños u otros problemas. Las copias de seguridad no son lo mismo que la replicación geográfica, y tienen diferentes propósitos y mitigan distintos riesgos.

Instancia administrada de SQL realiza automáticamente copias de seguridad completas, diferenciales y del registro de transacciones de las bases de datos. Para obtener más información sobre los tipos de copias de seguridad, su frecuencia, funcionalidades de restauración, costos de almacenamiento y cifrado de copia de seguridad, consulte Copias de seguridad automatizadas en SQL Managed Instance.

SQL Managed Instance proporciona copias de seguridad automatizadas integradas y también admite copias de seguridad de solo copia iniciadas por el usuario para bases de datos de usuario. Para más información, consulte Copias de seguridad de solo copia.

Replicación de copia de seguridad

Al configurar copias de seguridad automatizadas para la instancia administrada de SQL, puede especificar cómo se deben replicar las copias de seguridad. Las copias de seguridad configuradas para almacenarse en ZRS tienen un mayor nivel de resistencia. Se recomienda configurar las copias de seguridad para que usen uno de los siguientes tipos de almacenamiento:

- ZRS para la resistencia dentro de la región, si la región tiene zonas de disponibilidad
- GZRS para mejorar la resistencia de las copias de seguridad entre regiones si la región tiene zonas de disponibilidad y está emparejada con otra región
- GRS si la región no admite zonas de disponibilidad, pero tiene una región emparejada

Para más información sobre los distintos tipos de almacenamiento y sus funcionalidades, consulte Redundancia de almacenamiento de respaldo.

Restauración geográfica

La funcionalidad de restauración geográfica es una solución básica de recuperación ante desastres que permite restaurar copias de seguridad en otra región de Azure. La copia de seguridad geográfica normalmente requiere una cantidad significativa de tiempo de inactividad y pérdida de datos. Para lograr niveles más altos de capacidad de recuperación si se produce una interrupción regional, debe configurar grupos de conmutación por error.

Si usa la restauración geográfica, debe tener en cuenta cómo hacer que las copias de seguridad estén disponibles en la región secundaria:

- Si la región primaria está emparejada, use el almacenamiento de copia de seguridad de GZRS o GRS para admitir la restauración geográfica en la región emparejada.
- Si la región primaria no está emparejada, puede crear una solución personalizada para replicar las copias de seguridad en otra región. Considere la posibilidad de usar copias de seguridad de solo copia iniciadas por el usuario y almacenarlas en una cuenta de almacenamiento que use la replicación de objetos de blob para replicar en una cuenta de almacenamiento de otra región.

Resistencia al mantenimiento del servicio

Cuando SQL Managed Instance realiza el mantenimiento en la instancia, la instancia administrada de SQL permanece totalmente disponible, pero puede estar sujeta a reconfiguraciones breves. Las aplicaciones cliente pueden observar breves interrupciones de conectividad cuando se produce un evento de mantenimiento. Las aplicaciones cliente deben seguir las instrucciones de resistencia a errores transitorios para minimizar los efectos.

Instancia administrada de SQL permite especificar una ventana de mantenimiento que se usa generalmente para las actualizaciones del servicio y otras operaciones de mantenimiento. La configuración de una ventana de mantenimiento puede ayudarle a minimizar los efectos secundarios, como las conmutaciones automáticas por error, durante el horario comercial. También puede recibir notificaciones anticipadas del mantenimiento planeado.

Para obtener más información, consulte Ventana de mantenimiento en SQL Managed Instance.

Acuerdo de nivel de servicio

El contrato de nivel de servicio (SLA) para los servicios de Azure describe la disponibilidad esperada de cada servicio y las condiciones que la solución deberá cumplir para lograr esa expectativa de disponibilidad. Para obtener más información, consulte Acuerdos de Nivel de Servicio para servicios en línea.

Para SQL Managed Instance, el Acuerdo de Nivel de Servicio de disponibilidad solo se aplica cuando la red virtual de Azure está configurada correctamente para que no impida el tráfico de administración. Esta configuración incluye el tamaño de subred, los grupos de seguridad de red (NSG), las rutas definidas por el usuario (UDR), la configuración de DNS y otros recursos que afectan a la administración y el uso de recursos de red. Para obtener más información sobre la configuración de red necesaria para SQL Managed Instance, consulte Requisitos de red.

Contenido relacionado

- Introducción a la continuidad empresarial con SQL Managed Instance
- Lista de comprobación de alta disponibilidad y recuperación ante desastres
- Confiabilidad en Azure

Last updated on 22/01/2026

Fiabilidad en Azure Virtual Machine Scale Sets

Azure Virtual Machine Scale Sets es un recurso de proceso de Azure que puede usar para crear y administrar un grupo de instancias de máquina virtual (VM). El número de instancias de máquina virtual puede aumentar o disminuir automáticamente según la demanda, o de acuerdo a una programación definida. Los conjuntos de escalado de máquinas virtuales ayudan a que las aplicaciones sean de alta disponibilidad y resistentes mediante la distribución de máquinas virtuales entre varias zonas de disponibilidad y dominios de error.

Cuando se usa Azure, la confiabilidad es una responsabilidad compartida. Microsoft proporciona una variedad de funcionalidades para admitir resistencia y recuperación. Es responsable de comprender cómo funcionan esas funcionalidades dentro de todos los servicios que usa y de seleccionar las funcionalidades que necesita para cumplir los objetivos empresariales y los objetivos de tiempo de actividad.

En este artículo se describe cómo hacer que los conjuntos de escalado de máquinas virtuales sean resistentes a varias posibles interrupciones y problemas, incluidos errores transitorios, interrupciones de zona de disponibilidad, interrupciones de regiones, reconfiguración de máquinas virtuales y mantenimiento del servicio. También describe cómo puede usar copias de seguridad para recuperarse de otros tipos de problemas y resalta información clave sobre el acuerdo de nivel de servicio (SLA) de virtual Machine Scale Sets.

> **ⓘ Importante**
>
> Cuando considere la confiabilidad de un conjunto de escalado y sus máquinas virtuales, también debe tener en cuenta la confiabilidad de los discos, la infraestructura de red y las aplicaciones que se ejecutan en las máquinas virtuales. Mejorar la resistencia de las máquinas virtuales por sí solas podría tener un efecto limitado si los otros componentes no son igualmente resistentes. En función de los requisitos de resistencia, es posible que tenga que realizar cambios de configuración en varias áreas.

Recomendaciones de implementación de producción

Azure Well-Architected Framework proporciona recomendaciones para la confiabilidad, el rendimiento, la seguridad, el costo y las operaciones. Para obtener información sobre cómo estas áreas influyen entre sí y contribuyen a una solución de conjuntos de escalado de

máquinas virtuales confiables, consulte Procedimientos recomendados de arquitectura para Azure Virtual Machines y conjuntos de escalado en Well-Architected Framework.

Introducción a la arquitectura de confiabilidad

Un conjunto de escalado agrupa varias instancias de máquina virtual y aplica la configuración centralizada, las reglas de escalado automático y las actualizaciones graduales.

Los conjuntos de escalas admiten dos modos de orquestación distintos:

- *Los conjuntos de escalado flexibles* (recomendados) proporcionan más flexibilidad para implementar y administrar instancias de máquina virtual individuales.
- *Los conjuntos de escalado uniforme* implementan máquinas virtuales que tienen una configuración idéntica y las administra como un grupo.

Propagación de dominios de error

Los dominios de error son grupos de aislamiento de errores dentro de un centro de datos. Cada dominio de error es como un bastidor de servidor, que es una colección de nodos de hardware que comparten la misma alimentación, redes, enfriamiento y cronograma de mantenimiento de la plataforma. Dado que las instancias de máquina virtual de cada conjunto de escalado se distribuyen entre varios dominios de error, es probable que una interrupción planeada o no planeada que se produzca en un dominio de error no afecte a las instancias de máquina virtual en otros dominios de error.

Al implementar un conjunto de escalado, puede controlar el número de dominios de error entre los que se distribuyen las máquinas virtuales. Para la mayoría de los escenarios, use el comportamiento de *propagación máxima* , que usa tantos dominios de error como sea posible. Para obtener más información, consulte Elegir el Número Correcto de Dominios de Error para Virtual Machine Scale Sets.

En las regiones que tienen zonas de disponibilidad, cada zona tiene un conjunto distinto de dominios de error. Al crear un conjunto de escalado que abarca regiones, las instancias se distribuyen entre dominios de fallo en cada región que usa el conjunto de escalado.

Equilibrio de carga

Los conjuntos de escalado se pueden integrar con los servicios de equilibrio de carga de Azure, incluido Azure Load Balancer y Azure Application Gateway. Cuando el conjunto de escalado agrega o quita instancias, la integración integrada del equilibrador de carga actualiza automáticamente la configuración del equilibrador de carga. Para obtener más información, consulte Redes para conjuntos de escalado de máquinas virtuales.

Los conjuntos de escalado incluyen muchos otros controles y funcionalidades que afectan a la implementación, escalado, distribución y actualización de instancias. Para obtener más información, consulte Introducción a los conjuntos de escalado de máquinas virtuales.

Resistencia a errores transitorios

Los errores transitorios son errores breves e intermitentes en los componentes. Se producen con frecuencia en un entorno distribuido como la nube y son una parte normal de las operaciones. Los errores transitorios se corrigen después de un breve período de tiempo. Es importante que las aplicaciones puedan controlar errores transitorios, normalmente mediante el reintento de solicitudes afectadas.

Todas las aplicaciones hospedadas en la nube deben seguir las instrucciones de control de errores transitorios de Azure cuando se comunican con cualquier API, bases de datos y otros componentes hospedados en la nube. Para obtener más información, consulte Recomendaciones para controlar errores transitorios.

Las aplicaciones que se ejecutan en las máquinas virtuales deben implementar estrategias adecuadas de control de errores para asegurarse de que las interrupciones temporales del servicio no afecten a la carga de trabajo.

Resiliencia ante problemas de instancia

Cuando un conjunto de escalado inicia una tarea de creación o eliminación de una instancia de máquina virtual, es posible que se produzca un error en la operación. Para reintentar automáticamente las tareas de creación o eliminación de instancias de máquina virtual con errores, considere la posibilidad de usar la característica resistente de creación y eliminación de conjuntos de escalado de máquinas virtuales (versión preliminar).

Es posible que se produzcan problemas durante la ejecución de instancias. Por ejemplo, una instancia podría dejar de responder debido a bloqueos de aplicación o agotamiento de recursos. Use reparaciones automáticas de instancias para supervisar el estado de la aplicación y reiniciar, restablecer la imagen o reemplazar automáticamente una instancia de máquina virtual cuando sea necesario.

Resistencia a errores de zona de disponibilidad

Las zonas de disponibilidad son grupos físicamente independientes de centros de datos dentro de una región de Azure. Cuando una zona falla, los servicios pueden transferirse a una de las zonas restantes.

Virtual Machine Scale Sets admite zonas de disponibilidad en configuraciones zonales y de expansión de zona.

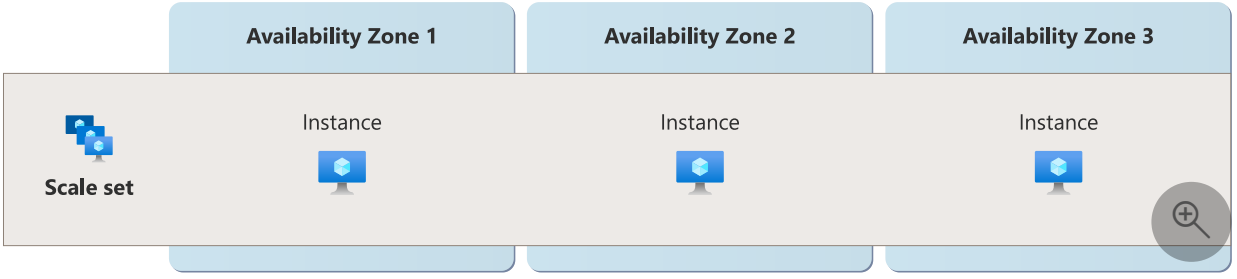
- Un conjunto de escalado que abarca zonas distribuye instancias en varias zonas de disponibilidad que seleccionas.

  Distribuir instancias de máquina virtual entre zonas de disponibilidad te ofrece el nivel más alto de SLA. Recomendamos utilizar conjuntos de escalado que abarcan zonas para la mayoría de las cargas de trabajo basadas en máquinas virtuales en Azure.

  En un conjunto de escala que abarca zonas, cada instancia de máquina virtual y sus discos están vinculados a una zona de disponibilidad específica. Cuando todas las zonas están en buen estado, las instancias pueden comunicarse entre zonas mediante una red de alto rendimiento y baja latencia. Si una zona experimenta un problema de interrupción o conectividad, las instancias de las otras zonas siguen sin verse afectadas.

  De forma predeterminada, el conjunto de escalado usa un enfoque de esfuerzo máximo para repartir equitativamente las instancias entre las zonas seleccionadas. Pero si necesita un equilibrio estricto, puede cambiar la configuración de equilibrio de zona.

  En el diagrama siguiente se muestra un conjunto de instancias escalables que abarca tres zonas, incluyendo una instancia en cada zona.



  La expansión de zona es similar a la redundancia de zona en otros servicios de Azure, pero los conjuntos de escalado no realizan automáticamente la replicación de datos entre zonas ni la conmutación por error cuando las zonas están inactivas. En algunas situaciones inusuales, un conjunto de escalado de expansión de zona también podría tener sus instancias implementadas en una sola zona, como cuando se conectan máquinas virtuales individuales de una zona a un conjunto de escalado flexible de expansión de zona que tiene una sola instancia en la misma zona.

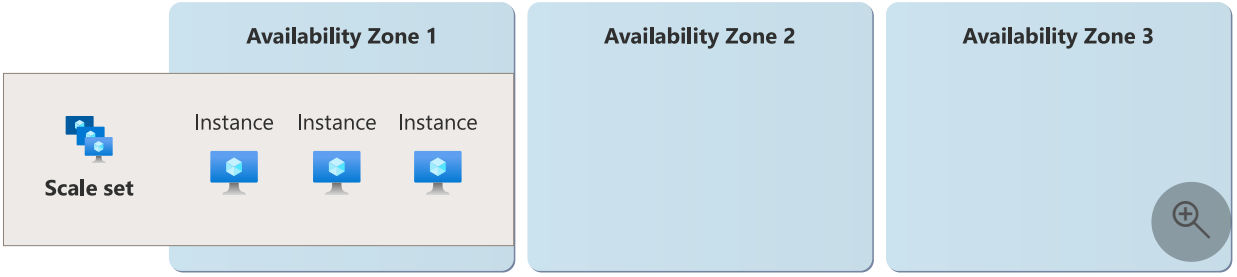
  > ⓘ **Nota:**
  >
  > Si usa el modo de orquestación flexible y adjunta, desasocia o quita máquinas virtuales individuales, asegúrese de que las máquinas virtuales se reparten entre

varias zonas. Si todas las máquinas virtuales están en una sola zona, es posible que el conjunto de escalado no sea resistente a una interrupción en dicha zona.

- Un conjunto de escala *zonal*, también llamado *alineado con la zona*, coloca todas sus instancias en una sola zona de disponibilidad que usted especifique. Cada máquina virtual y sus discos son zonales, por lo que se anclan a esa zona específica.

  > **ⓘ Importante**
  >
  > La asignación a una sola zona de disponibilidad solo se recomienda cuando la latencia entre zonas es demasiado alta para sus necesidades y después de comprobar que la latencia no cumple sus requisitos. Por sí mismo, un recurso zonal no proporciona resistencia a una interrupción de zona de disponibilidad. Para mejorar la resiliencia de un recurso zonal, debe desplegar explícitamente recursos independientes en múltiples zonas de disponibilidad y configurar el enrutamiento del tráfico y la conmutación por error. Para obtener más información, consulte Recursos zonales y resistencia de zona.

  En el diagrama siguiente se muestra un conjunto de escalado zonal en una sola zona, con tres instancias en la zona.



Si no especifica zonas de disponibilidad para el conjunto de escalado, es *no zonal* o *regional*. En este escenario, es posible que las instancias se coloquen en cualquier zona dentro de la región y no se distribuyan uniformemente ni se encuentren en la misma zona. Cuando se usa un conjunto de escalado no zonal, se garantiza la colocación de disco en la misma zona para discos Ultra y Premium v2. La coubicación se proporciona de la mejor manera posible para los discos Premium v1 y no se garantiza para los discos de SKU estándar, incluidos los discos de unidad de estado sólido (SSD) o de disco duro (HDD). Si se produce un error en alguna zona de la región, el conjunto de escalado podría experimentar tiempo de inactividad.

Requisitos

- **Soporte para regiones:** Puede implementar conjuntos de escalado que abarquen zonas y conjuntos zonales en cualquier región que soporte zonas de disponibilidad.

  Sin embargo, algunos tipos y tamaños de máquina virtual solo están disponibles en regiones específicas o zonas específicas dentro de una región. Para comprobar qué regiones y zonas admiten los tipos de máquina virtual que necesita, use los siguientes recursos:

  - Para comprobar los tipos de máquina virtual disponibles en cada región, consulte Productos disponibles por región.
  - Para comprobar los tamaños y tipos de máquina virtual admitidos dentro de cada zona de una región específica, consulte Comprobación de la disponibilidad de la SKU de máquina virtual.

  Si una SKU de máquina virtual específica no está disponible en ninguna de las zonas que seleccione para el conjunto de escalamiento, es posible que el conjunto de escalamiento no pueda escalar para satisfacer los requisitos de capacidad.

- **Hosts dedicados:** Las implementaciones de hosts dedicados de Azure no admiten conjuntos de escalado que abarcan zonas ni conjuntos de escalado zonal.

- **Tipos:** La compatibilidad con zonas de disponibilidad está disponible para todos los tipos de conjuntos de escalado, incluidos los conjuntos de escalado flexibles y uniformes.

Consideraciones

- **Propagación de dominios de error:** Cuando el conjunto de escalado usa zonas de disponibilidad, debe seleccionar entre enfoques específicos de propagación de dominios de error. Se recomienda usar la propagación máxima, que usa tantos dominios de error como sea posible, para la mayoría de las cargas de trabajo. Para obtener más información, consulte Elegir el Número Correcto de Dominios de Error para Virtual Machine Scale Sets.

- **Equilibrio de zona:** el equilibrio de zona determina si las instancias de máquina virtual de un conjunto de escalado se distribuyen uniformemente entre las zonas que seleccione. Un conjunto de escalado se considera equilibrado si cada zona tiene el mismo número de máquinas virtuales, más o menos una máquina virtual. Puede establecer el modo de balanceo de zona en mejor esfuerzo o estricto. Esta configuración controla si el conjunto de escalado puede escalar horizontalmente de forma desigual, incluidos los escenarios de interrupción de zona.

- **Grupos de selección de ubicación:** Para conjuntos de escalado uniformes, si configura varios grupos de selección de ubicación, Azure implementa varios grupos de selección de ubicación en cada zona que usa el conjunto de escalado.

Cost

No hay ninguna diferencia de costo entre un conjunto de escalado que abarca zonas, un conjunto zonal y uno no zonal que tengan el mismo número y tipo de instancias de máquinas virtuales.

Configurar soporte de zonas de disponibilidad

En esta sección se explica cómo configurar la compatibilidad con zonas de disponibilidad para el conjunto de escalas.

- **Cree un conjunto de escalado zonal o que abarque múltiples zonas.** Puede configurar zonas de disponibilidad al crear un nuevo conjunto de escalado. Para obtener más información, consulte Creación de un conjunto de escalado de máquinas virtuales que usa zonas de disponibilidad.

  > ⓘ **Nota:**
  >
  > Cuando selecciona qué zonas de disponibilidad usar, en realidad está seleccionando la *zona de disponibilidad lógica*. Si implementa otros componentes de la carga de trabajo en una suscripción de Azure diferente, podrían usar un número de zona de disponibilidad lógica distinto para acceder a la misma zona de disponibilidad física. Para obtener más información, consulte **zonas de disponibilidad física y lógica**.

- **Convierta los conjuntos de escalado existentes para usar zonas de disponibilidad.** Puede convertir un conjunto de escalado no zonal existente para usar zonas de disponibilidad. Para más información, consulte Actualización de conjuntos de escalado para agregar zonas de disponibilidad.

- **Cambie la configuración de zona de disponibilidad de un conjunto de escalado existente.** Puede agregar zonas a un conjunto de escalado existente, pero no puede quitar zonas. Para más información, consulte Actualización de conjuntos de escalado para agregar zonas de disponibilidad.

  > ⓘ **Importante**

> Al expandir un conjunto de escalado a más zonas, las instancias de máquina virtual originales no migran ni cambian inmediatamente. Al escalar, se crean nuevas instancias y se distribuyen uniformemente en las zonas de disponibilidad seleccionadas. Si necesita obtener datos de las instancias originales, será responsable de migrar esos datos a instancias en las nuevas zonas. Al escalar en el conjunto de escalado, las instancias regionales se priorizan primero para la eliminación. A continuación, las instancias se quitan en función de la directiva de escalado del conjunto de escalado. Para obtener más información, consulte **Cómo equilibrar manualmente el conjunto de escalado**.

Planeamiento y administración de capacidad

Para prepararse para errores de zona de disponibilidad, considere la *posibilidad de sobreaprovisionar* el número de instancias de máquina virtual del conjunto de escalado. Este enfoque permite a la solución tolerar cierta pérdida de capacidad y seguir funcionando sin degradar el rendimiento y garantiza que las zonas restantes tengan capacidad suficiente para controlar la carga de producción completa. Para obtener más información, consulte Administración de la capacidad mediante el aprovisionamiento excesivo.

Comportamiento cuando todas las zonas están en buen estado

En esta sección se describe qué esperar cuando los conjuntos de escalado están configurados con compatibilidad con la zona de disponibilidad y todas las zonas de disponibilidad están operativas.

- **Enrutamiento de tráfico entre zonas:** Es responsable del enrutamiento del tráfico entre máquinas virtuales del conjunto de escalado, incluidas las máquinas virtuales que se encuentran en distintas zonas de disponibilidad. Entre los enfoques comunes se incluyen Load Balancer y Application Gateway, que proporcionan integración integrada con conjuntos de escalado. Para obtener más información, consulte Redes para conjuntos de escalado de máquinas virtuales.

- **Replicación de datos entre zonas:** Es responsable de cualquier replicación de datos que necesite producirse entre máquinas virtuales, incluidas las máquinas virtuales de diferentes zonas de disponibilidad. Las bases de datos y otras aplicaciones con estado similares que se ejecutan en máquinas virtuales suelen proporcionar funcionalidades para replicar datos.

Comportamiento durante un fallo de zona

En esta sección se describe qué esperar cuando los conjuntos de escalado están configurados con soporte para zonas de disponibilidad y se produce una interrupción en estas.

- **Detección y respuesta:** Es responsable de detectar la pérdida de una zona de disponibilidad y decidir cómo responder.

  En el caso de los conjuntos de escalado que abarcan zonas, las instancias de máquina virtual en la zona afectada podrían no estar disponibles. Las instancias de las zonas saludables permanecen operativas.

  En el caso de los conjuntos de escalado zonales implementados en la zona afectada, es posible que todas las instancias de máquina virtual no estén disponibles. Debe planificar cómo responder a una falla de zona. Por ejemplo, puede redirigir el tráfico a otro conjunto de escalado en una zona o región diferente.

- **Notificación**: Microsoft no le notifica automáticamente cuando una zona está inactiva. Sin embargo, puede usar Azure Resource Health para supervisar el estado de un recurso individual y puede configurar alertas de Resource Health para notificarle problemas. También puede usar Azure Service Health para comprender el estado general del servicio, incluidos los errores de zona, y puede configurar alertas de Service Health para notificarle problemas.

- **Solicitudes activas:** Es probable que se finalicen las solicitudes activas u otro trabajo que se produzca en las máquinas virtuales de la zona de disponibilidad afectada.

- **Pérdida de datos esperada:** Es posible que los discos de máquina virtual zonales no estén disponibles durante un error de zona.

  Si usa discos de almacenamiento de zona redundante (ZRS) y una interrupción afecta a su máquina virtual, puede forzar la desasociación de los discos ZRS de la máquina virtual que ha fallado. Este enfoque permite conectar los discos ZRS a otra máquina virtual.

- **Tiempo de inactividad esperado:** Las máquinas virtuales de la zona afectada permanecen inactivas hasta que se recupere la zona de disponibilidad. Al usar conjuntos de escalado que abarcan múltiples zonas, las máquinas virtuales ubicadas en zonas saludables siguen funcionando.

- **Reenrutamiento del tráfico:** Es responsable de redirigir el tráfico a otras máquinas virtuales en zonas correctas.

  Si configura un equilibrador de carga resistente a zonas que realiza comprobaciones de estado, el equilibrador de carga suele detectar máquinas virtuales con errores y puede enrutar el tráfico a otras instancias de máquina virtual en zonas correctas.

- **Reemplazo de instancia:** No se garantiza que los conjuntos de escalado de máquinas virtuales agreguen automáticamente nuevas instancias a zonas saludables.

  Si tiene un conjunto de escalado que abarca zonas, puede escalar horizontalmente para agregar más instancias. Si el error de zona está restringido a conjuntos específicos de servidores dentro de la zona, la operación de escalabilidad horizontal podría agregar instancias correctas a la misma zona o podría agregar instancias a otras zonas. Sin embargo, si el conjunto de escalado usa un equilibrio de zona estricto, el conjunto de escalado bloquea las operaciones de escalado horizontal que provocan un desequilibrio.

  > **Sugerencia**
  >
  > Se recomienda configurar reglas de **escalado automático** en función del uso de CPU o memoria. Las reglas de escalado automático pueden permitir que el conjunto de escalado responda a una pérdida de las instancias de máquina virtual en una zona aumentando el escalado para agregar nuevas instancias en las zonas operativas restantes.

Recuperación de zona

Cuando la zona está en buen estado, las máquinas virtuales de la zona se reinician. Es responsable de cualquier procedimiento de recuperación de zona y sincronización de datos que requieran las cargas de trabajo.

Si agrega instancias temporales al grupo de escalado durante un fallo de zona, cuando se restaura la zona, es posible que tenga que reducir el tamaño del grupo de escalado a la capacidad original.

Prueba de fallos de zona

Puede usar Azure Chaos Studio para simular la pérdida de máquinas virtuales en una o varias zonas de disponibilidad como parte de un experimento. Chaos Studio proporciona errores integrados para conjuntos de escalado, incluida la capacidad de apagar máquinas virtuales en zonas específicas. Puede usar estas capacidades para simular fallos a nivel de zona y probar los procesos de conmutación por error.

Resistencia a errores en toda la región

Los conjuntos de escalado son recursos de una sola región. Si la región no está disponible, los conjuntos de escalado de la región tampoco están disponibles.

Soluciones personalizadas de varias regiones para la resistencia

Puede implementar varios conjuntos de escalado en diferentes regiones, pero debe implementar procesos de replicación, equilibrio de carga y conmutación por error. Por ejemplo, puede implementar conjuntos de escalado idénticos en varias regiones y usar Azure Front Door o Azure Traffic Manager, junto con sondeos de estado, para enrutar el tráfico. Es responsable de replicar el estado mediante mecanismos de aplicación o servicios de datos administrados.

Copias de seguridad y restauración

Azure Backup proporciona compatibilidad nativa con copias de seguridad para máquinas virtuales. Azure Backup crea y administra copias de seguridad y proporciona protección coherente con la aplicación para toda la máquina virtual, incluidos todos los discos conectados. Una solución de copia de seguridad de máquina virtual con Azure Backup es ideal cuando se necesita una copia de seguridad coordinada de varios discos o copias de seguridad compatibles con aplicaciones. Sin embargo, en el caso de las cargas de trabajo de base de datos, considere las soluciones de copia de seguridad específicas de la aplicación que proporcionan protección coherente con la transacción y opciones de recuperación más rápidas.

Con Azure Backup para vm, puede personalizar la frecuencia de copia de seguridad, la duración de retención y la configuración de almacenamiento para satisfacer sus necesidades. Para más información, consulte Azure Backup para VM.

La copia de seguridad también admite discos conectados a máquinas virtuales. Para obtener más información, consulte Información general sobre Azure Disk Backup.

Para la mayoría de las soluciones, no debe confiar exclusivamente en copias de seguridad. En su lugar, utilice las otras capacidades descritas en esta guía para apoyar los requisitos de resiliencia. Sin embargo, las copias de seguridad protegen contra algunos riesgos que otros enfoques no. Para más información, consulte ¿Qué son la redundancia, la replicación y la copia de seguridad?.

Resistencia a la reconfiguración de máquinas virtuales

Los conjuntos de escalado permiten controlar cómo se aplican los cambios de configuración a las máquinas virtuales, como cambiar la SKU de la máquina virtual, cambiar la imagen que usa

cada máquina virtual y agregar o quitar extensiones de máquina virtual. Puede controlar el *modo de directiva de actualización*, que determina cómo se aplican las actualizaciones. Para obtener más información, consulte modos de actualización de la política para conjuntos de máquinas virtuales a escala.

Algunos tipos de actualización requieren la reestructuración o la reimplementación de una instancia. Para excluir instancias específicas de las actualizaciones automáticas, considere la posibilidad de usar la protección de instancias. Puede excluir instancias que contengan estado que necesite conservar o configuración que no pueda replicar en otras instancias.

Resistencia al mantenimiento del servicio

Azure realiza periódicamente actualizaciones para mejorar la confiabilidad, el rendimiento y la seguridad de la infraestructura de host para las máquinas virtuales. Los conjuntos de escalado proporcionan varias maneras de comprender y controlar el mantenimiento planeado:

- Las notificaciones de mantenimiento planeado le indican cuándo se debe realizar el mantenimiento y le permiten controlar cuándo se produce el mantenimiento.
- Las configuraciones de mantenimiento permiten programar una ventana de mantenimiento a la vez que se adapte a sus necesidades empresariales.
- Scheduled Events for Linux VMs and for Windows VMs (Eventos programados para máquinas virtuales Linux y para máquinas virtuales Windows) proporciona tiempo a la aplicación para prepararse para el mantenimiento de máquinas virtuales. Proporciona información sobre los próximos eventos de mantenimiento, como un reinicio, para que la aplicación pueda prepararse para ellos y limitar la interrupción.

Acuerdo de nivel de servicio

El contrato de nivel de servicio (SLA) para los servicios de Azure describe la disponibilidad esperada de cada servicio y las condiciones que la solución deberá cumplir para lograr esa expectativa de disponibilidad. Para obtener más información, consulte Acuerdos de Nivel de Servicio para servicios en línea.

Los conjuntos de escalado de máquinas virtuales comparten el SLA de disponibilidad para las máquinas virtuales. Puede lograr un porcentaje de tiempo de actividad mayor para las máquinas virtuales mediante un conjunto de escalado que cumpla ambos criterios:

- El conjunto de escalado contiene dos o más instancias.
- El conjunto de escalado distribuye esas instancias en dos o más zonas de disponibilidad.

Contenido relacionado

- ¿Qué son las zonas de disponibilidad?
- ¿Qué son los conjuntos de escalado de máquinas virtuales?

Last updated on 14/02/2026

Confiabilidad en Azure Virtual Machines

Azure Virtual Machines proporciona recursos de proceso escalables a petición. Como servicio de infraestructura fundamental, está diseñado para ofrecer confiabilidad y disponibilidad de nivel empresarial para cargas de trabajo críticas.

Al usar Azure, relibilidad es una responsabilidad compartida. Microsoft proporciona una variedad de capacidades para admitir resiliencia y recuperación. Es responsable de comprender cómo funcionan esas funcionalidades dentro de todos los servicios que usa y de seleccionar las funcionalidades que necesita para cumplir los objetivos empresariales y los objetivos de tiempo de actividad.

En este artículo se describe cómo hacer que Virtual Machines sean resistentes a una variedad de posibles interrupciones y problemas, incluidos errores transitorios, interrupciones de zona de disponibilidad, interrupciones de regiones y mantenimiento del servicio. También se describe cómo puede usar copias de seguridad para recuperarse de otros tipos de problemas y se resalta cierta información clave sobre el acuerdo de nivel de servicio (SLA) de Virtual Machines.

> **ⓘ Importante**
>
> Al considerar la confiabilidad de una máquina virtual (VM), también debe tener en cuenta la confiabilidad de los discos, la infraestructura de red y las aplicaciones que se ejecutan en las máquinas virtuales. La mejora de la resistencia de la máquina virtual por sí sola podría tener un impacto limitado si los otros componentes no son igualmente resistentes. En función de los requisitos de resistencia, es posible que tenga que realizar cambios de configuración en varias áreas.

Recomendaciones de implementación de producción

Para obtener más información sobre cómo implementar máquinas virtuales para cumplir con los requisitos de confiabilidad de la solución y cómo la confiabilidad afecta a otros aspectos de la arquitectura, consulte Mejores prácticas de arquitectura para máquinas virtuales y conjuntos de escalado en el marco bien arquitectado de Azure.

Introducción a la arquitectura de confiabilidad

Las máquinas virtuales son la unidad de proceso fundamental en Azure, tanto si aprovisiona las máquinas virtuales usted mismo como si usa otros servicios de proceso Azure que aprovisionan y administran de forma transparente.

Una máquina virtual individual también se conoce como una *sola máquina virtual de instancia*. Se ejecuta en un host específico, que es un servidor físico. La mayoría de las máquinas virtuales comparten su host con otras máquinas virtuales.

Al crear las máquinas virtuales, puede influir en dónde se ejecutan dentro de la infraestructura subyacente. Normalmente, se toman decisiones de selección de ubicación en función de los requisitos de confiabilidad, latencia y aislamiento. Azure proporciona varias opciones de configuración que afectan a cómo se colocan las máquinas virtuales.

- **Region:** Puede seleccionar en qué Azure región que debe ejecutar la máquina virtual. Una región es un área geográfica que puede contener varios centros de datos, cada uno con un gran número de hosts.

- **Availability zone:**Availability zones son grupos de centros de datos físicamente independientes dentro de cada región de Azure. En regiones que admiten zonas de disponibilidad, puede seleccionar en qué zona se ejecuta la máquina virtual. Para obtener más información, consulte Resilience to availability zone failures (Resistencia a errores de zona de disponibilidad).

- **Availability sets:** Un conjunto de disponibilidad es una agrupación lógica de máquinas virtuales que permite a Azure comprender cómo se crea la aplicación para proporcionar redundancia y disponibilidad.

  Al usar conjuntos de disponibilidad, Azure distribuye un grupo de máquinas virtuales entre diferentes dominios de errores. Esta distribución minimiza el riesgo de errores de hardware localizados mediante la agrupación de máquinas virtuales que comparten una fuente de alimentación común y un conmutador de red.

  Los conjuntos de disponibilidad también pueden colocar diferentes máquinas virtuales en diferentes dominios de actualización, lo que controla cómo la plataforma de Azure gestiona la implementación de las actualizaciones de la plataforma. Mediante el uso de dominios de actualización, puede asegurarse de que solo se reinicia un subconjunto de las máquinas virtuales para las actualizaciones al mismo tiempo.

- **Grupos de colocación de proximidad:** Para las cargas de trabajo que necesitan lograr la menor latencia posible entre las máquinas virtuales, puede usar un grupo de proximidad

para asegurarse de que Azure coloque las máquinas virtuales físicamente cerca. Sin embargo, la ubicación de proximidad significa que una interrupción del centro de datos puede afectar a todas las máquinas virtuales del grupo. Para asegurar una fiabilidad elevada, es posible que tenga que aprovisionar varios grupos de colocación de proximidad en diferentes zonas de disponibilidad.

- **Dedicated hosts:** Puede usar Azure Dedicated Host para aprovisionar su propio servidor físico que ejecute una o varias máquinas virtuales, como para requisitos de cumplimiento estrictos. Sin embargo, al aprovisionar un host dedicado, una interrupción en su centro de datos puede afectar a todas las máquinas virtuales de ese host. Para lograr una alta confiabilidad, es posible que tenga que aprovisionar varios hosts dedicados en distintas zonas de disponibilidad.

Si crea un conjunto de máquinas virtuales que realizan funciones similares, considere la posibilidad de usar Azure Virtual Machine Scale Sets para crear y administrar las máquinas virtuales como un grupo. Los conjuntos de escalado también proporcionan más opciones de confiabilidad, como distribuir las máquinas virtuales entre varias zonas de disponibilidad.

Para obtener más información sobre la disponibilidad de las máquinas virtuales, consulte Opciones de disponibilidad para Virtual Machines.

Resistencia a errores transitorios

Los errores transitorios son errores breves e intermitentes en los componentes. Se producen con frecuencia en un entorno distribuido como la nube y son una parte normal de las operaciones. Los errores transitorios se corrigen después de un breve período de tiempo. Es importante que las aplicaciones puedan controlar errores transitorios, normalmente mediante el reintento de solicitudes afectadas.

Todas las aplicaciones hospedadas en la nube deben seguir las instrucciones de control de errores transitorios Azure cuando se comunican con cualquier API, bases de datos y otros componentes hospedados en la nube. Para obtener más información, consulte Recomendaciones para controlar errores transitorios.

Las aplicaciones que se ejecutan en las máquinas virtuales deben implementar estrategias adecuadas de control de errores para asegurarse de que las interrupciones temporales del servicio no afecten a la carga de trabajo.

Resistencia a errores de zona de disponibilidad

Availability zones son grupos de centros de datos físicamente independientes dentro de una región de Azure. Cuando una zona falla, los servicios pueden transferirse a una de las zonas restantes.

Una máquina virtual individual se puede implementar en una configuración *zonal* , lo que significa que está anclada a una sola zona de disponibilidad que seleccione. Por sí misma, una máquina virtual zonal no es resistente a interrupciones de zona. Sin embargo, puede crear varias máquinas virtuales y colocarlas en distintas zonas de disponibilidad y, a continuación, distribuir las aplicaciones y los datos en las instancias de máquina virtual. Como alternativa, puede usar conjuntos de escalado de máquinas virtuales para implementar un conjunto de máquinas virtuales en varias zonas de disponibilidad.

Si no configura una máquina virtual para que sea zonal, se considera *no zonal* o *regional*. Las máquinas virtuales no zonales se pueden colocar en cualquier zona de disponibilidad dentro de la región. Si alguna zona de disponibilidad de la región experimenta una interrupción, las máquinas virtuales nozonales podrían estar en la zona afectada y pueden experimentar tiempo de inactividad.

Requisitos

Compatibilidad con regiones: Las máquinas virtuales zonales se pueden implementar en cualquier región que admita zonas de disponibilidad.

Sin embargo, algunos tipos y tamaños de máquina virtual solo están disponibles en regiones específicas o zonas específicas dentro de una región. Para comprobar qué regiones y zonas admiten los tipos de máquina virtual que necesita, use los siguientes recursos:

- Para comprobar los tipos de máquina virtual disponibles en cada región, consulte Productos disponibles por región .
- Para comprobar los tamaños y tipos de máquina virtual admitidos dentro de cada zona de una región específica, consulte Comprobación de la disponibilidad de la SKU de máquina virtual.

Cost

No hay ninguna diferencia de costo entre una máquina virtual zonal y no zonal.

Configurar soporte de zonas de disponibilidad

En esta sección se explica cómo configurar la compatibilidad con la zona de disponibilidad para la instancia de máquina virtual.

> ⓘ **Nota:**
>
> Cuando selecciona qué zonas de disponibilidad usar, en realidad está seleccionando la *zona de disponibilidad lógica*. Si implementa otros componentes de carga de trabajo en una suscripción de Azure diferente, pueden usar un número de zona de disponibilidad lógica diferente para acceder a la misma zona de disponibilidad física. Para obtener más información, consulte **zonas de disponibilidad física y lógica**.

- **Cree una máquina virtual zonal.** Puede crear una máquina virtual zonal mediante las siguientes guías:
  - Crear máquinas virtuales en una zona de disponibilidad mediante el portal de Azure
  - Crear una máquina virtual en una zona de disponibilidad mediante el CLI de Azure
  - Crear una máquina virtual en una zona de disponibilidad mediante Azure PowerShell

- **Convierta las máquinas virtuales existentes en una configuración zonal.** Puede pasar de una máquina virtual no zonal a una máquina virtual zonal. Este proceso crea una nueva máquina virtual en la zona de disponibilidad de destino y requiere que la máquina virtual se detenga durante el proceso de traslado. Para obtener más información, consulte Mover máquinas virtuales de instancia única de Azure de zonas de disponibilidad regionales a zonales.

- **Cambie la zona de disponibilidad de una máquina virtual zonal existente.** Las máquinas virtuales zonales no se pueden mover a otra zona de disponibilidad. En su lugar, debe implementar una nueva máquina virtual zonal en la zona de disponibilidad deseada.

- **Convierta una máquina virtual zonal en una configuración no zonal.** Las máquinas virtuales zonales no se pueden convertir en una configuración no zonal. En su lugar, debe implementar una nueva máquina virtual no zonal.

Comportamiento cuando todas las zonas están en buen estado

En esta sección se describe qué esperar cuando las instancias de máquina virtual están configuradas con compatibilidad con la zona de disponibilidad y todas las zonas de disponibilidad están operativas.

- **Enrutamiento de tráfico entre zonas:** Es responsable del enrutamiento del tráfico entre máquinas virtuales, incluidas las máquinas virtuales que se encuentran en distintas zonas de disponibilidad. Entre los enfoques comunes se incluyen Azure Load Balancer y Azure Application Gateway. Para obtener más información, consulte Opciones de equilibrio de carga.

- **Replicación de datos entre zonas:** Es responsable de cualquier replicación de datos que necesite producirse entre máquinas virtuales, incluidas las máquinas virtuales de diferentes zonas de disponibilidad. Las bases de datos y otras aplicaciones con estado similares que se ejecutan en máquinas virtuales suelen proporcionar funcionalidades para replicar datos.

Comportamiento durante un fallo de zona

En esta sección se describe qué esperar cuando las instancias de máquina virtual están configuradas con compatibilidad con la zona de disponibilidad y hay una interrupción en sus zonas de disponibilidad.

- **Detección y respuesta:** Es responsable de detectar y responder a errores de zona que afectan a las máquinas virtuales.

- **Notification**: Microsoft no le notifica automáticamente cuando una zona está inactiva. Sin embargo, puede usar Azure Resource Health para supervisar el estado de un recurso individual y puede configurar Resource Health alertas para notificarle problemas. También puede usar Azure Service Health para comprender el estado general del servicio, incluidos los errores de zona, y puede configurar alertas de Service Health para notificarle problemas.

- **Solicitudes activas:** Es probable que se finalicen las solicitudes activas u otro trabajo que se produzca en la máquina virtual durante el error de zona.

- **Pérdida de datos esperada:** Es posible que los discos LRS zonales no estén disponibles durante un error de zona.

  Si usa discos de almacenamiento de zona redundante (ZRS) y una interrupción afecta a su máquina virtual, puede forzar la desasociación de los discos ZRS de la máquina virtual que ha fallado. Este enfoque permite conectar los discos ZRS a otra máquina virtual.

  Para obtener más información sobre cómo hacer que los discos sean resistentes a errores de zona de disponibilidad y otros problemas, consulte Reliability in Azure Disk Storage.

- **Tiempo de inactividad esperado:** Las máquinas virtuales permanecen inactivas hasta que se recupera la zona de disponibilidad.
- **Reenrutamiento del tráfico:** Es responsable de redirigir el tráfico a otras máquinas virtuales en zonas correctas.

  Si configura un equilibrador de carga resistente a zonas y realiza comprobaciones de estado, el equilibrador de carga normalmente detecta máquinas virtuales con errores y puede enrutar el tráfico a otras instancias de máquina virtual en zonas correctas.

Recuperación de zona

Cuando la zona está en buen estado, las máquinas virtuales de la zona se reinician. Es responsable de cualquier procedimiento de recuperación de zona y sincronización de datos que requieran las cargas de trabajo.

Prueba de fallos de zona

Puede usar Azure Chaos Studio para simular la pérdida de una máquina virtual como parte de un experimento. Chaos Studio proporciona errores integrados para máquinas virtuales, incluida la capacidad de apagar una máquina virtual. Puede usar estas capacidades para simular fallos a nivel de zona y probar los procesos de conmutación por error.

Soluciones personalizadas multizona para la resiliencia

Al implementar múltiples máquinas virtuales en distintas zonas, usted es responsable de configurar y administrar la replicación, el equilibrio de carga, la conmutación por error y los procesos de conmutación por recuperación.

Algunas aplicaciones proporcionan funcionalidades integradas que pueden ayudar al implementar en varias máquinas virtuales. Por ejemplo, SQL Server en máquinas virtuales Azure proporciona un conjunto de funcionalidades para simplificar los procesos de configuración y administración en todas las zonas de disponibilidad.

Puede considerar el uso de Azure Site Recovery recuperación ante desastres de zona a zona (DR) cuando su aplicación se ejecuta en una sola zona a la vez y no requiere un failover casi instantáneo entre zonas. La recuperación ante desastres de zona a zona tiene algunas limitaciones importantes, por lo que debe revisar los requisitos exhaustivamente.

Resistencia a errores en toda la región

Las máquinas virtuales son recursos de una sola región. Si la región deja de estar disponible, la máquina virtual tampoco está disponible.

Soluciones personalizadas de varias regiones para la resistencia

Puede implementar varias máquinas virtuales en diferentes regiones, pero debe implementar procesos de replicación, equilibrio de carga y conmutación por error.

Site Recovery es un servicio que permite la recuperación ante desastres mediante la replicación de máquinas virtuales y sus datos en una región secundaria. Puede seleccionar casi cualquier región de Azure como región secundaria, incluidas combinaciones de regiones no emparejadas. Para obtener más información, consulte la arquitectura de recuperación ante desastres de Azure a Azure.

Algunas aplicaciones crean clústeres u otras construcciones para replicar datos y distribuir el trabajo entre varias máquinas virtuales, incluidas en regiones diferentes. Estas aplicaciones pueden simplificar la configuración de una solución de varias regiones.

Para obtener una arquitectura de ejemplo que ilustra cómo usar máquinas virtuales en varias regiones, consulte Equilibrio de carga en varias regiones con Azure Traffic Manager, Azure Firewall y Application Gateway.

Resistencia al mantenimiento del servicio

Azure realiza un mantenimiento periódico normal en máquinas virtuales para garantizar la confiabilidad. Hay varias maneras de asegurarse de que las cargas de trabajo permanecen operativas durante las actividades de mantenimiento:

- Al usar conjuntos de disponibilidad o conjuntos de escalado de máquinas virtuales, puede configurar dominios de actualización. Los dominios de actualización ayudan a distribuir las actividades de mantenimiento en diferentes máquinas virtuales en momentos diferentes, por lo que las máquinas virtuales no se reinician simultáneamente.
- Puede personalizar el tiempo que se aplica el mantenimiento a las máquinas virtuales mediante *el control de mantenimiento*. Puede usar configuraciones de mantenimiento para programarla en un momento que se adapte a la carga de trabajo.
- Puede recibir notificaciones de las próximas actividades de mantenimiento.

Para obtener más información, consulte Información general sobre las actualizaciones de invitado y el mantenimiento del host.

Copias de seguridad y restauración

Azure Backup proporciona compatibilidad nativa de copia de seguridad con máquinas virtuales. Azure Backup crea y administra copias de seguridad y proporciona protección coherente con la aplicación para toda la máquina virtual, incluidos todos los discos conectados. Una solución de copia de seguridad de máquina virtual con Azure Backup es ideal cuando se necesita una copia de seguridad coordinada de varios discos o copias de seguridad compatibles con aplicaciones. Sin embargo, en el caso de las cargas de trabajo de base de datos, considere las soluciones de copia de seguridad específicas de la aplicación que proporcionan protección coherente con la transacción y opciones de recuperación más rápidas.

Con Azure Backup para máquinas virtuales, puede personalizar la frecuencia de copia de seguridad, la duración de retención y la configuración de almacenamiento para satisfacer sus necesidades. Para obtener más información, consulte Azure Backup for VMs.

La copia de seguridad también admite discos conectados a máquinas virtuales. Para obtener más información, vea Overview of Azure Disk Backup.

Para la mayoría de las soluciones, no debe confiar exclusivamente en copias de seguridad. En su lugar, utilice las otras capacidades descritas en esta guía para apoyar los requisitos de resiliencia. Sin embargo, las copias de seguridad protegen contra algunos riesgos que otros enfoques no. Para más información, consulte ¿Qué son la redundancia, la replicación y la copia de seguridad?.

Acuerdo de nivel de servicio

El acuerdo de nivel de servicio (SLA) para Azure servicios describe la disponibilidad esperada de cada servicio y las condiciones que la solución debe cumplir para lograr esa expectativa de disponibilidad. Para obtener más información, vea SLAs for servicios en línea.

Para Virtual Machines, el Acuerdo de Nivel de Servicio proporciona un nivel base de disponibilidad. El porcentaje de tiempo de actividad definido en el Acuerdo de Nivel de Servicio aumenta cuando tiene dos o más máquinas virtuales y realiza las siguientes acciones:

- Configure esas máquinas virtuales para que se implementen en dos o más zonas de disponibilidad.

- Configure esas máquinas virtuales para su implementación en un grupo de disponibilidad.

Pasos siguientes

- VMs en Azure
- Conjuntos de Escala de Máquinas Virtuales

Recurso relacionado

- Reliability in Azure Disk Storage
- Fiabilidad en Azure

Last updated on 08/04/2026

Responsabilidad compartida para la confiabilidad

En la plataforma de nube pública Azure, la confiabilidad es una responsabilidad compartida entre Microsoft y usted. Dado que hay diferentes niveles de confiabilidad en cada carga de trabajo que diseñe e implemente, es importante que comprenda quién tiene la responsabilidad principal de cada uno de esos niveles desde una perspectiva de resistencia.

Para ayudarle a comprender mejor cómo funciona la responsabilidad compartida, especialmente cuando se enfrenta a una interrupción o desastre, en este artículo se describe el modelo de responsabilidad compartida para la confiabilidad. Para obtener más información sobre cómo usar realmente este modelo para planear la recuperación ante desastres, consulte Recomendaciones para diseñar una estrategia de recuperación ante desastres.

Modelo de responsabilidad compartida para la confiabilidad

El modelo de responsabilidad compartida para la confiabilidad se compone de tres niveles:

- Confiabilidad de la plataforma principal. La plataforma Azure proporciona un nivel base de confiabilidad para todos los clientes y todos los servicios a través de la infraestructura, los servicios y los procesos subyacentes.
- Funcionalidades de mejora de la disponibilidad Azure ofrece un conjunto de características y servicios integrados que mejoran la confiabilidad, como el uso de zonas de disponibilidad, la implementación en varias regiones y la implementación de estrategias de copia de seguridad. Aunque Azure proporciona estas funcionalidades, es su responsabilidad evaluarlas y configurarlas para que se alineen con sus requisitos específicos. Los requisitos pueden incluir confiabilidad, costo, rendimiento y cumplimiento de los estándares normativos.
- Aplicaciones. Para hacer un uso eficaz de los demás niveles, la aplicación y la carga de trabajo deben diseñarse para la confiabilidad.

Diagrama que muestra el modelo de responsabilidad compartida para la confiabilidad: confiabilidad de la plataforma principal, funcionalidades de mejora de confiabilidad y aplicaciones.

Microsoft es el único responsable de la confiabilidad de la plataforma principal. Microsoft también es responsable de proporcionar funcionalidades de mejora de confiabilidad que puede usar. Usted es responsable de seleccionar y usar los componentes adecuados.

La elección de las categorías de servicio SaaS, PaaS o IaaS determina el tipo de decisiones que debe tomar. Por ejemplo, si usa un servicio SaaS, normalmente no es necesario participar en el uso de zonas de disponibilidad. Si usa servicios PaaS para el nivel de datos, es posible que tenga funcionalidades automatizadas para la copia de seguridad a su disposición. Si usa servicios IaaS, normalmente debe planear e implementar muchas funcionalidades de confiabilidad usted mismo.

> ⓘ **Nota:**
>
> Las categorías de servicio (SaaS, PaaS e IaaS) son útiles como una amplia agrupación de servicios, pero es importante comprender sus responsabilidades para cada servicio individual que use.
>
> Las guías de confiabilidad proporcionan información general sobre cómo funciona cada servicio desde una perspectiva de confiabilidad y le ayudan a tomar decisiones fundamentadas sobre cómo configurar los servicios para satisfacer sus necesidades.

También es responsable del diseño de la aplicación y de la carga de trabajo, y de definir los requisitos de confiabilidad, lo que le ayuda a decidir cómo diseñar y configurar la solución.

Confiabilidad de la plataforma principal

La plataforma Microsoft Cloud consta de una gran cantidad de infraestructura, hardware, software y procesos para admitir la implementación y administración del servicio. Cada componente está diseñado para ser altamente resistente, con varias redundancias para hardware y con procesos de software basados en investigación. Juntos, estos componentes componen el nivel de confiabilidad de la plataforma principal. Algunos ejemplos de cómo Microsoft proporciona una plataforma confiable incluyen lo siguiente:

- Las redes tienen vínculos redundantes y pueden omitir dinámicamente segmentos defectuosos.
- Dentro de cada región, los centros de datos se conectan a través de una red de baja latencia, lo que permite diversos enfoques de replicación de datos.
- Las instalaciones del centro de datos tienen conexiones de alimentación, refrigeración y red redundantes. Son operados por equipos in situ que los aseguran, supervisan y administran.
- El hardware, incluidos los clústeres y bastidores, tienen redundancia en varias capas.
- Las actualizaciones de clústeres de computación, bastidores y servidores siguen un proceso controlado. Usamos técnicas como hotpatching para reducir o eliminar el impacto en los hosts.

- Las actualizaciones y los cambios de configuración de la plataforma de software se aplican siguiendo nuestros procedimientos de implementación seguros.
- Microsoft audita proveedores externos críticos para asegurarse de que una interrupción de terceros no interrumpe los servicios de Azure.
- Cada servicio Azure debe tener un plan de recuperación ante desastres detallado. Realizamos exploraciones en profundidad de toda la región en regiones que coinciden con entornos de producción.

Todos los servicios Azure se benefician de estas funcionalidades principales de confiabilidad de la plataforma y con las mejoras en curso que Realiza Microsoft.

Funcionalidades de mejora de confiabilidad

Azure proporciona muchas funcionalidades de mejora de confiabilidad diferentes. Aunque Microsoft es responsable de proporcionar estas capacidades, usted es completamente responsable de seleccionar y usar las apropiadas para sus necesidades. Algunos ejemplos de estas funcionalidades son:

- **Regions.** Azure tiene más de 70 regiones y puede usar varias regiones en una única solución para lograr redundancia geográfica, satisfacer las necesidades de residencia de datos y habilitar la comunicación de baja latencia a los usuarios globalmente. Para obtener más información sobre las regiones, consulte ¿Qué son las regiones de Azure?.

- **Zonas de disponibilidad.** Muchas Azure regiones admiten zonas de disponibilidad, lo que le permite distribuir las cargas de trabajo entre varios conjuntos de centros de datos independientes. Azure servicios admiten zonas de disponibilidad de una manera que se adapte a su propósito previsto, normalmente admitiendo implementaciones zonales (ancladas a una sola zona) o implementaciones con redundancia de zona (distribuidas entre varias zonas). Para obtener más información sobre las zonas de disponibilidad, consulte ¿Qué son las zonas de disponibilidad?.

- **Niveles de servicio.** Los servicios proporcionan una variedad de ofertas y niveles que se ajustan a diferentes requisitos. Usted es responsable de elegir el nivel adecuado en función de sus objetivos de confiabilidad y otras necesidades. Por ejemplo, al crear una máquina virtual, puede elegir entre un disco Estándar, que proporciona una opción de bajo costo o un disco Premium para lograr un mayor nivel de disponibilidad.

- **Copias de seguridad.** Muchos Azure servicios que almacenan datos admiten copias de seguridad, que pueden ser automáticas, manuales o ambas. Con las copias de seguridad, puede proteger la carga de trabajo contra interrupciones, así como datos dañados y otros eventos de pérdida de datos. Es responsable de comprobar si las copias de seguridad están habilitadas y para configurarlas correctamente.

- **Gobernanza.** Las funcionalidades de plataforma, como Azure Policy, el control de acceso basado en roles y las funcionalidades de protección de identidades de Microsoft Entra ID, se pueden configurar para aplicar los requisitos de la organización de forma coherente. Con estos enfoques, puede proteger las cargas de trabajo frente a incidentes de seguridad y cambios accidentales que podrían provocar tiempo de inactividad u otros problemas con la carga de trabajo. Microsoft proporciona estas herramientas de gobernanza, pero es responsable de configurarlas y mantenerlas.

> **Importante**
>
> Es importante comprender los acuerdos de nivel de servicio *service* (SLA) para cada servicio Azure. Los SLA proporcionan información importante sobre el tiempo de actividad previsto del servicio y las condiciones que debe cumplir para poder acogerse a ellos. Usted es responsable de comprender y cumplir estas condiciones; Microsoft no supervisa ni aplica su idoneidad.
>
> Para obtener el SLA para cada servicio, consulte Acuerdos de Nivel de Servicio (SLA) para servicios en línea.

APLICACIONES

Es responsabilidad suya asegurarse de que las aplicaciones están diseñadas para ser resistentes a errores y seguir otros procedimientos recomendados de confiabilidad. Use los pilares Azure Well-Architected Framework para impulsar la excelencia arquitectónica en el nivel fundamental de una carga de trabajo. El pilar de confiabilidad se centra en cómo hacer que la carga de trabajo y las aplicaciones sean resistentes a distintos tipos de fallos y permitir la recuperación cuando se producen.

Pasos siguientes

El modelo de responsabilidad compartida se aplica a otras partes de la solución más allá de la confiabilidad. Para obtener más información sobre el modelo de responsabilidad compartida para la seguridad, consulte el Centro de confianza de Microsoft.

Last updated on 09/03/2026

¿Qué son la continuidad empresarial, la alta disponibilidad y la recuperación ante desastres?

En este artículo se definen y describen la continuidad empresarial y el planeamiento de la continuidad empresarial en términos de administración de riesgos mediante el diseño de alta disponibilidad y recuperación ante desastres. Aunque este artículo no proporciona instrucciones explícitas sobre cómo satisfacer sus propias necesidades de continuidad empresarial, le ayuda a comprender los conceptos que se usan en las instrucciones de confiabilidad de Microsoft.

Continuidad empresarial es el estado en el que una empresa puede continuar las operaciones durante errores, interrupciones o desastres. La continuidad empresarial requiere planificación proactiva, preparación e implementación de sistemas y procesos resistentes.

La planeación de la continuidad empresarial requiere identificar, comprender, clasificar y administrar riesgos. En función de comprender los riesgos y sus probabilidades, puede diseñar un plan de continuidad empresarial saludable para lograr una estrategia de alta disponibilidad (HA) y recuperación ante desastres (DR) que se adapte a sus necesidades empresariales.

La *alta disponibilidad* consiste en diseñar una solución que resista a los problemas diarios y que satisfaga las necesidades empresariales de disponibilidad.

La *recuperación ante desastres* consiste en planear cómo tratar riesgos poco comunes y las interrupciones catastróficas que puede haber como resultado.

Continuidad empresarial

En general, las soluciones en la nube están vinculadas directamente a las operaciones empresariales. El hecho de que una solución en la nube no esté disponible o experimente un problema grave puede afectar gravemente a las operaciones comerciales. Una consecuencia grave puede interrumpir la continuidad empresarial.

Entre las consecuencias graves en la continuidad empresarial, se incluyen las siguientes:

- Pérdida de ingresos empresariales.
- Imposibilidad de proporcionar un servicio importante a los usuarios.
- Incumplimiento de un compromiso con un cliente u otra parte.

Es importante comprender y comunicar las expectativas empresariales y las consecuencias de los errores a las partes interesadas importantes, incluidas aquellas que diseñen, implementen y

gestionen la carga de trabajo. Esas partes interesadas responden, entonces, compartiendo los costes que implica hacer realidad esa visión. Normalmente hay un proceso de negociación y revisiones de esa visión en función del presupuesto y otras restricciones.

Planeamiento de la continuidad empresarial

Para controlar o evitar completamente un impacto negativo en la continuidad empresarial, es importante crear de forma proactiva un *plan de continuidad empresarial*. Un plan de continuidad empresarial se basa en la evaluación de riesgos y en el desarrollo de métodos de control de esos riesgos a través de diversos enfoques. Los riesgos y enfoques específicos que se debe mitigar varían para cada organización y carga de trabajo.

Un plan de continuidad empresarial no solo tiene en cuenta las características de resistencia de la propia plataforma en la nube, sino también las características de la aplicación. Un plan de continuidad empresarial sólido también incorpora todos los aspectos de soporte técnico en la empresa, como personas, procesos manuales o automatizados relacionados con el negocio, y otras tecnologías.

El planeamiento de continuidad empresarial debe incluir los siguientes pasos secuenciales:

1. **Clasificación del nivel de importancia crítica**. Las cargas de trabajo se pueden clasificar en diferentes *niveles de importancia crítica* en función de su importancia para la empresa. Cada nivel tiene requisitos diferentes para la disponibilidad y, por tanto, requisitos diferentes para el planeamiento de la continuidad empresarial. Para determinar el nivel crítico de la carga de trabajo, consulte Well-Architected Framework: seleccione el nivel de importancia crítica.

2. **Identificación de riesgos**. Identifique los riesgos para la disponibilidad o la funcionalidad de una carga de trabajo. Los posibles riesgos podrían ser problemas de red, errores de hardware, errores humanos, interrupción de la región, etc. Comprenda el impacto de cada riesgo.

3. **Clasificación de riesgos**. Clasifique cada riesgo como un riesgo común, que se debe tener en cuenta en los planes de alta disponibilidad, o en un riesgo poco común, que debe formar parte del planeamiento de la recuperación ante desastres.

4. **Mitigación de riesgos**. Diseñe estrategias de mitigación de alta disponibilidad o recuperación ante desastres para minimizar o mitigar los riesgos, como la redundancia, la replicación, la conmutación por error y las copias de seguridad. Además, considere las mitigaciones y controles no técnicos y basados en procesos.

El planeamiento de la continuidad empresarial es un proceso, no un evento único. Cualquier plan de continuidad empresarial que se cree se debe revisar y actualizar periódicamente para

asegurarse de que sigue siendo relevante y eficaz, y que admite las necesidades empresariales actuales.

Identificación de riesgos

La fase inicial del planeamiento de la continuidad empresarial es identificar los riesgos para la disponibilidad o la funcionalidad de una carga de trabajo. Cada riesgo debe analizarse para comprender su probabilidad y su gravedad. La gravedad debe incluir cualquier posible tiempo de inactividad o pérdida de datos, así como si algún aspecto del resto del diseño de la solución podría compensar los efectos negativos.

La siguiente tabla es una lista no exhaustiva de riesgos, ordenados por probabilidad decreciente:

Expandir tabla

Riesgo de ejemplo	Description	Regularidad (probabilidad)
Problema de red transitorio	Error temporal en un componente de la pila de red, que se puede recuperar después de un breve tiempo (normalmente unos segundos o menos).	Normal
Reinicio de máquina virtual	Reinicio de una máquina virtual que usa usted o un servicio dependiente. Es posible que se produzcan reinicios porque la máquina virtual se bloquea o necesita aplicar una revisión.	Normal
Error de hardware	Error de un componente dentro de un centro de datos, como un nodo de hardware, un bastidor o un clúster.	Ocasional
Interrupción del centro de datos	Una interrupción que afecta a la mayoría o a todos los centros de datos, como un error de alimentación, un problema de conectividad de red o problemas con la calefacción y la refrigeración.	Inusual
Interrupción de la región	Una interrupción que afecta a toda una zona metropolitana o a una zona más amplia, como un desastre natural importante.	Muy inusual

El planeamiento de continuidad empresarial no es solo sobre la plataforma en la nube y la infraestructura. Es importante tener en cuenta el riesgo de errores humanos. Además, algunos riesgos que tradicionalmente podrían considerarse de seguridad, rendimiento o riesgos operativos también deben considerarse riesgos de confiabilidad porque afectan a la disponibilidad de la solución.

Estos son algunos ejemplos:

Expandir tabla

Riesgo de ejemplo	Description
Pérdida de datos o daños	Los datos se han eliminado, sobrescrito o dañado por un accidente o una vulneración de seguridad, como un ataque de ransomware.
Error de software	Una implementación de código nuevo o actualizado presenta un error que afecta a la disponibilidad o integridad, lo que deja la carga de trabajo en un estado que no funciona correctamente.
Implementaciones con errores	Error en la implementación de un nuevo componente o versión, lo que deja la solución en un estado incoherente.
Ataques por denegación de servicio	El sistema se ha atacado en un intento de evitar el uso legítimo de la solución.
Administradores no autorizados	Un usuario con privilegios administrativos ha realizado intencionadamente una acción perjudicial contra el sistema.
Entrada inesperada del tráfico a una aplicación	Un pico en el tráfico ha sobrecargado los recursos del sistema.

El *Análisis del modo de error* (FMA) es el proceso de identificar posibles formas en las que se puede producir un error en una carga de trabajo o sus componentes y cómo se comporta la solución en esas situaciones. Para obtener más información, consulte Recomendaciones para realizar el análisis del modo de error.

Clasificación de riesgos

Los planes de continuidad empresarial deben abordar riesgos comunes y poco comunes.

- *Se planean y se esperan riesgos comunes*. Por ejemplo, en un entorno en la nube es habitual que haya *errores transitorios* o *fallos*, incluidas interrupciones breves de red, reinicios del equipo debido a actualizaciones, tiempos de espera cuando un servicio está ocupado, etc. Dado que estos eventos se producen periódicamente, las cargas de trabajo deben poder superarlos.

  Una estrategia de alta disponibilidad debe tener en cuenta y controlar cada riesgo de este tipo.

- Los *riesgos poco comunes* suelen ser el resultado de un evento imprevisible, como un desastre natural o un ataque de red importante, que puede provocar una interrupción grave.

  Los procesos de recuperación ante desastres se ocupan de estos riesgos poco frecuentes.

La alta disponibilidad y la recuperación ante desastres están relacionadas entre sí, por lo que es importante planear estrategias para ambas a la vez.

La clasificación de riesgos depende de la arquitectura de la carga de trabajo y de los requisitos empresariales, y algunos riesgos se pueden clasificar como alta disponibilidad para una carga de trabajo y recuperación ante desastres para otra carga de trabajo. Por ejemplo, una interrupción completa de la región de Azure generalmente se consideraría un riesgo de recuperación ante desastres en las cargas de trabajo de esa región. Pero, para las cargas de trabajo que usan varias regiones de Azure en una configuración activa-activa con replicación completa, redundancia y conmutación automática por error de región, una interrupción de región se clasifica como un riesgo de alta disponibilidad.

Mitigación de riesgos

La mitigación de riesgos consiste en desarrollar estrategias de alta disponibilidad o recuperación ante desastres para minimizar o mitigar los riesgos de continuidad empresarial. La mitigación de riesgos puede basarse en la tecnología o el componente humano.

Mitigación de riesgos basados en la tecnología

La mitigación de riesgos basada en la tecnología usa controles de riesgo basados en cómo se implementa y configura la carga de trabajo; por ejemplo:

- Redundancia
- Replicación de datos
- Failover
- Backups

Los controles de riesgo basados en tecnología deben tenerse en cuenta dentro del contexto del plan de continuidad empresarial.

Por ejemplo:

- *Requisitos de tiempo de inactividad bajo.* Algunos planes de continuidad empresarial no pueden tolerar ninguna forma de riesgo de tiempo de inactividad debido a estrictos requisitos de alta disponibilidad. Hay ciertos controles basados en tecnología que pueden requerir tiempo para que un humano reciba la notificación al respecto y luego responda. Es probable que los controles de riesgo basados en tecnología que incluyan procesos manuales lentos no se ajusten a la inclusión en su estrategia de mitigación de riesgos.

- *Tolerancia a errores parciales.* Algunos planes de continuidad empresarial pueden tolerar un flujo de trabajo que se ejecuta en un *estado degradado*. Cuando una solución funciona en un estado degradado, es posible que algunos componentes estén deshabilitados o no

funcionen, pero las operaciones empresariales principales pueden seguir realizándose. Para más información, consulte Recomendaciones para la recuperación y conservación automáticas.

Mitigación de riesgos basada en humanos

La mitigación de riesgos basada en humanos usa controles de riesgo basados en procesos empresariales, como el siguiente:

- Desencadenar un cuaderno de estrategias de respuesta.
- Revertir a las operaciones manuales.
- Entrenamiento y cambios culturales.

> **Importante**
>
> Las personas que diseñan, implementan, gestionan y desarrollan la carga de trabajo deben ser competentes, animarse a hablar si tienen inquietudes y tener un sentido de responsabilidad por el sistema.

Dado que los controles de riesgo basados en humanos suelen ser más lentos que los controles basados en tecnología y más propensos a errores humanos, un buen plan de continuidad empresarial debe incluir un proceso de control de cambios formal para cualquier cosa que modifique el estado del sistema en ejecución. Por ejemplo, considere la posibilidad de implementar los siguientes procesos:

- Pruebe rigurosamente las cargas de trabajo de acuerdo con la importancia de la carga de trabajo. Para evitar problemas relacionados con los cambios, asegúrese de probar los cambios realizados en la carga de trabajo.
- Introduzca puertas de calidad estratégicas como parte de las prácticas de implementación seguras de la carga de trabajo. Para más información, consulte recomendaciones de procedimientos de implementación seguros.
- Formalice los procedimientos para el acceso a producción ad hoc y la manipulación de datos. Estas actividades, independientemente de lo menores que sean, pueden presentar un alto riesgo de causar incidentes de confiabilidad. Los procedimientos pueden incluir trabajar en conjunto con otro ingeniero, usar listas de verificación y obtener revisiones de pares antes de ejecutar scripts o aplicar cambios.

Alta disponibilidad

La alta disponibilidad es el estado en el que una carga de trabajo específica puede mantener su nivel de tiempo de actividad necesario día a día, incluso durante errores transitorios y

errores intermitentes. Dado que estos eventos se producen periódicamente, es importante que cada carga de trabajo esté diseñada y configurada para la alta disponibilidad de acuerdo con los requisitos de las expectativas específicas de la aplicación y del cliente. La alta disponibilidad de cada carga de trabajo contribuye al plan de continuidad empresarial.

Dado que la alta disponibilidad puede variar con cada carga de trabajo, es importante comprender los requisitos y las expectativas del cliente al determinar la alta disponibilidad. Por ejemplo, una aplicación que su organización usa para pedir suministros de oficina podría requerir un nivel de tiempo de actividad relativamente bajo, mientras que una aplicación financiera crítica podría requerir un tiempo de actividad mucho mayor. Incluso dentro de una carga de trabajo, diferentes *flujos* pueden tener requisitos distintos. Por ejemplo, en una aplicación de comercio electrónico, los flujos que admiten la exploración y la realización de pedidos pueden ser más importantes que los flujos de procesamiento de pedidos y de back-office. Para obtener más información sobre los flujos, consulte recomendaciones para identificar y clasificar flujos.

Normalmente, el tiempo de actividad se mide en función del número de "nueves" en el porcentaje de tiempo de actividad. El porcentaje de tiempo de actividad se relaciona con la cantidad de tiempo de inactividad que se permite durante un período de tiempo determinado. Estos son algunos ejemplos:

- Un requisito de tiempo de actividad del 99,9 % (tres nueves) permite aproximadamente 43 minutos de tiempo de inactividad en un mes.
- Un requisito de tiempo de actividad del 99,95 % (tres nueves y medio) permite aproximadamente 21 minutos de tiempo de inactividad en un mes.

Cuanto mayor sea el requisito de tiempo de actividad, menor tolerancia tendrá a las interrupciones y más trabajo tendrá que realizar para alcanzar ese nivel de disponibilidad. El tiempo de actividad no se mide por el tiempo de actividad de un solo componente como un nodo, sino por la disponibilidad general de toda la carga de trabajo.

> **ⓘ Importante**
>
> No diseñe en exceso su solución para alcanzar niveles de confiabilidad más altos que los justificados. Use los requisitos empresariales para guiar sus decisiones.

Alta disponibilidad: diseño

Para lograr los requisitos de alta disponibilidad, una carga de trabajo puede incluir una serie de elementos de diseño. Algunos de los elementos comunes se enumeran y se describen a continuación en esta sección.

> ⓘ **Nota:**
>
> Algunas cargas de trabajo son *críticas*, lo que significa que cualquier tiempo de inactividad puede tener consecuencias graves para la vida y la seguridad humanas, o pérdidas financieras importantes. Si va a diseñar una carga de trabajo crítica, hay cosas específicas que debe tener en cuenta al diseñar la solución y administrar la continuidad empresarial. Para obtener más información, consulte **Marco de buena arquitectura de Azure: cargas de trabajo críticas**.

Servicios y niveles de Azure que admiten alta disponibilidad

Muchos servicios de Azure están diseñados para ser de alta disponibilidad y se pueden usar para crear cargas de trabajo de alta disponibilidad. Estos son algunos ejemplos:

- Azure Virtual Machine Scale Sets proporcionan alta disponibilidad para máquinas virtuales (VM) mediante la creación y administración automáticas de instancias de máquina virtual y la distribución de esas instancias de máquina virtual para reducir el impacto de los errores de infraestructura.
- Azure App Service proporciona alta disponibilidad a través de una variedad de enfoques, incluido mover automáticamente los trabajos de un nodo incorrecto a un nodo correcto y proporcionar funcionalidades para la recuperación automática de muchos tipos de errores comunes.

Utilice cada guía de confiabilidad del servicio para comprender las capacidades del servicio, decidir qué niveles usar y determinar qué capacidades incluir en su estrategia de alta disponibilidad.

Revise los acuerdos de nivel de servicio (SLA) de cada servicio para comprender los niveles de disponibilidad esperados y las condiciones que necesita cumplir. Es posible que tenga que seleccionar o evitar niveles específicos de servicios para lograr determinados niveles de disponibilidad. Algunos servicios de Microsoft se ofrecen con el entendimiento de que no se proporciona ningún SLA, como los niveles de desarrollo o básicos, o que el recurso podría recuperarse de su sistema en ejecución, como las ofertas basadas en puntos puntuales. Además, algunos niveles han agregado características de confiabilidad, como la compatibilidad con zonas de disponibilidad.

Tolerancia a errores

La tolerancia a errores es la capacidad de un sistema de seguir funcionando, en alguna capacidad definida, en caso de error. Por ejemplo, una aplicación web podría diseñarse para seguir funcionando incluso aunque se produzca un error en un único servidor web. La

tolerancia a errores se puede lograr mediante redundancia, conmutación por error, creación de particiones, degradación correcta y otras técnicas.

La tolerancia a errores también requiere que las aplicaciones gestionen errores transitorios. Al compilar su propio código, es posible que tenga que habilitar el control de errores transitorios usted mismo. Algunos servicios de Azure proporcionan una gestión integrada de errores transitorios para algunas situaciones. Por ejemplo, de forma predeterminada, Azure Logic Apps reintenta automáticamente las solicitudes con errores a otros servicios. Para obtener más información, consulte Recomendaciones para el control de errores transitorios.

Redundancia

La redundancia es la práctica de duplicar instancias o datos para aumentar la confiabilidad de la carga de trabajo.

La redundancia se puede lograr mediante la distribución de réplicas o instancias redundantes de una de las maneras siguientes:

- Dentro de un centro de datos (*redundancia local*)
- Entre zonas de disponibilidad dentro de una región (*redundancia de zona*)
- Entre regiones (*redundancia geográfica*).

Estos son algunos ejemplos de cómo algunos servicios de Azure proporcionan opciones de redundancia:

- Azure App Service permite ejecutar varias instancias de la aplicación para asegurarse de que la aplicación permanezca disponible incluso aunque se produzca un error en una instancia. Si habilita la redundancia de zona, esas instancias se distribuirán entre varias zonas de disponibilidad en la región de Azure que use.
- Azure Storage proporciona alta disponibilidad mediante la replicación automática de datos al menos tres veces. Puede distribuir esas réplicas entre zonas de disponibilidad habilitando el almacenamiento con redundancia de zona (ZRS) y en muchas regiones también puede replicar los datos de almacenamiento entre regiones mediante el almacenamiento con redundancia geográfica (GRS).
- Azure SQL Database tiene varias réplicas para asegurarse de que los datos permanezcan disponibles incluso aunque se produzca un error en una réplica.

Para más información sobre cómo funciona la redundancia, consulte Redundancia, replicación y copia de seguridad. Para obtener información sobre cómo aplicar redundancia en la solución, consulte Recomendaciones para diseñar redundancia y recomendaciones para usar zonas de disponibilidad y regiones.

Escalabilidad y elasticidad

La escalabilidad y la elasticidad son las capacidades de un sistema para controlar la mayor carga mediante la adición y eliminación de recursos (escalabilidad) y para hacerlo rápidamente a medida que cambian los requisitos (elasticidad). La escalabilidad y la elasticidad pueden ayudar a un sistema a mantener la disponibilidad durante las cargas máximas.

Muchos servicios de Azure admiten escalabilidad. Estos son algunos ejemplos:

- Azure Virtual Machine Scale Sets, Azure API Management y otros servicios admiten la escalabilidad automática de Azure Monitor. Con el escalado automático de Azure Monitor, puede especificar políticas como "cuando mi CPU supere constantemente el 80 %, agregar otra instancia".
- Azure Functions puede aprovisionar dinámicamente instancias para atender las solicitudes.
- Azure Cosmos DB admite el rendimiento de escalabilidad automática, donde el servicio puede administrar automáticamente los recursos asignados a las bases de datos en función de las directivas que especifique.

La escalabilidad es un factor clave que se debe tener en cuenta durante un mal funcionamiento parcial o completo. Si una réplica o una instancia de proceso no están disponibles, es posible que los componentes restantes necesiten cargar más para controlar la carga que el nodo defectuoso controló anteriormente. Considere la posibilidad de realizar un *sobreaprovisionamiento* si su sistema no puede escalar lo suficientemente rápido como para manejar los cambios esperados en la carga.

Para obtener más información sobre cómo diseñar un sistema escalable y elástico, consulte Recomendaciones para diseñar una estrategia de escalado confiable.

Técnicas de implementación sin tiempo de inactividad

Las implementaciones y otros cambios del sistema presentan un riesgo significativo de tiempo de inactividad. Dado que el riesgo de tiempo de inactividad es un desafío para los requisitos de alta disponibilidad, es importante utilizar prácticas de implementación de tiempo de inactividad cero para realizar actualizaciones y cambios de configuración sin ningún tiempo de inactividad requerido.

Las técnicas de implementación sin tiempo de inactividad pueden ser las siguientes:

- Actualizar un subconjunto de los recursos a la vez.
- Controlar la cantidad de tráfico que llega a la nueva implementación.
- Supervisión de cualquier impacto en los usuarios o el sistema.

- Corregir rápidamente el problema, como revertir a una implementación conocida anterior.

Para obtener más información sobre las técnicas de implementación sin tiempo de inactividad, consulte Prácticas de implementación segura.

Azure usa enfoques de implementación sin tiempo de inactividad para nuestros propios servicios. Al compilar sus propias aplicaciones, puede adoptar implementaciones de tiempo de inactividad cero a través de una variedad de enfoques, como los siguientes:

- Azure Container Apps proporciona varias revisiones de la aplicación, que se pueden usar para lograr implementaciones sin tiempo de inactividad.
- Azure Kubernetes Service (AKS) admite una variedad de técnicas de implementación sin tiempo de inactividad.

Aunque las implementaciones de tiempo de inactividad cero suelen estar asociadas a las implementaciones de aplicaciones, también se deben usar para los cambios de configuración. Estas son algunas maneras de aplicar los cambios de configuración de forma segura:

- Azure Storage le permite cambiar sus claves de acceso de la cuenta de almacenamiento en varias fases, lo que evita el tiempo de inactividad durante las operaciones de rotación de claves.
- Azure App Configuration proporciona marcas de características, instantáneasy otras funcionalidades que le ayudarán a controlar cómo se aplican los cambios de configuración.

Si decide no implementar implementaciones con tiempo de inactividad cero, asegúrese de definir *ventanas de mantenimiento* para poder realizar cambios en el sistema en el momento en que los usuarios lo esperan.

Pruebas automatizadas

Es importante probar la capacidad de la solución de resistir las interrupciones y los errores que considere tener en el ámbito de la alta disponibilidad. Muchos de estos errores se pueden simular en entornos de prueba. Probar la capacidad de la solución para tolerar o recuperarse automáticamente de una variedad de tipos de errores se denomina *ingeniería de caos*. La ingeniería de caos es fundamental para las organizaciones maduras con estrictos estándares para la alta disponibilidad. Azure Chaos Studio es una herramienta de ingeniería de caos que puede simular algunos tipos de errores comunes.

Para obtener más información, consulte Recomendaciones para diseñar una estrategia de pruebas de confiabilidad.

Supervisión y alertas

La supervisión le permite conocer el estado del sistema, incluso cuando se realizan mitigaciones automatizadas. La supervisión es fundamental para comprender cómo se comporta la solución y para observar las señales tempranas de errores, como el aumento de las tasas de error o el consumo elevado de recursos. Con las alertas, puede recibir de forma proactiva cambios importantes en su entorno.

Azure proporciona una variedad de funcionalidades de supervisión y alertas, entre las que se incluyen las siguientes:

- Azure Monitor recopila registros y métricas de recursos y aplicaciones de Azure, y puede enviar alertas y mostrar datos en los paneles.
- Application Insights de Azure Monitor proporciona una supervisión detallada de las aplicaciones.
- Azure Service Health y Azure Resource Health supervisa el estado de la plataforma de Azure y los recursos.
- Los eventos programados avisan cuándo se planifica el mantenimiento de las máquinas virtuales.

Para obtener más información, consulte Recomendaciones para diseñar una estrategia de supervisión y alertas confiable.

Recuperación ante desastres

Un *desastre* es un evento distinto, poco común e importante que tiene un impacto mayor y más duradero que el que una aplicación puede mitigar a través del aspecto de alta disponibilidad de su diseño. Algunos ejemplos de desastres son:

- **Desastres naturales**, como huracanes, terremotos, inundaciones o incendios.
- **Errores humanos que producen un gran impacto**, como eliminar accidentalmente los datos de producción o un firewall mal configurado que expone datos confidenciales.
- **Incidentes de seguridad principales**, como ataques por denegación de servicio o ransomware que provocan daños en los datos, pérdida de datos o interrupciones del servicio.

La *recuperación ante desastres* consiste en planear cómo responde a estos tipos de situaciones.

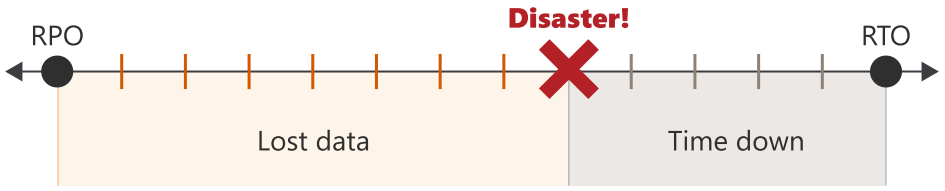
> **Nota:**
>
> Debe seguir los procedimientos recomendados en la solución para minimizar la probabilidad de estos eventos. Sin embargo, incluso después de una planeación proactiva

cuidadosa, es prudente planear cómo respondería a estas situaciones si surgen.

Requisitos de recuperación ante desastres.

Debido a la rareza y gravedad de los eventos de desastre, el planeamiento de recuperación ante desastres aporta diferentes expectativas para la respuesta. Muchas organizaciones aceptan el hecho de que, en un escenario de desastre, cierto nivel de tiempo de inactividad o pérdida de datos es inevitable. Un plan de recuperación ante desastres completo debe especificar los siguientes requisitos empresariales críticos para cada flujo:

- El **objetivo de punto de recuperación** (RPO) es la duración máxima de la pérdida de datos que es aceptable durante un desastre. El RPO se mide en unidades de tiempo, como "30 minutos de datos" o "cuatro horas de datos".
- El **objetivo de tiempo de recuperación (RTO)** es la duración máxima del tiempo de inactividad aceptable en caso de desastre, donde el "tiempo de inactividad" se define según su especificación. El RTO también se mide en unidades de tiempo, como "ocho horas de tiempo de inactividad".



Cada componente o flujo de la carga de trabajo puede tener valores de RPO y RTO individuales. Examine los riesgos del escenario de desastres y las posibles estrategias de recuperación al decidir cuáles son los requisitos. El proceso de especificar un RPO y RTO crea eficazmente requisitos de recuperación ante desastres para la carga de trabajo como resultado de sus preocupaciones empresariales únicas (costes, impacto, pérdida de datos, etc.).

> **Nota:**
>
> Aunque es tentador apuntar a un RTO y RPO de cero (sin tiempo de inactividad y sin pérdida de datos en caso de desastre), en la práctica es difícil y costoso de implementar. Es importante que las partes interesadas técnicas y empresariales analicen estos requisitos juntos y decidan los requisitos realistas. Para obtener más información, consulte Recomendaciones para definir objetivos de confiabilidad.

Planes de recuperación ante desastres

Independientemente de la causa del desastre, es importante crear un plan de recuperación ante desastres bien definido y probable. Ese plan se usará como parte de la infraestructura y el diseño de aplicaciones para admitirlo activamente. Puede crear varios planes de recuperación ante desastres para diferentes tipos de situaciones. Los planes de recuperación ante desastres suelen depender de controles de procesos e intervención manual.

La recuperación ante desastres no es una característica automática de Azure. Sin embargo, muchos servicios proporcionan características y funcionalidades que puede usar para admitir las estrategias de recuperación ante desastres. Debe revisar las guías de confiabilidad para cada servicio de Azure para comprender cómo funciona el servicio y sus funcionalidades y, a continuación, asignar esas funcionalidades al plan de recuperación ante desastres.

Un plan fuerte de recuperación ante desastres convierte la estrategia en una acción decisiva. Proporciona una hoja de ruta clara para responder a desastres, minimizar el tiempo de inactividad y garantizar la continuidad empresarial.

Para lograr esto, todos los planes de recuperación ante desastres deben documentarse para incluir un manual de operaciones claro, un plan de comunicación bien definido y una ruta de escalación estructurada. Para obtener más información sobre estos elementos del plan de recuperación ante desastres, consulte Well-Architected Framework : documente el plan de recuperación ante desastres.

En las secciones siguientes se enumeran algunos enfoques comunes en un plan de recuperación ante desastres y se describe cómo Azure puede ayudarle a lograrlos.

Conmutación por error y conmutación por recuperación

Algunos planes de recuperación ante desastres implican el aprovisionamiento de una implementación secundaria en otra ubicación. Si un desastre afecta a la implementación principal de la solución, el tráfico se puede *conmutar por error* al otro sitio. La conmutación por error requiere una planeación e implementación cuidadosas. Azure proporciona una variedad de servicios para ayudar con la conmutación por error, como:

- Azure Site Recovery proporciona conmutación por error automatizada para entornos locales y soluciones hospedadas en máquinas virtuales en Azure.
- Azure Front Door y Azure Traffic Manager admiten la conmutación por error automatizada del tráfico entrante entre diferentes implementaciones de la solución, como en diferentes regiones.

Normalmente, se tarda algún tiempo en un proceso de conmutación por error para detectar que se ha producido un error en la instancia principal y cambiar a la instancia secundaria. Asegúrese de que el RTO de la carga de trabajo esté alineado con el tiempo de conmutación por error.

También es importante tener en cuenta la *conmutación por recuperación*, que es el proceso por el que se restauran las operaciones en la región primaria después de recuperarse. La conmutación por recuperación puede ser compleja de planear e implementar. Por ejemplo, es posible que los datos de la región primaria se hayan escrito *después* de que se iniciara la conmutación por error. Deberá tomar decisiones empresariales cuidadosas sobre cómo controlar esos datos.

Para obtener más información, consulte Conmutación por error y conmutación por recuperación.

Backups

Las copias de seguridad implican realizar una copia de los datos y almacenarlos de forma segura durante un período de tiempo definido. Con las copias de seguridad, puede recuperarse de desastres cuando la conmutación automática por error a otra réplica no es posible o cuando se han producido daños en los datos.

Al usar copias de seguridad como parte de un plan de recuperación ante desastres, es importante tener en cuenta lo siguiente:

- *Ubicación de almacenamiento*. Cuando se usan copias de seguridad como parte de un plan de recuperación ante desastres, deben almacenarse por separado en los datos principales. Normalmente, las copias de seguridad se almacenan en otra región de Azure.
- *Pérdida de datos*. Dado que las copias de seguridad normalmente se realizan con poca frecuencia, la restauración de copias de seguridad suele implicar la pérdida de datos. Por este motivo, la recuperación de copia de seguridad debe usarse como último recurso y un plan de recuperación ante desastres debe especificar la secuencia de pasos e intentos de recuperación que deben realizarse *antes de* realizar una restauración desde una copia de seguridad. Es importante asegurarse de que el RPO de carga de trabajo esté alineado con el intervalo de copia de seguridad.
- *Tiempo de recuperación*. La restauración de copias de seguridad suele tardar tiempo, por lo que es fundamental probar las copias de seguridad y los procesos de restauración para comprobar su integridad y comprender cuánto tarda el proceso de restauración. Asegúrese de que el RTO de la carga de trabajo tenga en cuenta el tiempo necesario para restaurar la copia de seguridad.

Muchos servicios de almacenamiento y datos de Azure admiten copias de seguridad, como los siguientes:

- Azure Backup proporciona copias de seguridad automatizadas para discos de máquina virtual, cuentas de almacenamiento, AKS y una gran variedad de otros orígenes.

- Muchos servicios de base de datos de Azure, incluidos Azure SQL Database y Azure Cosmos DB, tienen una funcionalidad de copia de seguridad automatizada para las bases de datos.
- Azure Key Vault proporciona características para realizar copias de seguridad de los secretos, certificados y claves.

Para obtener más información sobre las estrategias de recuperación para la copia de seguridad y restauración, consulte Well-Architected Framework : estrategia de recuperación para la copia de seguridad y restauración.

Implementaciones automatizadas

Para implementar y configurar rápidamente los recursos necesarios por si se produce un desastre, use recursos de infraestructura como código (IaC), como archivos de Bicep, plantillas de ARM o un archivo de configuración de Terraform. El uso de IaC reduce el tiempo de recuperación y el potencial de error, en comparación con la implementación y configuración manual de recursos.

Pruebas y simulacros

Es fundamental validar y probar de forma rutinaria los planes de recuperación ante desastres, así como su estrategia de confiabilidad más amplia. Incluya todos los procesos humanos en sus simulacros y no se centre solo en los procesos técnicos.

Si no ha probado los procesos de recuperación en una simulación de desastre, es más probable que se enfrente a problemas importantes al usarlos en un desastre real. Además, al probar los planes de recuperación ante desastres y los procesos necesarios, puede validar la viabilidad de su RTO.

Para obtener más información, consulte Recomendaciones para diseñar una estrategia de pruebas de confiabilidad.

Contenido relacionado

- Use las guías de confiabilidad del servicio de Azure para comprender cómo cada servicio de Azure admite la confiabilidad en su diseño y para obtener información sobre las funcionalidades que puede crear en los planes de alta disponibilidad y recuperación ante desastres.
- Use el Marco de buena arquitectura de Azure: pilar de confiabilidad para obtener más información sobre cómo diseñar una carga de trabajo fiable en Azure.

- Use la perspectiva de Marco de buena arquitectura de los servicios de Azure para obtener más información sobre cómo configurar cada servicio de Azure para satisfacer sus requisitos de confiabilidad y los demás pilares del Marco de buena arquitectura.
- Para obtener más información sobre el planeamiento de la recuperación ante desastres, consulte Recomendaciones para diseñar una estrategia de recuperación ante desastres.

Last updated on 22/01/2026

¿Qué son la redundancia, la replicación y la copia de seguridad?

A menudo pensamos en la nube como un sistema omnipresente y distribuido globalmente. Sin embargo, en realidad, la nube se compone de hardware que se ejecuta en centros de datos. La resistencia requiere que tenga en cuenta algunos de los riesgos asociados a las ubicaciones físicas en las que se ejecutan los componentes hospedados en la nube.

En este artículo se proporciona una introducción general a la redundancia, la replicación y la copia de seguridad, que son métodos que se usan para crear cargas de trabajo resistentes a los riesgos físicos que provocan interrupciones del servicio, interrupciones o pérdida de datos.

La redundancia es la capacidad de mantener varias copias idénticas de un componente de servicio y usar esas copias de una manera que impida que cualquier componente se convierta en un único punto de error.

La replicación o la redundancia de datos es la capacidad de mantener varias copias de datos, denominadas réplicas.

La copia de seguridad es la capacidad de mantener una copia con marca de tiempo de los datos que se pueden usar para restaurar los datos que se han perdido.

Desde una perspectiva de confiabilidad, una manera importante de mitigar muchos riesgos es incluir alguna forma de redundancia, replicación o copia de seguridad en el planeamiento de continuidad empresarial.

> ⓘ **Nota:**
>
> En este artículo no se proporcionan instrucciones de diseño ni información detallada sobre los servicios individuales de Azure. Para obtener información sobre cómo funcionan los servicios específicos de Azure desde una perspectiva de confiabilidad, **consulte la guía de confiabilidad de cada servicio**.

Redundancia

La redundancia consiste en implementar varias *instancias* de un *componente*. Aunque la redundancia es sencilla de entender, en algunas situaciones, puede ser compleja de implementar.

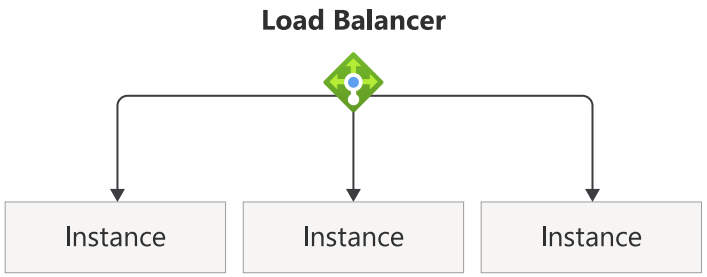
Al empezar a comprender la redundancia, es más fácil considerar la redundancia en relación con los componentes sin estado, que son componentes que no almacenan datos. Aunque la

mayoría de las soluciones del mundo real requieren administrar el estado, en esta sección se describe la redundancia en términos de una interfaz de programación de aplicaciones sin estado (API) de ejemplo. La API de ejemplo acepta la entrada, funciona en esa entrada y, a continuación, devuelve una salida, sin almacenar ningún dato. En la siguiente sección, Replicación: redundancia para los datos, agregaremos consideraciones para la redundancia de estado.

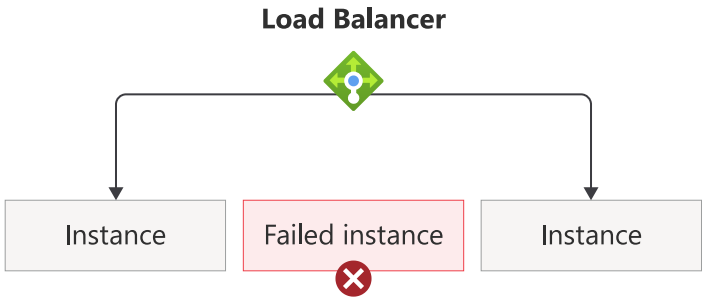
Escenario: redundancia sin estado

En este ejemplo, se implementa un componente de API sin estado en una máquina virtual. Para evitar el tiempo de inactividad del componente de API si se produce un error de hardware en el host físico, el ejemplo implementa una solución redundante que:

- Implementa varias copias de la instancia de API.
- Implementa un *equilibrador de carga* para distribuir solicitudes entre instancias de API.



El equilibrador de carga supervisa el estado de cada instancia para enviar solicitudes solo a instancias correctas.



Consideraciones sobre redundancia

Al implementar soluciones redundantes, como en el escenario anterior, es importante tener en cuenta lo siguiente:

- **Costos de recursos.** Por definición, la redundancia implica tener varias copias de algo, lo que aumenta el costo total para hospedar la solución.

- **Rendimiento.** Cuanto más amplio sea el área geográfica en la que distribuya las copias, más riesgos ayudará a mitigar. Sin embargo, las solicitudes de los clientes tardarán más tiempo en recorrer distancias más largas porque tienen que atravesar más infraestructura de red y esto aumenta la *latencia de red*.

  En la mayoría de las situaciones reales, la latencia de red es inconsecuente para distancias cortas, como dentro de un centro de datos o incluso entre centros de datos a través de una ciudad. Pero si distribuye copias a una larga distancia, la latencia de red puede ser más significativa.

- **Coherencia de instancias.** Es importante mantener las instancias coherentes entre sí y evitar configurar instancias individualmente, ya que puede introducir accidentalmente diferencias entre las instancias. Si hay diferencias entre instancias, es posible que las solicitudes se procesen de forma diferente en función de la instancia que las sirva. Esto puede provocar errores y comportamientos difíciles de diagnosticar.

- **Distribución del trabajo.** Cuando tenga varias instancias de un componente, debe distribuir el trabajo entre ellos. Puede distribuir el trabajo en todas las instancias de la misma manera, o puede enviar solicitudes a una sola instancia *principal* y usar solo una instancia secundaria si la instancia principal no está disponible.

  En el caso de los componentes que reciben solicitudes entrantes, los balanceadores de carga suelen utilizarse para dirigir esas solicitudes a instancias. Sin embargo, a veces los componentes funcionan de forma independiente y no reciben solicitudes de los clientes, por lo que, en estas situaciones, es posible que las instancias deban coordinar su trabajo entre sí.

- **Supervisión de estado.** El estado de cada instancia determina si esa instancia puede realizar su trabajo y la supervisión de estado es importante para habilitar la recuperación rápida si hay un problema.

  Los equilibradores de carga típicamente realizan el monitoreo del estado. En el caso de los componentes que no incluyen un equilibrador de carga, es posible que tenga un componente externo que supervise el estado de todas las instancias o que cada instancia pueda supervisar el estado de otras instancias.

Ubicaciones físicas en la nube

La necesidad de redundancia es clara cuando comprende que, incluso en un entorno de nube, una instancia o un fragmento de datos se almacenan en una ubicación física específica.

Por ejemplo:

- Una máquina virtual se ejecuta en un único lugar físico en cualquier momento.
- Los datos se almacenan en una ubicación física específica, como en una unidad de estado sólido (SSD) o en un disco duro conectado a los servidores.

Incluso si hay varias copias de un fragmento de datos o instancias de un componente, cada copia está vinculada al hardware físico en el que se almacena.

La ubicación física de un entorno en la nube se puede organizar en ámbitos físicos. En cada ámbito físico, hay posibles riesgos que podrían poner en peligro los componentes o los datos dentro de ese ámbito. Esta es una lista no exhaustiva de riesgos que se pueden definir en términos de ámbito físico:

Expandir tabla

Ámbito físico	Posible riesgo
Un elemento específico de hardware, como un disco o un servidor	Error de hardware
Un bastidor en un centro de datos	Interrupción del conmutador de red de la parte superior del bastidor
Un centro de datos	Problema con el sistema de refrigeración del edificio
Un grupo de centros de datos, que en Azure se denomina *zona de disponibilidad*	Tormenta eléctrica de toda la ciudad
Área geográfica más amplia en la que se encuentra el centro de datos, como una ciudad, que es una *región* de Azure.	Desastre natural generalizado

Desde una perspectiva de confiabilidad, una manera importante de mitigar los riesgos asociados a un ámbito físico es distribuir instancias de un componente en distintos ámbitos físicos. Los servicios de Azure con redundancia integrada pueden ofrecer una o varias de las tres maneras siguientes de implementar instancias redundantes:

- *La redundancia local* coloca instancias en varias partes de un único centro de datos de Azure y protege frente a errores de hardware que podrían afectar a una sola instancia. La redundancia local normalmente proporciona el menor costo y latencia. Sin embargo, un error del centro de datos podría significar que todas las instancias no están disponibles.

- *La redundancia de zona* distribuye instancias en varias zonas de disponibilidad en una sola región de Azure. La redundancia de zona protege frente a errores del centro de datos, además de errores de hardware.

- *La redundancia geográfica* coloca instancias en varias regiones de Azure y proporciona protección contra interrupciones regionales a gran escala. La redundancia geográfica conlleva un costo más alto y requiere tener en cuenta el diseño más amplio de la solución y cómo cambiará entre instancias de los componentes de cada región. También debe tener en cuenta la latencia, que se describe en Replicación sincrónica y asincrónica.

Funcionamiento de la redundancia en Azure: Servicios de proceso

La redundancia se emplea en casi todas las partes de Azure. Como ejemplo de cómo implementa Azure la redundancia, tenga en cuenta los servicios que ejecutan cargas de trabajo de cómputo.

En Azure, una máquina virtual individual se ejecuta en un único host físico en un centro de datos de Azure. Puede proporcionar instrucciones a Azure para asegurarse de que la máquina virtual se ejecuta en el lugar esperado, como la región y la zona de disponibilidad, y en algunas situaciones, es posible que incluso quiera colocarla en un host dedicado a usted.

Es habitual ejecutar varias instancias de una máquina virtual. En ese escenario, puede optar por administrarlos individualmente o mediante un conjunto de escalado de máquinas virtuales. Cuando se usa un conjunto de escalado, todavía puede ver las máquinas virtuales individuales que lo sustentan, pero el conjunto de escalado proporciona una serie de funcionalidades para ayudar a administrar máquinas virtuales redundantes. Estas funcionalidades incluyen la colocación automática de las máquinas virtuales en función de las reglas que especifique, como distribuirlas entre dominios de error dentro de la región o zona de disponibilidad.

Hay muchos otros servicios de proceso de Azure que dependen de máquinas virtuales para realizar tareas de proceso. Los servicios de proceso suelen ofrecer varias opciones de redundancia que determinan cómo se distribuyen las máquinas virtuales. Por ejemplo, un servicio podría proporcionar una configuración de redundancia de zona, que puede habilitar para distribuir automáticamente máquinas virtuales entre zonas de disponibilidad y configurar el equilibrio de carga.

Replicación: redundancia para los datos

La redundancia, cuando se aplica al estado (datos), se denomina replicación. Con la replicación, puede mantener varias copias en tiempo real o casi en tiempo real de datos activos, *denominadas réplicas*. Para mejorar la resistencia a los riesgos, puede distribuir réplicas en diferentes ubicaciones. Si una réplica deja de estar disponible, el sistema podría *conmutar por error* para que otra réplica tome el control de su función.

Hay diferentes tipos de replicación y cada uno coloca prioridades diferentes en la coherencia de los datos, el rendimiento y el costo. Cada tipo de replicación afecta a dos métricas clave que se usan en discusiones de continuidad empresarial: *objetivo de tiempo de recuperación* (RTO), que es la cantidad máxima de tiempo de inactividad que puede tolerar en un escenario de desastre y el objetivo de punto de *recuperación* (RPO), que es la cantidad máxima de pérdida de datos que puede tolerar en un escenario de desastre. Para obtener más información sobre estas métricas y cómo se relacionan con la continuidad empresarial, consulte ¿Qué son la continuidad empresarial, la alta disponibilidad y la recuperación ante desastres?.

Debido a la importancia de la replicación en cumplir los requisitos funcionales y de rendimiento, la mayoría de los sistemas de base de datos y otros productos y servicios de almacenamiento de datos incluyen algún tipo de replicación como una característica estrechamente integrada. Los tipos de replicación que ofrecen suelen basarse en su arquitectura y en las formas en que están diseñados para usarse. Para más información sobre los tipos de replicación admitidos por los servicios de Azure, consulte Guías de confiabilidad de los servicios de Azure.

> **ⓘ Importante**
>
> La replicación no es la misma que la copia de seguridad. La replicación sincroniza todos los cambios entre varias réplicas y no mantiene copias antiguas de datos.
>
> Supongamos que elimina accidentalmente algunos datos. La operación de eliminación se replica en cada réplica y los datos se eliminan en todas partes. En esta situación, es probable que tenga que restaurar los datos eliminados de una copia de seguridad. La copia de seguridad se describe más adelante en este artículo.

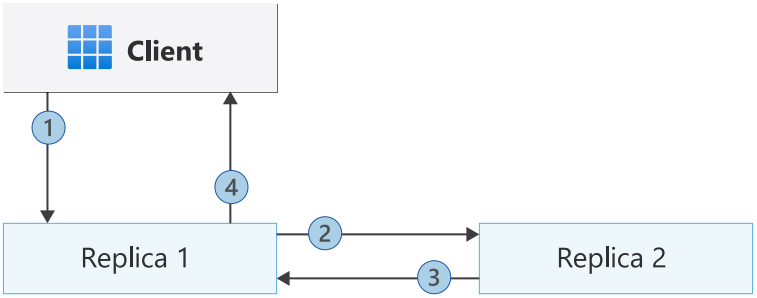
Replicación sincrónica y asincrónica

Al replicar datos, los cambios realizados en esos datos se deben mantener sincronizados entre las réplicas. Hay un par de desafíos principales al intentar mantener la coherencia cuando cambian los datos:

- *Latencia.* La actualización de una réplica lleva tiempo, y cuanto más separadas estén las réplicas, más tiempo lleva transmitir datos a través de la distancia entre ellas y recibir un acuse de recibo.
- *Administración de cambios*. Los datos pueden cambiar mientras se sincronizan las réplicas y, por tanto, administrar la coherencia de los datos puede ser compleja.

Para abordar estos desafíos, hay dos maneras principales de replicar los cambios de datos y administrar la coherencia de los datos:

- **La replicación sincrónica** requiere que se realicen actualizaciones en varias réplicas antes de que la actualización se considere completada. En el diagrama siguiente se muestra cómo funciona la replicación sincrónica:

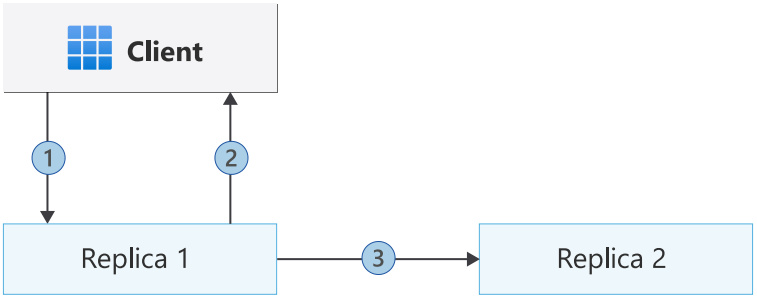


En este ejemplo, se produce la siguiente secuencia de pasos:

1. Un cliente cambia los datos y la solicitud se envía a la réplica 1, que procesa la solicitud y almacena los datos modificados.
2. La réplica 1 envía los cambios a la réplica 2, que procesa la solicitud y almacena los datos modificados.
3. La réplica 2 confirma el cambio a la réplica 1.
4. La réplica 1 confirma el cambio al cliente.

La replicación sincrónica puede garantizar la coherencia, lo que significa que puede admitir un RPO de cero. Sin embargo, esto conlleva el costo del rendimiento. Cuantos más separados sean las réplicas geográficamente y cuantos más saltos de red deban atravesarse, el proceso de replicación introducirá más latencia.

- **La replicación asincrónica** se produce en segundo plano. En el diagrama siguiente se muestra cómo funciona la replicación asincrónica:



En este ejemplo, se produce la siguiente secuencia de pasos:

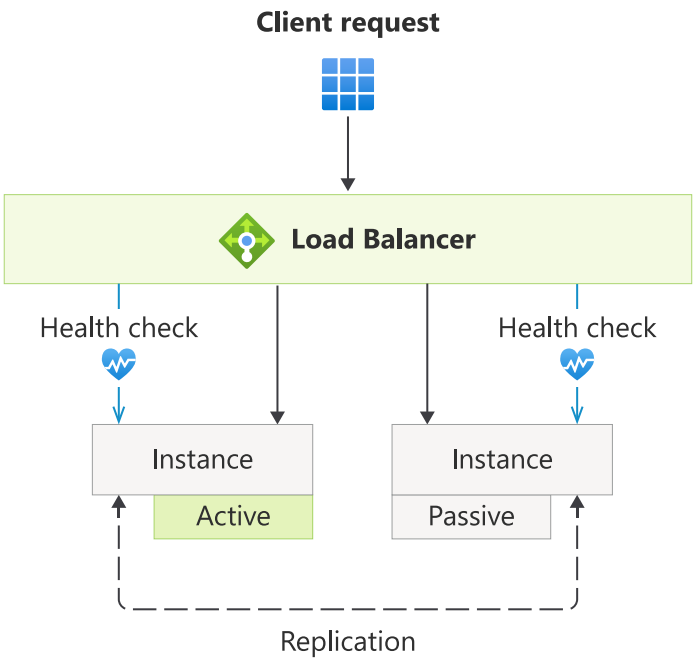
1. Un cliente cambia los datos y la solicitud se envía a la réplica 1.
2. La réplica 1 procesa la solicitud, almacena los datos modificados y confirma inmediatamente el cambio al cliente.
3. En algún momento posterior, la réplica 1 sincroniza el cambio en la réplica 2.

Dado que la replicación asincrónica se produce fuera de los flujos de transacción, quita la replicación como una restricción en el rendimiento de la aplicación. Sin embargo, si necesita realizar una conmutación por error a otra réplica, es posible que no cuente con los datos más recientes, por lo que su RPO debe ser mayor que cero. El RPO exacto que la replicación asincrónica puede admitir depende de la frecuencia de replicación.

Roles de réplica

En muchos sistemas de replicación, las réplicas pueden asumir diferentes roles, lo que ayuda a coordinar los cambios en los datos y reduce la posibilidad de conflictos. Hay dos tipos principales de roles, *activos* y *pasivos*. Hay dos formas comunes de distribuir réplicas con estas funciones:

- La replicación **activa-pasiva** significa que tiene una réplica *activa*, que es responsable de actuar como fuente de verdad. Los cambios realizados en los datos deben aplicarse a esa réplica. Cualquier otra réplica actúa en un rol *pasivo* , lo que significa que reciben actualizaciones de los datos de la réplica activa, pero no procesan los cambios directamente de los clientes. Las réplicas pasivas no se usan para el tráfico activo a menos que se produzca una *conmutación por error* y los roles de las réplicas cambien. En el diagrama siguiente se muestra un sistema activo-pasivo con una réplica pasiva:

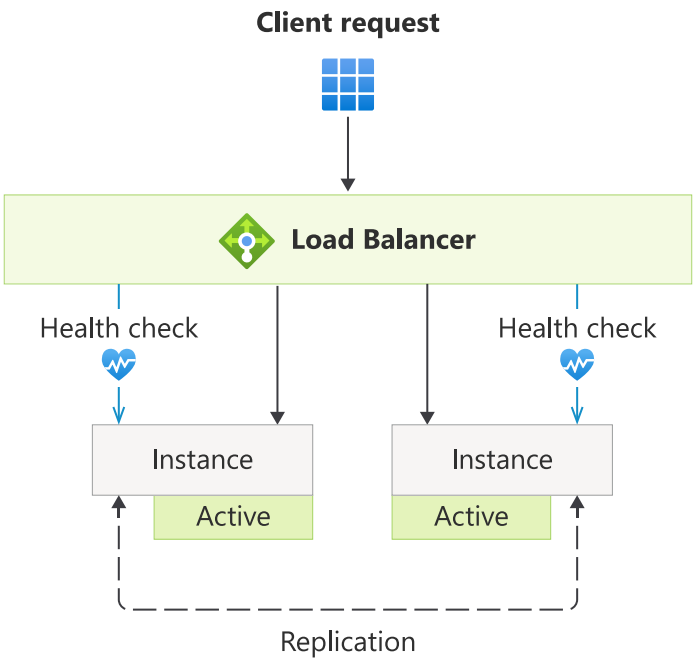


En un sistema activo-pasivo, el período de tiempo que se tarda en la conmutación por error determina el RTO. Normalmente, el RTO para un sistema activo-pasivo se mide en minutos.

Algunas soluciones de replicación también admiten *réplicas de solo* lectura, lo que permite leer (pero no escribir) datos de las réplicas pasivas. Este enfoque puede ser útil

para obtener más uso de las réplicas, como cuando necesite realizar análisis o informes sobre datos sin afectar a la réplica principal que usa para el trabajo transaccional de la aplicación. Varios servicios de Azure admiten réplicas de solo lectura, incluido Azure Storage con el tipo de replicación GRS de acceso de lectura (RA-GRS) y la replicación geográfica activa de Azure SQL Database.

- La replicación **activa-activa** permite usar varias réplicas activas para el tráfico activo simultáneamente y cualquiera de las réplicas puede procesar solicitudes:



  La replicación activa-activa permite un alto nivel de rendimiento porque el sistema puede usar los recursos de todas las réplicas. La replicación activa-activa puede admitir un RTO de cero en algunas situaciones. Sin embargo, estas ventajas tienen el costo de complicar la coherencia de los datos, ya que es posible que los cambios simultáneos en varias réplicas deban conciliarse de forma asincrónica.

Los servicios de datos complejos pueden combinar la replicación activa-activa y activa-pasiva. Por ejemplo, podrían implementar un conjunto de réplicas en una región de Azure y otra en otra región diferente. Dentro de cada región, una única réplica activa atiende solicitudes mientras que una o varias réplicas pasivas están en espera para el failover. Mientras tanto, ambas regiones operan en un modelo activo-activo, lo que permite distribuir el tráfico entre ellas.

Funcionamiento de la replicación en los servicios de datos de Azure

Cada servicio de Azure que almacena los datos ofrece alguna forma de replicación. Sin embargo, cada servicio puede usar un enfoque diferente específico de la arquitectura del

servicio y los usos previstos.

Por ejemplo, Azure Storage puede proporcionar replicación sincrónica y asincrónica a través de un conjunto de funcionalidades:

- Varias copias de los datos se replican sincrónicamente dentro de la región primaria. Puede elegir si colocar réplicas en hardware físico diferente en un único centro de datos en almacenamiento con redundancia local (LRS) o distribuirlas entre varias zonas de disponibilidad para el almacenamiento con redundancia de zona (ZRS).
- Si la región primaria está emparejada y habilita el almacenamiento con redundancia geográfica (GRS), los datos también se replican en la región emparejada. Dado que las regiones emparejadas están geográficamente distantes, esta replicación se produce de forma asincrónica para reducir el impacto en el rendimiento de la aplicación.
- Puede optar por usar almacenamiento con redundancia de zona y almacenamiento con redundancia geográfica simultánea mediante el nivel de almacenamiento con redundancia de zona geográfica (GZRS). Los datos de la región se replican de forma sincrónica y los datos entre regiones se replican de forma asincrónica.

Para más información, vea Redundancia de Azure Storage.

Otro ejemplo es Azure Cosmos DB, que también proporciona replicación. Todas las bases de datos de Azure Cosmos DB tienen varias réplicas. Al distribuir réplicas globalmente, se admite la escritura en varias regiones, donde los clientes pueden escribir en una réplica en cualquiera de las regiones utilizadas. Esas operaciones de escritura se replican sincrónicamente dentro de la región y, a continuación, se replican de forma asincrónica en otras regiones. Azure Cosmos DB proporciona un mecanismo de resolución de conflictos en caso de que haya conflictos de escritura en las distintas réplicas. Para más información, consulte Distribución global de datos con Azure Cosmos DB en segundo plano.

Si usa máquinas virtuales, puede usar Azure Site Recovery para replicar máquinas virtuales y sus discos entre zonas de disponibilidad o en otra región de Azure.

Al diseñar una solución de Azure, consulte las guías de confiabilidad de cada servicio para comprender cómo ese servicio proporciona redundancia y replicación, incluida en diferentes ubicaciones.

Backup

La copia de seguridad realiza una copia de tus datos en un momento específico en el tiempo. Si hay un problema, puede *restaurar* la copia de seguridad más adelante. Sin embargo, los cambios realizados en los datos que se produjeron después de realizar la copia de seguridad no estarán en la copia de seguridad y podrían perderse.

Mediante la copia de seguridad, puede proporcionar soluciones para realizar copias de seguridad y recuperar los datos dentro o en la nube de Microsoft Azure. La copia de seguridad puede protegerle de diversos riesgos, entre los que se incluyen:

- Pérdidas catastróficas de hardware u otra infraestructura.
- Corrupción y borrado de datos.
- Ciberataques, como ransomware.

> **Importante**
>
> Es fundamental que pruebe y compruebe los procesos de copia de seguridad y restauración con regularidad, junto con los demás pasos de recuperación. Las pruebas garantizan que las copias de seguridad sean completas y sin errores y que los procesos las restauren correctamente. Las pruebas también son importantes para asegurarse de que el equipo comprende los procesos que se deben seguir. Para más información, consulte Pruebas y simulaciones.

Cómo afecta la copia de seguridad a los requisitos

Cuando se usa como parte de una estrategia de recuperación ante desastres, las copias de seguridad suelen admitir un RTO y un RPO medidos en horas:

- RTO se ve afectado por el tiempo necesario para iniciar y completar los procesos de recuperación, incluida la restauración de una copia de seguridad y la validación de que la restauración se completó correctamente. Según el tamaño de la copia de seguridad y el número de archivos de copia de seguridad que se deben leer, es habitual tardar varias horas o incluso más tiempo en restaurar completamente una copia de seguridad.

- El RPO se ve afectado por la frecuencia del proceso de copia de seguridad. Si hace copias de seguridad con más frecuencia, significa que pierde menos datos si tiene que restaurar desde una copia de seguridad. Sin embargo, las copias de seguridad requieren almacenamiento y, en algunas situaciones, pueden afectar al rendimiento del servicio mientras se realizan copias de seguridad. Por este motivo, debe tener en cuenta la frecuencia de copia de seguridad y encontrar el equilibrio adecuado para los requisitos de su organización. La frecuencia de copia de seguridad debe considerarse en la planificación de la continuidad empresarial.

Algunos sistemas de copia de seguridad admiten requisitos de copia de seguridad más complejos, incluidos varios niveles de copia de seguridad con diferentes períodos de retención, y copias de seguridad diferenciales o incrementales que son más rápidas para ahorrar y consumir menos almacenamiento.

Copia de seguridad en servicios de Azure

Muchos servicios de Azure proporcionan funcionalidades de copia de seguridad para los datos.

Azure Backup es una solución de copia de seguridad dedicada para varios servicios clave de Azure, como máquinas virtuales, Azure Storage y Azure Kubernetes Service (AKS).

Además, muchas bases de datos administradas proporcionan sus propias funcionalidades de copia de seguridad como parte del servicio, como:

- Azure SQL Database proporciona copias de seguridad automatizadas.
- Azure Cosmos DB proporciona funcionalidades de copia de seguridad continuas y periódicas.
- Azure Key Vault le permite descargar una copia de seguridad de los datos del almacén.
- Azure App Service proporciona copias de seguridad automáticas y personalizadas para aplicaciones web y también puede realizar copias de seguridad de sus bases de datos.

Copia de seguridad frente a replicación

La copia de seguridad y la replicación le protegen contra distintos riesgos, y los dos enfoques son complementarios entre sí.

La replicación respalda la resiliencia diaria y se utiliza comúnmente en una estrategia de alta disponibilidad. Algunos enfoques de replicación requieren poco o ningún tiempo de inactividad o pérdida de datos y admiten un RTO y RPO bajo. Sin embargo, la replicación no le protege frente a riesgos que provocan pérdida de datos o daños.

En cambio, la copia de seguridad suele ser una última línea de defensa frente a riesgos catastróficos. Las copias de seguridad a menudo requieren un RTO y RPO relativamente altos, aunque la forma en que configures las copias de seguridad afecta exactamente el nivel de alto que tendrán. Una restauración total a partir de una copia de seguridad suele formar parte de un plan de recuperación ante desastres.

Preparación de componentes para redundancia

Al diseñar un sistema que use redundancia como parte de su arquitectura, también es importante tener en cuenta cómo:

- Configuración de recursos duplicada para la coherencia.
- Administre la capacidad durante los errores de instancia mediante el aprovisionamiento excesivo.

Configuración de recursos duplicados

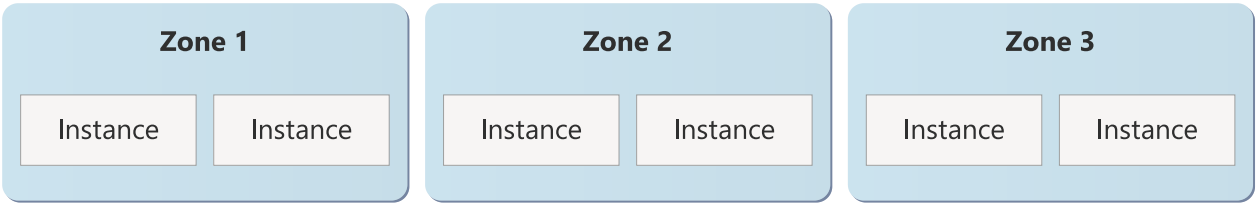
En entornos en la nube, la configuración de cada uno de los recursos es fundamental. Por ejemplo, cuando se crea un equilibrador de carga de red, se configuran diversas opciones que afectan a cómo funciona; y al implementar una función mediante Azure Functions, se configuran las opciones relacionadas con la seguridad, el rendimiento y las opciones de configuración de la aplicación. Cada recurso de Azure tiene algún tipo de configuración que impulsa su comportamiento.

Al administrar copias redundantes de recursos en diferentes lugares, es importante controlar su configuración. Muchas configuraciones deberán configurarse de la misma manera en cada copia, de modo que los recursos se comporten de la misma manera. Pero algunas opciones de configuración pueden ser diferentes entre cada copia, como referencias a la red virtual de una región específica.

Un enfoque común para mantener la coherencia en los recursos es usar la infraestructura como código (IaC), como Bicep o Terraform. Estas herramientas le permiten crear archivos que definen un recurso y puede reutilizar esas definiciones para cada instancia del recurso. Al usar IaC, puede reducir la carga de crear y administrar varias instancias de recursos con fines de resistencia y también hay muchas otras ventajas. Para más información, consulte ¿Qué es la infraestructura como código (IaC)? y Recomendaciones para usar la infraestructura como código.

Administración de la capacidad con sobreaprovisionamiento

Cuando se produce un error en una instancia, la capacidad general del sistema puede ser diferente de la capacidad necesaria durante las operaciones correctas. Por ejemplo, supongamos que normalmente tiene seis instancias de un servidor web para procesar el tráfico web entrante y esas instancias se distribuyen igualmente entre tres zonas de disponibilidad de Azure en una región:



Si una zona de disponibilidad experimenta una interrupción, puede perder temporalmente dos instancias y dejarse con solo cuatro instancias de servidor web. Si tu aplicación está típicamente ocupada y requiere las seis instancias para mantenerse al día con el tráfico normal, entonces funcionar a una capacidad reducida podría afectar al rendimiento de tu solución.

Para prepararse para errores, puede *sobreaprovisionar* la capacidad del servicio. El sobreaprovisionamiento permite a la solución tolerar cierto grado de pérdida de capacidad y seguir funcionando sin degradar el rendimiento.

Para sobreaprovisionar instancias del servidor web para tener en cuenta el error de una zona de disponibilidad, siga estos pasos:

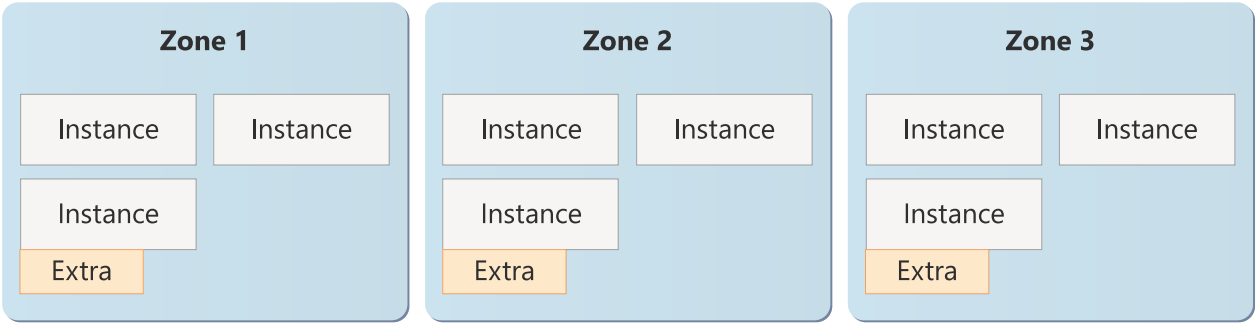
1. Determine el número de instancias que requiere la carga de trabajo máxima.
2. Recupere el recuento de instancias de sobreaprovisionamiento multiplicando el recuento máximo de instancias de carga de trabajo por un factor de [(zonas/(zonas-1)].
3. Redondea el resultado al número entero más cercano.

> **Nota:**
>
> En la tabla siguiente se asume que está utilizando tres zonas de disponibilidad y desea considerar la pérdida de capacidad de una de esas zonas. Si los requisitos son diferentes, ajuste la fórmula en consecuencia.

Expandir tabla

Número máximo de instancias de carga de trabajo	Factor de [(zonas/(zonas-1)]	Formula	Instancias para aprovisionar (redondeadas)
3	3/2 o 1,5	(3 x 1,5 = 4,5)	5 instancias
4	3/2 o 1,5	(4 x 1,5 = 6)	6 instancias
5	3/2 o 1,5	(5 x 1,5 = 7,5)	8 instancias
6	3/2 o 1,5	(6 x 1,5 = 9)	9 instancias
7	3/2 o 1,5	(7 x 1,5 = 10,5)	11 instancias
8	3/2 o 1,5	(8 x 1,5 = 12)	12 instancias
9	3/2 o 1,5	(9 x 1,5 = 13,5)	14 instancias
10	3/2 o 1,5	(10 x 1,5 = 15)	15 instancias

En el ejemplo anterior, la carga de trabajo máxima requiere seis instancias del servidor web, por lo que el aprovisionamiento excesivo requiere un total de nueve instancias:



Pasos siguientes

Aprenda sobre la conmutación por error y la conmutación de retorno.

Last updated on 22/01/2026

¿Qué son la conmutación por error y la conmutación por recuperación?

En este artículo se proporciona información general sobre cómo funcionan la *conmutación por error* y *la conmutación por recuperación* en un entorno de nube. Sin embargo, para comprender la conmutación por error, primero debe comprender la redundancia y la replicación. Para obtener información sobre esos conceptos antes de continuar con este artículo, consulte Redundancia, replicación y copia de seguridad.

Una razón habitual para mantener copias redundantes tanto de las aplicaciones como de las réplicas de datos es poder llevar a cabo una *conmutación por error*. Con la conmutación por error, puede redirigir el tráfico y las solicitudes de las instancias en estado incorrecto a instancias en estado correcto. A continuación, una vez que las instancias originales vuelvan a estar en buen estado de salud, puede realizar un *retorno* para volver a la configuración original.

Roles de instancia activos y pasivos

En el contexto de la conmutación por error, una *instancia* puede ser un único componente, como una base de datos o una colección de varios componentes que componen una implementación de servicio en una región. En una solución completa, puede conmutar por error diferentes partes de esa solución de maneras diferentes y en situaciones diferentes.

Un componente, o una colección de componentes configurados para la conmutación por error y la conmutación por recuperación, requiere varias instancias. Cada una de esas instancias asume un *rol* determinado:

- *Las instancias principales* o *activas* funcionan activamente, como atender las solicitudes entrantes de los clientes. Normalmente, hay solo una instancia principal a la vez.
- *Las* instancias *secundarias o pasivas* están inactivas, pero están listas para cambiarse para convertirse en la principal si es necesario. Puede haber varias instancias secundarias.

Hay varias maneras de configurar instancias pasivas. Cada manera implica inconvenientes entre el tiempo de recuperación y otros factores, como el costo y la complejidad operativa:

- *Sistemas de respaldo en caliente*, diseñados para estar listos para aceptar el tráfico de producción en cualquier momento.

- *Esperas activas, diseñadas* para estar casi preparadas para aceptar el tráfico de producción, pero esto puede requerir algunos cambios de configuración o operaciones de escalado para completarse antes de que puedan aceptar el tráfico.
- *Las esperas ligeras piloto*, que se implementan parcialmente en una configuración mínima, y requieren que se complete una preparación significativa antes de que puedan aceptar el tráfico de producción.
- *Soportes fríos*, que podrían no implementarse en absoluto, y confían en que los componentes se desplieguen antes de que puedan aceptar el tráfico de producción.

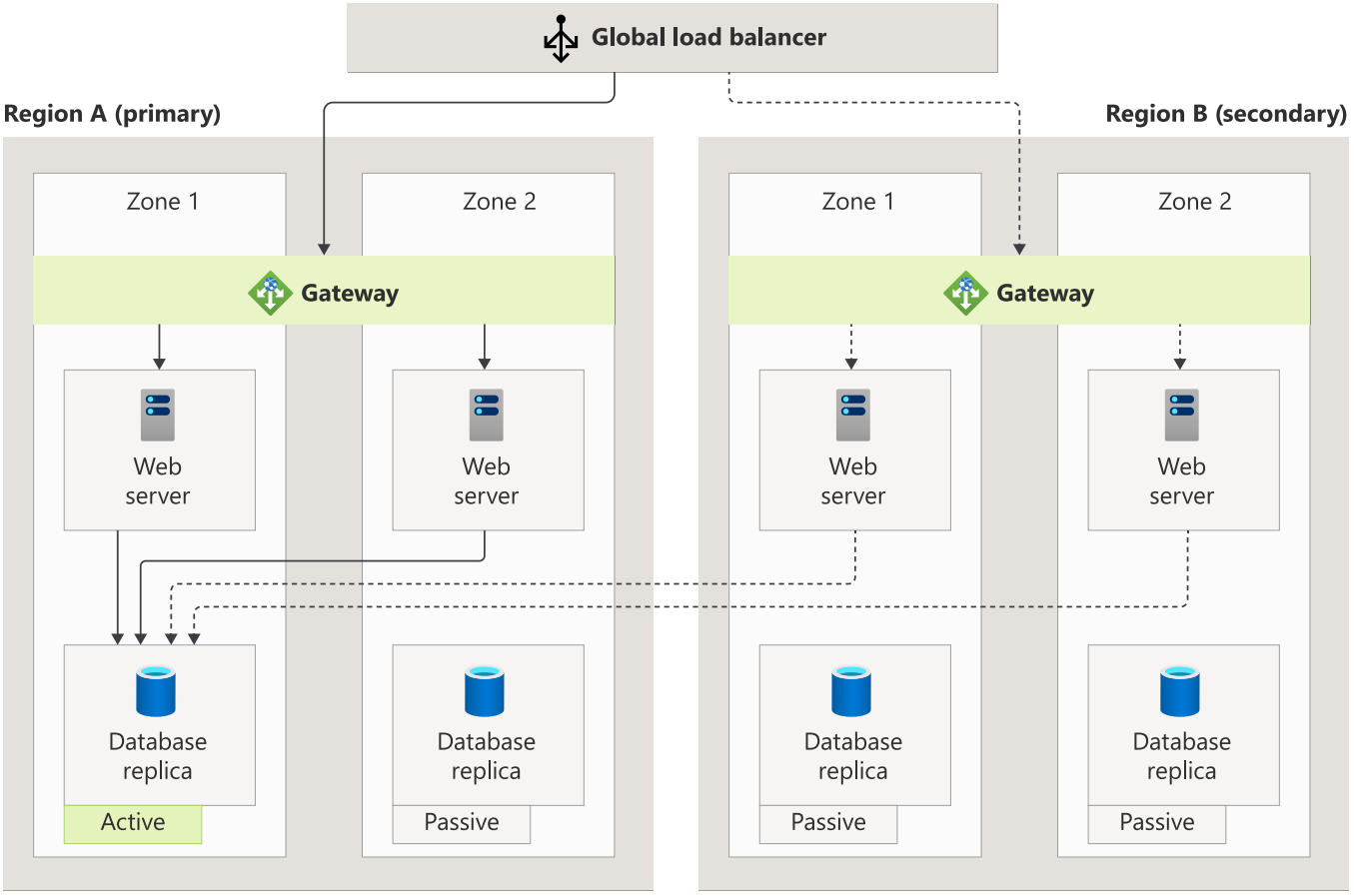
> **Sugerencia**
>
> Algunas soluciones se crean para usar un enfoque *activo-activo*, lo que significa que varias instancias atienden las solicitudes. Un sistema activo-activo no requiere conmutación por error, ya que todas las instancias atienden activamente las solicitudes en todo momento.

Ámbitos de conmutación por error

Las diferentes situaciones requieren estrategias de conmutación por error diferentes. Para ilustrar estas estrategias posibles, considere una solución de ejemplo que consta de una aplicación que accede a los datos de una base de datos. Configura la solución para la conmutación por error creando copias redundantes de tu servidor de aplicaciones y múltiples réplicas de la base de datos. A continuación, configure:

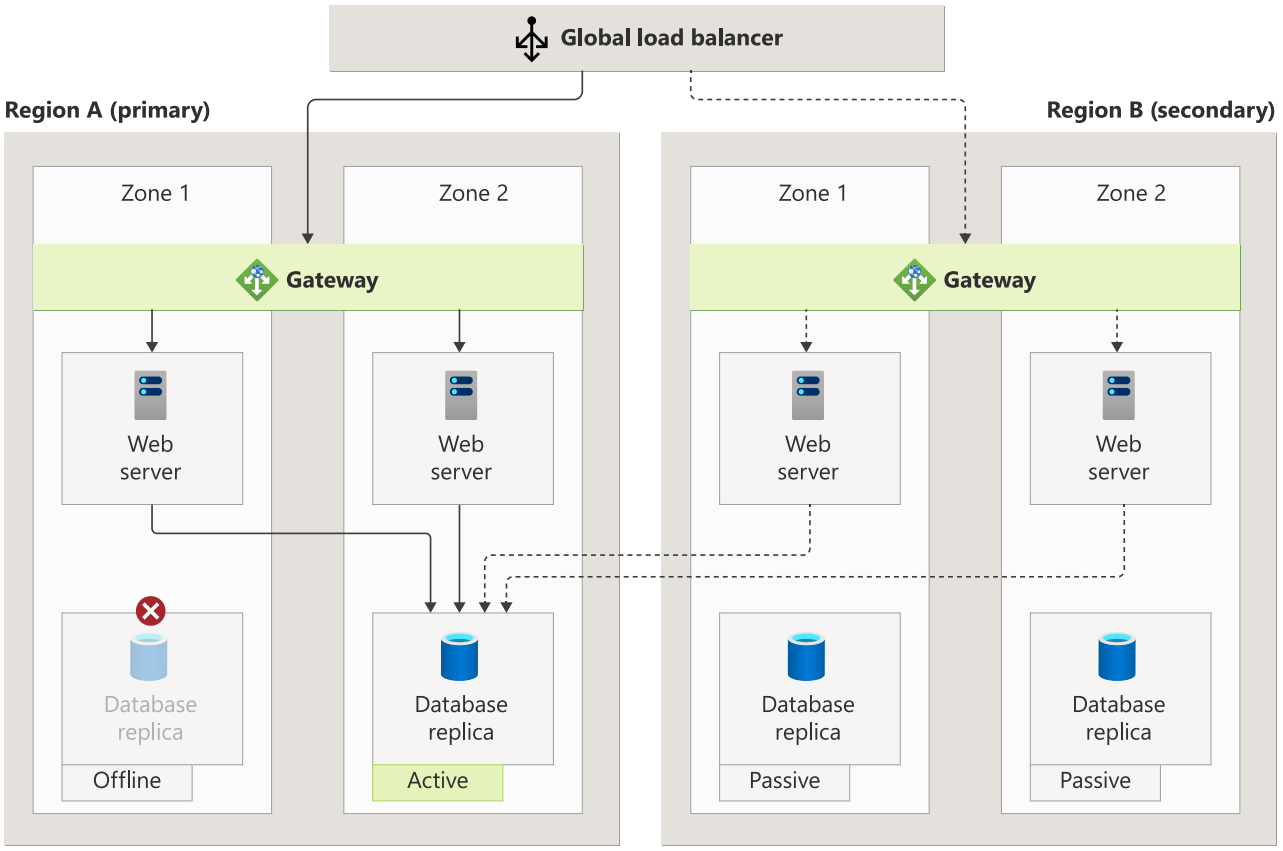
- Redundancia de zona colocando copias y réplicas en diferentes zonas de disponibilidad dentro de una región de Azure.
- Redundancia geográfica mediante un equilibrador de carga global para el cambio automático entre regiones.

Este es un diagrama simplificado que ilustra la arquitectura general en operaciones normales:

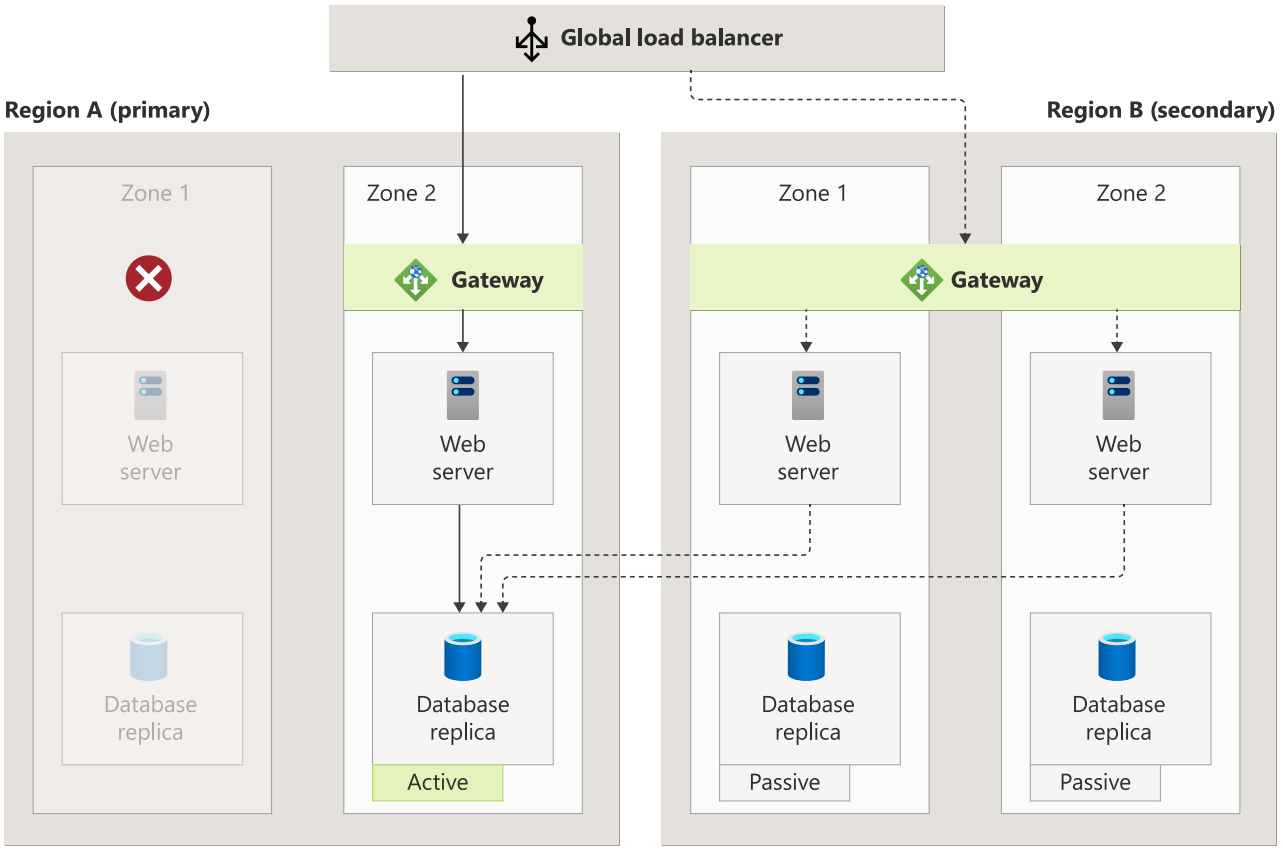


Diferentes situaciones pueden desencadenar eventos de conmutación por error diferentes en esta solución. Cada uno de ellos corresponde a un *ámbito de conmutación por error*, que representa la granularidad de los componentes que conmutan por error:

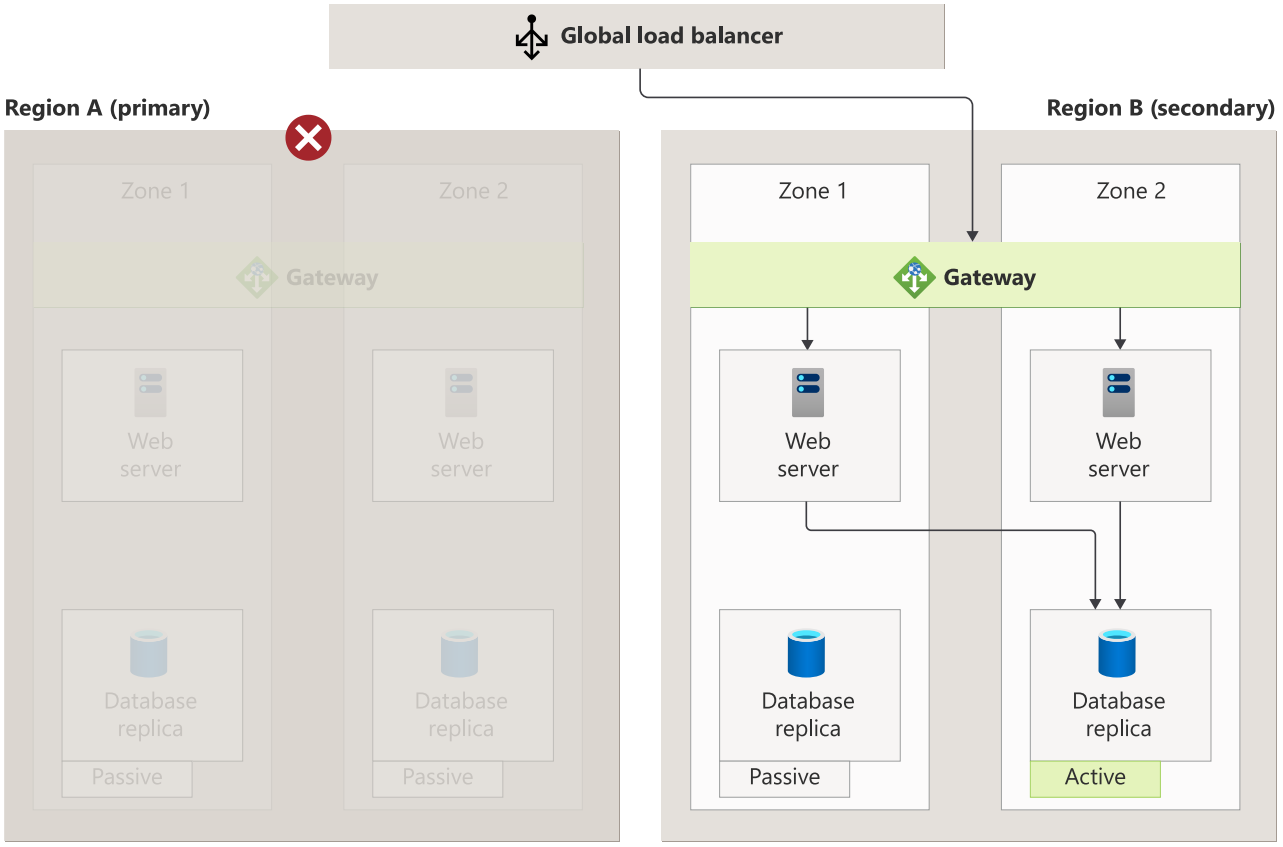
- Es posible que se produzca **una conmutación por error de réplica de base de datos** cuando la réplica de base de datos activa deje de estar disponible. La réplica pasiva se promueve para convertirse en la réplica activa. Normalmente, las aplicaciones pueden volver a enrutar rápidamente sus solicitudes a la nueva réplica activa:



- **Una conmutación por error de zona de disponibilidad** puede producirse si una zona de disponibilidad en su totalidad experimenta una interrupción. Este tipo de interrupción requiere que todo el tráfico se enrute al servidor web de la zona restante y también garantiza que la réplica de base de datos de la zona sobreviviente se convierta en la réplica activa si aún no lo es.



- **Un conmutación por error de región** podría ocurrir si hay una pérdida catastrófica de toda la región primaria de Azure.



Aunque cada uno de estos ámbitos proporciona un tipo de conmutación por error, es posible que tengan distintos requisitos y procesos de conmutación por error. Además, Microsoft puede

ser responsable de algunos ámbitos de conmutación por error, como cuando se usan servicios con redundancia de zona, mientras que es posible que sea responsable de la conmutación por error en ámbitos más amplios, como la conmutación por error entre regiones de Azure.

Planeación de la conmutación por error y continuidad empresarial

Una parte del plan de continuidad empresarial es diseñar sus estrategias de failover, que incluyen los diferentes ámbitos en los que puede realizar la conmutación.

Por lo general, los planes de continuidad empresarial deben incluir procedimientos de conmutación por error automatizados dentro o entre zonas de disponibilidad. Este tipo de conmutación por error forma parte de la estrategia de alta disponibilidad. Por ejemplo, si se produce un error en la réplica activa de una base de datos, un proceso automatizado puede promover una réplica pasiva para convertirse en la réplica activa. A continuación, los servidores web se comunican con la nueva réplica activa. De forma similar, si se produce un error en una zona de disponibilidad, muchas soluciones se compilan para recuperarse automáticamente mediante el uso de las zonas restantes.

Se emplean diferentes procedimientos de conmutación por error para escenarios de desastres, como en el poco probable caso de una interrupción total de la región. En el caso de una interrupción de región, puede redirigir las solicitudes web entrantes para utilizar una segunda región, así como desencadenar una conmutación por error de la base de datos a una réplica en la región secundaria.

Tenga en cuenta que, al incluir procedimientos de conmutación por error en el planeamiento de continuidad empresarial, es necesario realizar pruebas y diseño más detallados. Para obtener más información, consulte ¿Qué son la continuidad empresarial, la alta disponibilidad y la recuperación ante desastres?.

Conmutaciones por error planeadas y no planeadas

Las conmutaciones por error no planeadas son aquellas que se realizan durante una interrupción de un componente, de modo que pueda restaurar el servicio mediante otra instancia. Las conmutaciones por error no planeadas a veces producen tiempo de inactividad o pérdida de datos, en función de cómo se diseñe una solución. Las conmutaciones por error no planeadas requieren algo para detectar el error y tomar una decisión sobre cuándo desencadenar la conmutación por error.

En cambio, las *conmutaciones por error planeadas* son las que se desencadenan de forma proactiva. En previsión de que ocurra algo, como cuando una máquina virtual va a ser parcheada y reiniciada. Las conmutaciones por error planeadas pueden tener una tolerancia menor para el tiempo de inactividad y la pérdida de datos, ya que forman parte de los procedimientos de mantenimiento normales.

Cómo funciona la conmutación por error

Por lo general, el proceso de conmutación por error involucra los siguientes pasos, que se pueden realizar manualmente o mediante un sistema automatizado. Los detalles específicos de cada uno de estos pasos dependen del sistema concreto.

1. **Detectar un error (solo conmutaciones por error no planeadas).** Una conmutación por error automatizada requiere que algo detecte cuándo una instancia no está disponible, que normalmente se basa en algún tipo de comprobación de estado. Los diferentes servicios definen su estado de maneras diferentes. Por ejemplo, algunos servicios envían eventos *heartbeat* de forma proactiva entre instancias. Otros requieren un componente independiente para sondear cada instancia a intervalos regulares. A menudo se tarda algún tiempo en que los monitores de estado detecten un error en una instancia y a menudo es importante proporcionar un período de gracia en caso de que la instancia simplemente estaba ocupada y no pudo responder.

2. **Elija la conmutación por fallo.** En algún momento, se tomará la decisión de realizar una conmutación por error. Una herramienta automatizada podría tomar la decisión o manualmente. La tolerancia al riesgo de su organización puede afectar a la rapidez con la que se toma esta decisión. Si tiene una tolerancia baja al riesgo, puede optar por cambiar a modo de reserva rápidamente ante cualquier indicio de problema. Si tiene una mayor tolerancia al riesgo, puede optar por esperar para ver si el problema se puede resolver antes de realizar un traspaso.

3. **Seleccione una nueva instancia principal.** Una de las instancias restantes debe convertirse en la nueva primaria.

   En algunas situaciones, es posible que tenga predefinida qué instancia debería convertirse en la nueva principal o que solo disponga de una instancia para cambiar.

   En otras situaciones, hay un proceso automatizado por el que el sistema selecciona una nueva instancia principal. Hay varios *algoritmos de consenso* que se usan en la computación distribuida, incluido para las *elecciones de líderes*. Estos algoritmos se implementan dentro

de los servicios pertinentes, como las bases de datos. En algunos sistemas, es importante que cada instancia sea informada de la nueva réplica principal, por lo que los resultados de la selección se anuncian automáticamente a cada réplica.

4. **Solicitudes de redirección.** Configure el entorno para que las solicitudes se dirijan a las instancias correctas o a la nueva instancia principal.

   Para lograr esto, es posible que tenga que actualizar otros sistemas para que sepan dónde enviar solicitudes. Esto puede implicar la actualización de un sistema de equilibrio de carga para excluir la instancia incorrecta. En otras situaciones, el sistema de nombres de dominio (DNS) se usa normalmente como una manera de enviar solicitudes a una instancia activa de un sistema. Como parte del proceso de conmutación por error, normalmente debe actualizar los registros DNS para que las solicitudes se enruten a la nueva instancia principal. DNS tiene el concepto de tiempo de vida (TTL), que indica a los clientes con qué frecuencia deben comprobar los registros DNS actualizados. Si el TTL se establece en un valor largo, los clientes pueden tardar tiempo en recibir información sobre la conmutación por error y puede que sigan enviando solicitudes a la principal anterior.

Dado que los procesos de conmutación por error pueden incluir retrasos, es importante planear los procedimientos de conmutación por error para cumplir los requisitos de tiempo de inactividad (objetivo de punto de recuperación o RTO) y pérdida de datos (objetivo de punto de recuperación o RPO). Para más información, consulte ¿Qué son la continuidad empresarial, la alta disponibilidad y la recuperación ante desastres?.

Failback

Recuperación es el proceso de reestablecer y redirigir el tráfico de nuevo a la instancia principal original.

En algunas situaciones, no es necesario realizar la recuperación en absoluto porque cada instancia es igualmente capaz de actuar como primaria. Sin embargo, hay algunas situaciones en las que es importante realizar una conmutación por recuperación, como cuando es necesario ejecutar tus aplicaciones desde una región de Azure específica pero se ha realizado una conmutación por error a otra región de forma temporal debido a una interrupción regional.

A veces, la conmutación por recuperación se controla de la misma manera que una conmutación por error. Sin embargo, el retorno también puede ser más complejo que la conmutación por error, por varias razones:

- **Problemas de sincronización de datos.** Durante, e incluso después, una conmutación por error normal, es posible que la instancia principal anterior haya realizado algún trabajo o haya escrito algunos datos en un almacén de datos. Parte del proceso de retorno consiste en garantizar la coherencia y la integridad de los datos en toda su solución, incluida la administración de conflictos o duplicaciones entre las instancias principales y secundarias.

  Es habitual que los problemas de sincronización de datos requieran intervención manual. Si no necesita los datos en conflicto, puede optar por restablecer la base de datos u otro estado.

- **Pasos de corrección.** Si se intentaron pasos de remediación en la principal antes de que se produjera la conmutación por error, podrían haber dejado la instancia principal en un estado desconocido.

  Si existe un riesgo de que la instancia principal esté en un estado incoherente, es posible que tenga que destruir y volver a implementar la instancia principal para que esté en un estado correcto conocido antes de realizar la conmutación por recuperación.

- **Tiempo de inactividad adicional.** El tiempo de inactividad durante el proceso de conmutación por recuperación puede ser mayor que durante una conmutación por error debido a reconfiguraciones necesarias o operaciones para restaurar la coherencia de los datos.

  Puede mitigar este problema ejecutando procesos de retorno durante una ventana de mantenimiento o informando a los usuarios sobre el cambio con anticipación. Además, es posible que pueda realizar algunas de las operaciones preparatorias mientras el sistema está en línea y minimizar el tiempo de inactividad necesario.

- **Tolerancia al riesgo.** Si se produjo la conmutación por error debido a una interrupción, la tolerancia de la organización al tiempo de inactividad u otros riesgos durante la conmutación por recuperación puede ser menor.

  Las partes interesadas de la empresa deben mantenerse informados de la situación durante todo el proceso, y deben hacerse plenamente conscientes de la necesidad de la conmutación por recuperación y las consecuencias de los procedimientos de conmutación por recuperación. Es posible que pueda negociar un momento adecuado para realizar los cambios.

Conmutación por error y conmutación por recuperación en servicios de Azure

Aunque es importante comprender cómo funciona la conmutación por error en general, tenga en cuenta que cada servicio Azure puede abordar la conmutación por error y la conmutación por recuperación de forma diferente. Para obtener información sobre cómo funcionan los servicios Azure específicos desde una perspectiva de confiabilidad, consulte la guía de confiabilidad de cada servicio.

Muchos de los servicios de Azure controlan determinados tipos de conmutación por error automáticamente. Por ejemplo, cuando se usan los servicios de Azure configurados para que sean con redundancia de zonas, Microsoft realiza automáticamente la conmutación ante fallos entre zonas de disponibilidad. Para obtener más información, consulte ¿Qué son las zonas de disponibilidad? y las guías de confiabilidad del servicio Azure.

Si usa máquinas virtuales, Azure Site Recovery replica las máquinas virtuales y sus discos entre zonas de disponibilidad o hacia otra región de Azure, y puede realizar la conmutación por error para usted.

Al diseñar su propia solución que combina varios servicios Azure juntos, es posible que los requisitos de conmutación por error se vuelvan más complejos. Supongamos que va a diseñar una solución con un nivel de aplicación y una base de datos, y quiere crear una arquitectura activa/pasiva multiregión. Durante una interrupción en la región primaria, es importante que tanto la aplicación como la base de datos realicen una conmutación por error hacia la región secundaria. En función de los servicios exactos que use, es posible que deba planear su propia estrategia de conmutación por error para cambiar entre los despliegues de cada región. Azure proporciona enrutamiento de tráfico global y equilibrio de carga a través de Azure Front Door y Azure Traffic Manager, y puede seleccionar la tecnología que satisface sus necesidades de conmutación por error. Cada servicio admite la supervisión del estado de cada instancia regional de la aplicación y puede configurarlo para volver a enrutar automáticamente el tráfico a la instancia correcta.

Pasos siguientes

- Obtenga información sobre la responsabilidad compartida de confiabilidad.
- Obtenga información sobre recomendaciones para el diseño multiregión de alta disponibilidad en Azure Well-Architected Framework.

ⓘ**Nota:** El autor ha creado este artículo con ayuda de inteligencia artificial. Más información

---

Last updated on 13/08/2026

Confiabilidad y soberanía en Azure

Azure admite un espectro de opciones de implementación, incluidas las regiones de nube pública, los entornos híbridos y los entornos de nube soberana o nacional. Esta flexibilidad le ayuda a diseñar la confiabilidad al cumplir los requisitos normativos y jurisdiccionales.

En este artículo se explica cómo las consideraciones de soberanía influyen en las decisiones de diseño de confiabilidad y en qué planear en los puntos de decisión clave de la arquitectura.

Qué significa soberanía para una carga de trabajo confiable

La soberanía significa que conserva el control sobre sus datos y garantiza que los datos permanecen sujetos a las leyes de su jurisdicción. En la práctica, la soberanía afecta al diseño de confiabilidad de dos maneras clave:

- **Control de acceso:** Solo las partes autorizadas pueden acceder o mover datos. Los controles incluyen cifrado, claves administradas por el cliente, control de acceso basado en roles y controles operativos.
- **Control geográfico:** Los datos permanecen dentro de los límites geográficos especificados para que se rige por las leyes locales.

La soberanía se basa en los resultados. No requiere aislamiento de la nube global. Se pueden cumplir muchos requisitos de cumplimiento en entornos de nube pública cuando se aplican los controles adecuados, incluidos los controles de residencia de datos, el cifrado, la auditabilidad y las protecciones de acceso legal.

Capacidades de Azure que respaldan la soberanía

Azure proporciona varias funcionalidades que le ayudan a cumplir los requisitos de soberanía:

- **Controles de datos:** Las regiones de Azure se agrupan en geografías que definen límites de residencia de datos. Azure también admite el cifrado y las claves administradas por el cliente, incluida la administración centralizada de claves a través de Azure Key Vault.
- **Controles operativos:** Azure Policy puede restringir la ubicación de los recursos a regiones específicas, el control de acceso basado en roles puede limitar los permisos y Customer Lockbox puede regular el acceso del soporte técnico de Microsoft. También puede usar

arquitecturas de registro e gobernanza inmutables para aplicar límites de protección de cumplimiento a escala.

- **Opciones de implementación e infraestructura:** Azure admite varios modelos de implementación y aislamiento. Azure Government y Azure en China son nubes físicamente aisladas que operan independientemente de la Azure global. Microsoft Sovereign Cloud admite patrones de aislamiento lógico dentro de la nube pública de Azure. Las opciones híbridas, como Azure Local, admiten escenarios de operación controlados por el cliente y sin conexión.

Dónde se intersecan las decisiones de diseño de confiabilidad y soberanía

Al diseñar la confiabilidad de la solución, también debe tener en cuenta y planear cualquier problema de soberanía. En esta sección se resaltan algunas decisiones que normalmente necesita tomar al considerar la confiabilidad y la soberanía.

Selección de región y zona de disponibilidad

El lugar en el que coloca la infraestructura redundante determina tanto la confiabilidad como la posición de cumplimiento. Los diseños de varias regiones suelen mejorar la confiabilidad, pero las restricciones de soberanía pueden limitar la selección de regiones a límites geográficos o geográficos aprobados.

Si diseña una solución basada en recuperación ante desastres con varias regiones y la carga de trabajo debe permanecer dentro de una jurisdicción específica, elija una región de recuperación ante desastres en ese mismo límite. Si no hay ninguna región secundaria compatible disponible, use la confiabilidad de una sola región con la recuperación basada en copia de seguridad y documente los inconvenientes de recuperación relacionados.

Además de los patrones multiregión y de varias zonas en Azure nube pública, puede usar Azure Local para entornos controlados por el cliente, como centros de datos locales o sitios remotos. Azure Local admite modelos operativos conectados y desconectados, lo que puede ayudarle a mantener la continuidad empresarial de las cargas de trabajo locales cuando se interrumpe la conectividad en la nube.

Azure también proporciona ofertas soberanas para los clientes con requisitos jurisdiccionales y normativos específicos. Entre los ejemplos se incluyen Azure Government, Azure en China y Microsoft nube soberana (anteriormente Microsoft Cloud for Sovereignty). Estos entornos tienen

límites operativos y de cumplimiento distintos, por lo que debe confirmar que la arquitectura de confiabilidad se alinea con los requisitos del entorno de nube específico que use.

> **Sugerencia**
>
> Las zonas de disponibilidad proporcionan una capa de confiabilidad adicional dentro de una región sin cruzar los límites geográficos. Use la arquitectura de varias zonas donde se admita.

Ubicaciones de copia de seguridad, replicación y recuperación ante desastres

Los destinos de copia de seguridad y replicación deben seguir los mismos límites normativos que la carga de trabajo que protegen. Algunos servicios admiten la replicación en regiones que elija, mientras que otros usan pares de regiones definidos por Azure.

Al almacenar copias de seguridad o réplicas fuera del límite principal, asegúrese de que los datos están cifrados y la colocación de claves está planeada cuidadosamente. Si las claves de cifrado solo se almacenan en la región primaria, es posible que no estén disponibles durante una interrupción de la región.

Cuando se usan claves administradas por el cliente con replicación en otra región de Azure que no es la región emparejada, alinee la ubicación de las claves con la ubicación de los datos. Azure Key Vault HSM administrado admite la distribución de claves entre regiones para ayudar a mantener la disponibilidad de claves alineada con los datos protegidos.

Acceso operativo y auditabilidad

Las operaciones de confiabilidad deben permanecer controladas y auditables. Las prácticas de implementación segura de Azure despliegan los cambios de forma incremental y aíslan los dominios de fallo para reducir el riesgo de impacto multirregional.

Para las cargas de trabajo reguladas, Customer Lockbox ayuda a garantizar que los ingenieros de soporte técnico de Microsoft no puedan acceder a los datos de los clientes sin una aprobación explícita. Use Azure Monitor y Azure Activity Log para conservar los registros de conmutación por error y de recuperación para informes de auditoría y cumplimiento normativo.

Contenido relacionado

- ¿Qué son las regiones de Azure?
- ¿Qué son las zonas de disponibilidad?
- Pares de regiones de Azure y regiones que no están emparejadas
- Azure Government
- Azure en China
- Nube soberana de Microsoft
- Información general de Azure Local
- Azure Key Vault Managed HSM: información general
- Lockbox de cliente para Microsoft Azure
- Introducción a Azure Backup
- Residencia de datos en Azure .

ⓘ**Nota:** El autor ha creado este artículo con ayuda de inteligencia artificial. Más información

---

Last updated on 13/08/2026

Cómo leer un acuerdo de nivel de servicio (SLA)

Un acuerdo de nivel de servicio (SLA) es uno de los documentos a los que se hace referencia con más frecuencia en la arquitectura en la nube, pero también es uno de los más frecuentes incomprendidos. Muchos ingenieros tratan un Acuerdo de Nivel de Servicio como una garantía de disponibilidad simple, pero es un documento contractual que incluye definiciones, condiciones y exclusiones precisas que dan forma a su significado. Si no lee detenidamente un Acuerdo de Nivel de Servicio, puede tomar decisiones arquitectónicas basadas en suposiciones no admitidas.

En este artículo se explica cómo interpretar los SLA de forma metódica para que pueda usarlos como entradas de ingeniería al diseñar cargas de trabajo resistentes. En este artículo se describe cómo:

- Explicar qué garantiza realmente un Acuerdo de Nivel de Servicio.
- Identifique cómo los proveedores de servicios tecnológicos suelen definir y medir la disponibilidad.
- Reconocer condiciones y exclusiones que afectan a la cobertura.
- Utilice acuerdos de nivel de servicio para informar las decisiones arquitectónicas, no para reemplazar la robustez que diseña.

La mayoría de los Acuerdos de Nivel de Servicio definen compromisos en torno al tiempo de actividad o la disponibilidad, pero los proveedores de servicios también pueden definir acuerdos de nivel de servicio en torno a otros requisitos, como el rendimiento, el tiempo de recuperación o la durabilidad de los datos.

> **Importante**
>
> En este artículo se explica cómo *interpretar* acuerdos de nivel de servicio. No proporciona asesoramiento legal ni orientación sobre cómo disputar con proveedores de servicios o reclamar créditos de servicio. No es específico de los SLA de Microsoft. Los ejemplos y principios se aplican ampliamente para ayudarle a leer el Acuerdo de Nivel de Servicio de cualquier proveedor de servicios de forma crítica.

Cómo se relacionan los SLA con los SLO y los objetivos de confiabilidad

Antes de leer un Acuerdo de Nivel de Servicio, comprenda dónde encaja junto con otros conceptos de confiabilidad:

- Un **Acuerdo de Nivel de Servicio** es un compromiso contractual entre un proveedor de servicios y un cliente, con consecuencias definidas para los objetivos no establecidos.
- Un **objetivo de nivel de servicio (SLO)** es un objetivo de confiabilidad interno que establece para su propia carga de trabajo.

Un Acuerdo de Nivel de Servicio describe la garantía *mínima* a la que se compromete el proveedor de servicios. Un SLO refleja la confiabilidad que *los usuarios* realmente necesitan. Utilice acuerdos de nivel de servicio (ANS) para informar su objetivo de nivel de servicio (SLO), pero no adopte simplemente el porcentaje de un ANS como meta de confiabilidad de su carga de trabajo. Tenga en cuenta su propio código, dependencias, procesos operativos y tolerancia de los usuarios para el tiempo de inactividad. A veces, necesita objetivos de nivel de servicio más estrictos que los acuerdos de nivel de servicio combinados de sus proveedores de servicios. En otras ocasiones, la carga de trabajo puede tolerar problemas que superen lo que el proveedor de servicios se compromete financieramente.

Para obtener más información sobre cómo definir los SLO y las metas de confiabilidad, consulte Recomendaciones para definir metas de confiabilidad.

Por qué debe leer acuerdos de nivel de servicio

Los Acuerdos de Nivel de Servicio son señales de ingeniería, no garantías de disponibilidad estrictas. Describen las condiciones en las que un proveedor de servicios se compromete a un nivel de servicio definido. También describen las consecuencias financieras si el proveedor de servicios no cumple esos compromisos. Los Acuerdos de Nivel de Servicio no garantizan que la aplicación sea confiable, que los usuarios eviten tiempo de inactividad o que el proveedor de servicios impida todos los errores.

Las suposiciones sobre los Acuerdos de Nivel de Servicio conducen a decisiones arquitectónicas deficientes. Los arquitectos a veces tratan el porcentaje de tiempo de actividad de un Acuerdo de Nivel de Servicio como un indicador directo de la confiabilidad del servicio. Este enfoque suele pasar por alto los detalles críticos, incluido cómo se define el tiempo de inactividad, qué condiciones deben cumplirse y qué escenarios se excluyen. El diseño de la carga de trabajo resultante podría no tolerar los modos de error que implica el Acuerdo de Nivel de Servicio.

Una revisión cuidadosa de los Acuerdos de Nivel de Servicio descubre señales importantes para diseñar sistemas resistentes:

- **Se esperan errores transitorios:** Los proveedores de servicios esperan que se produzcan errores de corta duración. Los Acuerdos de Nivel de Servicio hacen que esta expectativa sea explícita mediante definiciones de tiempo de inactividad y requisitos de reintento. Diseñe la arquitectura para controlar los errores transitorios como una condición normal, no una excepcional.
- **Los servicios no son equivalentes:** Los servicios varían en sus ámbitos, condiciones y exclusiones del Acuerdo de Nivel de Servicio.
- **Los porcentajes de tiempo de actividad no se aplican uniformemente:** Un servicio con un Acuerdo de Nivel de Servicio 99.99% para sus operaciones principales podría ofrecer un Acuerdo de Nivel de Servicio inferior o ningún Acuerdo de Nivel de Servicio en absoluto para características específicas de las que dependa.
- **Los Acuerdos de Nivel de Servicio no cubren todos los factores de confiabilidad:** Los Acuerdos de Nivel de Servicio por sí solos no abordan todas las consideraciones de confiabilidad. Revise también los límites de servicio, las cuotas, los comportamientos de restricción, las limitaciones conocidas y las orientaciones operativas en la documentación del producto.

Qué es y qué no es un Acuerdo de Nivel de Servicio

Un Acuerdo de Nivel de Servicio es un conjunto de compromisos condicionales. Indica que *si* un cliente cumple ciertos *requisitos, el* proveedor de servicios se compromete a un nivel de servicio definido. *Si* el proveedor de servicios no cumple ese *nivel, el* cliente recibe una compensación, normalmente en forma de créditos de servicio.

Los Acuerdos de Nivel de Servicio solo se aplican cuando se cumplen condiciones previas específicas, como requisitos de configuración, topologías de implementación o comportamientos operativos. El Acuerdo de Nivel de Servicio especifica exactamente lo que cuenta como un error. Si un problema no cumple la definición documentada del tiempo de inactividad, el Acuerdo de Nivel de Servicio no lo cubre.

No cree un *SLA compuesto* que multiplique los porcentajes de SLA para calcular la disponibilidad de un sistema que tenga múltiples dependencias. Este cálculo puede proporcionar un límite máximo aproximado, pero no es confiable en la práctica. Se supone que todos los servicios producen un error de forma independiente, lo que rara vez sucede. También omite las definiciones, condiciones y exclusiones únicas de cada Acuerdo de Nivel de Servicio. Los cálculos compuestos suelen hacer que un sistema parezca más confiable de lo que es y no reemplazan la definición de SLO adecuada o el análisis del modo de error (FMA) para la carga de trabajo. Para obtener más información, vea Recomendaciones para definir objetivos de fiabilidad.

Un modelo práctico para leer acuerdos de nivel de servicio

No lea un Acuerdo de Nivel de Servicio de principio a fin. En su lugar, use un enfoque estructurado de cinco pasos que extraiga la información que importa para la arquitectura.

Paso 1: Definiciones

Comience con la sección de definiciones. Las definiciones controlan cómo el Acuerdo de Nivel de Servicio mide la disponibilidad y qué errores cuentan.

- **Las definiciones importan tanto como porcentajes:** Un Acuerdo de Nivel de Servicio de 99.99% suena impresionante, pero su valor depende completamente de lo que el Acuerdo de Nivel de Servicio cuenta como un error y qué unidades mide. Cuando la definición de tiempo de inactividad es estrecha, la cobertura efectiva es más estrecha que el porcentaje que sugiere.
- **Términos clave:** Los términos como *la solicitud válida*, *el tiempo de inactividad* y *los minutos facturables* definen a qué se compromete el proveedor de servicios. Si una operación no se califica como una solicitud válida en el Acuerdo de Nivel de Servicio, no afecta al cálculo de disponibilidad.

  Identifique qué operaciones cubre el Acuerdo de Nivel de Servicio, qué componentes excluye explícita o implícitamente, y si se aplica a cada instancia de servicio o a través de un grupo de instancias.

> **Sugerencia**
>
> Un Acuerdo de Nivel de Servicio de 99.9% no garantiza que el servicio siempre permanezca disponible. Tampoco significa que debería esperar solo 8,7 horas de tiempo de inactividad al año. La definición del SLA del tiempo de inactividad determina lo que se va a medir. Las definiciones estrechas significan que el servicio puede experimentar problemas que afectan a los usuarios sin infringir el Acuerdo de Nivel de Servicio.

Paso 2: Medida

Comprender cómo el SLA calcula la disponibilidad:

- **Medición basada en el tiempo frente a la medida basada en solicitudes:** La medición basada en tiempo normalmente realiza un seguimiento de si el servicio estaba disponible durante cada intervalo de tiempo. La medida basada en solicitudes realiza un seguimiento de la proporción de respuestas correctas a las solicitudes válidas totales. La misma interrupción puede producir porcentajes de disponibilidad diferentes en función del método que use el Acuerdo de Nivel de Servicio.
- **Reintentos y agregación:** Algunos Acuerdos de Nivel de Servicio requieren que vuelva a intentar las solicitudes con errores durante un período especificado antes de que se cuente como errores. Otras SLA agregan disponibilidad durante todo el período de facturación, por lo que es posible que las interrupciones breves no reduzcan la disponibilidad calculada suficiente para desencadenar un crédito.
- **Ámbito de agregación:** Determine si el Acuerdo de Nivel de Servicio mide la disponibilidad por instancia de servicio individual o entre todas las instancias de un ámbito, como una cuenta o una suscripción. Cuando el Acuerdo de Nivel de Servicio calcula la disponibilidad en varias instancias, las instancias correctas pueden desplazar las instancias con errores. Este cálculo podría producir una disponibilidad notificada más alta que cualquier instancia única realmente experimentada.

> **Sugerencia**
>
> Los Acuerdos de Nivel de Servicio suelen medir el tiempo de actividad durante un período de facturación, no en tiempo real. Un servicio puede experimentar una interrupción que dura varios minutos y sigue cumpliendo su

Acuerdo de Nivel de Servicio durante el mes.

Paso 3: Requisitos previos, variaciones y exclusiones

Identifique las condiciones que requiere el Acuerdo de Nivel de Servicio:

- **Configuraciones y comportamientos necesarios:** Muchos acuerdos de nivel de servicio requieren configuraciones de implementación específicas o comportamientos operativos, como implementar lógica de reintento o permanecer dentro de las cuotas documentadas. Por ejemplo, un Acuerdo de Nivel de Servicio de almacenamiento solo se puede aplicar cuando se usan unidades de estado sólido (SSD) en lugar de unidades de disco duro (HDD) porque el compromiso de tiempo de actividad del proveedor de servicios depende de las características de rendimiento de ese tipo de almacenamiento.

- **La configuración afecta a los términos del Acuerdo de Nivel de Servicio:** Examine cómo la configuración del servicio cambia el Acuerdo de Nivel de Servicio. Los niveles de precio superior a veces ofrecen compromisos de Acuerdo de Nivel de Servicio diferentes, pero la diferencia no siempre es un porcentaje de tiempo de actividad mayor. Los niveles superiores pueden tener definiciones diferentes, métodos de medición diferentes o menos exclusiones. Un Acuerdo de Nivel de Servicio de nivel Premium podría definir un tiempo de inactividad más amplio (mejor cobertura) en lugar de aumentar el porcentaje de tiempo de actividad. No confíes solo en el porcentaje. Lea también las definiciones.

  Los modelos de implementación también afectan a los términos del Acuerdo de Nivel de Servicio. Para el mismo servicio, una implementación de instancia única podría tener un Acuerdo de Nivel de Servicio menor que una implementación de varias instancias que proporcione redundancia.

- **Exclusiones ambientales y operativas:** Los Acuerdos de Nivel de Servicio suelen excluir el tiempo de inactividad del mantenimiento planeado, fuerza mayor, acciones iniciadas por el cliente, configuración incorrecta, cuotas superadas, solicitudes con formato incorrecto y características en versión preliminar o no admitidas.

Sugerencia

Una implementación correcta no garantiza la cobertura del Acuerdo de Nivel de Servicio. El Acuerdo de Nivel de Servicio no se aplica si no cumple otras condiciones, como implementar la lógica de reintento necesaria o si un evento excluido provoca el error. Compruebe si las características de las que depende están dentro del ámbito del Acuerdo de Nivel de Servicio. Los Acuerdos de Nivel de Servicio suelen abarcar solo operaciones, puntos de conexión o niveles específicos. Pueden excluir características en versión preliminar, determinadas API y versiones de nivel inferior.

Paso 4: Implicaciones de diseño

Para usar el servicio de forma más eficaz, tenga en cuenta lo que revela el Acuerdo de Nivel de Servicio sobre cómo se comporta el servicio y cómo debe diseñar la solución en torno a ella:

- **Modos de error esperados:** Las definiciones y condiciones del Acuerdo de Nivel de Servicio indican qué tipos de error trata el proveedor de servicios como normales y esperados. La arquitectura debe tolerar estos errores.

- **Prioridades del proveedor de servicios:** La forma en que el proveedor de servicios define la disponibilidad, la mide y excluye determinados escenarios revela su compromiso principal y lo que considera su responsabilidad.

- **Decisiones de diseño:** Si el Acuerdo de Nivel de Servicio excluye determinados modos de error, decida si aceptar el riesgo, agregar redundancia o usar servicios alternativos.

Paso 5: Cómo reclamar créditos de servicio

El Acuerdo de Nivel de Servicio define créditos de servicio como compensación por tiempo de inactividad, pero los proveedores de servicios normalmente no los aplican automáticamente. No supervisan de forma proactiva las violaciones del SLA ni presentan reclamaciones por usted. Debe detectar interrupciones que infringen el Acuerdo de Nivel de Servicio, documentarlas y enviar notificaciones dentro de la fecha límite. Siga estos pasos para prepararse y realizar ese proceso.

Paso 1: Saber qué supervisar

Antes de que se produzca un incidente, identifique qué datos necesita recopilar para justificar una reclamación. Revise las definiciones y el método de medición del Acuerdo de Nivel de Servicio para determinar:

- **Qué métricas se asignan al Acuerdo de Nivel de Servicio:** Si el Acuerdo de Nivel de Servicio mide la tasa de éxito de la solicitud, supervise los códigos de respuesta HTTP de los puntos de conexión que cubre el Acuerdo de Nivel de Servicio. Si mide el tiempo de actividad en minutos, supervise la disponibilidad del servicio a intervalos regulares.
- **Qué instancias están en el ámbito:** Los Acuerdos de Nivel de Servicio a menudo se aplican a niveles, versiones o configuraciones de implementación específicos. Supervise solo las instancias que califican para la cobertura del Acuerdo de Nivel de Servicio.
- **Lo que el proveedor considera un error:** Algunos Acuerdos de Nivel de Servicio definen el tiempo de inactividad de forma estrecha. Por ejemplo, el Acuerdo de Nivel de Servicio podría requerir que todas las solicitudes produzcan un error durante uno o varios minutos consecutivos, o podría requerir una interrupción documentada en una lista pública. Si es necesario, configure la supervisión para capturar las condiciones de error específicas que define el Acuerdo de Nivel de Servicio.

Paso 2: Mantener registros durante y después de un incidente

Cuando se produce una interrupción o degradación, recopile y conserve evidencias que se alinean con las definiciones del Acuerdo de Nivel de Servicio:

- **Marcas de tiempo:** Registre la hora de inicio y finalización del incidente, incluidas las zonas horarias. Tenga en cuenta la duración del impacto y si era continua o intermitente.
- **Detalles del error:** Capture los códigos de error específicos, los mensajes de error o los comportamientos de tiempo de espera que observe. Estos detalles deben corresponder a las condiciones de error que define el Acuerdo de Nivel de Servicio.
- **Registros de solicitudes:** Si el Acuerdo de Nivel de Servicio requiere reintentos antes de contar un error, conserve los registros para mostrar que cumple ese requisito. Incluya el número de reintentos, los intervalos de retroceso y el resultado de cada intento.
- **Evidencia de configuración:** Si el Acuerdo de Nivel de Servicio requiere una configuración de implementación específica, como una cantidad específica de redundancia o un nivel Premium, documente cómo las instancias cumplen esos requisitos previos en el momento del incidente.
- **Comunicaciones del proveedor de servicios:** Guarde las notificaciones de incidentes, las actualizaciones de página de estado o la correspondencia de soporte técnico del proveedor de servicios que confirmen el problema.

Paso 3: Determinar si el incidente infringe el Acuerdo de Nivel de Servicio

Antes de presentar una notificación, calcule si el incidente cumple el umbral de un crédito:

1. Aplique la fórmula de disponibilidad del Acuerdo de Nivel de Servicio mediante los datos registrados. Use las mismas unidades, como minutos, solicitudes u otra métrica, que especifica el Acuerdo de Nivel de Servicio.

2. Resta cualquier tiempo que se encuentre bajo exclusiones documentadas, como ventanas de mantenimiento programado, tipos de eventos excluidos o períodos en los que la configuración no cumplió con los requisitos previos del Acuerdo de Nivel de Servicio.

3. Compara el resultado con el porcentaje comprometido del Acuerdo de Nivel de Servicio (SLA). Si la disponibilidad calculada está por debajo del umbral, podría aplicarse un crédito.

Paso 4: Presentar la reclamación dentro del plazo requerido

La mayoría de los Acuerdos de Nivel de Servicio imponen una fecha límite para enviar notificaciones después de un incidente. Si pierde la fecha límite, perderá el crédito independientemente de la gravedad de la interrupción.

1. **Busque la ventana de reclamación:** Consulte el documento del Acuerdo de Nivel de Servicio para la fecha límite de envío. Los períodos de tiempo comunes oscilan entre 30 y 60 días después del final del mes de facturación cuando se produce el incidente, pero cada proveedor de servicios establece su propia fecha límite.

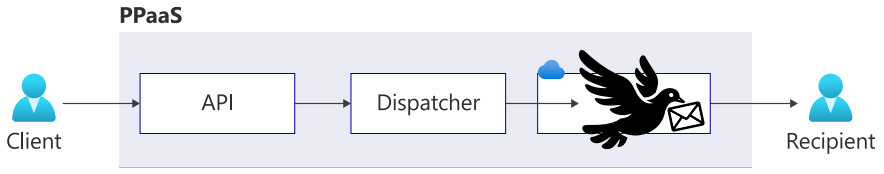
2. **Incluya pruebas complementarias:** Adjunte los registros recopilados en el paso 2, incluidas las marcas de tiempo, los registros de errores, la evidencia de reintento y los detalles de configuración. Los proveedores de servicios procesan las notificaciones bien documentadas más rápidamente.

3. **Envíe a través del canal correcto:** Normalmente, los proveedores de servicios requieren que se realicen notificaciones a través de un portal de soporte técnico o un proceso de solicitud formal. Siga el procedimiento documentado para evitar el rechazo de reclamaciones.

> ⓘ **Nota:**
>
> Los créditos de servicio suelen ser un porcentaje de la tarifa mensual del servicio afectado. Si su gasto mensual en un servicio es modesto, un crédito de 10% puede ser una pequeña suma, independientemente de la gravedad del impacto en su negocio. **Los créditos no cubren los ingresos perdidos, la intrición del cliente ni los daños de reputación.**

Un ejemplo guiado: Descodificar un Acuerdo de Nivel de Servicio ficticio

Para ilustrar cómo leer un Acuerdo de Nivel de Servicio en la práctica, en esta sección se usa un servicio ficticio, *Pigeon Post como servicio (PPaaS)*. PPaaS es un servicio de mensajería administrado hipotético que entrega mensajes a través de palomas portadoras entrenadas. Una API acepta un mensaje y lo envía a un distribuidor responsable de liberar la paloma y enviar el mensaje al destinatario.



PPaaS es intencionadamente absurdo, pero la estructura de su Acuerdo de Nivel de Servicio refleja muchos acuerdos de nivel de servicio reales. En este ejemplo se muestra el enfoque de cinco pasos.

> ⓘ **Nota:**
>
> PPaaS es completamente ficticio. Cualquier similitud con servicios reales, de aves o de otro tipo, es pura coincidencia.

Acuerdo de Nivel de Servicio de PPaaS

El Acuerdo de Nivel de Servicio de PPaaS incluye las siguientes definiciones:

- **Solicitud válida:** Un mensaje enviado a través de la API de PPaaS que se ajusta al esquema de mensajes documentado.
- **Tiempo de inactividad:** Un período en el que todos los intentos de enviar una *solicitud válida* a un punto de conexión PPaaS aprovisionado producen un error o no reciben ninguna respuesta.
- **Total de minutos facturables:** Número total de minutos del mes de facturación durante el cual se intentó al menos una operación de envío *de solicitud válida* .
- **Porcentaje de tiempo de actividad mensual**, calculado como:

$$\text{Porcentaje de tiempo de actividad mensual} = \frac{\text{Minutos Totales Facturables} - \text{Tiempo de Inactividad}}{\text{Minutos Totales Facturables}} \times 100$$

El Acuerdo de Nivel de Servicio de PPaaS incluye los siguientes créditos de servicio.

Expandir tabla

Nivel	Porcentaje de tiempo de actividad mensual	Crédito de servicio
Standard Coop	Menos de 99,9%	10 %
Standard Coop	Menos de 99%	25%
Basic Coop	No es aplicable	No es aplicable

El Acuerdo de Nivel de Servicio de PPaaS incluye las siguientes condiciones y exclusiones:

- El Acuerdo de Nivel de Servicio no cubre los errores causados por condiciones meteorológicas graves.
- El Acuerdo de Nivel de Servicio excluye hasta 15 minutos por semana del mantenimiento programado de la cooperación.
- Debe implementar la lógica de reintento con al menos tres reintentos y retroceso exponencial antes de que una solicitud se considere fallida.
- El Acuerdo de Nivel de Servicio no se aplica al nivel Basic Coop.

Lea el Acuerdo de Nivel de Servicio de PPaaS paso a paso.

Los siguientes pasos aplican el enfoque de cinco pasos al Acuerdo de Nivel de Servicio de PPaaS. Cada paso expone información que debe dar forma a la arquitectura.

Paso 1: Descripción de las definiciones

Examine lo que significa cada término definido:

- **Solicitud válida:** Mensajes que se ajustan al esquema documentado y no superan el límite de tamaño. Puede comprobar el límite en la documentación del producto. Los mensajes de tamaño excesivo que fallan no se miden en función del Acuerdo de Nivel de Servicio.
- **Total de minutos facturables:** Minutos cuando intenta enviar un mensaje. Si no envía mensajes durante parte del mes, esos minutos de inactividad no se factorizarán en el cálculo, incluso si el servicio tiene un problema durante ese tiempo.
- **Cómo se define la disponibilidad:** El Acuerdo de Nivel de Servicio cubre el tiempo para poner en cola un mensaje, no el tiempo de vuelo, la confirmación de entrega o el tránsito de retorno. Si tu paloma tarda tres días en entregar un mensaje, eso está fuera del ámbito del Acuerdo de Nivel de Servicio.

Paso 2: Comprender cómo se mide la disponibilidad

El Acuerdo de Nivel de Servicio de PPaaS revela las siguientes observaciones clave sobre la medición:

- **Requisitos de reintento:** Los errores solo cuentan si se vuelve a intentar con retardo exponencial. Sin reintentos, el Acuerdo de Nivel de Servicio no mide el error.
- **Agregación en un período de facturación:** El Acuerdo de Nivel de Servicio calcula la disponibilidad durante todo el mes. Es posible que una interrupción de 10 minutos no infrinja el Acuerdo de Nivel de Servicio si el servicio se ejecuta continuamente el resto del mes.

Paso 3: Identificación de condiciones y exclusiones

Varias condiciones limitan la aplicabilidad del Acuerdo de Nivel de Servicio:

- **Condiciones meteorológicas graves:** Compruebe la documentación del producto o el proveedor de servicios para determinar cómo definen condiciones meteorológicas graves. Esta exclusión elimina una categoría completa de errores que pueden afectar con frecuencia a un servicio basado en palomas.
- **Mantenimiento programado de la cooperativa:** El Acuerdo de Nivel de Servicio excluye hasta 15 minutos por semana de mantenimiento planeado. Esta ventana de mantenimiento no se contabiliza en el Acuerdo de Nivel de Servicio, aunque el servicio esté indisponible durante ese tiempo. Es posible que no tenga control sobre cuándo se produce ese mantenimiento. Compruebe la documentación del producto sobre cuándo se produce el mantenimiento y planee en consecuencia.
- **Compromisos específicos de nivel:** El nivel Basic Coop no tiene ningún Acuerdo de Nivel de Servicio. Si usa el nivel Básico, no tiene ningún compromiso de disponibilidad contractual.

Paso 4: Extracción de señales arquitectónicas

El Acuerdo de Nivel de Servicio de PPaaS revela varias señales importantes para la arquitectura:

- **Lo que PPaaS considera un fallo:** Solo cuentan los fallos de envío completos y sostenidos que duren al menos un minuto completo. El Acuerdo de Nivel de Servicio espera que la carga de trabajo gestione breves interrupciones, despachos lentos y fallos de solicitudes individuales.
- **Lo que no garantiza:** El Acuerdo de Nivel de Servicio solo cubre si puede enviar correctamente un mensaje a la API de PPaaS y aceptarlo para su envío. No garantiza el tiempo de entrega, el éxito de entrega o la integridad del mensaje después de liberar la paloma. La entrega confirmada se encuentra fuera del Acuerdo de Nivel de Servicio por completo.

- **Donde los arquitectos deben introducir resiliencia:** Diseñen para fallos de despacho intermitentes mediante la implementación de la lógica de reintento necesaria. Agregue un mecanismo de reserva para situaciones en las que las garantías de entrega sean importantes porque el Acuerdo de Nivel de Servicio no cubre la entrega real. El Acuerdo de Nivel de Servicio excluye las interrupciones relacionadas con el tiempo y el mantenimiento programado, por lo que la arquitectura debe controlar esos escenarios a través de rutas alternativas o tolerancia a la falta de disponibilidad planeada que se produce fuera del control.

Paso 5: Preparación para reclamar créditos de servicio

Para estar listo para justificar una reclamación si PPaaS no cumple su Acuerdo de Nivel de Servicio, prepare los siguientes requisitos:

- **Mantener registros detallados:** Realice un seguimiento del éxito o error de envío con una granularidad de un minuto. Conserve los registros para mostrar que cada intento erróneo incluía los tres reintentos necesarios con retroceso exponencial. Conserve también los datos meteorológicos independientes de la región de implementación. PPaaS podría atribuir tiempo de inactividad a condiciones meteorológicas graves en lugar de un problema de servicio.

- **Presentar dentro de la ventana de reclamación:** Los SLA verdaderos normalmente requieren reclamaciones en un plazo de 30 a 60 días después del final del mes de facturación. Agregue la fecha límite al calendario para que no pierda una reclamación válida.

Pasos siguientes

- Revise los Acuerdos de Nivel de Servicio de los servicios de los que depende. Para los servicios de Microsoft, comience con los Acuerdos de Nivel de Servicio para servicios en línea. La carga de trabajo también puede depender de los servicios de otros proveedores de servicios. Aplique el mismo enfoque de cinco pasos a esos Acuerdos de Nivel de Servicio.
- Compare las garantías del Acuerdo de Nivel de Servicio con los objetivos de confiabilidad de la carga de trabajo para identificar brechas.
- Obtenga información sobre cómo definir SLO que van más allá de los SLA del proveedor de servicios. Vea Recomendaciones para definir objetivos de fiabilidad.

Last updated on 01/04/2026

Qué hacer durante una interrupción del servicio Azure

En Microsoft, hacemos todo lo posible para garantizar que nuestros servicios estén siempre disponibles cuando los necesite. Sin embargo, se producen interrupciones del servicio no planeadas. En este artículo se explica lo que sucede cuando lo hacen.

Microsoft proporciona un Acuerdo de Nivel de Servicio (SLA) para muchos de sus servicios como compromiso para el tiempo de actividad y la conectividad. El Acuerdo de Nivel de Servicio para servicios de Azure individuales se puede encontrar en Azure Contratos de nivel de servicio.

Descripción del ámbito del incidente

Cuando hay un incidente, si comprende el ámbito de ese incidente, puede determinar el mejor curso de acción.

Para comprender el ámbito de un incidente, siga estos pasos:

1. Vaya a Azure Service Health, que proporciona:
    - Información sobre problemas o eventos que podrían afectar a los servicios.
    - Alertas automáticas de actualización de incidentes para que pueda recibir una notificación de las actualizaciones de estado de incidente automáticamente. Cuando Microsoft entiende el ámbito de un incidente, actualizamos el estado del incidente. Puede configurar estas actualizaciones de estado para ir a un grupo de acciones de Azure Monitor, que puede enviar alertas a una variedad de lugares como una dirección de correo electrónico o a su propio sistema de administración de incidentes.
2. Si tiene problemas para acceder al portal de Azure, vaya a estado de Azure.

    > **Nota:**
    >
    > Azure status solo muestra problemas generalizados que cumplen ciertos criterios. Dado que la página estado de Azure no sabe qué suscripciones e inquilinos administra, no puede proporcionar una vista precisa de los problemas más pequeños que pueden afectar a los servicios.

3. Si hay problemas con la página de estado, compruebe si hay publicaciones de @AzureSupport en la plataforma social X.

Incidentes de zona de disponibilidad y centro de datos

Muchos problemas se limitan a una sola zona de disponibilidad. Las zonas de disponibilidad representan centros de datos o grupos de centros de datos que están aislados de otras zonas de disponibilidad de la misma región. Cuando una zona de disponibilidad experimenta un problema, el impacto que ve depende de la forma en que se implementa un servicio:

- Los servicios zonales anclados a la zona de disponibilidad afectada pueden verse afectados.
- Es poco probable que los servicios con redundancia de zona se vean afectados. No debería necesitar realizar ninguna acción correctiva para los recursos con redundancia de zonas.
- Los servicios regionales (no zonales) pueden verse afectados porque pueden usar la zona de disponibilidad afectada.

Para obtener más información sobre la compatibilidad con zonas de disponibilidad en servicios de Azure y las diferencias entre los servicios zonales, con redundancia de zona y regionales (no zonales), consulte Azure servicios con compatibilidad con zonas de disponibilidad.

Si hay algún problema con los recursos zonales o regionales implementados en la zona de disponibilidad afectada, considere la posibilidad de iniciar los planes de continuidad empresarial y recuperación ante desastres (BC/DR).

Zonas de disponibilidad lógicas frente a físicas

Cada Azure suscripción ve una lista diferente de zonas de disponibilidad. Las zonas lógicas usadas por cada suscripción pueden corresponder a diferentes zonas físicas . Puede relacionar las zonas lógicas con las zonas físicas para confirmar qué recursos se ejecutan en la zona de disponibilidad física afectada. Para obtener más información, consulte Zonas de disponibilidad físicas y lógicas.

Incidentes en toda la región

En ocasiones, los problemas afectan a toda una región. Los problemas de toda la región pueden producirse cuando una región no tiene zonas de disponibilidad. Cuando se produce un incidente en toda la región, es posible que tenga que considerar la posibilidad de iniciar el plan de recuperación ante desastres. Su plan podría incluir la conmutación por error a otra región.

Priorizar la continuidad empresarial

Cuando se enfrenta a un incidente, la máxima prioridad es mantener las operaciones empresariales en funcionamiento. Demasiado enfoque en aislar o corregir la causa de un problema puede desviar la atención de restaurar las operaciones de la solución y mantener la continuidad empresarial.

Los siguientes factores presentan situaciones en las que no es necesario hacer nada necesariamente para conservar la continuidad empresarial:

- El impacto real del problema en los recursos de producción y en los usuarios o cargas de trabajo. Por ejemplo, un problema que se produce fuera del horario comercial estándar podría no requerir que haga nada inmediatamente.
- *Ámbito del incidente*. En el caso de problemas aislados en una sola zona de disponibilidad, es posible que no tenga que hacer nada si usa recursos con redundancia de zona o si los recursos que usa se encuentran en una zona de disponibilidad física no afectada.
- Tiempo de resolución estimado, si está disponible. Microsoft se esfuerza por proporcionar escalas de tiempo claras para la recuperación tan pronto como podamos. Si los procedimientos de recuperación tardan un período de tiempo significativo en funcionar, considere si se espera que el problema se resuelva en el momento en que se completen.
- Los objetivos de nivel de servicio (SLO) establecidos con los usuarios de la carga de trabajo afectada, si los objetivos existen. Los SLO están allí para guiar la toma de decisiones en este tipo de situación. Por ejemplo, en algunas situaciones, puede ser posible cambiar a modos de operación manual hasta que sus servicios estén en buen estado, y esta decisión podría reflejarse en el SLO del sistema. Para obtener más información sobre los Objetivos de Nivel de Servicio (SLO) y cómo definirlos, consulte Recomendaciones para definir objetivos de confiabilidad en el Marco de Azure Well-Architected.

Sin embargo, si la continuidad empresarial requiere alguna forma de acción y tiene un plan de recuperación ante desastres, el siguiente paso es tener en cuenta si se debe iniciar ese plan.

Considere el plan de recuperación ante desastres.

Un plan de recuperación ante desastres explica lo que espera hacer en caso de que se produzca un incidente importante. Los planes de recuperación ante desastres suelen incluir acciones como las siguientes:

- Para una interrupción aislada dentro de una zona de disponibilidad, escale horizontalmente el recurso si es posible.

- Para una interrupción de zona de disponibilidad cuando se usan servicios zonales, despliegue en otra zona de disponibilidad y conmute por error a otra zona de disponibilidad.
- En el caso de una interrupción de zona de disponibilidad cuando se usan servicios con zonas redundantes, es posible que no necesite hacer nada. Si observa problemas de rendimiento, considere la posibilidad de escalar horizontalmente el recurso para que las instancias de las zonas restantes puedan procesar el aumento de tráfico que reciben.
- Para una interrupción regional, implemente en otra región y conmute por error.

> **Importante**
>
> No realice ninguna acción sin pensar en ellas. Las decisiones apresuradas a veces pueden empeorar las cosas. Si ya ha desarrollado un plan de recuperación ante desastres que cubre el escenario, suele ser mejor usarlo en lugar de crear un plan en el momento.

La conmutación por error puede ser un proceso complejo. Solo debe desencadenar una conmutación por error cuando pueda justificar los costos y los riesgos. En algunas situaciones, podría dar lugar a una pérdida de datos, como si los cambios recientes no se replicaran entre regiones antes de que se iniciara ningún tiempo de inactividad. También podría experimentar tiempos de inactividad, especialmente si necesita redirigir el tráfico a una implementación en otra región. En función del modo en que se diseñe la solución, es posible que tenga que actualizar los registros DNS y esperar a que se propaguen.

Conmutación por error iniciada por Microsoft

Algunos servicios de Azure cambian automáticamente a regiones alternativas durante incidentes. Microsoft gestiona el proceso de conmutación por error para esos servicios. Sin embargo, la conmutación por error iniciada por Microsoft normalmente se realiza como último recurso y después de que se haya invertido un tiempo significativo en los intentos de recuperación. En general, la directiva de Microsoft consiste en minimizar la pérdida de datos incluso si esto significa un tiempo de recuperación más largo. No debe confiar exclusivamente en la conmutación por error iniciada por Microsoft para sus propias soluciones, especialmente si necesita minimizar el tiempo de recuperación. Si puede, use la conmutación por error iniciada por el cliente para servicios como Azure Storage.

Soporte de Azure

Si el incidente de servicio ya se está comunicando en Azure Service Health, toda la información más reciente se proporciona allí y no es necesario abrir una solicitud de soporte técnico.

Sin embargo, puede considerar la posibilidad de abrir un caso de soporte técnico cuando:

- Estás viendo incidencias que no se abordan en la descripción del incidente en Azure Service Health.
- Necesita ayuda de Microsoft como parte de los esfuerzos de recuperación.

> **Sugerencia**
>
> Si usted se ve afectado por una interrupción del servicio, podrá generar un ticket de soporte técnico gratuito durante un máximo de 72 horas tras la mitigación del problema para ayudar con los pasos que pueda necesitar para la recuperación.

Al abrir un caso de soporte técnico, explique claramente los recursos afectados y el impacto del problema. Especifique el identificador de suscripción Azure y la región que está experimentando un problema. Establezca la gravedad del caso de soporte técnico en función del impacto en su empresa. Tenga en cuenta que muchos clientes pueden ponerse en contacto de forma reactiva con Soporte técnico de Azure durante una interrupción del servicio Azure fuera de estas condiciones descritas. Esta carga adicional sobre los recursos de soporte de Azure podría desafortunadamente retrasar la atención a sus necesidades de soporte técnico.

Después de un incidente

1. Para comprender lo que Microsoft ha aprendido del incidente, lea la revisión posterior a incidentes (PIR) del panel Historial de estado de Azure Service Health o a través de alertas de Service Health configuradas por el cliente. Diferentes incidentes pueden recibir varios tipos de PIR. Normalmente, los PIR preliminares se publican unos días después de un incidente y los PIR más completos siguen unas semanas más tarde.
2. En el caso de los incidentes mayores que se enumeraron en nuestra página de estado público, únase a una transmisión en vivo de la Retrospectiva de Incidentes de Azure para obtener respuestas a cualquier pregunta o ver la grabación.
3. Si cree que puede ser elegible para un crédito del Acuerdo de Nivel de Servicio, cree una nueva solicitud de soporte técnico con un tipo de problema "Solicitud de reembolso" e incluya el identificador de seguimiento de incidentes.
4. Tenga en cuenta lo que puede aprender del incidente para mejorar su propia resistencia y procesos. Tenga en cuenta preguntas como:
    - ¿Cómo funciona el plan de recuperación ante desastres? ¿Hay alguna mejora que pueda realizar? Para obtener más información, consulte la guía de Azure Well-

Architected Framework sobre estrategias de recuperación ante desastres.

- ¿La respuesta al incidente es adecuada para su gravedad? Si no es así, ¿hay maneras de estar mejor informados y tomar mejores decisiones sobre qué hacer?
- ¿Existen guías de confiabilidad del servicio de Azure para los servicios que usted utiliza? Las guías de confiabilidad proporcionan información sobre cuántos servicios de Azure se pueden configurar para satisfacer los requisitos de resistencia.
- ¿Existe alguna concesión de diseño que pueda hacer para mejorar su resiliencia en el futuro para abordar este tipo de problema? Para obtener más información, consulte el pilar de relibilidad del Azure Well-Architected Framework.
- ¿El SLO o el Acuerdo de Nivel de Servicio (SLA) que se ofrecen a los usuarios siguen siendo adecuados ahora que ha experimentado esta interrupción no planificada? Ahora es un buen momento para volver a consultar los compromisos que está realizando en la base de usuarios para alinear las expectativas con lo que ha aprendido de este incidente.
- ¿Debe configurar alertas de Azure Service Health para recibir notificaciones automáticas de cualquier incidente futuro?

5. Si tiene comentarios sobre cómo podemos mejorar nuestra respuesta a incidentes, háganoslo saber a través de su representante de Microsoft asignado o a través del sitio de comentarios de Azure.

Last updated on 09/03/2026

Servicios disponibles por categorías y tipos de región

La disponibilidad de los servicios en Azure regiones depende del tipo de una región. Hay dos tipos de regiones en Azure: *recommended* y *alternate*.

- **Las regiones recomendadas** proporcionan la gama más amplia de funcionalidades de servicio y actualmente admiten zonas de disponibilidad. En el portal de Azure, las regiones recomendadas se designan como **Recommended**.
- **Alternate** las regiones amplían la superficie de Azure dentro de un límite de residencia de datos donde existe actualmente una región recomendada. Las regiones alternativas ayudan a optimizar la latencia y proporcionan una segunda región para las necesidades de recuperación ante desastres, pero no admiten zonas de disponibilidad. Azure realiza evaluaciones periódicas de regiones alternativas para determinar si deben ser regiones recomendadas. En el portal de Azure, las regiones alternativas se designan como **Other**.

Categorías de servicio entre tipos de regiones

Azure servicios se agrupan en tres categorías: *foundational*, *mainstream* y *strategic*. Azure directiva general sobre la implementación de servicios en cualquier región determinada se basa principalmente en el tipo de región, las categorías de servicio y la demanda de clientes.

- **Fundamentales**: están disponibles en todas las regiones recomendadas y alternativas cuando la región está disponible con carácter general, o bien en un plazo de 90 días después de que un nuevo servicio fundamental esté disponible con carácter general.
- **Estándar**: disponibles en todas las regiones recomendadas en un plazo de 90 días a partir de la disponibilidad general de una región. Controlados por la demanda en regiones alternativas y muchos ya están implementados en un gran subconjunto de regiones alternativas.
- **Estratégicos** (anteriormente Especializados): ofertas de servicio dirigidas, a menudo centradas en el sector o respaldadas por hardware personalizado. Disponibilidad controlada por la demanda entre regiones, y muchos ya están implementados en un gran subconjunto de regiones recomendadas.

Para ver los servicios que se implementan en una región y la futura hoja de ruta de la versión preliminar o la disponibilidad general de los servicios de una región, consulte Productos

disponibles por región.

Si una oferta de servicio no está disponible en una región, póngase en contacto con el representante de ventas de Microsoft para obtener más información y explorar las opciones.

Expandir tabla

Tipo de región	No regional	Fundamental	Estándar	Estratégico	Zonas de disponibilidad	Residencia de datos
Recommended	✓	✓	✓	Basado en la demanda	✓	✓
Alternate	✓	✓	Basado en la demanda	Basado en la demanda	N/A	✓

Servicios disponibles por categoría de región

Azure asigna categorías de servicio como fundamentales, estándar y estratégica a disponibilidad general. Normalmente, los servicios comienzan como un servicio estratégico y se actualizan a estándar y fundamental a medida que crecen la demanda y el uso.

Azure servicios se presentan en las siguientes listas por categoría. Tenga en cuenta que algunos servicios no son regionales, lo que significa que están disponibles globalmente independientemente de la región. Para obtener información y una lista completa de servicios no regionales, consulte Productos disponibles por región.

Servicios fundamentales

- Azure Application Gateway
- Azure Backup
- Azure Blob Storage (Servicio de almacenamiento de blobs de Azure)
- Azure Cosmos DB (la base de datos de Azure Cosmos)
- Azure Data Lake Storage Gen2
- Azure Disk Storage
- Azure Event Hubs
- Azure ExpressRoute
- Archivos de Azure
- Azure Key Vault
- Azure Kubernetes Service (AKS)

- Azure Load Balancer
- Puerta de enlace NAT de Azure
- Azure Service Bus
- Azure Service Fabric
- Azure Site Recovery
- Azure SQL Database
- Azure SQL Managed Instance
- Azure Stream Analytics
- Azure Virtual Machine Scale Sets (Conjuntos de Escala de Máquinas Virtuales de Azure)
- Azure Virtual Machines
  - Av2-series
  - Serie Bs
  - Series Ddv5 y Ddsv5
  - Serie Dv2 y DSv2
  - Serie Dv3 y DSv3
  - Serie Dv5 y DSv5
  - Series Edv5 y Edsv5
  - Series Ev3 y Esv3
  - Series Ev5 y Esv5
- Red virtual de Azure
- Azure VPN Gateway
- Direcciones IP públicas

Servicios estándar

- Búsqueda de Azure AI
- Azure API Management
- Azure App Configuration
- Azure App Service
- Azure Bastion
- Azure Batch
- Azure Container Instances
- Azure Container Registry
- Azure Data Explorer
- Azure Data Factory
- Azure Database for MySQL
- Azure Database for PostgreSQL

- Protección contra DDoS de Azure
- resolución privada de Azure DNS
- Azure Event Grid
- Azure Firewall
- Azure Firewall Manager
- Azure Functions
- Azure HDInsight
- Azure IoT Hub
- Azure Logic Apps
- Azure Managed Redis
- Registros de Azure Monitor
- Azure Monitor: Application Insights
- Azure Network Watcher
- Azure Private Link
- Azure Storage: Blob Storage Premium
- Azure Virtual Machines
  - Serie Ddsv4
  - Serie Ddv4
  - Serie Dsv4
  - Serie Dv4
  - Serie Edsv4
  - Serie Edv4
  - Serie Esv4
  - Serie Ev4
  - Serie Fsv2
  - Serie M
- Azure Virtual WAN
- Servicios de dominio de Microsoft Entra

☀ Servicios estratégicos

- Servicios de Azure AI
- Azure Analysis Services
- Pruebas de aplicaciones de Azure
- Azure Attestation
- Azure Automation
- Azure Chaos Studio

- Azure Container Apps
- Azure Data Share
- Azure Database Migration Service
- Azure Databricks
- Azure Dedicated HSM
- Azure Digital Twins
- SAN elástico de Azure
- Azure HPC Cache
- HSM administrado por Azure Key Vault
- Azure Administrador de flotas de Kubernetes
- Azure Machine Learning
- Azure Managed Grafana
- Azure Instancia administrada para Apache Cassandra
- Azure NetApp Files
- Red Hat OpenShift en Azure
- Azure Remote Rendering
- Azure SignalR Service
- Azure Storage: Almacenamiento de archivo
- Azure Storage: Sincronización de Archivos de Azure
- Azure Synapse Analytics
- Azure Ultra Disk Storage
- Azure Virtual Machines
    - Serie Bsv2
    - Series Dasv5 y Dadsv5
    - Series Dav4 y Dasv4
    - Serie DCsv2
    - Serie Easv5 y Eadsv5
    - Series Eav4 y Easv4
    - Serie FX
    - Serie HBv2
    - Serie HBv3
    - Serie HCv1
    - Serie Lsv2
    - Serie Lsv3
    - Serie Lsv4, Lasv4 y Laosv4
    - Serie Mv2
    - Serie NCasT4_v3

  - Serie NCv3
  - Serie NDasrA100_v4
  - La Serie NDm_A100_v4
  - Serie NDv2
  - Serie NP
  - Serie NVv3
  - Serie NVv4
  - SAP HANA en Azure (instancias grandes)
- Azure Virtual Network Manager
- Azure VMware Solution
- Azure Web PubSub
- Programador de tareas duraderas
- Microsoft Purview
- SQL Server en máquinas virtuales de Azure

Las generaciones anteriores de servicios o máquinas virtuales no aparecen en la lista. Para más información, consulte Generaciones anteriores de tamaños de máquina virtual.

Para más información sobre los servicios en versión preliminar que aún no están en disponibilidad general y para ver una lista de estos servicios, consulte Productos disponibles por región. Para obtener una lista completa de los servicios que admiten zonas de disponibilidad, consulte Azure servicios que admiten zonas de disponibilidad.

Contenido relacionado

- Azure servicios y regiones que admiten zonas de disponibilidad

ⓘ**Nota:** El autor ha creado este artículo con ayuda de inteligencia artificial. Más información

---

Last updated on 11/06/2026

Estrategias de arquitectura para pruebas de confiabilidad

Se aplica a la siguiente recomendación de la lista de comprobación de fiabilidad del marco de trabajo Azure Well-Architected:

Expandir tabla

RE:08	Pruebe los escenarios de resistencia y disponibilidad aplicando los principios de ingeniería de caos. **Use pruebas de confiabilidad para comprobar que la carga de trabajo puede soportar errores, escalar bajo demanda y recuperarse dentro de los destinos definidos.**

Las pruebas de confiabilidad detectan debilidades arquitectónicas antes de provocar interrupciones. Sin realizar pruebas deliberadas mediante escenarios de fallo, no se puede saber si los patrones de resiliencia realmente funcionan o si la carga de trabajo se recupera dentro de los objetivos definidos.

Puede reforzar la confiabilidad de la carga de trabajo al probar los riesgos de seguridad que afectan a la disponibilidad, los problemas de rendimiento que hacen que el sistema no se pueda usar y las brechas operativas que limiten la respuesta a los incidentes. Use las estrategias de este artículo para establecer una cadencia de pruebas que valide regularmente la carga de trabajo en función de sus modos de error. Evoluciona las pruebas a medida que la arquitectura cambia y los incidentes revelan nuevas debilidades. Genere confianza en que su carga de trabajo puede resistir fallos, escalar para satisfacer la demanda y recuperarse dentro de sus objetivos de RTO y RPO, mediante la creación de un bucle de retroalimentación que refuerce su postura de fiabilidad con el tiempo.

Las estrategias clave de este artículo se basan en las prácticas de prueba fundamentales descritas en Estrategias de arquitectura de OE:09 para pruebas. Revise primero ese artículo. Las recomendaciones de esta guía se limitan a la confiabilidad y se centran en lograr la confianza en la capacidad de la carga de trabajo para resistir errores y recuperarse dentro de los objetivos.

En la tabla siguiente se definen los términos de confiabilidad clave que se usan en este artículo.

Definiciones

Expandir tabla

Término	Definición
Availability	Cantidad de tiempo que se ejecuta una carga de trabajo de aplicación en un estado correcto sin tiempo de inactividad significativo.
Ingeniería de caos	La práctica que incorpora de forma segura las pruebas de caos en la cultura del equipo, las prácticas de ingeniería y el ciclo de vida de desarrollo para impulsar la mejora continua a la resistencia del sistema.
Pruebas de caos	Experimento controlado que valida una hipótesis de resistencia específica sobre cómo debe comportarse una carga de trabajo en condiciones perjudiciales.
Inyección de errores	Técnica para introducir deliberadamente errores en un componente, dependencia o ruta de acceso del sistema.
Capacidad de recuperación	La capacidad de restaurar las operaciones normales después de una interrupción dentro del tiempo de recuperación acordado (RTO) y los objetivos de punto de recuperación (RPO).
Resiliency	La capacidad de una carga de trabajo para soportar errores (como errores transitorios, interrupciones de infraestructura, picos de demanda) y seguir funcionando dentro de una experiencia de usuario aceptable.
Failover	Proceso de cambio a un componente secundario tras el error del componente principal.
Failback	Proceso mediante el cual se restauran las operaciones en la región primaria después de que esta se haya recuperado.
Presupuesto de errores	Nivel máximo aceptable de error para un sistema durante un período definido, derivado de objetivos de nivel de servicio (SLO).

Definición del ámbito de pruebas de confiabilidad en función de los tipos de servicio

Al definir el ámbito de las pruebas de confiabilidad, considere el modelo de responsabilidad compartida de los servicios que usa. Cada tipo de servicio (IaaS, PaaS, SaaS) incluye diferentes garantías de confiabilidad y diferentes niveles de control sobre el control de errores. Centre sus pruebas en los aspectos de confiabilidad que posee.

- **Adapte la profundidad de las pruebas a su nivel de responsabilidad.** En el caso de los servicios de infraestructura (IaaS), su equipo posee la mayoría de las decisiones de confiabilidad, por lo que invierte en una validación exhaustiva a través de la ingeniería de caos y la inyección de errores. Para los servicios de plataforma (PaaS) y software (SaaS), el proveedor administra gran parte de la confiabilidad subyacente. Céntrese en cómo interactúa la carga de trabajo con esos servicios, como cómo controla las conmutaciones

por error de infraestructura en PaaS, la limitación, la degradación del servicio o los cambios en los patrones de carga.

- **Tenga en cuenta las cargas de trabajo de servicio mixto.** Cuando la carga de trabajo abarca varios tipos de servicio, las responsabilidades de prueba varían entre los componentes. Es posible que pruebe la conmutación por error de los componentes de la infraestructura basada en máquinas virtuales durante una interrupción de la zona de disponibilidad, pero confíe en garantías de proveedor para una base de datos PaaS diseñada para alta disponibilidad. Identifique dónde están esos límites y asegúrese de que las pruebas cubren las brechas entre ellos.

Prueba con los objetivos de confiabilidad de un extremo a otro

Los objetivos de confiabilidad, como los SLO, los RTO y los RPO, definen cómo debe comportarse la carga de trabajo en condiciones de error. Úselos como criterios de paso y error en flujos críticos completos, no solo componentes individuales.

- **Valide la recuperación a lo largo de todo el flujo.** Un solo componente puede restaurarse dentro de su RTO, pero el tiempo total de recuperación puede superar su objetivo cuando las dependencias posteriores también deben restaurarse. Tenga en cuenta el tiempo total de recuperación en todo el flujo, incluido el tiempo para detectar el problema y responder.

- **Defina el ámbito de prueba con SLO y presupuestos de errores.** El presupuesto de errores representa la inversión que puede realizar en la inyección de errores. Limite las pruebas de caos para mantenerse dentro de sus SLO y use sus objetivos de recuperación del flujo para definir los límites de cada prueba.

Crear escenarios de confiabilidad a partir de flujos críticos y modos de error

Comience con los flujos críticos de la carga de trabajo y los modos de error que pueden afectarlos. Use el análisis del modo de error para identificar los escenarios de error más impactantes y crear pruebas que validen la resistencia y las estrategias de recuperación.

Priorice por impacto y probabilidad. No todos los modos de error garantizan la misma inversión de prueba. Céntrese primero en escenarios con el mayor impacto potencial del usuario y la mayor probabilidad de producirse. Deje que el análisis del modo de error impulse esta priorización.

Validación de los mecanismos de tolerancia a errores y recuperación

Céntrese en escenarios que probablemente causen tiempo de inactividad o degradan la experiencia del usuario. Use las estrategias de esta sección para crear pruebas que validen la capacidad de la carga de trabajo para controlar errores y recuperarse de forma eficaz.

Copias de seguridad y restauración

Las pruebas de copia de seguridad y restauración deben validar que el enfoque de protección de datos cumple los objetivos de recuperación.

- **Establecer una cadencia de pruebas.** Determine con qué frecuencia necesita probar las restauraciones en función de la frecuencia con la que cambia la configuración de copia de seguridad, el esquema de datos o la infraestructura. Los cambios más frecuentes requieren pruebas de restauración más frecuentes.
- **Establecer objetivos de recuperación.** Mida los tiempos reales de restauración en comparación con sus objetivos de RPO y RTO para confirmar que su estrategia de copia de seguridad cumple sus objetivos de recuperación.
- **No dé por sentado que la copia de seguridad esté completa.** Las copias de seguridad se pueden configurar de forma incorrecta para capturar solo un subconjunto de los datos. Valide la integridad y la exhaustividad de los datos, no solo si una operación de restauración se completa con éxito.
- **Prueba de forma aislada.** Valide las restauraciones en un entorno independiente de producción para poder ejecutar comprobaciones exhaustivas sin interrumpir las cargas de trabajo activas.

Errores transitorios

Los errores transitorios, como las interrupciones de red momentáneas, la falta de disponibilidad del servicio breve y los tiempos de espera de conexión, son riesgos comunes de confiabilidad. Compruebe que la carga de trabajo controla estos errores sin afectar a los usuarios. Para obtener más información, consulte Recomendaciones para controlar errores transitorios.

- **Céntrese en lo que posee.** Si los SDK o los servicios de plataforma controlan los reintentos y la interrupción del circuito automáticamente, pruebe lo que sucede cuando se agoten

esos mecanismos integrados, como cuando se produzcan errores en todos los reintentos o se abra un disyuntor.

- **Valide las configuraciones de control de errores.** Evalúe las configuraciones de control de errores de la carga de trabajo. Compruebe que las directivas de reintento, los umbrales del disyuntor y los valores de tiempo de espera se comportan según lo previsto en condiciones de error realistas.

- **Pruebe el límite entre errores transitorios y persistentes.** Compruebe que la carga de trabajo pasa correctamente del comportamiento de reintento a la reserva o degradación cuando un error persiste más allá de los umbrales esperados.

- **Tenga en cuenta los errores transitorios de las características de resistencia.** La redundancia de zona y diseños similares suelen producir errores transitorios durante las operaciones de conmutación por error normales. Por ejemplo, una base de datos con redundancia entre zonas podría provocar fallos transitorios de conexión mientras el tráfico se redirige a una zona en buen estado durante una interrupción. Pruebe si la carga de trabajo puede controlar estos errores transitorios esperados sin un impacto significativo en el usuario.

Respuesta ante la carga y la escalabilidad

Compruebe que la carga de trabajo mantiene la confiabilidad durante los cambios de demanda, tanto los picos repentinos como los aumentos graduales. Para obtener más información, consulte Estrategia de escalado. Para obtener instrucciones sobre pruebas de carga y esfuerzo, consulte Recomendaciones para pruebas.

- **Pruebe tanto el escalado horizontal como el escalado vertical.** Compruebe que la nueva capacidad entra en línea con la suficiente rapidez y que el escalado de reducción no pierda solicitudes ni deje recursos huérfanos.

- **Tenga en cuenta el retraso de escalado.** Siempre hay un retraso entre cumplir las condiciones para desencadenar el escalado y tener lista la nueva capacidad. Determine si la carga de trabajo puede controlar la demanda durante esa brecha o si necesita aprovisionar previamente capacidad adicional.

Fallos de dependencia

Es probable que la carga de trabajo dependa de los servicios fuera del control directo, como api de terceros, servicios de plataforma administrada o servicios internos compartidos. Compruebe

que la carga de trabajo controla los errores en esas dependencias sin interrupciones significativas para los usuarios.

- **Clasificar las dependencias por importancia.** No todas las dependencias garantizan la misma inversión de prueba. Priorice las pruebas de las dependencias que están en sus flujos críticos y que carecen de redundancia integrada o vías alternativas.
- **Pruebe el comportamiento de recuperación para cada dependencia.** Cuando una dependencia deja de estar disponible, la carga de trabajo debe revertir a una ruta de acceso o comportamiento alternativo en lugar de generar errores por completo. Compruebe que cada mecanismo de respaldo se activa correctamente y que la funcionalidad no relacionada sigue funcionando.
- **Tenga en cuenta los errores parciales y en cascada.** Las dependencias rara vez fallan de forma binaria. No solo pruebe las interrupciones completas. Cubra los aumentos de latencia, los errores intermitentes y la disponibilidad parcial de los datos.
- **Valide el aislamiento y la contención del radio de impacto.** Compruebe que un fallo de una única dependencia no se propague en cascada a funcionalidades no relacionadas.

Autoconservación y recuperación

Valide cómo responde su diseño de recuperación automática y autoconservación a errores de funcionamiento, incluido si la carga de trabajo puede seguir funcionando en una capacidad reducida cuando la recuperación completa no es inmediata.

- **Pruebe la recuperación automatizada de un extremo a otro.** Evalúe sus modelos de estado para asegurarse de que incluyen las comprobaciones adecuadas. Compruebe que esas comprobaciones detecten errores con precisión, desencadenen la corrección automatizada según lo previsto y devuelvan el sistema a un estado correcto en períodos de tiempo aceptables.
- **Validar procedimientos manuales de recuperación.** La recuperación automatizada no abarca todos los escenarios. Pruebe los procedimientos operativos manuales en condiciones realistas para asegurarse de que los operadores pueden ejecutarlos bajo presión y dentro de sus objetivos de tiempo de recuperación.
- **Valide el comportamiento de la degradación controlada.** Cuando un componente falla, la carga de trabajo debería degradarse de forma gradual en lugar de fallar por completo. Compruebe que puede funcionar en un modo reducido, como poner en cola las solicitudes

para su revisión manual, y que la experiencia degradada es aceptable para los usuarios. Confirme que el equipo sabe cómo operar la carga de trabajo en este estado y cómo restaurar la funcionalidad completa.

Incidentes y recuperación ante desastres (DR)

Cuando se produce un incidente o desastre, la capacidad de detectarlo rápidamente y responder de forma eficaz es crítico. Pruebe los planes y los procesos para asegurarse de que abordan errores catastróficos y incidentes importantes. Use un entorno dedicado para las pruebas de recuperación ante desastres para evitar afectar a las cargas de trabajo de producción. Para más información, consulte plan de recuperación ante desastres.

- **Valide los mecanismos de detección de incidentes.** Simulación de incidentes para comprobar que los registros de supervisión capturan la información necesaria y que las alertas se desencadenan correctamente. Por ejemplo, compruebe con qué rapidez se detectan los aumentos en las tasas de error de las solicitudes y con qué frecuencia se muestrean los datos de supervisión.

- **Pruebe los procesos de administración de incidentes.** Confirme que el equipo puede seguir los procedimientos de respuesta a incidentes de forma eficaz. Para obtener más información, consulte respuesta a incidentes.

- **Pruebe la conmutación por error completa y la restauración tras la conmutación por error.** Probar componentes individuales por separado puede pasar por alto fallos de coordinación que solo aparecen durante una conmutación real. Valide la secuencia de conmutación por error completa, incluida la conmutación de DNS, la restauración de copias de seguridad y la coherencia de la replicación de datos y la reconexión del cliente. Pruebe también la reversión por error al despliegue principal, que a menudo es más compleja que la conmutación inicial.

- **Valide la capacidad en el entorno de conmutación por error.** Asegúrese de que su entorno de conmutación por error tenga suficiente capacidad preaprovisionada para absorber el tráfico inmediatamente después de la conmutación por error sin colapsarse por la carga. Compruebe que el entorno puede sostener las operaciones mientras se activan los mecanismos de escalado y valide su estrategia de escalado. Para obtener más información, consulte Respuesta de carga y escalado.

- **Mida sus resultados con respecto a sus objetivos.** Si una prueba de recuperación ante desastres no cumple con el RTO o el RPO, analice las brechas y actualice el plan de recuperación ante desastres en consecuencia.

- **Validar personas y procesos.** Las pruebas de recuperación ante desastres deben verificar los canales de comunicación, confirmar que los operadores disponen del acceso y de los permisos necesarios para ejecutar los procedimientos de recuperación y garantizar que pueden encontrar rápidamente los procedimientos específicos de recuperación ante desastres en situaciones de presión.

Pon a prueba y evalúa tu plan con ejercicios de simulación

Los ejercicios de tableta le ayudan a encontrar brechas en las pruebas de confiabilidad antes de que un incidente real los exponga. Al simular escenarios de error con el equipo, puede identificar condiciones no probadas y validar que los procedimientos de respuesta funcionan según lo previsto.

- **Simular incidentes realistas.** Recorra un escenario de error, como una interrupción regional o una implementación dañada, y haga que el equipo describa los pasos que se deben seguir para detectar, responder y recuperarse de ella. Estas discusiones suelen revelar suposiciones sobre el comportamiento del sistema que no se han validado a través de las pruebas.

- **Convierta los resultados en casos de prueba.** Use los huecos y desconocidos que aparecen durante el ejercicio para crear nuevas pruebas de confiabilidad. Si el equipo detecta que nadie sabe cómo se comporta la carga de trabajo cuando se produce un error en una dependencia específica, es un escenario para agregar a la estrategia de pruebas.

Aprovechar las interrupciones planeadas y no planeadas

Cuando la carga de trabajo no está disponible debido a un mantenimiento planificado o a una interrupción imprevista, tiene una oportunidad única para realizar pruebas y entender mejor su carga de trabajo.

Mantenimiento planeado

Durante las ventanas de mantenimiento para actualizaciones o revisiones, pruebe componentes y flujos que no intervienen en el trabajo de mantenimiento. Puede ejecutar pruebas sin riesgo de degradar inesperadamente la carga de trabajo o desconectarla. Si tiene suficiente tiempo, pruebe también los componentes implicados en el mantenimiento una vez completado el trabajo.

Interrupción no planeada

Cada interrupción no planeada es una oportunidad para reforzar la estrategia de pruebas de confiabilidad. Después de restaurar el servicio y completar la revisión posterior al incidente, use los resultados para mejorar las pruebas.

- **Pruebe las medidas de mitigación.** Si ha aplicado una solución alternativa o una corrección temporal, valide que puede controlar la carga esperada antes de que se aplique la corrección permanente.
- **Busque puntos débiles similares.** Revise otros componentes para conocer los mismos patrones de configuración o diseño que han contribuido a la interrupción. Agregue pruebas para esos componentes antes de que produzcan un error de la misma manera.

Combinar diferentes tipos de pruebas

Al ejecutar diferentes tipos de pruebas juntas, puede revelar problemas de confiabilidad que no son visibles cuando cada prueba se ejecuta de forma aislada. Una carga de trabajo puede soportar una carga específica en condiciones normales, pero esa misma carga provoca fallos cuando se introduce un fallo.

- **Reutilizar la infraestructura de prueba existente.** Si ya tiene infraestructura y un arnés de pruebas para pruebas de carga, úselo para ejecutar pruebas de caos simultáneamente. Combinar pruebas de carga e inyección de fallos en una misma ejecución de prueba muestra cómo se comporta su carga de trabajo en condiciones realistas, donde la demanda y los fallos se producen al mismo tiempo.
- **Comience en entornos que no son de producción.** Comience las pruebas de confiabilidad en entornos de menor riesgo, donde puede explorar de forma segura los modos de error. Pase a realizar pruebas en producción solo después de haber aprovechado al máximo la información obtenida en entornos no productivos y de contar con mecanismos de protección para limitar el alcance del impacto y revertir los cambios rápidamente.

Use la inyección de fallos y la ingeniería del caos

La inserción de errores y la ingeniería de caos le ayudan a crear confianza en la resistencia de la carga de trabajo mediante la introducción deliberada de errores y la observación de la respuesta. Asegúrese de que tiene estrategias de mitigación en vigor antes de ejecutar experimentos.

- **Ejecute experimentos de caos con regularidad.** Evalúe el ámbito de prueba. Inserte errores en componentes y flujos que se supone que son confiables. A medida que cambia la arquitectura, los nuevos modos de error emergen y es posible que las suposiciones

anteriores ya no se mantengan. Ejecute experimentos con una cadencia regular para detectar regresiones, validar nuevas dependencias y confirmar que los cambios recientes no han introducido puntos débiles.

- **Utilice su análisis de modos de fallo para enfocar los experimentos.** Cada experimento debe tener como destino un error específico o un conjunto de errores y tener una hipótesis clara, como probar la capacidad de un flujo determinado para resistir la pérdida de un componente determinado. A medida que los experimentos revelan nuevos modos de error o dependencias no detectadas, actualice el análisis del modo de error para mantenerlo actualizado.

- **Limita el impacto durante los experimentos.** Identifique los componentes objetivo que pueda recuperar rápidamente y establezca expectativas fundamentadas sobre el efecto de cada inyección de fallos. Si un experimento va más allá del ámbito o genera resultados inesperados, deténgalo. Equilibre la recopilación de datos significativos con la minimización del efecto en los usuarios.

- **Medición respecto a líneas de base.** Establezca métricas de confiabilidad y rendimiento coherentes para los flujos y componentes de cada experimento. Compare las métricas de estado degradadas con estas líneas base para comprender el efecto completo del error y determinar si el diseño de resistencia cumple sus objetivos.

- **Vuelva a incorporar los resultados a la estrategia de pruebas.** Use los resultados del experimento para impulsar nuevas pruebas, actualizar planes de recuperación e informar a los elementos de trabajo pendiente de corrección. A medida que surgen comportamientos inesperados, cree pruebas dirigidas para esos comportamientos y estrategias de corrección de diseño.

> **Compensación**: las pruebas de inyección de errores en producción pueden ser perjudiciales y pueden provocar tiempo de inactividad. Ser transparente con las partes interesadas sobre esta posibilidad. Coloque las medidas de seguridad en su lugar para que pueda detener los experimentos y revertirlos rápidamente y mantenerse flexible sobre cuándo ejecutar pruebas o evitarlos durante períodos confidenciales. Para protegerse frente a interrupciones no deseadas, planee una redundancia suficiente y asegúrese de que las partes interesadas comprendan el equilibrio de costos.

> **Riesgo:** Las pruebas de confiabilidad pueden expandir la cobertura en muchos modos de error, pero debe detenerse una vez que ya no proporciona un valor significativo. Si ya

> tiene un trabajo pendiente de problemas de confiabilidad conocidos, priorice la corrección de esos problemas en lugar de agregar más pruebas.

Facilitación con Azure

Azure Test Plans es una solución de administración de pruebas basada en explorador que proporciona todas las funcionalidades necesarias para las pruebas manuales planeadas, las pruebas de aceptación de usuarios, las pruebas exploratorias y la recopilación de comentarios de las partes interesadas.

Azure Chaos Studio es un servicio administrado que usa pruebas de caos para ayudarle a medir, comprender y mejorar la resistencia de la aplicación y el servicio en la nube.

Pruebas de aplicaciones de Azure es un servicio que permite ejecutar pruebas funcionales con Espacios de trabajo de dramaturgo y pruebas de rendimiento mediante Azure Load Testing para identificar problemas en sus aplicaciones.

Azure Service Health tiene un panel de mantenimiento planeado que es una sección dedicada del portal de Azure que le mantiene informado sobre las próximas actividades de mantenimiento. Resalta los eventos que pueden afectar a los recursos de Azure, lo que le ayuda a prepararse de antemano.

El monitor de conexión es una característica de Azure Network Watcher que puede usar para la supervisión sintética y las pruebas en tiempo real de la conectividad de red en entornos híbridos y de Azure. En escenarios de ingeniería de caos, el monitor de conexión se puede usar para insertar errores en componentes de red de destino y medir continuamente métricas clave, como la latencia y la pérdida de paquetes. Esta funcionalidad permite a los equipos observar cómo afectan las interrupciones a la confiabilidad de la red, tanto para los componentes de flujo individuales como para las rutas de acceso de red de un extremo a otro. Para evaluar el impacto de los errores e identificar las áreas de mejora, compare la telemetría del monitor de conexión con las métricas de línea de base.

Vínculos relacionados

- Copia de seguridad y recuperación ante desastres para aplicaciones de Azure
- Lista de comprobación para pruebas de confiabilidad
- Prueba de aplicaciones para disponibilidad y resistencia

Lista de comprobación de confiabilidad

Consulte el conjunto completo de recomendaciones.

Lista de comprobación de confiabilidad

Last updated on 01/06/2026

Vínculos rápidos de confiabilidad

Aplique instrucciones de confiabilidad a la arquitectura para que la carga de trabajo sea resistente a un funcionamiento incorrecto y asegúrese de que vuelve a un estado de funcionamiento completo después de producirse un error.

Puntos clave de aprendizaje

INICIO RÁPIDO

Principios de diseño

Checklist

Tradeoffs

Modelo de madurez

Patrones de confiabilidad

Evaluación de la revisión de Azure Well-Architected

CURSOS

Reliability

VIDEO

Dentro de la arquitectura del centro de datos de Azure con Mark Russinovich

Revisar principios de diseño

CONCEPTO

Diseño para los requisitos empresariales

Diseño para resistencia

Diseño para recuperación

Diseño para operaciones

Manténgalo sencillo

Establecimiento y medición de objetivos de confiabilidad

GUÍA PASO A PASO

Identificación de flujos

Realizar análisis del modo de error

Establecimiento de destinos de confiabilidad

Diseñar una estrategia de supervisión y alertas

Lograr objetivos de confiabilidad

GUÍA PASO A PASO

Diseño para redundancia

Escalado confiable

Diseño para la recuperación automática y conservación

Diseño con simplicidad

Prueba y realización de simulacros

GUÍA PASO A PASO

Diseño de una estrategia de prueba

Diseño de una estrategia de recuperación

Explorar recursos relacionados

REFERENCIA

Azure Advisor: Recomendaciones de confiabilidad

Documentación sobre confiabilidad de Azure

Documentación acerca de la ingeniería de confiabilidad de sitios