Estado actual de las tecnologías de bodega de datos y OLAP aplicadas a bases de datos espaciales

Current data warehousing and OLAP technologies' status applied to spatial databases

Diego Orlando Abril Frade[1] y José Nelson Pérez Castillo[2]

RESUMEN

Las organizaciones requieren de una información oportuna, dinámica, amigable, centralizada y de fácil acceso para analizar y tomar decisiones acertadas y correctas en el momento preciso. La centralización se logra con la tecnología de bodega de datos. El análisis lo proporcionan los sistemas de procesamiento analítico en línea, OLAP *(On Line Analytical Processing*). Y en la presentación de los datos se pueden aprovechar tecnologías que usen gráficos y mapas para tener una visión global de la compañía y así tomar mejores decisiones. Aquí son útiles los sistemas de información geográfica, SIG, que están diseñados para ubicar espacialmente la información y representarla por medio de mapas. Las bodegas de datos generalmente se implementan con el modelo multidimensional para facilitar los análisis con OLAP. Uno de los puntos fundamentales de este modelo es la definición de medidas y de dimensiones, entre las cuales está la geografía. Diversos investigadores del tema han concluido que en los sistemas de análisis actuales la dimensión geográfica es un atributo más que describe los datos, pero sin profundizar en su parte espacial y sin ubicarlos en un mapa, como si se hace en los SIG. Visto de esa manera, es necesaria la interoperabilidad entre SIG y OLAP (que ha recibido el nombre de *Spatial OLAP* o *SOLAP*) y diversas entidades han adelantado varios trabajos de investigación para lograrla.

Palabras clave: OLAP, OLAP espacial, bodega de datos, bodega de datos espacial, SIG.

ABSTRACT

Organisations require their information on a timely, dynamic, friendly, centralised and easy-to-access basis for analysing it and taking correct decisions at the right time. Centralisation can be achieved with data warehouse technology. On-line analytical processing (OLAP) is used for analysis. Technologies using graphics and maps in data presentation can be exploited for an overall view of a company and helping to take better decisions. Geographic information systems (GIS) are useful for spatially locating information and representing it using maps. Data warehouses are generally implemented with a multidimensional data model to make OLAP analysis easier. A fundamental point in this model is the definition of measurements and dimensions; geography lies within such dimensions. Many researchers have concluded that the geographic dimension is another attribute for describing data in current analysis systems but without having an in-depth study of its spatial feature and without locating them on a map, like GIS does. Seen this way, interoperability is necessary between GIS and OLAP (called spatial OLAP or SOLAP) and several entities are currently researching this. This document summarises the current status of such research.

Keywords: OLAP, spatial OLAP, data warehouse, spatial data warehouse, GIS.

Recibido: agosto 17 de 2006
Aceptado: marzo 5 de 2007

Introducción

La forma en que se presenta la información que analizan los directivos de las empresas tiene una gran importancia en los resultados de los análisis efectuados. Para analizarla convenientemente el usuario debe ver los datos, explorarlos y entenderlos, por lo que es importante que las herramientas usadas, aparte de mostrarle los datos, le permitan hacer análisis sobre ellos para luego sí proceder a tomar decisiones.

Para mostrar la información de tal manera que el usuario pueda explorarla, reorganizarla y entenderla, es necesario usar herramientas que permitan llevar a cabo procesos

[1] Ingeniero de sistemas. Candidato a M.Sc. en Teleinformática, Universidad Distrital Francisco José de Caldas, Bogotá, correo electrónico: . doabrilf@estudiante.udistrital.edu.co

[2] Ingeniero de sistemas. M.Sc., en Teleinformática. Ph.D., en Informática, Universidad de Alcalá, España. Investigador y profesor, Universidad Distrital Francisco José de Caldas, Bogotá. nelsonp@udistrital.edu.co

rápidos de organización, presentación y análisis. Estas herramientas, conocidas como herramientas OLAP, se caracterizan por ser amigables, de tal forma que el usuario no requiere tener un conocimiento demasiado técnico de la estructura en que la información está organizada y pueda disponer así de los datos de una manera fácil de navegar, es decir, que permite ir de conceptos generales a específicos o viceversa. Para lograr esta navegación, hay que definir las características que describen y clasifican los datos y su agrupación mediante conceptos jerárquicos. Es necesario, además, que los datos y sus características se presenten no solo de modo alfanumérico, sino que se haga uso de otros medios que permitan visualizar fácilmente la información, como gráficos, tablas y mapas. Para ello es necesario el uso de elementos que los representen, tales como los rótulos en los gráficos, los encabezados en las tablas o las leyendas en los mapas. Estas representaciones hacen entendible al usuario la información clave, a través de la visualización de los datos y de las particularidades que los describen en un solo marco de referencia, en el que las características están en ejes de visualización y los datos están referenciados (y descritos) por tales ejes.

Las características que se pueden presentar con tablas y diversos tipos de gráficos (como tortas o *"pies"* y diagramas de barras) son todas aquellas que cuentan con listas de valores, ojalá cortas. Entre ellas está el tiempo, ya sea en rangos de días, en horas o incluso en periodos más cortos. La única que se puede representar en mapas es la ubicación geográfica.

Las herramientas también deben permitir que la presentación de los datos sea dinámica, es decir, que se actualicen automáticamente al momento en que el usuario haga cualquier cambio en la selección de los consultados. Para el caso particular de los mapas, la mejor forma es usar un SIG para ubicarlos espacialmente.

Las tecnologías de análisis de datos (OLAP) y de información geográfica (SIG) han sido desarrolladas para atender problemas o requerimientos específicos y su implementación se ha hecho de manera independiente, pero desde hace varios años se han adelantado diversos trabajos de investigación que buscan la integración de dichas tecnologías para aprovechar las características analíticas de OLAP y las de presentación de los SIG.

En la sección dos del presente artículo se resumen los conceptos básicos y antecedentes de las tecnologías bodega de datos, OLAP y SIG. En la sección tres se mencionan algunas de las propuestas que ya se han hecho con respecto a la integración de OLAP y datos espaciales. Por último, en la sección cuatro se concluye y propone actividades que se pueden llevar a cabo para continuar con este trabajo.

Conceptos

Bodega de datos

La tecnología de bodega de datos surgió a finales de los años ochenta como respuesta a la necesidad de facilitar la consolidación de información en los sistemas de soporte a la toma de decisiones, llamados sistemas de apoyo a la decisión o *DSS* (*Decision Support Systems*) (Daniel, 1999). Estos sistemas se crearon para cambiar la idea de que los datos y la tecnología debían ser usados exclusivamente por las personas del área técnica. A cambio debían ser más amigables al usuario final, para lo cual se simplifica el modelo de datos, pasando del relacional (Batín; Ceri, S. y Navathe, 1992) al multidimensional (Codd, Codd y Salley, 1993) y se proporcionan herramientas que permitan al usuario consultar los datos sin necesidad de conocer de manera técnica el modelo (Codd, Codd y Salley, 1993).

De acuerdo con William H. Inmon, una bodega de datos o *Data Warehouse* es *"una colección de datos, orientados a hechos relevantes del negocio, integrados, que incluyen el tiempo como característica importante de referencia y no volátiles para el proceso de toma de decisiones"* (Inmon, 1997). Según esta definición, es un sistema de información (y no solamente la base de datos) donde los datos de toda la empresa son recolectados, organizados y agrupados con respecto a los hechos o las actividades del negocio. Además, el uso del atributo tiempo permite mantener y referenciar información tanto histórica como reciente, y es no volátil, porque después de que los datos son cargados a la bodega, los cambios sobre ellos son poco frecuentes y se pueden mantener por largos períodos de tiempo.

Inmon define la arquitectura de una bodega de datos con cuatro componentes: 1) los sistemas fuente, donde se gestiona la información relevante de la operación de la organización; 2) el área intermedia (o *staging area*), en la cual se hace la integración, unificación y limpieza de los datos que vienen de los diferentes sistemas fuente; 3) el área de almacenamiento, conformada por dos elementos: el repositorio y los metadatos; y 4) el área de acceso a los datos a través de diferentes herramientas de consulta, tales como publicación en la web, generadores de reportes dinámicos y predefinidos, herramientas de minería de datos y OLAP. En la Figura 1 se presenta la arquitectura general de una bodega de datos.

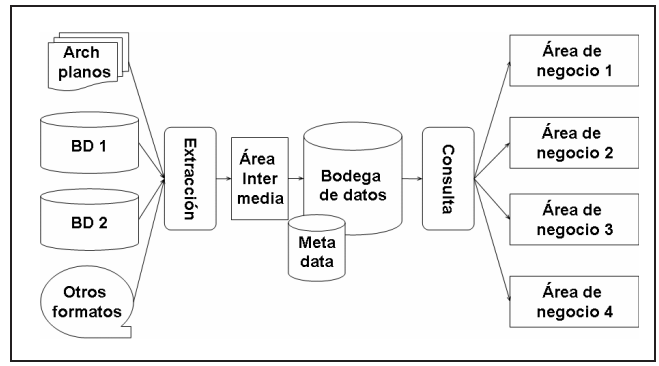
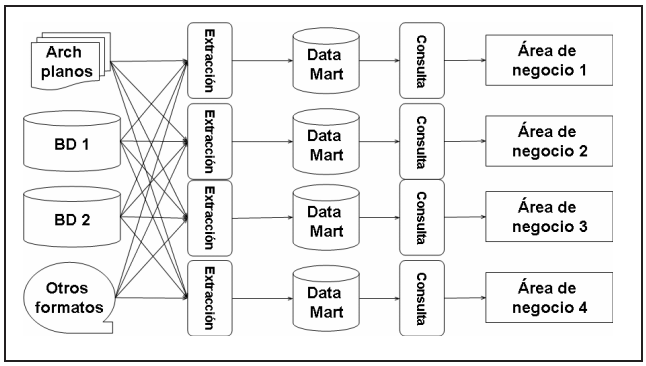


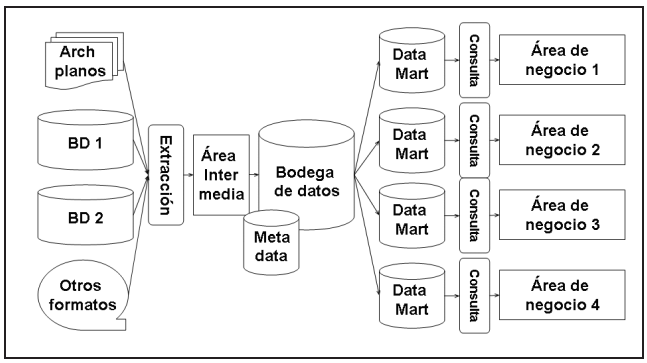
Figura 1. Arquitectura de una bodega de datos, adaptado de Inmon, 1996

Se han propuesto algunas variaciones a la arquitectura anterior (Barclay, Slutz y Gray, 2000; Bédard, Merrett y

Han, 2000; Corey y Abey, 1997; Inmon, 1996; Kimball, 1996; Kimball, *et al*., 1998), entre las que se incluye la creación de *datamarts*, los cuales se pueden considerar pequeñas bodegas porque sólo contienen información de un tema o área de negocio en particular (a diferencia de la bodega corporativa, que abarca toda la empresa). Los *datamarts* pueden ser de dos tipos: independientes, si cada uno, actuando como una bodega, obtiene su propia información, como los mostrados en la parte a) de la Figura 2, o dependientes, si obtienen su información de una bodega, como los mostrados en la parte b) de la misma figura.



a



b

Figura 2. Arquitectura de un esquema de datamarts. a. independientes y b. dependientes

En cualquier caso, una bodega de datos siempre conserva la característica de consolidar y almacenar la información en una estructura que facilite los procesos de análisis, implementada en una base de datos optimizada para responder rápidamente a las consultas. Esta base de datos es conocida como multidimensional, soportada en el modelo multidimensional de datos. La estructura de la base multidimensional se caracteriza por la presencia de una gran tabla central normalizada, llamada tabla de hechos o *fact table* y un conjunto de tablas pequeñas, generalmente desnormalizadas y llamadas comúnmente dimensiones, las cuales contienen las descripciones de las características de los datos (Agrawal, Gupta y Sarawagi, 1997; Corey y Abey, 1997; Devillers, Bédard y Jeansoulin, 2005; Dodge y Gorman, 1998; Inmon, 1996; The Data Warehousing Institute, 1999).

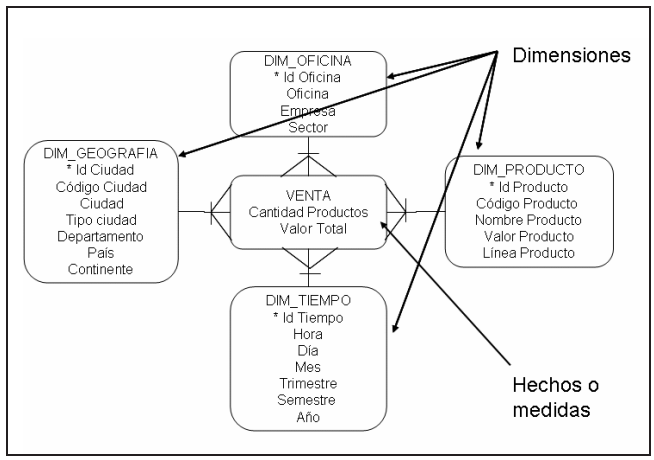
La tabla de hechos sólo contiene columnas de dos tipos: medidas y claves foráneas a las dimensiones. Las medidas son columnas que almacenan datos numéricos, por ejemplo: la cantidad de productos vendidos o el valor total de ventas. Las demás columnas actúan como claves foráneas hacia las tablas de descripciones, verbigracia: el código del almacén donde se hicieron las ventas y la fecha. El nombre de "tabla de hechos" se debe a que cada uno de los registros representa un hecho que realmente sucedió, así: el día 1 de febrero en el almacén 1 se vendieron cinco productos por un total de $10.000. Los datos PRODUCTOS=5, TOTAL=10.000, FECHA=1 de febrero o ALMACÉN=5 no significan nada por sí solos, pero su combinación sí tiene significado. La tabla de hechos está normalizada porque tales columnas no deben tener ningún tipo de dependencia entre ellas, garantizando las formas normales (Batín; Ceri, S. y Navathe, 1992), por ejemplo: la existencia de un almacén 5 no influye en la existencia del día 1 de febrero ni en que se hagan ventas por $10.000.

Las dimensiones se consideran desnormalizadas porque generalmente contienen estructuras redundantes jerárquicas (del tipo maestro-detalle) de varios niveles en el mismo registro, donde el nivel 0 es el más general y el nivel N el más detallado, esto se puede ver en la Figura 3. El elemento de una dimensión en un nivel en particular es denominado miembro y agrupa algunos de los miembros del nivel inmediatamente inferior. La redundancia se debe a la dependencia que hay entre los datos de un nivel y los niveles superiores (Batín; Ceri, S. y Navathe, 1992). Esta desnormalización se implementa con el fin de reducir el número de asociaciones entre tablas (*joins*) y, de esta manera, agilizar las consultas. También permite la construcción de consultas *ad-hoc* por parte de los usuarios finales sin necesidad de conocer la complejidad de los lenguajes de consulta de datos.

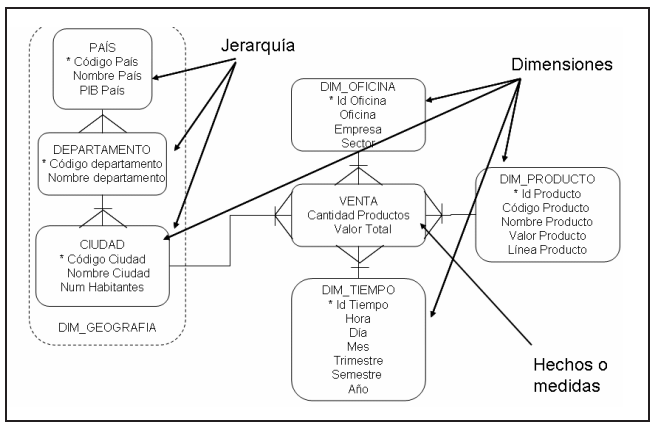
Los elementos descritos (tablas, columnas, claves foráneas) se definen para el modelo multidimensional tal como se definen en una base de datos relacional (Eisenberg, *et al*., 2004); (Eisenberg y Melton, 2000; Eisenberg y Melton, 1999; ISO, 1992; ISO, 2003).

El modelo conceptual que representa estas características multidimensionales se denomina "modelo en estrella" (Bédard y Paquette, 2005; Corey y Abey, 1997; Dodge y Gorman, 1998; Harjinder y Rao, 1996; Inmon, 1997; Kimball, 1996; Rao, *et al*., 2003; Tory y Moller, 1998). Una modificación a ese modelo se obtiene cuando una o más dimensiones se normalizan para separar la jerarquía de niveles en tablas maestro-detalle, y es conocida como modelo en "copo de nieve" o "*snowflake*". Los dos modelos se muestran en la Figura 3, donde las estructuras son equivalentes, pero el modelo en copo de nieve presenta normalizada la dimensión GEOGRAFÍA. Para consultar la descripción del nivel PAÍS de esa dimensión en el modelo en estrella, sólo requiere de una asociación entre la tabla de hechos y la dimensión, mientras que en el caso del copo de nieve, la misma consulta requiere de tres asociaciones: una entre la

tabla de hechos y la dimensión y dos más para la asociación entre las tres tablas de la dimensión (Batini, Ceri y Navathe, 1992, Chelghoum y Zeitouni, 2004; Kimball, 1996).



a



b

Figura 3. a. Modelo en estrella y b. Modelo copo de nieve

OLAP y análisis multidimensional

El término OLAP fue presentado en 1993 (Codd, Codd y Salley, 1993), publicado por Codd y asociados y apoyado por *Arbor Software Corporation,* compañía que creó ESSBASE una de las primeras herramientas OLAP que aparecen en el mercado, adquirida luego por *Hyperion* Software. Según la definición que le dio Codd, OLAP es un tipo de procesamiento de datos que se caracteriza, entre otras cosas, por permitir el análisis multidimensional (Kimball, *et al*., 1998; Koperski, Han y Stefanovic, 2001). Dicho análisis consiste en modelar la información en medidas, dimensiones y hechos. Las medidas son los valores de un dato, en particular, las dimensiones son las descripciones de las características que definen dicho dato y los hechos corresponden a la existencia de valores específicos de una o más medidas para una combinación particular de dimensiones. Este modelo se representa vectorialmente: los hechos se ubican lógicamente en una celda que queda en la intersección de ciertas coordenadas según el modelo (x, y, z, ...), donde cada una de las coordenadas de la celda representa una dimensión (Agrawal, Gupta y Sarawagi, 1997; Tory y Moller, 1998). Para materializar el análisis multidimensional en una base de datos se usa la correspondencia entre los elementos del modelo (hechos y coordenadas) y los de la base (tabla de hechos y dimensiones). Como la tabla de hechos y las dimensiones se implementan en tablas en una base de datos, se puede usar el lenguaje SQL (ISO, 1992) para la definición de las tablas de un modelo multidimensional en una base de datos relacional, pero se hizo necesario extender el modelo relacional con el fin de dar soporte a las funcionalidades propias de análisis multidimensional (Agrawal, Gupta y Sarawagi, 1997; Bédard y Paquette, 2005; Codd, Codd y Salley, 1993; Harjinder y Rao, 1996; Roddick, *et al*., 2004). Tales funcionalidades incluyen: 1. la declaración de dimensiones y jerarquías, ya que en el modelo relacional no se manejaban dichos conceptos; 2. un acceso más rápido a los datos, a través de métodos de generación de índices para datos espaciales desde el punto de vista multidimensional (Papadias, *et al*., 2002); 3. el cálculo de valores preagregados que optimicen las consultas, como lo propusieron Pedersen y Tryfona, 2001; Rao, *et al*., 2003; y 3. la definición de operaciones de navegación en las dimensiones y de agrupación de medidas (Bédard, Merrett y Han, 2000), tales como:

Slice-and-dice, corresponde a seleccionar sólo la información de un miembro en particular de una dimensión, es decir, se trabaja con un subconjunto del total de los datos para un valor determinado de un nivel en una dimensión.

Drill-down, permite ver la información del nivel inferior de la dimensión actual en una jerarquía definida, o sea, se muestran los datos detallados que en conjunto conforman el valor actual.

Roll-up, se encarga de pasar al nivel superior de la dimensión actual en una jerarquía definida, es decir, se consolidan los datos del nivel actual y se muestra el valor consolidado, correspondiente al nivel inmediatamente superior de la dimensión. También se conoce como *drill-up*.

Pivot, permite cambiar la posición de las dimensiones que caracterizan los datos presentados, esto es, se cambia el punto de vista con el que se presentan los datos, pero sin cambiar los datos o las dimensiones. También se conoce como *swap*.

Drill-across, habilita la visualización de la información de otro modelo multidimensional, o dicho de otro modo, no detalla ni consolida la información, sino que cambia el modelo multidimensional que se esta consultando. La condición para efectuar está operación es que los dos modelos compartan al menos una de las dimensiones.

Drill-through, de manera similar a la operación de *drill-down*, permite consultar la información detallada del nivel inferior en la dimensión actual. La diferencia radica en

que *drill-down* navega solamente dentro del modelo multidimensional y *drill-through* permite navegar por fuera del modelo multidimensional, es decir, establece un enlace entre el modelo multidimensional y el sistema fuente de datos (a partir del cual se construyó el modelo multidimensional) para consultar los datos del nivel detallado directamente sobre el sistema fuente. Esta operación depende de que se establezcan el acceso al sistema fuente desde el sistema OLAP.

Con respecto a las operaciones de agrupación que consolidan las medidas de un nivel, estas generan un nuevo valor que contiene un resumen de los elementos del nivel inferior. El nuevo valor consolidado se obtiene a través de la aplicación de operaciones básicas aritméticas o estadísticas de agregación sobre los datos de detalle. Entre tales operaciones están suma, cuenta (o conteo), promedio, máximo, mínimo o desviación estándar.

Según la propuesta original de Codd, el modelo multidimensional no necesariamente se tiene que almacenar previamente en una base de datos multidimensional, sino que plantea que puede acceder directamente a múltiples fuentes de información, como bases de datos (relacionales o multidimensionales), archivos planos, hojas de cálculo, e incluso algunos datos pueden ser introducidos por usuarios finales. Una vez adquiridos los datos, se consolida y organiza la información en el modelo lógico multidimensional, para luego presentarla al usuario. La arquitectura así definida (Codd, Codd y Salley, 1993) es muy similar a la de una bodega de datos, pero con la diferencia de que el modelo de Codd se concentra en el procesamiento de los datos en memoria, para el cual propone el uso de matrices multidimensionales y, aunque menciona el modelo físico, no profundiza en el tema de almacenamiento de los datos en tablas. Para el almacenamiento se puede usar tanto una base de datos relacional (*Relational* OLAP o ROLAP) como una multidimensional (*Multidimensional* OLAP o MOLAP).

Cambiando la idea de Codd donde propuso que se pueden obtener datos de múltiples fuentes, la experiencia ha permitido concluir que el análisis OLAP tiene un mejor desempeño si la información reposa consolidada en una sola fuente, y más aún si esa fuente es una base de datos multidimensional como la de una bodega de datos. Esta característica ha llevado a pensar que OLAP depende de las bodegas de datos o que OLAP y bodega de datos son la misma tecnología, pero aunque ambas usan el concepto de modelo multidimensional, su origen fue independiente y se han integrado para facilidad de los análisis (*The OLAP Council*, 1997).

Una de las propuestas para implementar el procesamiento OLAP fue hecha por Microsoft, cuando propuso el lenguaje MDX (*MultiDimensional eXpressions*). Con este lenguaje se define la estructura de datos multidimensionales y se lleva a cabo el intercambio de datos entre aplicaciones (Microsoft, 1997).

Con base en esta especificación, se han implementado diversas aplicaciones OLAP, las cuales coinciden en incluir, además de los datos multidimensionales, un esquema de metadatos en el que se define la estructura multidimensional, o sea, especifica cómo están puntualizadas las medidas y las dimensiones, describe sus características y las operaciones que se pueden llevar a cabo sobre ellas. Además, precisa los niveles, miembros y jerarquías para cada una de las dimensiones y especifica como se va a navegar a través de la información. También enuncia cómo se consolidan y organizan los datos de tal manera que la aplicación pueda procesarlos rápidamente y presentarlos al usuario. Los esquemas que definen la estructura multidimensional de las diferentes herramientas OLAP, implementan la especificación MDX, lo cual permite el análisis y exploración de los datos. Tales esquemas se caracterizan porque:

En algunos casos, el manejo interno de los metadatos ha sido personalizado para optimizar el desempeño, tal el caso de las compañías de *software* OLAP, como Hyperion, Cognos o Business Objects, y de bases de datos como DB2, Oracle o Microsoft SQLServer.

En otros casos son modelos demasiado genéricos que reducen las capacidades de análisis.

En otros más, son propuestas de investigación que se implementan internamente en las entidades que llevan a cabo la investigación, como Kheops (KHEOPS 2006) y *Environmental Systems Research Institute* (ESRI, 1998).

En otros se divulgan los resultados ante la comunidad académica, como GeoMiner (Han, Koperski y Stefanovic, 1997) y el Centro para la investigación en Geomática (Rivest, *et al*., 2005), pero no se profundiza en cómo se llevó a cabo la implementación.

Sistemas de Información Geográfica, SIG

Se define un SIG como un sistema de información computarizado que cuenta con herramientas para modelar, capturar, almacenar, recuperar, analizar y desplegar información referenciada geográficamente (Bosque Sendra, 1992); (Chang, 2002; Mitasova y Neteler, 2002). La información referenciada geográficamente o información geográfica, es aquella que se gestiona en un SIG acerca de sitios o elementos de la superficie de la Tierra. Dicha información está compuesta por la información espacial y los atributos. La información espacial hace referencia a la localización de los elementos con respecto a un sistema de coordenadas y los atributos son las características que describen tales elementos. Los atributos también se conocen como datos alfanuméricos.

La característica más destacable de los SIG es el despliegue de la información gestionada a través del uso de mapas, por esto se asocian inmediatamente los SIG a los sistemas para imprimir mapas, pero son mucho más que eso. Son sistemas que estructuran, organizan y dejan disponible la información de los elementos de la superficie terrestre para que sea consultada y utilizada en aquellas actividades

que requieren de la ubicación geográfica como parte de la información (Bosque Sendra, 1992; Chang, 2002).

Los SIG surgieron en 1964 en el Departamento de Agricultura de Canadá, con el *Canadian Geographic Information System,* y a finales de los 60 como parte de los sistemas de gestión de suelos en los Estados Unidos: *Land Use and Natural Resources Information System* de Nueva York, *Minnesota Land management Information System* y otros (Bosque Sendra, 1992). En todos los casos se inició como un proceso para levantar y gestionar información del territorio, y luego surgieron desarrollos orientados a: 1) optimizar los modelos de datos que soportaban dichos sistemas; 2) gestionar mejor la información descriptiva, conectando el SIG a bases de datos relacionales en las cuales se almacenan los atributos de manera normalizada; y 3) crear o mejorar la interfaz gráfica de los programas de visualización y mejorar los mapas generados, mejorando los dispositivos de salida. Recientemente se han venido estudiando nuevas aplicaciones para los SIG y la información espacial (Egenhofer, 1999) y se ha visto la necesidad de extender el modelo relacional para soportar las funcionalidades de la información espacial (Egenhofer, 1994).

Evolución de las bodegas de datos espaciales y de OLAP espacial

La tecnología de bodegas de datos ha evolucionado constantemente para mejorar las labores de recolección e integración de datos, optimizar el almacenamiento masivo de la información, agilizar su recuperación e incluir soporte a nuevos tipos de datos, tales como información no estructurada y de grandes dimensiones (básicamente archivos en formato binario, imágenes y formatos propietarios). Estos nuevos tipos incluyen imágenes satelitales, fotografías, información geográfica y documentos digitalizados (Barclay, Slutz y Gray, 2000; Bédard, *et al*., 2004; Bédard, Merrett y Han, 2000; Camara, Vieira Monteiro *et al*., 2003; Malinowski y Zimanyi, 2005; Papadias, *et al*., 2001), todo esto debido a que se ha incrementado la cantidad de información en los sistemas de información y ha aumentado la complejidad de la misma.

También se han hecho propuestas para optimizar las interfaces gráficas de los sistemas de análisis (Camara, Modesto de Souza *et al*., 2003) y han sido desarrolladas y mejoradas diversas herramientas que, al integrarse con las bodegas de datos, facilitan las condiciones en que los analistas toman decisiones (Egenhofer, 1999 ; ESRI, 1998; KHEOPS, 2006; Matias y Moura-Pires, 2001; The Data Warehousing Institute; Tsois, Karayannidis y Sellis, 1999). Pero para esto es necesario lograr que toda esa nueva información también pueda ser modelada mediante dimensiones (The Metadata Coalition, 2001), y que se defina toda la arquitectura necesaria para poder incluirla dentro de los análisis. Dentro de las características necesarias para dicha arquitectura hay que considerar cómo optimizar el acceso a los datos y su correspondiente manipulación al momento de hacer los análisis. Para esto se han propuesto diversos algoritmos que buscan formas de acceder más rápido a datos en particular o métodos para precalcular las operaciones de consolidación y así reducir los tiempos de respuesta (Bédard, Merrett y Han, 2000; Rivest *et al*., 2005; Chelghoum y Zeitouni, 2004; Han y Fu, 1997; Han, Stefanovic y Koperski, 1998; Koperski, Han y Stefanovic, 2001; López, Snodgrass y Moon, 2005; Papadias, *et al*., 2002; Papadias *et al*., 2001; Pedersen y Tryfona, 2001; Rao *et al*., 2003; Yin, Hui y Chee, 1998; Yu *et al*., 2005). El acceso se logra a través de índices y el precálculo de los valores consolidados genera un nuevo valor, el cual puede guardarse en la base de datos o dejarse para calcular cada vez que se consolide la información de un nivel a otro. Como los índices y los valores precalculados de todas las posibles consolidaciones no existían en la base de datos, hay que tener en cuenta que para guardarlos se necesita de un mayor espacio de almacenamiento. En este punto es necesario definir si se va a llevar a cabo la materialización de todos los datos calculados, de solo algunos o de ninguno, para lo cual se han estudiado varios algoritmos y procedimientos (Bédard, Merrett y Han, 2000; Rivest *et al*., 2005; Chelghoum y Zeitouni, 2004; Han y Fu, 1997; Han, Stefanovic y Koperski, 1998; Koperski, Han y Stefanovic, 2001; López, Snodgrass y Moon, 2005; Papadias *et al*., 2001; Pedersen y Tryfona, 2001; Yu *et al*., 2005).

En cuanto a los SIG, se mencionó que son muy fuertes en la gestión de las características geográficas de la información y en su presentación gráfica, pero para el análisis de la información no espacial requieren de desarrollos adicionales, ya que dichos sistemas no están diseñados para efectuar análisis especializados. Inclusive el almacenamiento de los datos alfanuméricos en una base de datos relacional fue una de las mejoras implementadas en la evolución de este tipo de sistemas de información (Bosque Sendra, 1992).

Integración de bodega de datos y OLAP con SIG

Aunque en las bodegas de datos y los sistemas OLAP se definen las maneras para gestionar dimensiones, varios autores (Agrawal, Gupta y Sarawagi, 1997; Bédard, Merrett y Han, 2000; Rivest *et al*., 2005; Camacho, 2001; Chelghoum y Zeitouni, 2004; ESRI, 1998; Fidalgo, Barros *et al*., 2003; Fidalgo, Salgado *et al*., 2003; KHEOPS, 2006 Malinowski y Zimanyi, 2005; Pedrosa *et al*., 2003) concluyen que la dimensión geográfica en estos sistemas es sólo un atributo que describe los datos evaluados, pero sin profundizar en su concepto espacial y sin llegar a su representación en mapas, lo cual sí se hace en los SIG. Debido a esto, se han propuesto diversos modelos de bodegas de datos espaciales (Camara, Vieira Monteiro *et al*., 2003; ESRI, 1998; Jensen *et al*., 2004; Malinowski y Zimanyi, 2005; Pedrosa *et al*., 2003) que agregan el componente espacial para referenciar los datos geográficamente. También se han desarrollado herramientas OLAP con soporte para el componente espacial, agregando funcionalidades de SIG a las bodegas de datos o a OLAP (González, 1999; KHEOPS, 2006).

Tales funcionalidades incluyen la gestión de los tipos de datos espaciales, la inclusión de nuevas operaciones de consolidación, que son usadas específicamente en los datos espaciales, y la definición de las formas de navegación entre los diferentes elementos espaciales, características que complementan el modelo multidimensional no espacial para que las funcionalidades SIG sean aprovechadas en el mundo multidimensional. Entre esas funcionalidades se mencionan:

Merge: la consolidación de dos o más elementos geográficos da como resultado un nuevo elemento geográfico conformado por la combinación de los elementos iniciales.

Unión: la consolidación de dos o más elementos geográficos da como resultado un elemento geográfico conformado por la unión de los elementos iniciales (los elementos siguen teniendo sus propias características, pero se comportan como un conjunto).

La consolidación de los atributos alfanuméricos puede calcularse a partir de una operación aritmética de consolidación de los valores alfanuméricos de los elementos originales, por ejemplo: suma o promedio.

Trabajos relacionados

La necesidad de integración de bodegas de datos y OLAP con información espacial ya ha sido estudiada en diferentes instituciones, donde se han tratado los diversos aspectos que conforman el amplio tema del OLAP en bases y bodegas de datos espaciales (Rengifo, 2004); entre ellos se ha tocado el modelo multidimensional, las bodegas de datos espaciales, los análisis de datos sobre información espacial, el modelo de metadatos, la definición de jerarquías en dimensiones espaciales, etc. Se hace referencia a los siguientes:

El centro de investigación de Microsoft, en San Francisco, propuso una bodega de datos espacial, proyecto al que se le llamó "TerraServer" (Barclay, Slutz y Gray, 2000). Este proyecto definió la arquitectura de almacenamiento, indexación y acceso a un gran volumen de imágenes de satélite y datos topográficos de los Estados Unidos, para dejarlos disponibles a consulta a través de la web y hacer *zoom* sobre dichas imágenes, identificar y marcar sitios de interés, pero no hace análisis OLAP sobre la información espacial almacenada.

El Centro de Informática de la Universidad Federal de Pernambuco, en Brasil, ha trabajado en la especificación del lenguaje de definición de datos espaciales y de intercambio de información espacial entre sistemas llamado *Geography Markup Language for Analysis* (GMLA). Este lenguaje ha sido desarrollado dentro del proyecto *Geographic On-Line Analytical Processing Architecture* (Golapa), que propone una arquitectura para hacer OLAP sobre datos geográficos (Fidalgo, Barros *et al*., 2003; Fidalgo, Salgado *et al*., 2003; Fidalgo, Souza *et al*., 2003; Fidalgo, Times y Souza, 2001).

En la Universidad Simon Fraser de Burnaby, Canadá, el proyecto GeoMiner (Han, Koperski y Stefanovic, 1997), que es un prototipo para hacer *Data Mining* espacial, contó con diferentes frentes, entre los cuales está la óptima selección y clasificación de la información espacial (Han y Fu, 1997; Han, Stefanovic y Koperski, 1998; Koperski, Han y Stefanovic, 2001), y se involucró el trabajo de Stefanovic (Stefanovic, 1997), quien propuso el esquema para implementar OLAP sobre una base de datos espacial y diferentes maneras para optimizar el almacenamiento de medidas espaciales para luego proporcionarlas a los módulos de *data mining* de GeoMiner.

El Centro de Investigación en Geomática de la Universidad Laval de Québec, Canadá, tiene varios proyectos que tratan sobre las bodegas de datos espaciales (Bédard *et al*., 2004; Rivest, Bédard y Marchand, 2001; Marchand, *et al*., 2003), el análisis multidimensional de datos espaciales (Bédard y Paquette, 2005; Bédard, Merrett y Han, 2000; Devillers, Bédard y Jeansoulin, 2005), y tecnología SOLAP (Spatial OLAP), que propone la integración de bases de datos espaciales y SIG con OLAP (Rivest, Bédard y Marchand, 2001; Rivest *et al*., 2005). En estos trabajos se menciona que para la integración SIG y OLAP se requiere de la redefinición de las dimensiones y las medidas usadas normalmente en OLAP para incluir dimensiones y medidas espaciales, y tipifica los sistemas en que hacen la integración. Además, se presentan propuestas de posibles arquitecturas para implementar dicha integración.

En Colombia se encuentran varios proyectos de grado que proponen la implementación de sistemas de información geográfica para efectuar análisis de datos, pero que no hacen referencia a OLAP, sino a la gestión de información espacial para propósitos específicos. Entre esos proyectos de grado se mencionan "El rol de la información geográfica en las bases de datos como un nuevo horizonte en el proceso de toma de decisiones" (Camacho, 2001), y el "Estudio y aplicación de la minería de datos en bases de datos espaciales" (Rengifo, 2004). En ambos casos se introduce la necesidad de realizar análisis sobre la información geográfica, pero no se propone explícitamente OLAP como solución al problema planteado.

Resumiendo las propuestas anteriores, para la integración SIG y OLAP se precisa de la redefinición de las dimensiones y las medidas usadas normalmente en OLAP para incluir dimensiones y medidas espaciales.

Según los autores anteriores (Bédard, Merrett y Han, 1998; Han, Stefanovic y Koperski, 1998), las dimensiones que incluyen el componente espacial son de tres tipos:

Dimensiones no geográficas: son las que no tienen representación en los mapas y sus diferentes jerarquías tienen miembros no geográficos, es decir, son las dimensiones en las que ninguno de sus niveles se puede representar en un mapa. Son las dimensiones usadas en las bodegas de datos

y herramientas OLAP tradicionales, por ejemplo: la dimensión tiempo o una de productos, las cuales no tienen una representación espacial.

Dimensiones geográficas: son las que tienen representación en los mapas y para las que se definen jerarquías de consolidación geográficas, esto es, son las dimensiones en las que en todos sus niveles tienen representación en un mapa, el ejemplo principal es la dimensión geográfica.

Dimensiones mixtas: son aquellas que tienen representación en un mapa para unos niveles, pero para otros no. Por ejemplo: se puede definir una dimensión almacén, que tiene los almacenes en el nivel más detallado (y cada almacén se ubica en un mapa de la ciudad) y que en el nivel inmediatamente superior se agrupan en zonas de mercadeo (las cuales también tienen representación en un mapa), pero en un nivel superior se agrupan dentro de la vicepresidencia comercial y las vicepresidencias no tienen representación en una mapa.

Según los mismos autores, las medidas usadas en análisis espacial son de dos tipos:

Medidas numéricas: aquellas que almacenan sólo valores numéricos, estas son las usadas en las bodegas de datos y herramientas OLAP tradicionales, verbigracia: el número de artículos vendidos o el valor de dichos artículos.

Medidas geográficas, conformadas por elementos geográficos resultantes de las operaciones de análisis, sea el caso: una medida aritmética puede ser el número de artículos vendidos por almacén y, al consolidarse por Zona de Mercadeo, se puede definir la medida zona de mayor número de ventas (que se va a marcar en un mapa como un elemento de tipo geográfico).

Adicionalmente, los proyectos mencionados definen tres tipos de sistemas OLAP espaciales:

Los que inicialmente son SIG y se les agregan algunas características de OLAP, llamados SIG-Dominantes.

Los que inicialmente son OLAP y que se les implementa la representación de la dimensión geografía en SIG, llamados OLAP-Dominantes, y

Los que integran SIG y OLAP y que se caracterizan por tener tanto representación SIG como funcionalidades OLAP, llamados Sistemas Integrados.

Conclusiones y trabajos futuros

Las diversas aplicaciones usadas en la actualidad se implementan para atender requerimientos específicos de las organizaciones y se especializan en resolver ciertos problemas, lo que hace que tengan fortalezas en determinados campos, pero es necesario integrarlas con otras tecnologías, con el fin de aprovechar las fortalezas de cada una y llegar a una solución integrada.

La interoperabilidad entre SIG y OLAP está en proceso de desarrollo y no existe una estructura definitiva de cómo aplicar OLAP sobre la información de un SIG. Se han llevado a cabo avances en el tema que definen las características de cómo se deberían hacer e implementan ciertas arquitecturas del modelo.

La integración de SIG y OLAP depende de que el sistema resultante tenga las características analíticas y de presentación gráfica con el dinamismo de OLAP y que además agregue a ese análisis la presentación gráfica de mapas de los SIG y que esa presentación también tenga el dinamismo de OLAP. Para contar con tales características es necesario que OLAP pueda manejar los datos espaciales y consolidarlos de manera acorde a este tipo de datos, y también que los datos con componente geográfico tengan su correspondencia en un SIG para poderse representar en mapas.

Este trabajo se puede continuar y complementar en el futuro a través del desarrollo de aplicaciones en las que se implementen las características de la tecnología de OLAP espacial. Las aplicaciones deben definir una arquitectura que soporte las características presentadas en el artículo, tales como:

La arquitectura ha de incluir un esquema que permita la definición de los diferentes elementos de las bases de datos multidimensionales y OLAP (tales como hechos, dimensiones, miembros, niveles, jerarquías, operaciones de navegación, operaciones de consolidación) para los tipos de datos espaciales. Así mismo, permitir la definición del componente espacial que se usa en la representación SIG sobre los datos referenciados por la dimensión geográfica.

Con base en la definición anterior, la arquitectura debe permitir llevar a cabo operaciones de navegación y consolidación de los datos para hacer análisis OLAP e integrarse con un SIG para aprovechar las funcionalidades de representación en mapas de dicha tecnología.

Bibliografía

Agrawal, R., Gupta, A. and Sarawagi, S., Modeling Multidimensional Databases., In: ICDE '97: Proceedings of the 13th International Conference on Data Engineering, IEEE Computer Society, 1997.

Barclay, T., Gray, J. and Slutz, D., TerraServer: A Spatial Data Warehouse., Proceedings of the 2000 ACM-SIGMOD Conference, 2000.

Batini, C., Ceri, S. and Navathe, S. Conceptual Database Design., Benjamin Cummings, 1992.

Bédard, Y., Devillers, R., Gervais, M. and Jeansoulin, R., Towards multidimensional user manuals for geospatial datasets: legal issues and their considerations., 3rd symposium on spatial data quality, Austria, 2004.

Bédard, Y., Merrett, T. and Han, J., Fundamentals of spatial data warehousing for geographic knowledge discovery., In: H. Miller and J. Han (Editors), Geographic data mining and knowledge discovery, 2000.

Bédard, Y. and Paquette, F., Extending Entity/Relationship formalism for spatial information systems., Laboratory for spatial information systems, Laval University, 2005.

Bosque-Sendra, J. Sistemas de Información Geográfica., Ediciones Rialp, Madrid, España, 1992

Camacho, J. E., El rol de la información geográfica en las bases de datos como un nuevo horizonte en el proceso de toma de decisiones., Tesis presentada a la Universidad de los Andes, para optar al título de Ingeniero de Sistemas y Computación, Bogotá, 2001.

Camara, G., Modesto de Souza, R. C, Vierira-Monteiro, A.M., Paiva, J. A. and Pinto de Garrido, J. C., Handling Complexity in GIS interface Design., National Institute for Space Research, Sao Jose dos Campos, Brasil, 2003.

Camara, G., Vieira-Monteiro, A. M., Paiva, J. A., Gomes, J. and Velho, L., Towards a unified framework for geographical data models., Nacional Institute for Space Research, Sao Jose dos Campos, Brasil, 2003.

Chang, K. T., Introduction to Geographic Information Systems., Mc Graw Hill, New York, 2002.

Chelghoum, N. and Zeitouni, K., Spatial Data Mining Implementation., Prism Laboratory, Université of Versailles, Versailles, France, 2004.

Codd, E. F., Codd, S. B. and Salley, C. T., Providing OLAP to user-analysts: An IT mandate., E. F. Codd and Associates, 1993.

Corey, M. J. and Abey, M., Oracle Data Warehousing., McGraw Hill, 1997.

Daniel, J., What is a decision support system?., In: At http://dssresources.com/papers/whatisadss/index.html.

Devillers, R., Bédard, Y. and Jeansoulin, R., Multidimensional management of geospatial data quality information for its dynamic use within GIS., American Society for Photogrammetry and R S., 2005.

Dodge, G. and Gorman, T., Oracle 8 Data Warehousing. John Wiley & Sons inc. USA., 1998.

Egenhofer, M. J., Spatial Information Appliances: A Next Generation of Geographic Information Systems., National Center for Geographic Information and Analysis, University of Maine, In: I Brazilian Symposium of Geoinformatics GEOINFO 99, Brasil, 1999.

Egenhofer, M. J., Spatial SQL: a query and presentation language., IEEE transactions on knowledge and data engineering, Vol. 6, No. 1, 1994.

Eisenberg, A., Melton, J., Kulkarni, K., Michels, J-E. and Zemke, F., SQL:2003 Has Been Published., IBM y Oracle Corp, En: ACM SIGMOD Record, Vol. 33, No. 1, 2004, pp. 119-126.

Eisenberg, A. y Melton, J., SQL Standardization: The Next Steps., Progress Inc. y Oracle Corp, En: ACM SIGMOD Record, Vol. 29, No. 1, 2000, pp. 63-67.

Eisenberg, A. y Melton, J., SQL:1999, formerly known as SQL3., Sybase y Oracle Corp., En: ACM SIGMOD Record, Vol. 28, No. 1, 1999, pp. 131-138.

ESRI., Environmental Systems Research Institute Inc., Spatial data warehousing white paper, 1998.

Fidalgo, R. N., Barros, R., da Silva, J. and Times, V., Towards a Web service for geographic and multidimensional processing., GEOINFO, 2003.

Fidalgo, R. N., Salgado, A, da Silva, J. and Times, V., Propondo uma linguagem de consulta geografica multidimensional., Universidade federale de Pernambuco, GEOINFO, 2003.

Fidalgo, R. N., da Silva, J., Times, V., Souza, F. F. and Barros, R., GMLA: A XML Schema for Integration and Exchange of Multidimensional-Geographical Data., GEOINFO, 2003.

Fidalgo, R. N., Times, V. C. and Souza, F. F., GOLAPA: Uma Arquitetura Aberta e Extensível para Integração entre SIG e OLAP., In Proc., GEOINFO, 2001.

Gonzalez, M. L., Spatial OLAP: Conquering Geography. DB2 magazine., Online, at (http://www.db2mag. com/db_area/archives/1999/q1/99sp_gonz.shtml) , 1999.

Han, J. and Fu, Y., Discovery of Multiple Level Association Rules from Large Databases., 1997

Han, J., Koperski, K. and Stefanovic, N., GeoMiner: a system prototype for spatial data mining., Proc. ACM SIGMOD, Int. Conf. on Management of Data, Tucson, Arizona, 1997

Han, J., Stefanovic, N. and Koperski, K., Selective Materialization: An Efficient Method for Spatial Data Cube Construction., In: Proceedings of the 2nd Pacific-Asia Conference on R&D in Knowledge Discovery and Data Mining, Springer-Verlag, 1998, pp. 144--158.

Harjinder, G. and Rao, P. C., Data Warehousing,, Prentice-Hall, Hispanoamericana S. A., 1996.

Inmon, W. H., Building the Data Warehouse., 2nd edition, John Wiley & Sons, 1997.

Inmon, W. H., Using the Data Warehouse., John Wiley & Sons, 1996.

ISO, International Standard Organization., ISO/IEC 9075/1992 Information technology – database languages - Structured Query Language (SQL) Geneva, Switzerland, 1992.

ISO, International Standard Organization., ISO/IEC 9075- 1:2003 Information technology - Database languages - SQL - Part 1: Framework (SQL/Framework) Geneva, Switzerland, 2003.

Jensen, C. S., Multidimensional data modeling for LBS's., The VLDB Journal, 13, 1, 2004.

KHEOPS Technologies., JMap Spatial OLAP: innovative technology., En: http://www.kheops-tech.com/en/jmap/WP_JMap_SOLAP.pdf, March, 2006.

Kimball, R., The Data Warehouse Toolkit: practical techniques for building dimensional data warehouses., John Wiley & Sons, Inc, 1996.

Kimball, R., Reeves, L., Ross, M. and Thornthwaite, W., The Data Warehouse Lifecycle Toolkit: expert methods for designing, developing and deploying data warehouses., John Wiley & Sons, Inc, 1998.

Koperski, K., Han, J. and Stefanovic, N., An Efficient two-step method for classification of Spatial Data, Simon Fraser University, 2001.

Lopez, I. F., Snodgrass, R. T. and Moon, B., Spatiotemporal Aggregate Computation: A Survey., IEEE Transactions on Knowledge and Data Engineering, 17, 2, 2005.

Malinowski, E. and Zimanyi, E., Spatial Hierarchies and Topological Relationships in the Spatial MultiDimER

model., In: 22nd British National Conference on Databases, Springer, 2005.

Marchand, P, Brisebois, A., Bédard, Y. and Edwards, G., Implementation and evaluation of a hypercube-based method for spatiotemporal exploration and analysis., Center for Research in Geomatics (CRG) & Geomatics for Informed Decisions (GEOIDE) - Université Laval, En: Journal of the International Society of Photogrammetry and Remote Sensign (ISPRS. Vol. 59, No. 1-2, 2003.

Matias, R. and Moura-Pires, J., SOLAP: a tool to analyze the emission of pollutants in industrial installations., Instituto politécnico de Leiria y Universidad Nueva de Lisboa, Portugal, 2001.

Microsoft Corp., Multidimensional expressions MDX., specification of OLEDB for OLAP, 1997.

Mitasova, H. and Neteler, M., Open Source GIS: A GRASS GIS Approach., Kluwer Academia Publishers, Boston, 2002.

Papadias, D, Tao, Y., Kalnis, P. and Zhang, J., Indexing Spatio-Temporal Data Warehouses., In: Proceedings. 18th International Conference on Data Engineering. IEEE Computer Society, 2002.

Papadias, D., Kalnis, P., Zhang, J. and Tao, Y., Efficient OLAP Operations in Spatial Data Warehouses., International Symposium on Spatial and Temporal Databases, (SSTD), Springer-Verlag, 2001.

Pedersen, T. B. and Tryfona, N., Pre-aggregation in Spatial Data Warehouses., In: SSTD '01: Proceedings of the 7th International Symposium on Advances in Spatial and Temporal Databases, 2001.

Pedrosa, B., Cámara, G., Fonseca, F., Carneiro, T., Modesto de Souza, R. C., TerraML: a language to support spatial dynamic modeling., Nacional Institute for Space Research, Sao Jose dos Campos, Brasil, 2003.

Rao, F., Zhang, L., Yu, X. L., Li, Y. and Chen, Y., Spatial hierarchy and OLAP-favored search in spatial data warehouse., In: DOLAP '03: Proceedings of the 6th ACM international workshop on Data warehousing and OLAP, ACM Press, 2003.

Rengifo, J. A. Estudio y aplicación de la minería de datos en bases de datos espaciales. Tesis presentada a la Universidad de los Andes para optar al título de magíster en Ingeniería de Sistemas y Computación, Bogotá. 2004.

Rivest, S., Bédard, Y. and Marchand, P., Toward Better Support for Spatial Decision Making: Defining the Characteristics of Spatial., On-Line Analytical Processing (SOLAP), Geomatica, 55, 4, 2001.

Rivest, S. Bédard, Y., Proulx, M. J., Nadeau, M., Hubert, F. and Pastor, J., SOLAP technology: Merging business intelligence with geospatial technology for interactive spatio-temporal exploration and analysis of data., Journal of International Society for Photogrammetry and Remote Sensing, Vol. 60, 2005.

Roddick, J. F., Egenhofer, M. J., Hoel, E., Papadias, D. and Salzberg, B., Spatial, Temporal and Spatio-Temporal Databases - Hot Issues and Directions for PhD Research., ACM SIGMOD, Record 33, 2004.

Stefanovic, N., Design and implementation of on-line analytical processing (OLAP) of spatial data., Tesis presentada a la Simon Fraser University, para optar al título de Magíster, in Computer Sciences, Burnaby, Canada, 1997.

The Data Warehousing Institute., http://www.dw-institute.com/

The Metadata Coalition., Open information Model (OIM), En: http://www.mdcinfo.com/oim/index.html, 2001.

The OLAP Council., OLAP Council White paper. En: http://www.symcorp.com/downloads/OLAP_CouncilWhitePaper.pdf, 1997, pp. 5.

Tory, M. and Moller, T., A model-based visualization taxonomy., School of computing science, Simon Fraser University, Burnaby, Canada, 1998.

Tsois, A., Karayannidis, N. and Sellis, T., MAC: Conceptual Data Modeling for OLAP., Knowledge and Database Systems Laboratory, Nacional Technical University of Athens, Greece, 1999.

Yin, S., Hui, L. and Chee, F. W., Integration of Web-based GIS and Online Analytical Processing., Departments of Geography and Computer Science, University of Hong Kong, 1998.

Yu, H., Pei, J., Tang, S. and Yang, D., Mining most general multidimensional summarization of probable groups in data warehouses., Technical Report TR 2005-03, School of Computing Science Simon Fraser University Burnaby, BC, Canada, February, 2005.