Data Governance: The Way Forward

Rupa Mahanti

Data Governance and Data Management

Contextualizing Data Governance Drivers, Technologies, and Tools

Springer

Data Governance and Data Management

"R. Mahanti's book offers both students and life-time learners a broad description of data governance in one book. The compilation of information in one place is welcome, saving a lot of time, when starting a DG program and you need to decide how you want it to work. Mahanti's perspective has been formed by many academic and professional sources which she draws upon when forming her conclusions. Readers will appreciate her passion on the topic for many years to come."

—Dan Myers, *Principal Info Quality Educator, DQMatters, Founder of the Conformed Dimensions of Data Quality*

"Going well beyond compliance drivers, this volume illustrates what is needed to embed data governance in the fabric of your organization and maximize the value of your data."

—Jeannine Siviy, *Director, Healthcare Solutions, SDLC Partners, Co-author of CMMI and Six Sigma*

"The second volume of Rupa Mahanti's ambitious three volume project on Data Governance adds to the growing body of knowledge around improving data governance practices. Mahanti puts data governance work in the overall context of data management. She discusses the levels of governance for different types of data (master data, reference data, transactional data), with a particular emphasis on the relation of DG to data quality and data security. Her comprehensive and practical approach presents the 'Whys' while also getting at the 'How's' by describing the facilitation of DG activities through appropriate tooling. Data Governance and Data Management: Contextualizing Data Governance Drivers, Technologies, and Tools will enable organizations in a wide range of industries to improve the effectiveness of their Data Governance initiatives."

—Laura Sebastian-Coleman, *Author of Navigating the Labyrinth: An Executive Guide to Data Management*

"The fact that data is an asset is probably the understatement of the century. Though in order to be treated as such, it needs to be governed. Before going deeper onto the data governance path, I recommend understanding it better and learning how it helps, what its relationship is to data management functions, and data specific tools and technologies. Read Rupa's book to understand these aspects."

—George Firican, *Founder of LightsOnData*

"Data Governance and Data Management: Contextualizing Data Governance Drivers, Technologies, and Tools is a great resource for helping your organization become truly 'data-driven.' A lack of understanding about the relationship between technology and data

often becomes a stumbling block in achieving this goal. This book clearly explains how data operations and data governance can only be successful when viewed in the full context of the processes, technology, and people in an organization."

—Dr. John R. Talburt, *Acxiom Chair of Information Quality at the University of Arkansas at Little Rock, and Lead Consultant for Data Governance and Data Integration with Noetic Partners*

"I can hear readers who open this book for the first time thinking, 'This book will help me figure out how to start our data governance project.' I'm hoping that by the time they reach the end, these same readers will be fortified by the experiential knowledge that Rupa Mahanti and her band of data experts have imparted throughout these pages. Data has moved beyond a byproduct of the systems that generate it to become the major driving force behind business, and indeed our own evolution. It's more than a project, it could very well be our salvation."

—Jill Dyché, *Author and Data Strategy Consultant*

Rupa Mahanti

Data Governance and Data Management

Contextualizing Data Governance Drivers, Technologies, and Tools

Springer

Rupa Mahanti
Strathfield, NSW, Australia

ISBN 978-981-16-3582-3 ISBN 978-981-16-3583-0 (eBook)
https://doi.org/10.1007/978-981-16-3583-0

© The Editor(s) (if applicable) and The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2021
This work is subject to copyright. All rights are solely and exclusively licensed by the Publisher, whether the whole or part of the material is concerned, specifically the rights of reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.
The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.
The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Singapore Pte Ltd.
The registered company address is: 152 Beach Road, #21-01/04 Gateway East, Singapore 189721, Singapore

Data Governance: The Way Forward

This mini-series of three volumes looks at different aspects of data governance. The volumes do not assume any prior or specialist knowledge in data governance. Data Governance is an important component of the data management discipline, and these books should appeal aspiring information management candidates, aspiring/current BI and reporting analysts, project managers heading information management projects, business and IT consultants, especially those involved with data, university students, professionals in organisations from all industries and sectors who wish to gain a comprehensive understanding of data governance.

To my parents,
for their unwavering support, dedication,
love, and encouragement.

To my teachers and mentors,
for their guidance and patience.

Foreword

The second book in this series starts out with a discussion of "Data and Its Governance". The data explosion, where the amount of data worldwide is doubling roughly every 1.5–2 years, has led to growing expectations from corporate management, regulators, and the general public on how data is gathered, managed, and used by corporations and government entities.

For some companies, the data is their business. E-commerce companies such as Amazon, AirBnB, Netflix, and Expedia, and social media companies like Facebook, Twitter, and LinkedIn, wouldn't be in business without the Internet and sophisticated methods of gathering, managing, and acting on large volumes of data.

Unfortunately, for people as individuals, it's become clear that if you're not buying the product, then you are the product. There's serious concern among federal and state legislators on these large companies' data practices and what those practices have done to the ability of foreign governments and other non-state actors to influence elections, erode privacy rights, take advantage of cybersecurity vulnerabilities, and harm the financial and reputational standing of many law-abiding organizations and people.

The book continues with a helpful discussion of the history and basic concepts of data and its organization, database management systems (DBMS), the relational model, data warehousing, and the growing trend for organizations to treat their data as an asset, even though most companies do not yet list data on their balance sheets.

The book introduces the concept of the data asset life cycle, and how data governance spans that life cycle, from creation & acquisition, to capture, processing, storing and maintenance, through to archiving, purging, and retirement. The author also outlines the issues when data is not treated as an asset. Unfortunately for most companies, their data is their single biggest asset, but many Fortune 500 CEOs don't fully appreciate that fact (Andrew Lo, Financial Economist and Professor of Finance at MIT). The book also explores the question of why data is not consistently treated as an enterprise asset, and in fact, data sometimes degrades and turns from an asset to a liability.

The book outlines different ways to categorize data, assign it to domains or subject areas, and track its physical locations, uses, criticality, format, and sensitivity. The book also looks at the role of business, process, and technical metadata in better understanding master data, transactional data, and reference data.

The author goes on to describe the definitions and dimensions of data quality, and the role of data governance in the enterprise, in order to govern data as a strategic enterprise asset.

Many companies are going through "digital transformations", only to find out that their underlying data assets are not in good enough shape to support that transformation. Informal rather than formal data governance, under-investment in data governance, and data management technology (including data quality) and loose or poorly established data policies and practices can all be contributors to a state of "data anarchy."

The author outlines a number of indicators and warning signals that organizations need to move from informal to formal data governance, distinguishes between a "light" and "heavy" approach to data governance, and describes a spectrum from "lightly governed" to "tightly governed" for different types of data, ranging from public data, to transactional data, to confidential and compliance-related data. There's also a great section on the benefits of formal data governance.

There's also a great discussion of "Data Governance Technology and Tools". The usual three-way approach of "people, process, and technology" applies to data governance. Once the "why" and "what" are defined, people want to quickly move on to the "how" question. However, data governance is not primarily a technology-driven discipline. Here, the "people" and "process" aspects truly are more important, and the business initiatives carry more weight than the technology initiatives.

Data governance is fundamentally an organizational change management exercise, and technology can't and doesn't serve as a "silver bullet" for deep-rooted attitudes toward and behaviors around data. Technology at best acts as a force multiplier, to make governing and managing a large amount of data across a complex enterprise less unwieldy and frustrating.

Metadata management is a key area of data governance, but some organizations never go beyond that to other important areas such as managing the organizational structure itself, enhancing data security and privacy, automation through tools such as workflow, improved collaboration and communication, and data quality functions such as data profiling, data cleansing, parsing, standardization, matching, linking, merging, and enrichment.

Other important tools include vendor offerings for master data management and reference data management, as well as analytics tools that can create governance-related metrics and scorecards, which are helpful in tracking compliance with data governance policies.

An interesting observation by the author is that at this time, there is no single platform or tool that has all of the capabilities and features needed for large organizations' data governance business requirements. Part of evaluating any product should be to understand how easily it can be integrated with other, related data governance, quality and management products.

Hingham, USA Dan Power

Preface

Data is the life blood of an organization and a critical enterprise asset needed for informed decision making, compliance, regulatory reporting, and insights into trends, behaviors, performance, and patterns.

Since data forms the basis of your decisions, your decisions are as good as your data or in the words of Charles Babbage, "Garbage In, Garbage Out."

Good data is the key to staying ahead in a competitive market. With enterprises capturing and storing exponential volumes of data, and considering the business impact and uses of data, data needs to be given priority, and there needs to be an adequate management around the data to derive the best value.

Data management is no longer a simple discipline like it was in the early days of computing, when limited data needed to be managed and data management was all about inputting values through punch cards or storing data in magnetic tapes. In the present age of computing, data management is a multifaceted discipline comprising several closely interacting sub-disciplines or functions including data governance.

Data governance is a very important function in the data management space. It is one of the core data management functions/components that ties together and guides all the other data management functions.

However, data governance is often overlooked, misunderstood or confused with other terminologies, and data management functions. Given the pervasiveness and importance of data, it is imperative to understand the business drivers and benefits for data governance, the interactions of data governance function with other data management functions such as data governance tools and technologies.

Effective data governance is essential for attaining high-quality data. Data quality and data governance have a symbiotic relationship. When I was writing my first book on data quality, I realized I had a lot to convey to readers about data governance that needed a book in its own right. What started out as one book, is now a data governance trilogy:

- **Data Governance and Compliance**. This first book in the trilogy sets the stage in terms of evolution of corporate governance, laws, regulations, other forms of governance; how data governance interacts with other corporate governance

sub-disciplines and goes on to explain how data governance helps in achieving compliance.

- **Data Governance and Data Management**. This second book, which is the book you are now reading, provides the big picture of data, and data governance (data as an asset, different technical aspects of data, data governance drivers, and benefits). It also covers interactions with different data management disciplines, initiatives, data governance technology, and tools.
- **Data Governance Success**. This final book discusses the different perceptions that individuals and organizations have in relation to data governance, the challenges in implementing data governance, and the key factors that should be considered to ensure data governance success.

These books share the combined knowledge related to data and data governance that I have gained over the years of working in different industrial and research programs, and projects associated with data, processes, and technologies.

I have interacted with professionals all over the world and have read many books and articles, most of which are listed in the references. The books in the data governance trilogy will be highly beneficial for IT students, academicians, information management and business professionals, and researchers to enhance their knowledge, and get guidance on their own specific data governance projects.

This series is written primarily for information management professionals, risk management professionals, compliance professionals, data quality practitioners, information management researchers, and students looking forward to a career in data management or governance.

In addition, this series will be useful for aspiring information management candidates, aspiring Business Intelligence (BI) and reporting analysts, project managers heading information management projects, business consultants, IT consultants, university students, researchers, and professionals in organizations from all industry sectors who want to gain an understanding of data governance.

Data governance is often overlooked, and its value is grossly underestimated. A lot of people are highly skeptical about data governance. In order to retain interests of such an audience, I have conducted interviews with 11 thought leaders, researchers, and professors. Their interview responses have been included in the appendix section of the first book of the series, Data Governance and Compliance, with an intent to share their challenges and experiences with data governance.

Each book has a slight overlap in terms of people, processes, and metrics-related aspects of data governance, though it is the third book that covers these aspects in detail. The technical aspect is a major component of this book and has been moderately discussed in the first book in relation to compliance and will be minimally discussed in the third book.

Whenever I read a book on a particular subject, from my student days to this day, I find that a book containing a balance of concepts and examples, and illustrations are easier to understand and relate to, and I have tried to do the same while writing this series. Each book has a large number of visual illustrations (figures and tables) which makes it easy for the reader to retain information.

This book covers the proliferation of data and key terms related to the organization of data, as well as data management topics such as the classification of data, data as an asset, the data asset life cycle, the need for data governance, formal versus informal data governance, the business drivers, and use cases for data governance, and data governance benefits. Compliance is one of the most important drivers and was discussed in the first book of the series—*Data Governance and Compliance*. The other drivers of data governance such as data security, data quality, big data, and data analytics will be discussed in this book.

The data management discipline has several components with data governance being identified as one of the core components. Data governance ties together other data management components and initiatives, including data architecture management, data modeling and design, master data management, reference data management, data warehousing, data migration, business intelligence, data quality management, metadata management, data security management, data storage, data operations, document management, content management, data integration, and data interoperability, big data, data lake, and analytics.

The data governance function guides all the other data management functions and data initiatives by defining, reviewing, communicating, and enforcing policies, processes, metrics, rules, and standards in each of the other data management functions and initiatives, and establishing formal roles, responsibilities, and accountabilities.

A successful data governance program involves a combination of people, processes, as well as tools and technology. Tools and technology provide support, and enables the people and process aspects of data governance through automation, scaling, and augmentation.

It is essential to explore the various components and aspects of data governance that can be facilitated by technology and tools, the distinction between data management tools and data governance tools, the readiness checks to perform before exploring the market to purchase a data governance tool, the different aspects that you need to keep in mind for comparing and selecting appropriate data governance technologies and tools from large number of options available in the marketplace, and the different market players that provide tools for supporting data governance.

In case you have any questions or want to share your feedback about the book please feel free to e-mail me at rupa.mahanti0@gmail.com. Alternatively, you can contact me on LinkedIn at—https://www.linkedin.com/in/rupa-mahanti-62627915.

Strathfield, Australia Rupa Mahanti

Acknowledgments

Writing this book was an enriching experience and given me great pleasure and satisfaction, but has been more time consuming and challenging than I thought. I owe a debt of gratitude to many people who have directly or indirectly helped me on my data governance journey.

I am extremely grateful to the many leaders in the field of data governance and related fields who have taken the time to write articles and/or books so that I and many others could gain knowledge. The bibliography in one or more books in the data governance book series shows the extent of my appreciation to those who have made that effort; special thanks to Anne-Marie Smith, Boris Otto, Chisolm Malcolm, Carlo Batini, Dan Myers, Dan Power, Dannette McGilvray, David Loshin, David Marco, David Plotkin, Doug Laney, Dylan Jones, Evan Levy, George Firican, Gwen Thomas, Hubert Österle, John Ladley, John R. Talburt, Jill Dyché, Kelle O'Neal, Larissa Moss, Larry P. English, Laura Sebastian-Coleman, Lowell Fryman, Majid Abai, Monica Scannapieco, Neera Bhansali, Nicola Askham, Peter Aiken, Philip Russom, Prashanth H. Southekal, Robert F. Smallwood, Robert Seiner, Richard Wang, Sid Adelman, Steve Sarsfield, Sunil Soares, Thomas C. Redman, Todd Harbour, Tony Fisher, Wayne Eckerson, and Yinle Zhou.

I had quite a few questions in relation to book publishing, and I am extremely grateful to Bill Hefley, Jill Dyché, Karl Wiegers, Laura Sebastian-Coleman, Nicole Radziwill, Sandeep Nagar, and Victor Squires for answering some of them. Many thanks to Satish Gawade for helping me with understanding the terms of the book contract.

Many thanks to Andres Perez, Christopher Butler, George Firican, Jill Dyché, John Talburt, John Zachman, Laura Sebastian-Coleman, Phil Watt, Shannon Fuller, Stan Rifkin, and Tony Epler for agreeing for an interview and sharing their unique perspectives.

I would like to thank the many clients and colleagues who have challenged and collaborated with me on so many initiatives over the years. I appreciate the opportunity to work with such high-quality people.

I am very grateful to Springer for giving me an opportunity to publish this book. I am particularly thankful to Anushangi Weerakan for her continued cooperation and support for this project. She was extremely patient and flexible in accommodating my requests. Her unique perspective and feedback made this book so much better. I would like to thank Umamagesh Perumal and Saranya Kalidoss for making the production process smooth for me. Thanks to the Springer team for helping me make this book a reality. The Springer team made the process and the experience very easy and enjoyable. I am also thankful to the reviewers for their time and constructive feedback that helped improve the quality of this book. Many thanks to Jeannine Siviy for her feedback and helpful suggestions that helped make this a better book.

I am grateful to my teachers at Sacred Heart Convent, DAV JVM and Birla Institute of Technology, where I received education that created opportunities that have led me where I am today. Thanks to all my English teachers and special thanks to Miss Amarjeet Singh because of whose efforts I have acquired good reading and writing skills. My years in Ph.D. research have played a key role in my career and personal development, and I owe a special thanks to my Ph.D. guides—Dr. Vandana Bhattacherjee and Late Dr. S. K. Mukherjee and my teacher and mentor Dr. P. K. Mahanti who supported me during this period. Though miles way, Dr. Vandana Bhattacherjee and Dr. P. K. Mahanti still provide me with guidance and encouragement and I will always be indebted to them. I am also thankful to my students whose questions have enabled me think more and find a better solution.

Last but not least, many thanks to my parents for their unwavering support, encouragement, and optimism. They have been my rock throughout my life, even when they are not near me and hence share credit for every goal I achieve. Writing this book took most of my time outside of work hours. I would not have been able to write the manuscript without they being so supportive and encouraging. They were my inspiration and fueled my determination to finish this book.

About This Book

In the present age of computing, data management is a multifaceted discipline comprising of several closely interacting sub-disciplines or functions. Data governance is a very important function in the data management space and one of the core data management functions/components that ties together with other data management functions and initiatives. However, it is often overlooked, misunderstood, or confused with other terminologies and data management functions. Given the pervasiveness of data and the importance of data, it is imperative to understand the business drivers and benefits for data governance, the interactions of data governance function with other data management functions, various components and aspects of data governance that can be facilitated by technology and tools, the distinction between data management tools and data governance tools, the readiness checks to perform before exploring the market to purchase a data governance tool, the different aspects that you need to keep in mind for comparing and selecting the appropriate data governance technologies and tools from large number of options available in the marketplace, and the different market players that provide tools for supporting data governance. *Data Governance and Data Management: Contextualizing Data Governance Drivers, Technologies, and Tools* is the second book in the Data Governance Series consisting of three books and discusses these very aspects. This book shares the combined knowledge related to data and data governance that the author has gained over the years of working in different industrial and research programs and projects associated with data, processes, and technologies and unique perspectives of thought leaders and data experts through interviews conducted. This book will be highly beneficial for IT students, academicians, information management, and business professionals, and researchers to enhance their knowledge and get guidance on how to implement data governance in conjunction with their own specific data initiatives. This book contains a balance of concepts and examples and illustrations making it easy for the readers to understand and relate to their own specific data projects. While this book does discuss technical aspects, the reader does not need to have knowledge of any specific technology to understand the concepts discussed in the book.

Contents

1 Introduction to Data, Data Governance, and Data Management . . . 1
1.1 Evolution of Data . . . 1
1.2 Data and Its Governance . . . 2
1.3 Data Governance and Data Management . . . 2
1.4 Concluding Thoughts . . . 2
Reference . . . 3

2 Data and Its Governance . . . 5
2.1 The Data Deluge . . . 5
2.2 About Data and the Organization of Data . . . 8
2.3 Data as an Asset and Governance . . . 14
2.3.1 Is Data an Asset? . . . 14
2.4 Data Asset Life Cycle . . . 21
2.5 Common Problems with Data Not Being Treated as an Asset . . . 23
2.6 Classification of Data . . . 26
2.6.1 Entities . . . 27
2.6.2 Varieties of Data . . . 31
2.6.3 Acquisition/Creation of Data . . . 32
2.6.4 Data Domains or Data Subject Areas . . . 33
2.6.5 Time . . . 33
2.6.6 Uses of Data . . . 33
2.6.7 Data Criticality Based on Integrity and Availability . . . 33
2.6.8 Location of Data . . . 35
2.6.9 The Sensitivity of the Data and the Level of Protection the Data Requires . . . 35
2.7 Data Quality and Data Quality Dimensions . . . 39
2.7.1 Data Quality Dimensions . . . 40
2.8 Need for Good Data Governance . . . 44
2.9 Informal Versus Formal Data Governance . . . 48
2.9.1 Warning Signs that Indicate, You Need Formal Data Governance . . . 50

2.10 Data Governance is not the Same as Data Management or Data Quality . . . 53
2.10.1 Data Governance and Data Management . . . 53
2.10.2 Data Governance and Data Quality . . . 55
2.11 Data Governance Goals . . . 57
2.12 Data Governance—The Key Elements . . . 58
2.12.1 People . . . 58
2.12.2 Processes . . . 60
2.12.3 Tools and Technology . . . 60
2.13 Key Data Governance Business Drivers and Uses Cases . . . 60
2.13.1 Compliance . . . 62
2.13.2 Improving Customer Satisfaction . . . 64
2.13.3 Reputation Management . . . 64
2.13.4 Better Decision Making . . . 65
2.13.5 Data Security and Privacy . . . 66
2.13.6 Improving Data Quality . . . 66
2.13.7 Analytics . . . 67
2.13.8 Big Data . . . 68
2.13.9 Revenue Growth . . . 69
2.13.10 Improving Operational Efficiency . . . 70
2.13.11 Mergers and Acquisitions . . . 70
2.13.12 Partnering and Outsourcing . . . 71
2.14 Key Benefits of Data Governance . . . 72
2.14.1 Common Understanding of Data . . . 72
2.14.2 Greater Collaboration . . . 74
2.14.3 Improved Data Discovery . . . 74
2.14.4 Increased Confidence in Data . . . 74
2.14.5 Improved Brand Protection . . . 75
2.14.6 Improved Decision Making . . . 75
2.14.7 Competitive Advantage . . . 75
2.14.8 Improved Data Management . . . 75
2.14.9 Improved Risk Mitigation . . . 76
2.14.10 Cost Savings . . . 76
2.14.11 Support Impact Analysis . . . 76
2.14.12 Business and IT Partnership . . . 77
2.15 Concluding Thoughts . . . 77
References . . . 78

3 Data Governance and Data Management Functions and Initiatives . . . 83
3.1 Data Governance and Data Management . . . 83
3.2 Data Management Functions and Initiatives . . . 85
3.3 Data Architecture, Data Modeling, Design, and Data Governance . . . 87

3.4 Data Governance, Data Integration, and Data Interoperability ... 89
3.4.1 Stakeholder Engagement and Management ... 90
3.4.2 Establish Governance Policies, Processes, and Best Practices ... 91
3.4.3 Metadata Management and Data Lineage ... 92
3.4.4 Security and Privacy ... 92
3.4.5 Data Sharing Agreements ... 93
3.4.6 Data Integration Metrics ... 93
3.5 Data Governance and Reference Data Management ... 93
3.5.1 What is Reference Data? ... 94
3.5.2 Reference Data Categories ... 94
3.5.3 Reference Data Governance ... 95
3.6 Data Governance and Master Data Management ... 96
3.6.1 Agreement and Management of Critical Master Data Elements ... 97
3.6.2 Defining and Enforcing Data Policies, Processes, Rules, and Standards ... 98
3.6.3 Roles, Responsibilities, and Accountabilities ... 99
3.6.4 Agreement on Metrics ... 99
3.6.5 Agreement on All Associated Reference Data ... 99
3.7 Data Governance, Data Warehousing, and Business Intelligence ... 100
3.8 Data Governance and Data Migration ... 103
3.9 Data Governance and Metadata Management ... 104
3.10 Data Governance, Document, and Content Management ... 109
3.10.1 Document Management ... 110
3.10.2 Content Management ... 111
3.10.3 Document Management System (DMS) Versus Content Management System (CMS) ... 112
3.11 Data Governance and Data Security Management ... 113
3.11.1 Define a Data Classification Policy ... 116
3.11.2 Discover Sensitive Data, Establish Data Ownership, and Data Stewardship ... 123
3.11.3 Classify Data ... 123
3.11.4 Use the Data Classification Results to Improve Security and Compliance ... 125
3.12 Data Governance, Data Storage, and Operations ... 127
3.13 Data Governance and Data Quality Management (DQM) ... 128
3.14 Big Data and Data Analytics ... 132
3.14.1 What is Big Data? ... 132
3.14.2 How is Big Data Different from Data or Traditional Data? ... 132
3.14.3 Data Analytics ... 134

3.15 Big Data, Analytics, Data Lake, and Data Governance 135
3.16 Concluding Thoughts 139
References 140

4 Data Governance Technology and Tools 145
4.1 Data Governance and Technology 145
4.2 Data Governance Tools Versus Data Management Tools 147
4.3 Data Governance Elements That Can Be Supported By Tools . . . 148
4.3.1 Managing Data Artifacts 149
4.3.2 Metadata Management 151
4.3.3 Governance Organizational Structure 152
4.3.4 Data Security and Privacy 153
4.3.5 Program Management and Workflow Management 153
4.3.6 Data Stewardship Activities 154
4.3.7 Business Alignment 154
4.3.8 Communication and Collaboration 155
4.3.9 Data Management Activities and Data Quality 155
4.3.10 Master Data Management (MDM) and Reference Data Management 157
4.3.11 Data Governance Metrics 157
4.3.12 Data Policy Management 158
4.3.13 Data Issue Resolution 158
4.3.14 Managing Other Artifacts 159
4.4 Data Governance Tool Readiness, Selection, and Acquisition . . . 159
4.5 Data Governance Tool Vendors 162
4.6 Conclusion and Final Thoughts 165
References 168

5 Data Governance and Data Management—Concluding Thoughts and Way Forward 169
5.1 Data and Its Governance 169
5.2 Data Governance Stakeholders 169
5.3 Data Governance and Data Management 170
5.4 Data Governance—The Way Forward 173
References 173

Appendix A: Restricted Data 175

Appendix B: Glossary of Terms 179

Appendix C: Bibliography 187

Index 189

About the Author

Rupa Mahanti is a Business and Information Management consultant with extensive and diversified consulting experience in different solution environments, industry sectors, and geographies (USA, UK, India, and Australia). She has expertise in different information management disciplines, business process improvement, regulatory reporting, and more. Her research interests include quality management, information management, software engineering, empirical study, environmental management, simulation and modeling, and more. With a work experience that spans industry, academics, and research, Rupa has guided a doctoral dissertation, published a large number of research articles, and is the author of the books—*Data Quality: Dimensions, Measurement, Strategy, Management and Governance* (Quality Press), *Data Governance and Compliance: Evolving to Our Current High Stakes Environment, and Thoughts: A Collection of Inspirational Quotes*. She is an Associate Editor with the journal *Software Quality Professional* and a reviewer for several international journals.

Acronyms and Abbreviations

AML	Anti-money laundering
ANSI	American National Standards Institute
BI	Business intelligence
CDE	Critical data element
CDO	Chief data officer
CEO	Chief executive officer
CFO	Chief financial officer
CIO	Chief information officer
COREP	Common Reporting Framework
CPO	Chief privacy officer
CPU	Central processing unit
CRM	Customer relationship management
CSV	Comma-separated value
DA	Data analytics
DAMA	Data Management Association
DBA	Database administrator
DBMS	Database management system
DCAM	Data Management Capability Assessment Model
DCMI	Dublin Core Metadata Initiative
DG	Data governance
DGF	Data governance framework
DGI	Data Governance Institute
DGO	Data governance office
DII	Data integration and interoperability
DMBOK	Data Management Body of Knowledge
DQ	Data quality
DW	Data warehouse
EBA	European Banking Authority
ECM	Enterprise content management
EDW	Enterprise data warehouse

ePHI	Electronic Protected Health Information
ERP	Enterprise resource planning
ETL	Extract, transform, and load
EU	European Union
EY	Ernst & Young
FASB	Financial Accounting Standards Board
FATCA	Foreign Account Tax Compliance Act
FCPA	Foreign and Corrupt Practices Act
FERPA	Family Educational Rights and Privacy Act
FFI	Foreign financial institutions
FI	Financial institutions
FINREP	Financial Reporting Framework
FMV	Fair market value
GAAP	Generally Accepted Recordkeeping Principles®
GDPR	Global Data Protection Regulation
GRC	Governance, risk, and compliance
HIPAA	Health Insurance Portability and Accountability Act
IASB	International Accounting Standards Board
IBM	International Business Machines Corporation
IFRS	International Financial Reporting Standards
IIHI	Individually identifiable health information
IIS	InfoSphere Information Server
IoT	Internet of Things
IP	Internet Protocol
ISO	International Organization for Standardization
IT	Information technology
JSON	JavaScript Object Notation
KYC	Know your customer
MDM	Master data management
MiFID	Markets in Financial Instruments Directive
NACE	Nomenclature des Activités Économiques dans la Communauté Européenne
NISO	National Information Standards Organization
OLAP	Online analytical processing
OLTP	Online transaction processing
PCI DSS	Payment Card Industry Data Security Standard
PCI	Payment card industry
PHI	Protected Health Information
PII	Personally identifiable information
POC	Proof of concept
RACI	Responsible, accountable, consulted, and informed
RDBMS	Relational database management system
RDM	Reference data management
ROI	Return on investment
SCM	Supply chain management

SIC	Standard Industry Classification
SLA	Service level agreement
SME	Subject matter expert
SoX	Sarbane Oxley
SSN	Social security number
TSV	Tab-separated value
URL	Universal resource locator
US	United States
XML	Extensible Markup Language

List of Figures

Fig. 1.1 Data governance tying together the data management functions . . . 3
Fig. 2.1 The data hierarchy . . . 12
Fig. 2.2 Unique properties of data . . . 16
Fig. 2.3 Challenges of listing data in the balance sheet . . . 18
Fig. 2.4 The data asset life cycle . . . 22
Fig. 2.5 Data asset versus data liability . . . 26
Fig. 2.6 Different ways of classifying data . . . 28
Fig. 2.7 Master data groups . . . 29
Fig. 2.8 Entity—data categories . . . 30
Fig. 2.9 Data classification based on data criticality . . . 34
Fig. 2.10 Data classification by data sensitivity and protection (Mahanti 2021) . . . 36
Fig. 2.11 Data inconsistency example . . . 42
Fig. 2.12 Issues arising from ineffective data governance or absence of data governance . . . 46
Fig. 2.13 Light versus heavy governance . . . 52
Fig. 2.14 Data governance tying together the data management functions . . . 55
Fig. 2.15 People aspect of data governance . . . 59
Fig. 2.16 Data governance business drivers and use cases . . . 61
Fig. 2.17 Key benefits of data governance . . . 73
Fig. 3.1 Data governance and data management functions . . . 85
Fig. 3.2 Data governance tying together the other data management functions . . . 86
Fig. 3.3 Alignment between data governance and data architecture . . . 90
Fig. 3.4 Role of data governance in DII solutions . . . 91
Fig. 3.5 Role of data governance in master data management . . . 98
Fig. 3.6 Data warehouse—a high level view . . . 100
Fig. 3.7 Different sources of metadata . . . 106

Fig. 3.8 Metadata repository characteristics (adapted from Marco 2017) . . . 108
Fig. 3.9 Data classification process . . . 116
Fig. 3.10 Data classification policy components . . . 117
Fig. 3.11 Data classification schemes . . . 119
Fig. 3.12 Data classification by data sensitivity and protection (Mahanti 2021) . . . 120
Fig. 3.13 Different types of analytics . . . 134
Fig. 4.1 Data governance—combination of people, process, and technology . . . 146
Fig. 4.2 Data governance activities and artifacts supported by common toolkits like Excel, Word, Wikis, SharePoint and document and content management systems . . . 149
Fig. 4.3 Data governance activities that can be supported by vendor tools . . . 150
Fig. 4.4 Data governance technology and tool readiness—aspects to consider . . . 160
Fig. 4.5 Vendor tools—aspects to consider . . . 162
Fig. 4.6 Categories of data governance tool vendors as grouped by Forrester research . . . 163
Fig. 5.1 Data governance in a page . . . 172

List of Tables

Table 2.1 Difference between data warehouse and data lake 13
Table 2.2 Properties of fixed assets versus data assets 17
Table 2.3 Duplicate record for customer—"Monalisa King" 42
Table 2.4 Difference between informal and formal data governance (Adapted from Mauzy et al. 2016) 49
Table 3.1 Potential impact definitions for each security objective—confidentiality, integrity, and availability (Evans et al. 2004) 122
Table 3.2 Data classification document template 123
Table 4.1 Data governance functionalities and corresponding vendors and tools 166

Chapter 1
Introduction to Data, Data Governance, and Data Management

Data is a precious thing that will last longer than the system themselves.
—Tim Berners-Lee, Computer Scientist and Inventor of World Wide Web

Abstract This chapter introduces the audience to the evolution of data, importance of data, data governance, and data management.

1.1 Evolution of Data

Technology, civilizations, and culture have evolved over history. However, what has not changed are facts. The way they have been stored, maintained, and transferred has evolved. With the passage of time, and evolution of technologies, civilizations, and culture, the methodologies used to capture, store, process, and use these facts have evolved. Similarly, data, that is, a representation of facts has had its own evolution cycle.

Until the advent of computers, limited facts were documented, given the expense and scarcity of resources, and effort to store and maintain them. In ancient times, it was not uncommon for knowledge to be transferred from one generation to another by the process of oral learning, in contrast to our current digital age, which has elaborate document and content management systems that store knowledge in the form of documents.

With the advent of computers and subsequent innovations in computing and industrial automation, a marked shift in data processing has resulted in the electronic recording and processing of data to support business operations. While electronic storage and processing of data started at the end of the nineteenth century, owing to the cost and limitations of storage, the amount of data that could be stored was relatively less, and data management as a discipline was less complex. Technology was seen as a means to reduce manual overhead to generate correct reports, and data was seen as a by-product.

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2021

R. Mahanti, *Data Governance and Data Management*,
https://doi.org/10.1007/978-981-16-3583-0_1

However, with the advancement in technology, and decreasing cost of hardware, increasing volume of data could be stored, and with the internet age, it has culminated into what we can call an explosion of data.

1.2 Data and Its Governance

We are currently living in the digital age and data is the key differentiator. While physical and financial assets are valued, appear on an organization's balance sheet, and are usually adequately governed, data assets are often the worst governed, least understood, and most poorly utilized key asset in most organizations. This is due to its abundance, difficulty in assigning a financial value to data, and its faulty perception that data is viewed as a considerable expense (Pierce 2007). This is despite the fact that data outlives systems and applications, and can be disruptive, if not governed effectively. However, with data being the driving force behind decisions and activities in most organizations in today's age, data is an asset, and needs to be treated as such.

Organizations that have learned to value data and its governance have often learned it the hard way—when something goes wrong. An effective data governance system ensures that the data is treated as an enterprise asset by overseeing their use, aiding data discovery, and optimizing processes around their collection, protection, privacy, access, and usage.

1.3 Data Governance and Data Management

Data management is no longer restricted to the capture, processing, and storage of data for producing reports by technical teams but instead, it is a complex cross functional enterprise wide program and discipline, having several intertwined subdisciplines such as data security management, data architecture, data quality, master data management, reference data management, and data governance. Data governance is the adhesive tying together all these different data management sub-disciplines as shown in Fig. 1.1.

1.4 Concluding Thoughts

Organizations have lots of data which can be grouped by similar characteristics—master data, reference data, metadata, and transactional data. Data can also be grouped in terms of restricted data, confidential data, private data, internal data, and public data. The different categories of data need to be managed adequately, and organizations generally have a lot of data initiatives running in parallel, for example

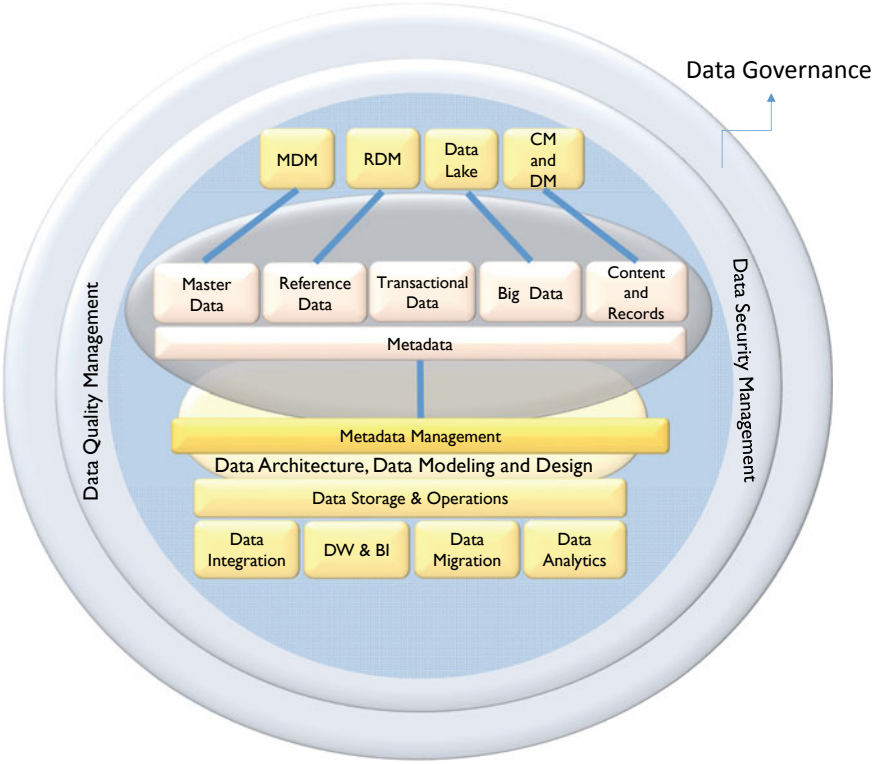

Fig. 1.1 Data governance tying together the data management functions

metadata management to manage the metadata, master data management to manage master data, data security initiatives to ensure data is secure, data quality initiatives, and so on. These initiatives need to be aligned, dependencies need to be understood, and a working rhythm needs to be established. Data governance provides oversight to these initiatives, and helps in establishing data policies, roles, responsibilities, decision rights, processes, and metrics which facilitate implementation of good data management practices.

Reference

Pierce EM (2007) Designing a data governance framework to enable and influence IQ strategy. In: Proceedings of the MIT 2007 information quality industry symposium. http://mitiq.mit.edu/IQIS/Documents/CDOIQS_200777/Papers/01_08_1C.pdf. Last accessed 12 Dec 2018

Chapter 2
Data and Its Governance

The world is one big data problem.
—Andrew McAfee

Abstract This chapter introduces to the audience, the basics concepts related to data and discusses the proliferation of data, key terms related to data, and the organization of data. We also discuss the concept of assets, why data can be considered as an asset, the unique properties of data, the data asset life cycle, how data differs from other fixed assets, and the challenges of listing data in the balance sheet. While data is an enterprise asset it is not always treated as an asset which causes problems. Organizations need to classify their assets to manage them effectively. Data can be classified in different ways and these will be discussed in this chapter. Data governance is confused with other data related terms. Some of these are discussed in this chapter. We also discuss in detail the business drivers for data governance and data governance benefits. The people, process, and technology components of data governance will be discussed at a high level in this chapter.

2.1 The Data Deluge

We are living in the digital age which is characterized by sophisticated technology and huge volumes of data. Technology is creating or enabling the creation of more and more data to the extent that, now we are experiencing what we can call a data explosion. While data has always been collected and used, the mode, volume, entity characteristics being captured, and the purposes for which data are used have evolved over ages. The evolution of data can be divided into three eras:

- **Before the advent of computer and databases**. Limited data related to transactions, events, entities, and individuals were captured and stored based on their criticality and need on a future date. Information and records were typically documented in paper files or registers, which were then filed in cabinets, but manual search and retrieval of information from these files was a time taking and tiresome process.

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2021

R. Mahanti, *Data Governance and Data Management*,
https://doi.org/10.1007/978-981-16-3583-0_2

- **Post advent of computer and electronic storage prior to the Internet era**. Data was collected and stored in electronic files and databases based on business requirements. The advancement of technology enabled increasing volumes of data to be captured, processed, and stored for operational, analysis, and reporting purposes.
- **The Internet era onwards**. The Internet era was characterized by a further progress of information technologies, declining cost of disk hardware, and availability of cloud storage. Electronic capture, processing, and storage of large volumes of data through multiple channels became common. The advancement in technologies enabled sourcing and processing massive amounts of data from heterogeneous sources using various software tools and technologies (example, data warehousing tools, big data technologies, and reporting tools).

The amount of data being created, captured, processed, and stored within IT infrastructures is growing exponentially. The sheer volume of data that exists in the new economy is enabled by exponential improvements in our ability to collect, store, transport, and analyze it.

In the last half century, while the cost of digital storage has gone down by fifty percent every two years, the storage density has increased 50-million-fold. Our ability to process data has doubled every 1.5 years (Pwc 2014; Turnbull). Data is being captured in a variety of formats and at very high velocities. Social media, mobile computing, Internet of Things (IoT), cloud, big data are some catchphrases associated with the varieties and volumes data that organizations are capturing and managing.

What has also evolved is the capture of metadata. In the words of Dr. John Talburt, Acxiom Chair of Information Quality at the University of Arkansas at Little Rock, and Lead Consultant for Data Governance and Data Integration with Noetic Partners Inc., "In the early days there was a great deal of focus on making data as compact as possible. The result was the decoupling of metadata from the data. Whereas the data was in a machine-readable file, the metadata describing the file was typically on a sheet of paper if it existed at all." This is because in the early days, both processing and storage was expensive (Mahanti 2021). However, now more metadata is captured and stored.

The building blocks of today's digital organizations are data. Organizations are dealing with more data than ever and the volume of data will keep on increasing at exponential rates. As volume increases, the expectations around the data are also increasing, and data is making a greater impact on the business. Regulators expect appropriate use of the data to comply with regulatory and compliance mandates, the customers expect organizations to use data to better meet their needs, and the internal executives expect data to be used to facilitate strategic decisions.

Furthermore, data supports and provides insights into an organization's operational processes, facilitates strategic decision making, provides competitive advantage, and generates shareholder value.

Many organizations such as LinkedIn, Apple, Alphabet, Microsoft, Google, Facebook, Airbnb, Netflix, eBay, and Amazon are data technology companies.

They are heavy on data assets to the extent that data is the heart of their business, and have changed the terms of competition in their respective industries. For some organizations, data has become a product in itself, that can be sold or bought. As an example, The Kroger Co, generates $100 million in incremental revenue per year by selling its inventory and point-of-sale data and "making that available as a syndicated data provider," according to Search CIO TechTech (Saxena 2019). In 2016, Microsoft Corp. acquired the online professional network LinkedIn Corp. for $26.2 billion. LinkedIn had 433 million registered users and approximately 100 million active users per month prior to the acquisition. The value of the data held by Linkedin contributed to the price of acquisition (Short and Todd 2017).

With enterprises capturing and storing exponential volumes of data, data needs to be given priority, and there needs to be adequate management around the data to derive the best value. Data management is no longer a simple discipline that existed in the early days of computing, when data management was all about inputting values through punch cards and storing data in magnetic tapes. In the present age of computing, data management is a multifaceted discipline comprising of several closely interacting sub-disciplines or functions, and with data governance being one of the core functions tying together all the other data management functions.

In today's rapidly changing digital world, with massive data growth, new regulations being introduced, and existing regulations being revised, good data becomes a key differentiator for organizations to gain competitive advantage—robust data governance is not optional but essential. In the absence of effective data governance, organizations have data quality issues, no or inadequate protection of sensitive data, and data ownership issues.

According to Dr. Peter Aiken, "Most organizations have no idea what data they have, they have no idea how good their people are at using data, and therefore they have no idea how their organization is using data to support their strategies." He adds that data governance addresses this big gulf by telling management what needs to be done (Knight 2017).

Data governance enables you to both reactively and proactively manage data assets effectively throughout the enterprise, and ensure that they are fit to use, by providing guidance in the form of policies, standards, processes, and rules. Data governance also demarcates roles and responsibilities to define who will do what, with respect to data. This is reinforced by Tony Epler, Chief Data Strategist, PricewaterhouseCoopers in his interview statement (Mahanti 2021):

> Data governance is critical to ensuring that an organization can put forth clean, rightly protected data. Data governance establishes rules, processes and controls to help an organization to know where its data is, what it is, and what is the highest use of the data at any point in time. Data governance establishes that data issues are most often business process issues, rather than technical issues.

This chapter starts with a discussion of basic concepts related to data and data-related terms. We discuss the concept of data as an asset, the data asset life cycle, the relation to governance, common problems with data not being treated as asset, classification of data, data quality dimensions, and the need for data governance.

Data governance is often confused with data management, and this topic is discussed in the next section in this chapter. Data governance goals, data governance business drivers, use cases, and data governance benefits are discussed in the subsequent sections.

2.2 About Data and the Organization of Data

The term *data* is actually the plural version of the rarely used term *datum*. The origins of the term *datum* lie in philosophy and dates back to mid—seventeenth century to the Latin noun and means "thing given." However, the term has a strong link with numbers, measurement, statistics, accounting, mathematics, records, economics, finance, science, and technology. The seventeenth-century philosophers used the term to refer to "things known or assumed as facts, making the basis of reasoning or calculation" (Sebastian-Coleman 2012). In today's world and age, data is used in the English language both as a plural noun as well as a singular mass noun.

The Cambridge Dictionary defines data as "information, especially facts or numbers, collected to be examined and considered, and used to help decision-making." From an IT perspective, the Cambridge Dictionary defines data as "information in an electronic form that can be stored and used by a computer."

The International Organization for Standardization (ISO) defines data as "re-interpretable representation of information in a formalized manner suitable for communication, interpretation, or processing" (ISO 11179). A singular datum provides "a fixed starting point of a scale or operation." Data are representation of facts related to entities, where an entity can be a concept, object, event, phenomena, party or location. It stands for things other than itself (Chisholm 2010; Orr 1998). Data can even represent other data, that is, data about data or metadata (Sebastian-Coleman 2012).

Data is both an interpretation of the objects it represents and an object that must be interpreted. In other words, in order for data to be understood, we need to have a context for data. This includes the context of creation of data, the characteristics of the entities it is supposed to represent, the context of usage of data, and the conventions of representation the data employ to convey meaning. These conventions of representation translate features of entities into numbers, identifiers, codes, or other symbols based on decisions made by people. To understand the meaning of a piece of data, one must understand not only what the data are supposed to represent, but also the conventions of representation they employ to convey meaning (Sebastian-Coleman 2012). These conventions can be thought of as data about data or metadata.

Data represents selective features of entities. Since there are usually more than one way to represent entities and their respective features, different people often make different choices when representing entities and their corresponding features.

This results in different representations of the same entities, and data taking on different forms, types, and sizes.

Data in organizations are usually stored in a standard structure in ***tables*** in a multitude of ***database systems***. Organizations have a large number of database and file systems which in turn have large number of tables and files respectively.

Databases are a collection of related tables. The software or program used to create and maintain database is called ***database management system (DBMS).*** These databases are focused on departments or business units, business functions or processes, such as sales, marketing, products, human resources, inventory, billing, order processing, payroll, and asset performance management depending on the industry sector and services the organization provides.

There are different types of databases—relational databases, hierarchical databases, network databases, object-oriented databases, and NOSQL databases with relational databases being the most popular database for storing structured data. NOSQL databases are primarily used to store data that are not structured and big data, though they can store structured data too.

The software system used to maintain relational databases is called ***relational database management system (RDBMS)***. The database structure is attained by appropriately organizing the data with the aid of a database model. A ***database model*** determines the logical structure of a database, and the manner in which the data can be stored, organized, and manipulated (Wiki-Database model).

Some of the common database models include relational models, hierarchical models, flat file models, object-oriented models, entity relationship models, network models, and NoSQL models. ***Database schema*** is the structure in a database management system that holds the physical implementation of a data model. The ***relational model*** is the theoretical basis of relational databases which is a framework of organizing data using two-dimensional tables called relations ("relational tables" or just "tables").

Depending on how the data is structured, information and attributes of an entity are stored in one or more related tables with relationship between the tables being maintained by a common field or set of fields. An entity may be a real life object, place, individual, event, concept or phenomena. A ***table*** is a collection of rows called records or tuples and columns called fields to represent a set of characteristics of a particular entity. Tables generally store a huge amount of data in them. Each ***row or record*** corresponds to an instance of the set of entity characteristics captured in the table whereas, the ***fields or columns*** define the data that the table holds. Often the tables are organized into ***data domains or subject areas*** containing tables with related content (Sebastian-Coleman 2012).

A ***key*** is a field or set of fields used to establish and identify relationships within tables and between tables in a database. A ***primary key*** is a single column, or a set of columns, which can be used to uniquely identify a row of data (or records) in a given table. Sometimes a single column is not sufficient to uniquely identify a record in table, in which case more than one column is required for the same. When a primary key consists of a set of columns, then it is called a ***composite primary key***. In the relational model, a ***foreign key*** is a column or a set of columns from one

relational table whose values must match another relational table's primary key, thus establishing a link between the two tables.

Organizations generally create specialized databases for recording transactions and storing operational data resulting from an organization's daily operational activities. In addition, organizations also set up systems to collect additional data beyond conventional transactions. These classes of systems are known as the OLTP (Online Transaction Processing) systems.

For example, enterprise software such as customer relationship management (CRM) systems are often used to manage and record data related to relationships and interactions with existing and prospective customers and to authorize staff to track and store the data at practically every customer contact point across different channels for subsequent analysis and decision making. Enterprise software is not only limited to CRM systems but also systems that touch every aspect of the value chain, including supply chain management (SCM) and enterprise resource planning (ERP) systems that facilitate combining data (Gallaugher 2009) across business units, departments, functions, and transforming data into a form that can be used by business users.

While operational or transactional databases and other OLTP systems are built for quick inserts, updates, deletes, and very fast query processing, they are not designed for supporting analytics, complex query processing, and decision making. This is where OLAP (Online Analytical Processing), data warehouse, and datamarts come into the picture.

A data warehouse is a database repository that sources, integrates, and consolidates data from multiple heterogeneous sources including OLTP databases and covers a wide range of subject areas or data domains, provides a centralized view of data across the organization, and supports complex query processing, analytics, and reporting. In contrast to an OLTP database that contains current data, a data warehouse contains historical data. Data warehouses generally store huge volumes of data. Data warehouses can contain detailed data, lightly summarized data, and highly summarized data, all formatted for analysis and decision support (Inmon 2005).

A data warehouse (DW) is also known as an enterprise data warehouse (EDW) and is a trusted source for integrated enterprise data. Data marts are essentially, specialized, sometimes local databases or custom built data warehouse offshoots that store data related to individual business units (e.g. sales and marketing) or specific subject areas (e.g. product, customer, asset, and event) or for addressing concerns of a particular business problem (e.g. increasing sales) (Mahanti 2019).

Warehouse data can be designed according to different meta models (star schemas, snowflake schemas, hierarchy, and fact/dimension) that mandate differing degrees of normalization.

Normalization, also known as data normalization is a refinement process and an efficient method of restructuring data in a database so as to eliminate or minimize the extent of unnecessary duplication and ensure that the data dependencies are logical without compromising the integrity of the stored data (Mahanti 2019).

In all the cases, data is understood in terms of tables and columns. The warehouse's data model provides a visual representation of the data content of a data warehouse (Sebastian-Coleman 2012). ***Dimensional modeling*** is a data structure technique that aids business users to query data stored in the data warehouse and was developed to be oriented to improve the query performance and ease of use (Zentut).

A dimensional model comprises of fact and dimension tables, and is designed to read, summarize, and analyze numeric information like values, balances, counts, weights, etc. in a data warehouse. ***Facts*** are the measurements/metrics, and these are generally numerical facts from an organization's business process. Fact tables also contain the keys which link to associated dimension tables. ***Dimension*** provides the perspective in relation to a specific business process event, and contains descriptive details about the fact, such as who, what, where, and/or when of a fact. Each dimension has several attributes.

In addition, to tables and databases, data are stored in a variety of other digital means as follows:

- Excel files which stores data in cells in a tabular fashion.
- XML (Extensible Markup Language) files which store data that can be easily read by other programs.
- JSON files that store simple data structures in JavaScript Object Notation (JSON) format, which is a standard data interchange format. It is largely used for transmission of data between a web application and a server.
- Fixed width text files are special cases of text files where the format is specified by data field widths specified in characters, pad character, and left/right alignment. The data field widths determine the maximum amount of data it can contain and is the same for all rows in the file. Each row in the file should also use the same pad character and the same alignment. Column headers are sometimes included as the first line, and each consequent line is a row of data.
- Delimited text files containing one record per row, with fields separated by characters like commas, tabs, pipes or other characters as delimiters, and rows delimited or separated from each other by carriage return character or new line character. Text files that have fields separated from one another by comma are called CSV (Comma Separated Value) files. Text files that have fields separated from one another by tab are called TSV (Tab Separated Value) files. Column headers are sometimes included as the first line, and each consequent line is a row of data.

To summarize, at the lowest level, are data elements or fields that represent the characteristics of the entity and define the data that a table or file hold. Data elements values are grouped together to form a data record, row or tuple, and represents a single instance of the entity or a set of entity characteristics that the table or file keeps track of. The table or file contains a set of records. A database contains a set of related tables and a file system contains a set of related files. The hierarchy has been summarized in Fig. 2.1.

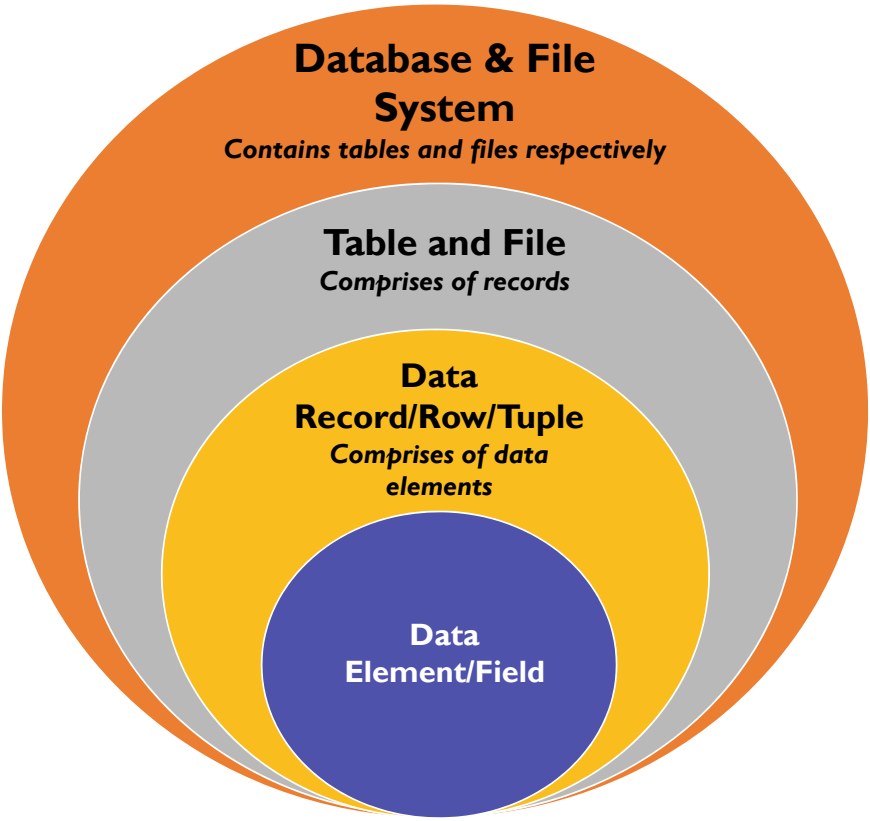

Fig. 2.1 The data hierarchy

In addition to the above terms, there is another term—"data set" that we will use in relation to data in this book. A ***data set*** is a set of records (both rows and columns) extracted from data files or database tables, for a specific purpose.

Data warehouses and data marts store structured data (data that complies to certain format and created using a predefined schema). A data warehouse applies schema-on-write, so structure and format of the data needs to defined, before one can save it. Thus, a lot of preparation is needed before the data can be stored. However, organizations have a lot of unstructured data (data that has no predefined format) and semi structured data (which lies between structured data and unstructured data). Unlike data warehouses, that stores structured data in tables and databases, data lakes can store any data (structured data, unstructured data and/or semi structured data), including data types not supported by a data warehouse, in their raw and native format at any scale. The data structure is only defined, and data is transformed as per the end user's requirements, when the data is needed for consumption. This is called "schema-on-read," because data is kept in its native

Table 2.1 Difference between data warehouse and data lake

Data warehouse	Versus	Data lake
Stores structured data in a predefined format	Data varieties and format	Stores any data (structured, unstructured, and semi-structured) in its native form
Requires speciality hardware	Hardware	Does not require speciality hardware
Schema-on-write Data retrieval is faster Operates on Extract, Load Transform or ELT strategy	Processing	Schema-on-read Data retrieval is slower Operates on Extract, Transform, Load or ETL strategy
Less agile, fixed configuration	Agility	Highly agile, configure and reconfigure as needed Adapts to changes easily
Data usage may be limited to a few pre-defined operational purposes Data is not loaded into a data warehouse until the uses are defined	Usage	Data usage is not pre-defined
Expensive for large data volumes	Storage	Designed for low cost storage
More mature	Security	Less mature
Business analysts and data analysts	Users	Data scientists and data engineers

form until it is ready to be used. Table 2.1 summarizes the differences between data warehouse and data lake.

James Dixon, founder and former CTO of Pentaho, who coined the the term "Data Lake" in October, 2010, uses the following statements to compare Data Lake and Data Mart (Foote 2020):

> If you think of a Data Mart as a store of bottled water, cleansed and packaged and structured for easy consumption, the Data Lake is a large body of water in a more natural state. The contents of the Data Lake stream in from a source to fill the lake, and various users of the lake can come to examine, dive in, or take samples.

Each data element in a data lake is assigned a unique identifier, and tagged with a set of extended metadata tags. When an user executes a business query based on a certain metadata, all the data tagged is then analyzed for the query (Patrizio 2018).

In contrast to the data warehouse, which stores data in an underlying database, data lakes use a flat file system to store data. With a database, the data elements need to be mapped to columns in a table before data can be written into the database, which is time consuming exercise. However, retrieving data from a database is much faster than retrieving data from a data lake, which has to process the data as it is read.

Jake Stein, CEO of Stitch, an ETL service that connects multiple cloud data sources, echoed the future-proofing sentiment. "If you're not sure when you're going to use the data and it's not important to have subsecond access and want to store it in a low-cost form, the data lake is the right format. It's often a case of if you

don't capture the data now, you will never get it again, so it's important to future proof yourself in that aspect (Patrizio 2018)."

2.3 Data as an Asset and Governance

What is an asset? The Cambridge English Dictionary defines an asset as,

> a useful or valuable quality; skill or person; or something of value belonging to an individual or organization that could be used to pay off debts.

According to the International Accounting Standards Board (IASB),

> an asset is a resource controlled by the enterprise as a result of past events, and from which future economic benefit is expected to flow to the enterprise.

The Financial Accounting Standards Board defines asset as,

> probable future economic benefit obtained or controlled by a particular entity as a result of past transactions or events.

American Institute of CPAs defines assets as,

> any economic resources (tangible/intangible) that can be owned or produce value. Assets have a positive economic value.

All these definitions allude to the fact that an asset can be described as any entity that has value, creates and maintains that value through its use, and has the ability to add value through its future use (Polikoff 2018). Value as a term relates to the usefulness, quality, importance, price, and worth of a subject (Fred 2017; Zeithaml 1988).

The term data asset is used in data governance to describe any data element or a data structure that has value to an organization and enables the organization to perform its functions and operational activities. A data asset may be a database, data file, or a dataset. It may also be a data structure (e.g., a table or a view) within a larger data asset, down to a specific column or field (Polikoff 2018).

2.3.1 Is Data an Asset?

The Financial Accounting Standards Board outlines three important characteristics of an asset (FASB 1980).

1. It can be managed for the benefit of an enterprise, either singularly or in combination with other assets.
2. A particular enterprise can obtain the benefit of the asset, and control access to it by other enterprises.

3. The transaction or event giving rise to the enterprise's right to control the asset has already occurred; that is, the investment has already been made and the asset is available for use.

Data has all the above characteristics, and hence can be considered as an asset.

However, in contrast to fixed assets such as factories, equipment, land, properties, and money, that are tangible (physical in nature), rivalrous (can be used by only one party at a time), fungible (one unit can be substituted by another unit), can be seen, are non-contextual, cannot be shared or cannot be easily shared, cannot be replicated, transformed, or destroyed easily, and are subject to wear and tear, data is a unique type of asset called circulating asset, and has properties that differ from other assets (Vincent 1990; Smallwood 2014).

Data is intangible (not physical in nature), non-rivalrous (that is the same data can be used by multiple users and multiple applications at the same time), non-fungible (one data cannot be substituted with another data), accumulative (data can be combined with other data and transformed into additional data assets at will), is contextual (who, what, when, where, why, and how—the metadata, or the data about the data), usually out of sight, is subject to change, moves through and across systems, applications, inter-departmental boundaries, sometimes across organizational boundaries and hence is spread across the enterprises, can be easily copied and replicated, and can grow very rapidly so that you can have so much data that is difficult to segregate the data that has business value from the data that has no business value.

Data can also be easily shared, transformed, and deleted, is not depleted, and can be reused, as data is not consumed as it is used—in fact its value increases with share and use. Figure 2.2 summarizes the unique properties of data. The relevance and value of data may increase or decrease or remain a constant with the passage of time. While data can be transformed, if maintained and governed properly, it does not wear out. Table 2.2 compares the properties of fixed assets with data assets.

There is a "fitness for purpose" aspect to data that does not apply to a lot of other assets. However, with data assets, it is important to ensure that the data and its underlying quality is fit for purpose. Data fit for one purpose, might not be fit for another purpose. For example, the customer data with missing addresses will not be a deterrent in sales reporting but will not be fit for marketing purposes. However, a lot of data is collected prior to understanding the purpose or over a period of time change in business processes no longer warrant the need of data, but data still continue to be collected and stored resulting in the "drowning in data" problem. Also, degree of data aggregation sometimes transcends the "fit for purpose" requirement.

The key to understanding the value of data is understanding (Redman 1996):

- how it is used (the contexts of usage),
- the value brought by its usage (the operational outcomes that the data drives), and,
- adverse impacts if data is not fit for purpose (for example, financial impacts, regulatory impacts, reputational damage, and productivity impacts).

Fig. 2.2 Unique properties of data

In contrast to other business assets that are present on the company's balance sheet which is a statement of the assets, liabilities, and capital of a company, data is not listed on the balance sheet. This means that the true value of the company is not captured in the balance sheet. Interestingly, on October 12, 2016, Mark Weinberger, the Global Chairman and CEO of EYshared an update on LinkedIn around the same lines:

> In 1975, 83% of the average firm's value was counted on the balance sheet. Today, it is only 16%. Quarterly reports just don't capture most of a company's true, long-term value. And, as result, they push companies to focus on short-term value over long-term investment. We need to overcome this and find a new, standardized way of reporting the full value of a company beyond the basic financials. #InclusiveCapitalism28.

Table 2.2 Properties of fixed assets versus data assets

Fixed assets	Data assets
Tangible	Intangible
Cannot be shared or not easily shared	Easily shared
Cannot be replicated, transformed or deleted easily	Easily replicated, transformed, and deleted
Cannot be copied	Can be copied
Cannot be in multiple places at the same time	Spread across organization
Consumed with use	Not consumed with use
Not accumulative	Accumulative
Non-contextual	Contextual
Rivalrous	Non-rivalrous
Fungible	Non-fungible
Wears out	Does not wear out
Cannot be sold without being given away	Can be sold without being given away (Birchler and Bütler 2007)

According to a report from the UK Treasury, the world's five most valuable companies (Apple, Amazon, Alphabet, Microsoft, and Facebook) are together worth £3.5 trillion, yet their balance sheets report just £172 billion of tangible assets. The discrepancy is because the remaining £3.3 trillion of value (their intangible assets) is well, missing in action (EverEdge 2019).

Prashant Southekal in his blog article states five key reasons that make it challenging for data to find a place in the balance sheet as follows (Southekal 2018) and summarized in Fig. 2.3.

Costing. Any asset listed on a company's balance sheet must have a fair market value (FMV). However, the Financial Accounting Standards Board (FASB) is yet to update its accounting rules to estimate the fair value of data assets. While many organizations talk about data as a monetizable asset, owing to the abstract and heterogeneous nature of data sets, they find it challenging to put a financial figure both on the cost of data management in the data lifecycle (from origination to consumption) and the benefits that data brings to the organization, as there is no standardized approach to value data.

As the Wall Street Journal reported, "Assigning a value to a physical asset like a store or equipment is relatively easy. But, in the murky world of intangible assets, the calculations are squishy. The problem of how to value such assets has vexed accountants for decades (D'Onofrio 2017)."

Depreciation. Tangible assets depreciate with age. But when data assets age, they can lose or gain value depending on the volatility and usage of the data. Time can affect the value of the data positively or negatively. Master data such as asset data (specifically in utilities companies) remains relatively constant with age, and gains value as it is used again and again.

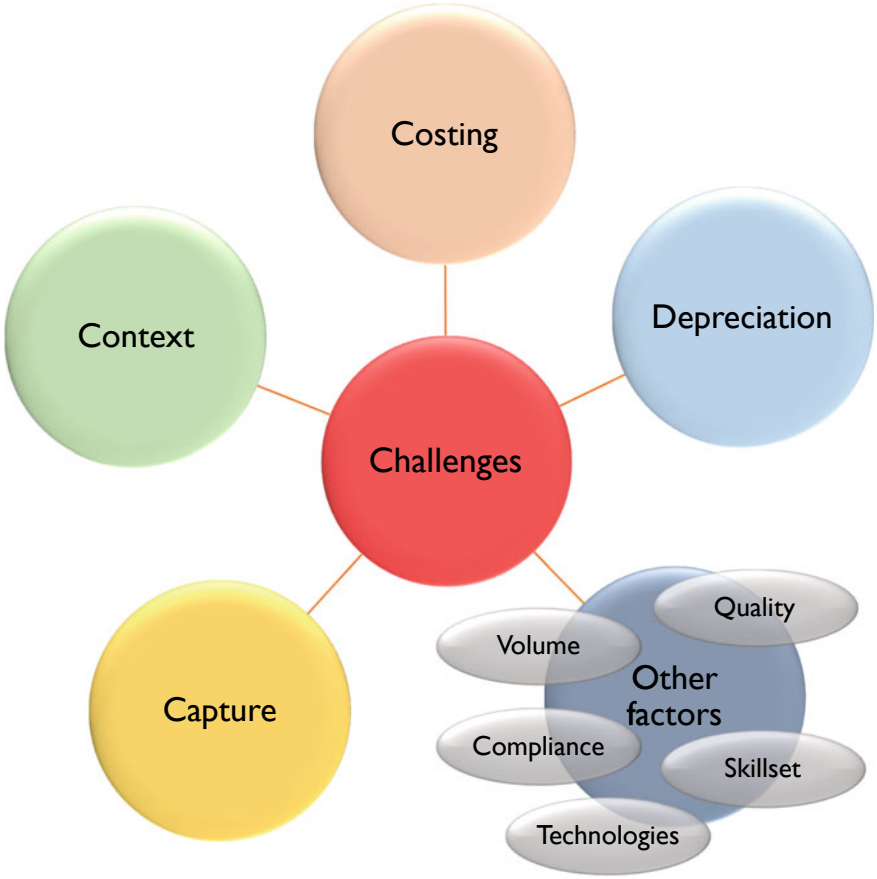

Fig. 2.3 Challenges of listing data in the balance sheet

Some data such as clinical and climate data have additive value, that is, the value of the original data increases as more data is amassed (Cebr 2013).

Other master data, like contact data does change and hence is affected negatively by time.

Transactional data (which represent business events such as purchase orders or patient appointment schedule information and account transactions such as deposits and withdrawal) age more quickly, losing its usability and relevance, and hence its value also diminishes with time.

In addition, if the data is incorrect (for example, social media data) and/or its useful life is just a few days or weeks, (for example, stock exchange data), then data cannot be properly depreciated or amortized. This requires determining the depreciation cycles, specific formulas, and regulations to calculate depreciation of data over the depreciation cycles.

Context. Assets listed on a company's balance sheet such as furniture, buildings, vehicles, and machinery are usually utilized in a similar mode at all times. However, since data are abstract representation of facts, you need to understand the context to make sense of it, use it, and there can be different uses of the same data depending on the context in hand. Hence, data is not used in the same fashion as the physical assets (Southekal 2018).

Capture. As per IFRS and GAAP accounting principles, any asset listed on a company's balance sheet should be an acquired or captured asset for which money was actually exchanged to procure the asset. Since data is consistently produced by a business and is often not purchased, it typically is not recorded on the balance sheet (D'Onofrio 2017; Southekal 2018), unless it is attained as part of a merger or acquisition or purchased.

Also, some organizations try to understand the price of their data and what the market will pay for it. For example, ISO data standards (for example, ISO-4217 currency standard and IS0-3166-1 country standard) can be bought by organizations. However, the worth of data does not always reside in selling. For example, for utilities companies the location of underground asset data is extremely valuable for its operations and monitoring performance of the underground assets which in turn impacts the quality of service. Technology organizations like Amazon and eBay use data to derive insights into consumers shopping behaviors and marketing trends, as well as use data to optimize their processes.

Other Factors. There are a number of other factors, both internal and external to an organization that can affect the value of data. Data can transform itself from an asset to a massive liability if it is not fit for purpose, that is, its quality is not up-to the mark (example, incorrect or outdated data, missing records or data that is not timely), or data has poor security and privacy controls resulting in data breaches and data hacking, or does not meet compliance requirements.

Also, with new regulations coming into picture and old regulations being revised, data requirements are also changing. Data associated with compliance is critical and if a data that was not required earlier, is now needed for compliance with regulations, the value of the data changes.

While tangible assets like machinery, land, and buildings can also become a liability for a company, the rate of change in asset to liability conversion in an intangible asset like data is significantly higher and more volatile when compared to a tangible asset. For example, Cambridge Analytics, an information rich business, filed for insolvency and closed operations in May, 2018 within 2 months of the Facebook data breach issue (Southekal 2018).

Consumer and competitor behaviors can also have an impact on the value of data (Cebr 2013). Organizations have a huge volume of different varieties of data-structured, unstructured, and semi-structured data, some that organizations are not even aware of. The value of the data is highly dependent on a combination of human intellect, skills, and technologies to discover, understand, and then unlock the value by using technologies. If you are not able to do this, it's like you are sitting on the top of a hidden gold mine.

Looking at the challenges of listing data on the balance sheet, the next question that comes in mind is that—Can data as an asset be subjected to valuation?

Doug Laney, VP and Distinguished Analyst at Gartner believes that asset valuation boils down to three things (Woodie 2016):

1. An asset is a thing that can be owned and controlled.
2. An asset is something that's exchangeable for cash.
3. An asset is something that generates probable future economic benefits.

Since data meets the above requirements, it can be subject to valuation. Doug Laney in his book Infonomics: How to Monetize, Manage, and Measure Information as an Asset for Competitive Advantage has discussed several information valuation models.

However, data asset valuation is subjective. Koutroumpis and Leiponen (2013) note that information goods are hard to value, especially due to their feature as an experience good and the subjective nature of valuation. The use of data assets affects their business and economic value. More the data is used, greater is its value. If data is not used, it becomes a liability because there is a cost involved in the collection/acquisition, storage, and maintenance of data. Data assets will be used if they are of high quality or fit for purpose. Yet, technically high-quality data is not automatically valuable (Fred 2017).

As Fred (2017) notes, value is relative to individual's perception and context, therefore, it is subjective. Data quality is multidimensional. There are various objective and subjective characteristics known as data quality dimensions that affect the quality of data, which in turn drives its usage, which in turn drives the valuation of data. Some key dimensions of data quality will be discussed in the section —*"Data Quality and Data Quality Dimensions"* in this chapter. It must be understood that not all data are of the same value to the business.

Assets need to be managed with care. Organizations have standard processes, practices, and policies in place for governing and managing their fixed assets. The procurement department has a set of processes and policies to support the ordering and purchase of office supplies; the HR department has policies and guidelines for hiring, managing, and terminating staff; and the finance or accounting department follows accounting rules and principles to manage the organization's financial assets.

However, while there does exist set of universal policies and principles for data, they need to be tailored for an organization. In order to be able to utilize data as an asset, a mix of soft skills, are required to build sustainable strategies and manage change, and hard skills (technology) to apply portfolio of data management tools and techniques to ensure the delivery of data that is fit for purpose and aligned with business strategies and initiatives (Eckerson 2011).

Data governance establishes the policies, processes, guidelines, and standards for managing, accessing, sharing data, and resolving conflict when the processes do not work, and drives the organization towards a more data driven culture. However, owing to the fact that data has unique properties very different from other assets,

data cannot be governed in the same way as fixed assets, and the unique properties of data need to be kept in mind when managing, controlling data, and designing data governance.

2.4 Data Asset Life Cycle

The data asset life cycle consists of the following phases as shown in Fig. 2.4.

- **Plan**—This phase involves creating the business, technical, operational, and security requirements relating to the data asset, and getting agreement and sign off from the stakeholders. Data governance should ensure that data producers, data consumers, data ownership and stewardship are established.
- **Create or Acquire**—This phase involves building, creating or acquiring the data asset and ensuring all the data quality requirements are met. This can involve manual data entry or electronic data capture or acquisition of data assets from external suppliers. Data governance should ensure that data quality dimensions and their respective thresholds are established for critical data assets which are created. Data governance should ensure that necessary controls and metrics are in place to gauge the quality of critical data elements, and the standards and processes to collect or acquire data are robust and followed.
- **Process and Store**—This phase involves processing the data asset as per business rules and testing against the requirements, preparing the data asset for use, classifying the data asset as per guidelines and policies, establishing metadata, creating supporting documentation about the data asset and its use, and ensuring that information about the data asset and its location can be easily retrieved by the users, as well as ensuring that the data assets are stored in secure data repositories. Data governance should ensure that data lineage is documented and maintained, metadata are captured, and rules and standards are in place to ensure data is processed such that its quality and integrity are preserved.
- **Access and Distribute**—This phase involves ensuring that authorized users can access the data asset as well as ensuring that security requirements are met with protocols established for secure, safe sharing, and re-use of the data asset. Data governance should ensure that approval authorities are established for reviewing and granting access to authorized individuals, processes and controls are in place to review and revoke access when necessary, and controls are in place to prevent unauthorized access, and that data is distributed to authorized individuals.
- **Maintain**—This phase involves the application of appropriate management strategies to the data asset, based on results of the assessments, for its optimization, rationalization, and enhancement (to improve the condition of the data asset and promote reuse to maximize the value of the data asset), replacement, archival or purging. Data needs care and attention. For example, some data are more volatile than others. For examples while date of birth of a person remains

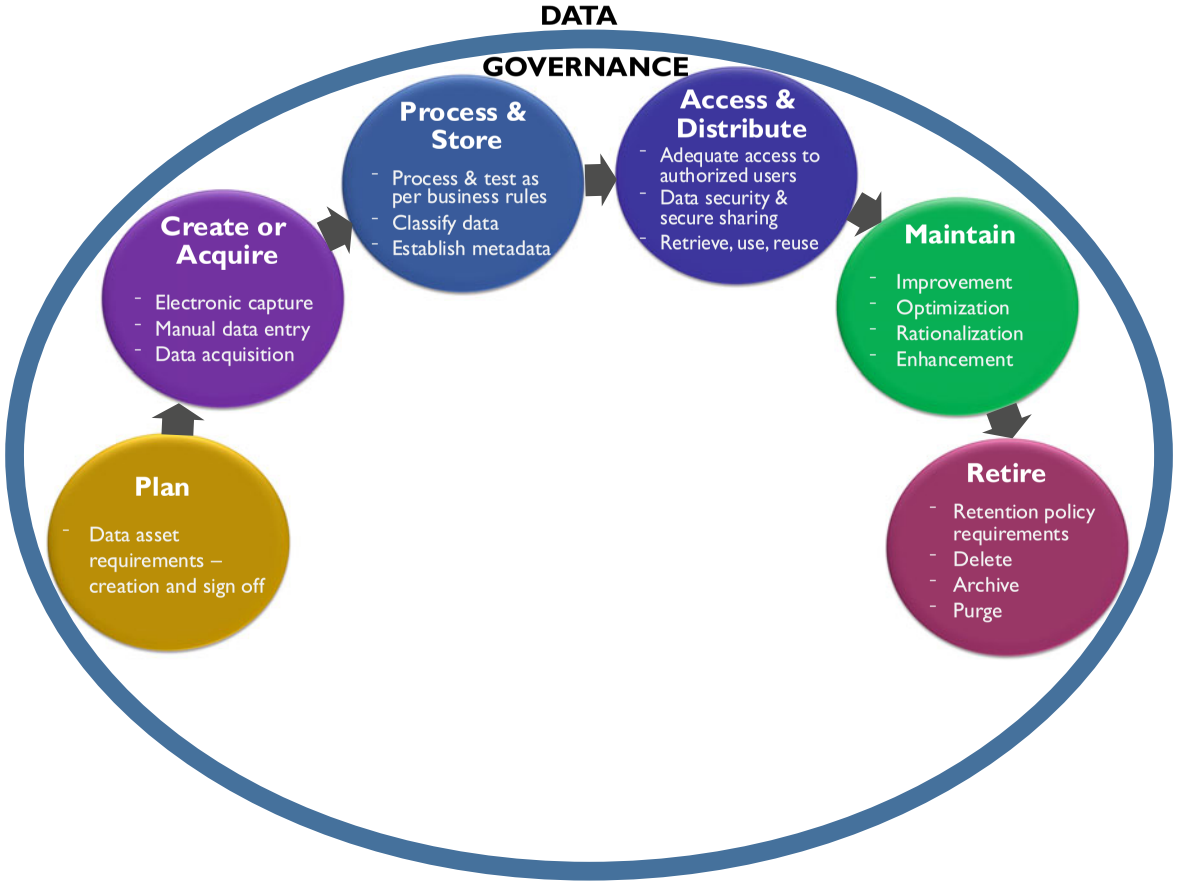

Fig. 2.4 The data asset life cycle

constant, the passport number generally, changes every 5 years or 10 years. Data governance should ensure that processes are in place to review and update data so that the data stays current.

- **Retire**—When the business need for using the data asset has significantly evolved or no longer exists then data asset has reached the end of its useful life, and it is archived and/or purged based on retention policies. Data governance should ensure processes are in place to archive or purge data in accordance with policies.

Data governance improves visibility in each phase of the data asset life cycle.

It is important to evaluate the current condition of the data asset, the cost involved in the maintenance of the data asset, its current business value, and opportunities for extended and future use. The scope of governance spans the life cycle of data assets from creation/acquisition and capture, through processing and storing, access and distribution, maintaining and archiving, deleting and purging, as well as meeting requirements around quality, security, accessibility, and disclosure.

2.5 Common Problems with Data Not Being Treated as an Asset

While most organizations recognize in principle that their data is an asset, they only pay lip-service to the way the concept is applied in practice (Glue Reply) resulting in data issues in organizations. In other words, there is a huge gap between preaching and practice. Below statistics illustrate this gap (i-scoop 2016):

- Only 4% of businesses can extract full value from their information.
- 68% of businesses plan to undertake an organizational transformation, only a small fraction has fully digitized their content-centric business processes.
- According to Forrester, around 60–73% of all data within enterprises goes unused, when it could and should have been used for analytics, driving innovation or creating new data monetization models (The Apec 2019).
- By 2020, the digital universe will reach 44 trillion gigabytes, a tenfold increase over 2013. However, most organizations are still at the beginning of their journey to extract value from information.

People have different perceptions in regards to the value of data and the view is reinforced by the following statement made by the financial economist Andrew Lo:

> Perceptions about the value of data vary widely from industry to industry.

Below are some statements that illustrate views regarding the value of data:

1. There is no division where you can't add value by using data—Davide Cervellin, Head of EU Analytics, eBay (Experian Data Quality 2014).

2. If you haven't yet made data a priority it could be the key factor that slows you down - so many organizations are too slow to react—Jora Gill, Chief Digital Officer, The Economist (Experian Data Quality 2015).
3. Computing hardware used to be a capital asset, while data wasn't thought of as an asset in the same way. Now, hardware is becoming a service people buy in real time, and the lasting asset is the data—Erik Brynjolfsson, Director, MIT Initiative on the Digital Economy (MIT + Oracle 2016).
4. For most companies, their data is their single biggest asset. Many CEOs in the Fortune 500 don't fully appreciate this fact—Andrew W. Lo, Financial Economist, Charles E. and Susan T. Harris, Professor of Finance at the MIT Sloan School of Management and director of the MIT Laboratory for Financial Engineering (MIT + Oracle 2016).
5. Most organizations have no idea what data they have, they have no idea how good their people are at using data, and therefore they have no idea how their organization is using data to support their strategies—Dr. Peter Aiken (Knight 2017).
6. Some financial services companies still don't seem to understand that they're sitting on a gold mine, and that if they ignore it, the gold mine can just turn into a trash heap. Specifically, some financial services institutions gather, but don't save, valuable demographic data about their clients and their activities. They're literally throwing away pearls of wisdom because nobody is looking at the data, and because it's taking up space—Andrew W. Lo, Financial Economist, Charles E. and Susan T. Harris, Professor of Finance at the MIT Sloan School of Management and director of the MIT Laboratory for Financial Engineering (MIT + Oracle 2016).

While the first statement establishes the value of using data, the second statement recognizes data as a priority. The third and fourth statements establish that data is an asset, and while the fourth statement reinforces the fact that while data is an asset, it often does not get treated as such. The third statement indicates that data was not treated as an asset in the same way as computing hardware in the past. The fifth statement establishes that data is not recognized as an asset. The sixth statement specifically alludes to some financial service companies not realizing the value of data, and hence, not treating it as an asset.

Organizations that treat data as an asset can derive great financial value. For example, WeWork captures and uses its data on the use of space, light, heat, and sound from more than 300 buildings in 75 countries to do cost reduction (The Apex 2019).

In response to the question—Why is data not treated as an enterprise asset? Below were the some of the answers, (Mahanti 2021):

Dr. Stan Rifkin: "All enterprises have difficult identifying their assets, and data, being intangible, and possibly fungible, is invisible, so evades notice and attention."
Dr. John Talburt: "I think this has changed significantly over the past couple of years. I agree traditionally data was thought of as a commodity and all of the focus was on programming and systems. This is probably a natural evolution. It has taken a long time to tame the program development cycle and produce re-usable, open source code. With the advent of distributed computing and cloud platforms, we are seeing a reversal of this attitude. Programs and systems now are becoming more of a commodity, and organizations are realizing their competitive edge lies is in the use of their own data resources."

"Now, organizations are actively trying to manage data as an asset and a resource, thus, we are now seeing a rising interest in data governance as an important part of the solution. You can see a growing number of organizations beginning to embrace data governance and the concept of data as an asset".
Dr. Laura Sebastian-Coleman: "Many factors limit the ability of organizations to treat data as an asset. First, in many organizations, data is created as a by-product of transactional processes, rather than something that has value for other uses (analysis, planning, etc.). While lots of organizations now claim to be or strive to be "data-driven", most still do not think about data production in terms of the quality of the data. Secondly, people have trouble abstracting a monetary value from data. If asked the question "How much is our data worth?" most CEOs probably would not be able to produce a consistent answer. Finally, there are still misconceptions about data itself. Even very smart people stumble around when trying to think about data outside of technology. This means that conversations that should be about data end up being about tools and IT processes, rather than data."

As highlighted by Dr. Laura Sebastian-Coleman above, organizations make the mistake of treating data as a technology asset given that technology has a huge role in the capture, storage, processing, access, and maintenance of data, when in reality, it is an enterprise asset, and can, and should be tied to enterprise objectives like revenue increase and cost reduction, to deliver competitive advantage.

Different business units and departments capture and maintain similar data in silos in different formats in the absence enterprise wide data definition, within their independent systems for business processing. This is because different units follow their own standards and definitions as per convenience with inadequate governance in place. The result is duplication and inconsistency of data as well as data integrity issues. For example, customer records may be stored in multiple systems in different formats. In the absence of clear data definitions, some systems store the customer's home address, the others store the customers' office address. When a customer changes jobs or house, there are chances that the address gets updated in one system whereas the other systems still reflect the old address. Hence, there are multiple addresses for the same customer. Many organizations are unable to maintain unique identifiers which further complicates matters. With unique identifiers, for example, you can match customers across their interactions with the organization; without them, it becomes very hard to get a full view of the customer experience.

When data is not treated as asset, you do not have adequate management and governance of data which in turn results in data quality and ownership issues. In some cases, it may lead to inability or difficulty to locate and access data.

Hence, data is not leveraged to its full potential. Also, in the absence of policies around data classification, protection and security, there are no adequate security and access controls around sensitive data which opens an organization to a huge risk in case the data is compromised.

In short, when you do not treat data as an asset, you do not manage and govern data adequately, resulting in degradation of data and it becoming a liability to the organization as illustrated in Fig. 2.5.

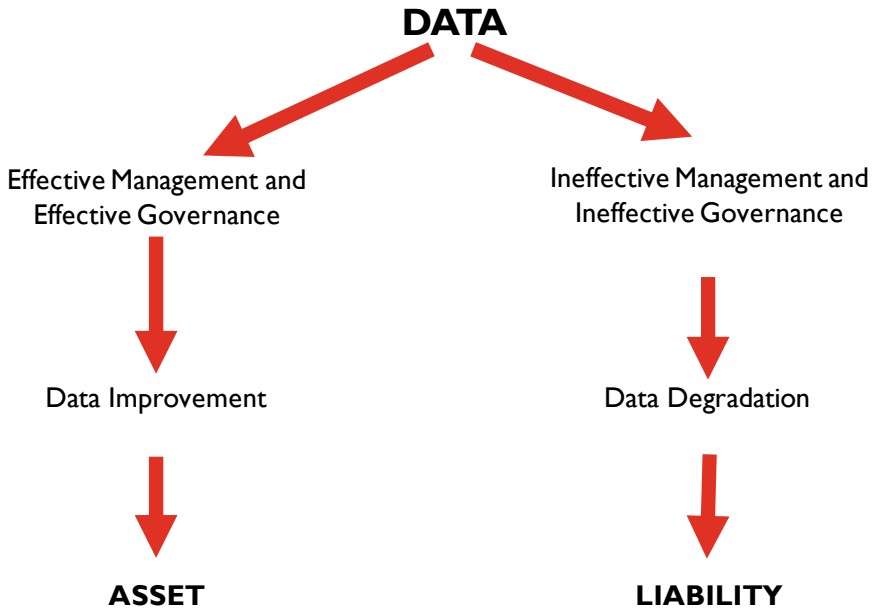

Fig. 2.5 Data asset versus data liability

In order to view data as an asset, organizations need to think about data production in terms of data quality in relation to the business needs that the data fulfils and not let conversations related to data to digress to technology, tools, and IT processes. It is important for organizations to see data as a useful resource rather than a by-product. Tony Epler, Chief Data Strategist, PricewaterhouseCoopers in his interview statement highlights the importance that PricewaterhouseCoopers places on data as an asset and the role data governance plays, when he says (Mahanti 2021):

> …. mature information governance is a three-legged stool – the information has to be managed, it has to be protected, and value must be derived from the information, otherwise it's of no use. Efficiently managing information could enable PwC to better meet legal and regulatory requirements while providing enhanced benefit and value to employees and customers. For PwC to become a more data-driven organization, it was important to develop leading-edge capabilities in business intelligence, data mining, data analytics and knowledge management in order to monetize data, provide more value to customers and successfully grow the business.

2.6 Classification of Data

Organizations need to classify their assets to manage them effectively. The practice of classification enables meaningful grouping of assets. Data classification is the process of organizing data with similar characteristics into different groups for its most effective storage, protection, retrieval, distribution, and usage.

A well-planned data classification system makes essential data easy to manage, locate, and retrieve with the same standards being applied to a data sub-group within a group and the data within the sub-group are also treated similarly.

This can be of particular importance for design, risk management, ensuring data security, privacy, audit, and compliance. For example, "data sensitivity and protection" is a group which has subgroups—restricted data, confidential data, private or internal data, and public data. The same governance processes, policies, and standards would apply to any one subgroup. Data can be classified in different ways as summarized in Fig. 2.6.

2.6.1 Entities

Organizations have data that can be divided into the following categories:

- Master data,
- Transactional data,
- Reference data, and
- Metadata.

2.6.1.1 Master Data

Master data is the consistent and uniform set of identifiers and extended attributes that describe the core entities of the enterprise, and are used across multiple business processes (Gartner 2016). Master data can be grouped by places, parties, and things as shown in Fig. 2.7.

2.6.1.2 Transactional Data

As the name suggests, transactional data is the data that describes the changes resulting from transactions or an event and always has a time dimension associated with it. An event or transaction involves participation of entities like places, things, and parties. Hence, master data participates in transactional data. Common examples of transaction data are:

- Reservations,
- Payments,
- Orders,
- Returns,
- Credits,
- Debits,
- Billing,
- Registrations, and
- Lending.

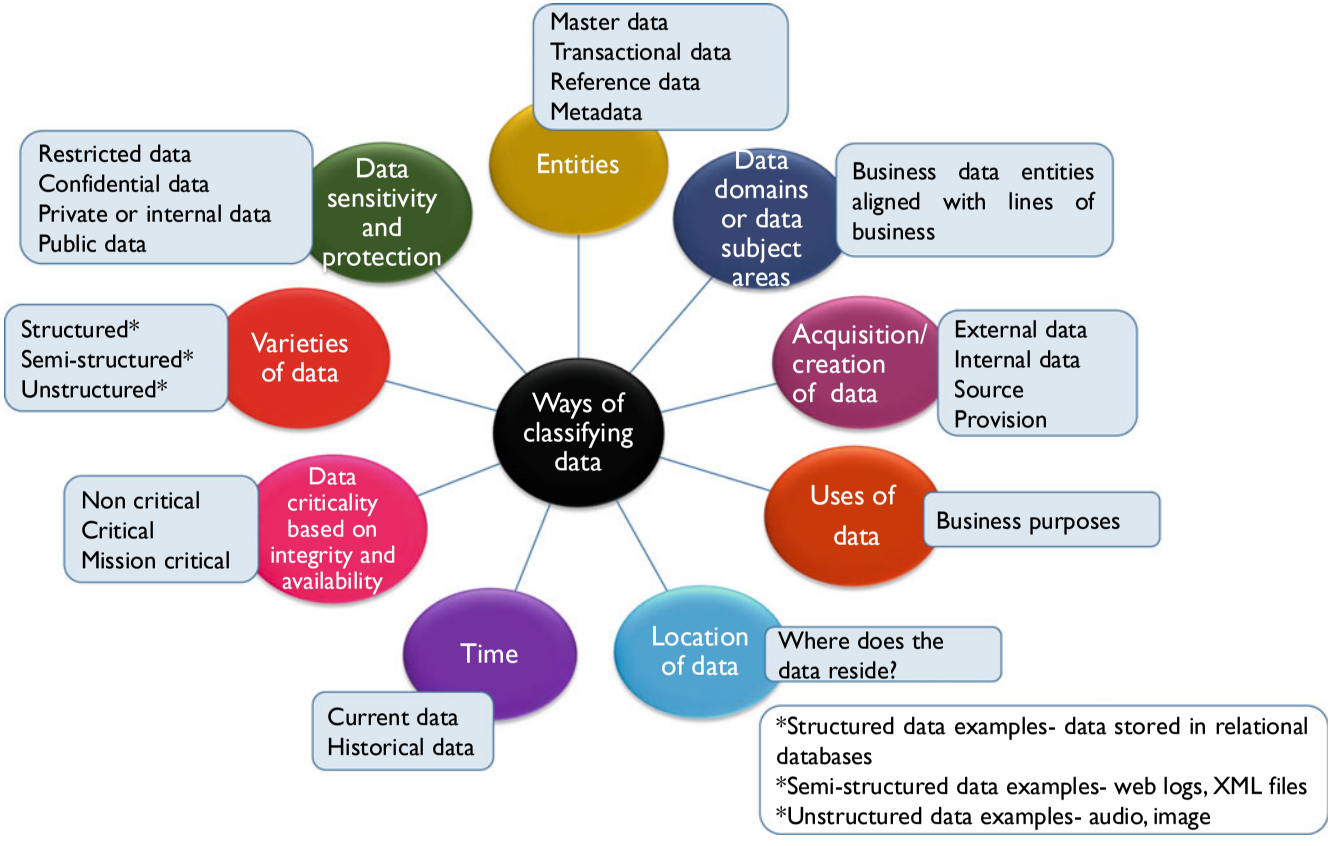

Fig. 2.6 Different ways of classifying data

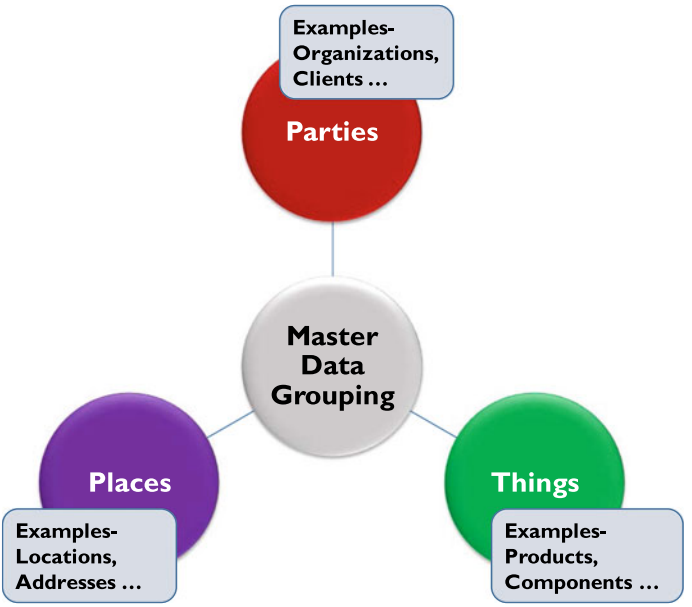

Fig. 2.7 Master data groups

2.6.1.3 Reference Data

Reference data is a known set of permissible values that are referenced and shared by other data like master data, transactional data, and systems with an aim to create a standard vocabulary, structure, and format across different systems. Reference data can be quite diverse. They are relatively stable and change infrequently.

Reference data can be used to categorize data, internal to an organization as well as outside the boundaries of an organization. External reference data are set by bodies outside the organization. Some examples of reference codes internal to the organization are:

- type codes,
- status codes, and
- role codes.

Some examples of reference codes external to organization are:

- currency codes,
- country codes, and
- time zones.

2.6.1.4 Metadata

Meta is a prefix that in most information technology usages means "an underlying definition or description (Rouse)."

Metadata is the data or additional information that describes other data, like master data, transactional data, and reference data. It describes characteristics such as content, format, location, date created/modified, and author/contact information. Metadata helps business users to easily locate, understand, and use the other data. It also enables different users to have a common understanding of the data's meaning and representation.

Figure 2.8 summarizes the different categories of data with the metadata common across the other types of data.

As Jenn Riley observes in her primer, *Understanding Metadata* (NISO 2017), "Metadata are pervasive in information systems, and come in many forms. [...] Metadata is key to the functionality of the systems holding the content, enabling users to find items of interest, record essential information about them, and share that information with others."

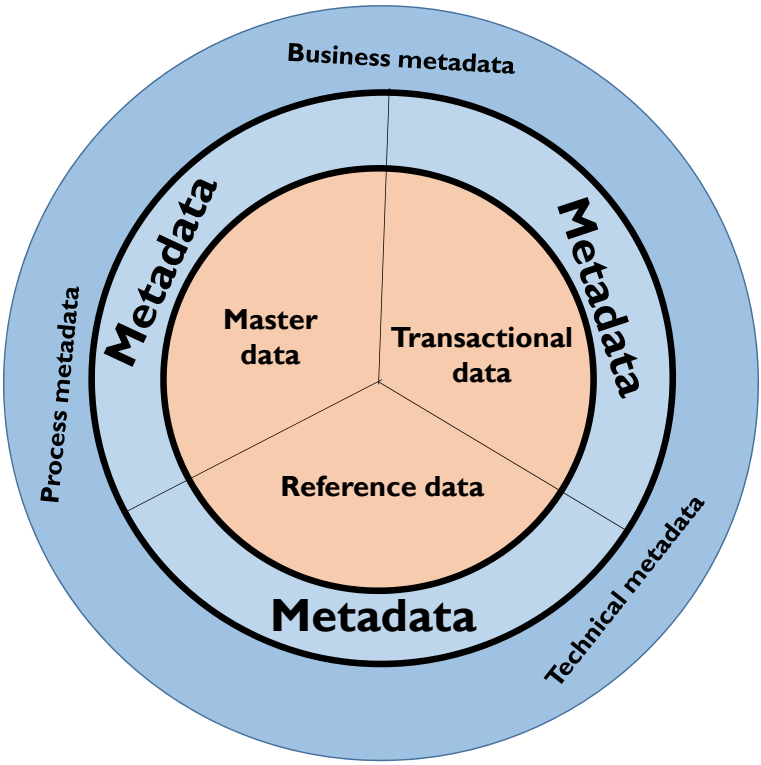

Fig. 2.8 Entity—data categories

Metadata can be classified into three groups:

Technical Metadata captures the technical aspects of the data captured and stored in the organization's data repositories including the names and description of the database tables and field names in the tables, data sizes, type of data (text, numbers, boolean, etc.), and information about content of the fields. They are used by the technical team to process data, develop, and support the applications and systems.

Business Metadata captures what the data means to the end user to make data fields easier to find and understand, for example, business names, descriptions, tags, quality, and masking rules. These tie into the business attributes definition so that everyone is consistently interpreting the same data by a set of rules and concepts that is defined by the business users (Gidley and Castanedo 2017).

Process Metadata is used to describe the results of various IT operations that create and deliver the data. For example, in an ETL process, data from tasks in the run time environment such as, scripts that have to be used to create, update, restore, or otherwise access data, start time, end time, CPU seconds used, disk reads/source table read, disk writes/ target table written, and rows read from the target, rows processed, rows written to the target, are logged on an execution (Mahanti 2019). Process metadata is usually used to troubleshoot technical issues and application failures.

2.6.2 Varieties of Data

Variety refers to the assortment of data types and data sources, and can be grouped into three types:

- Structured data,
- Unstructured data, and
- Semi-structured data.

Traditional data are structured data that can be stored in relational databases. Most traditional data sources fall in the structured data domain (NTNU).

2.6.2.1 Structured Data

Structured data are data that have clearly defined data types, and are structured by predefined data models and schema. Structured data can be stored in relational database tables in rows and columns, are predictable, and are easier to search. Structured Query Language commonly known as SQL are used to search structured data stored in relational databases. Data stored in relational databases are the most common examples of structured data.

2.6.2.2 Unstructured Data

Unstructured data are data that have no identifiable structure and are not structured via pre-defined data models or schema, and hence are not stored in relational database tables. Unstructured data are usually stored in non-relational databases such as NoSQL databases. Unstructured data are complex and are difficult to search and process. Examples of unstructured data are text (documents, social media data, and so on), audio, images, and video streams.

Unstructured data can be human generated (such as word documents, social media data like texts posted on LinkedIn, Facebook, and presentations) and/or machine generated (such as sensor data and satellite data).

2.6.2.3 Semi-structured Data

Semi-structured data lies between structured and non-structured data. Semi-structured data is data that may be erratic or incomplete, have a structure that may change rapidly or unpredictably, and organizational structures that make it easier to analyze. It does not have the same level of structure and predictability as structured data but generally has some structure such as tags to separate data elements and help to identify the data for later retrieval. However, semi-structured data do not conform to a fixed schema. Web logs, XML files, and JSON files are examples of semi-structured data.

2.6.3 Acquisition/Creation of Data

Data can be classified as external or internal data depending on how the data is acquired or created, that is the source/provenance versus the method of acquisition.

If the data is acquired by an organization from external vendors, that is, through third party feeds, then the data is external data. An example of such data is customer credit risk data acquired by financial organizations.

The data that is directly collected or created by an organization is internal data. Examples include patient and physician scheduling data in hospitals, financial product data, customer personal and financial details data captured by banks.

In case of external data, data is acquired from producers external to the organization. Data acquired is stored in an organization's systems, with these systems provisioning the data to downstream systems and consumers. In this case, it is not possible for the organization to control the original quality of the data. However, governance should have appropriate processes, roles and responsibilities to ensure external data is not transformed so that its quality is adversely impacted after ingestion. Also, processes should be there in place to ensure that data is acquired through appropriate secure mechanisms depending on the sensitivity of the data.

Even for data directly collected or created within the organization, while data might be collected by one system, the data might be provisioned by a different system to the wider group of consumers.

2.6.4 Data Domains or Data Subject Areas

A *Data domain* or *data subject area* is used to identify a particular group of data assets aligned with an operational business unit or lines of business such as account transaction (a data subject area within events), and employee (a data subject area within human resources (HR)).

Data subject areas are group classifications of business data entities at their highest level of data object abstraction and are not impacted by project related changes.

The granularity of data subject areas can vary among organizations, industry sectors, and data subject areas can also have sub areas. The number of subject areas would be driven by the depth and complexity of the enterprise. Greater the depth and complexity of your enterprise data, the greater the number of data subject areas that need to be defined (Hodgson 2017).

2.6.5 Time

Some data have a time dimension associated with it. The most common data which has a time dimension associated with them are transactional data describing an event. While the event is still ongoing, the corresponding data is current but once the event is complete, it becomes historical. Master data (like names and addresses) and reference data (like country codes, industry codes, and currency codes) remain current as long as they do not change. Historical data is specifically important for predicting trends.

2.6.6 Uses of Data

Uses of data refers to the specific business purpose or purposes that it is used for. For example: generating sales lead, tracking potential customers, financial reporting, analyzing consumer behaviour, determining marketing trends, and research.

2.6.7 Data Criticality Based on Integrity and Availability

Data criticality is a reflection of how important is the data to an organization's mission, functions, and processes in terms of integrity and availability, and can be divided into three categories (see Fig. 2.9).

Fig. 2.9 Data classification based on data criticality

2.6.7.1 Non-critical

Non-critical data are data whose loss of integrity or availability would result in a small delay or degradation of services and functions. While non-critical data are necessary for an organization to be able to operate efficiently, loss of integrity or availability of non-critical data would only have a minimal short-term impact on business continuity or operational effectiveness.

2.6.7.2 Critical

Critical data are important for an organization to be able to operate efficiently. Loss of integrity or availability of critical data may result in considerable disruption, delay, or degradation of vital services or functions, and would generally have a moderate short-term impact on business continuity or operational effectiveness.

2.6.7.3 Mission Critical

Mission critical data are vital for an organization to be able to operate efficiently. Loss of integrity or availability of mission critical data would result in extreme delay or degradation of vital services or functions to the extent that they may not be able to deliver, and would have significant short-term impact, and possible long-term impact on business continuity or operational effectiveness. The extended loss of mission critical data may threaten the ability of an organization to recover (Secure UD 2018).

2.6.8 Location of Data

Location of the data refers to where the data resides—e.g. data store, file names, table names, and server.

2.6.9 The Sensitivity of the Data and the Level of Protection the Data Requires

Data classified from a perspective of data security, is the classification of data based on its level of sensitivity and the impact to the organization if the data is disclosed, altered, accessed, transported, stolen or destroyed without authorization as shown in Fig. 2.10. Greater the sensitivity of the data, greater is the level of protection, security, and access controls that it requires.

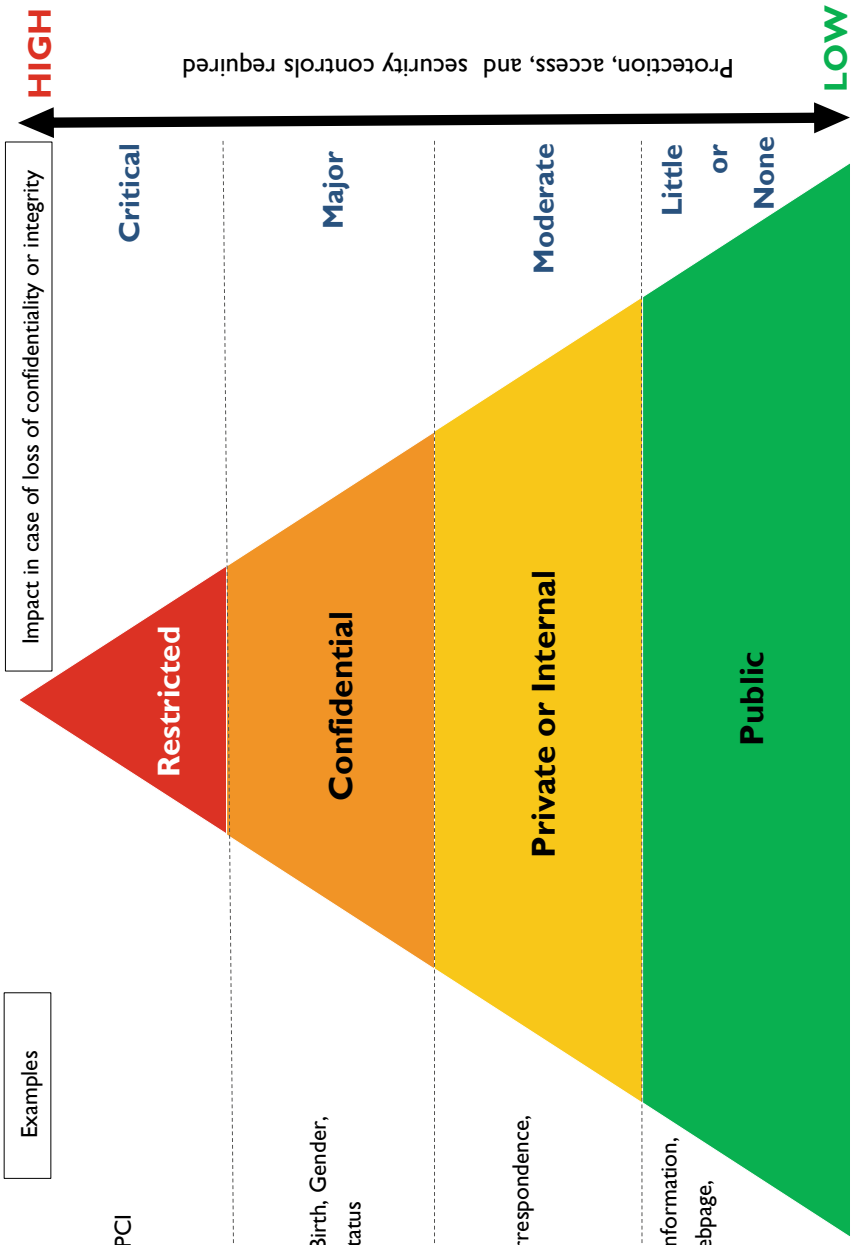

Fig. 2.10 Data classification by data sensitivity and protection (Mahanti 2021)

2.6.9.1 Restricted Data

Data should be classified as restricted when the unauthorized disclosure, alteration or destruction of that data could cause an extremely high level of risk—compliance risk, reputation risks or other risks to the organization.

The loss of confidentiality or integrity of this category of data can lead to harm to an individual including identity theft, bad press coverage/publicity, lawsuits, significant reputational damage, and costs in terms of fines and penalties.

Restricted data is super-sensitive and highly confidential data. It requires the highest level of privacy and security protection for storage, access, and transmission. Hence, the highest level of security, access controls, and protection should be applied to restricted data.

Special authorization should be required for collection, storage, processing, use, and transmission of restricted data. Records containing restricted data elements should be stored securely. Data records containing restricted data are generally not available for public inspection, and can only be accessed by and shared with authorized individuals who have legitimate purpose of accessing that data via secure authentication mechanisms, and secure mode of transport respectively. Any instance of unauthorized access or disclosure of restricted data should be notified.

Data protected by global, local, national privacy regulations, and confidentiality agreements are restricted data. Examples of restricted data are as follows:

- Personally identifiable information (PII),
- Protected Health Information (PHI) protected by Federal HIPAA legislation,
- Electronic protected health information (ePHI) protected by Federal HIPAA legislation,
- Personally Identifiable Education Records,
- Credit card data regulated by the Payment Card Industry (PCI),
- Passport number,
- Username and passwords,
- Social security numbers (SSN) or last four digits of the SSN, and
- Financial information.

Some of the above examples of restricted data will be discussed in detail in Appendix A.

2.6.9.2 Confidential Data

Confidential data is moderately sensitive information that needs to be protected from unauthorized access, and is only intended for limited dissemination.

The difference between restricted data and confidential data is that the likelihood, prospect, duration, and the level of harm is lesser in case of confidential data than restricted data.

The unauthorized access or disclosure of confidential data can have adverse impacts, even though such impacts are less severe than in case of restricted data.

A high-level data protection, access restriction, and security controls are required to prevent unauthorized access; however the rigor is less than that in case of restricted data. Examples include,

- home address and phone,
- birth date,
- gender,
- religious or sexual orientation,
- other non-restricted personal information,
- student records,
- grade evaluations, and
- recommendation letters.

Data classified as confidential could potentially become classified as restricted, if, in aggregation, such data could be reconstructed to reveal personally identifiable information (PII).

2.6.9.3 Private or Internal Data

Data should be classified as *private* or *internal* when the unauthorized disclosure, alteration or destruction of that data could result in a moderate level of risk to the organization. By default, data that is not explicitly classified as restricted, confidential or public data should be treated as private or internal data. As the name suggests, private or internal data is meant for internal use in the organization and is not meant for public release. While confidentiality of the private data is preferred, the data might be subject to open records disclosure. Examples include,

- Email communications,
- Earnings,
- Budget plan, and
- Project related information.

An affordable level of access and security controls should be applied to private data, though the rigor is much less than those in case of confidential data.

2.6.9.4 Public Data

Data should be classified as *public* when the unauthorized disclosure, distribution, modification, usage or destruction of that data would result in negligible or no risk to the organization, and have minimal adverse impacts. Examples of public data include,

- Press releases,
- Public directories,
- Public web pages, and
- Research publications.

Public data is the least sensitive data, has the lowest security requirements, and can be released without restriction. However, there is still a matter of who releases the data. Little or no controls are required to safeguard the privacy of public data. However, control is required to prevent its unauthorized release, modification or destruction. For example, processes, accountabilities (approval authority), responsibilities, and decision rights need to be established related to the release of public data as the release of incorrect data or wrong timing can tarnish an organization's image.

Usually these groups can be placed within a hierarchy to determine which classification should take precedence over others. Higher the classification, greater should be the controls.

The same data can also fall under multiple groups, which adds another layer of complexity as to how the data should be managed (Firican 2018). For example, the personal information like person's name, data of birth, and social security number are restricted data as well as master data and since restricted data has stricter controls than master data, they will need to be managed as restricted data and controls applicable for restricted data will take precedence over those for master data.

2.7 Data Quality and Data Quality Dimensions

Data quality is the fitness of data for use in a given context or a set of specific tasks in hand.

The general definition of data quality is "fitness for use," or more specifically, to what extent some data successfully serve the purposes of the user (Tayi and Ballou 1998; Cappiello et al. 2003; Lederman et al. 2003; Watts et al. 2009).

Despite the fact that, fitness for use or purpose does capture the principle of quality in a holistic sense, it is intangible and not measurable, hence, making it impossible to measure or track the quality of data using the concept of "fitness for use or purpose."

However, there are different aspects or dimensions– both objective and subjective, that can be measured quantitatively and qualitatively respectively, that make data fit for use or purpose. Hence, the data quality is multidimensional and measurable. The different aspects or dimensions of data quality need to be taken into consideration when accounting for the quality of data (Mahanti 2019).

2.7.1 Data Quality Dimensions

The different dimensions used to characterize and measure the quality of data are called data quality dimensions. Each data quality dimension captures a particular measurable aspect of data quality. The data quality dimensions can be used to quantify the levels of data quality of data element/field, data record, datasets, database tables, data stores, and can be used to identify the gaps and opportunities for data quality improvement across different systems within the organization. Data elements can relate to master data, reference data or transactional data (Mahanti 2019).

There is no universal agreement on the number, grouping, and definition of data quality dimensions, with several researchers, practitioners, authors, and organizations approaching data quality dimensions and/or their categorization differently. Some of them are D. P. Ballou, H. L. Pazer, Richard Y. Wang, Diane M. Strong, Thomas C. Redman, Larry English, Yang W. Lee, Leo L. Pipino, David Loshin, DAMA, Dan Myers, Malcolm Chisholm, and Rupa Mahanti (Ballou and Pazer 1985; Redman 1997; English 1999; Pipino et al. 2002; Loshin 2011; DAMA 2013; McGilvray 2008; Myers 2017; Mahanti 2019).

Some of the key data quality dimensions are described as follows:

- **Completeness**—This data quality dimension relates to the completeness of data – data values, data records or entire datasets to meet business needs. Completeness relates to whether data values are present or not. Completeness is important because missing data can have a significant effect on the inferences and decisions that are drawn from the data.
  In a relational database, 'data present' implies non-blank values in a data field in a table or presence of one or more records in the table; 'data absent' implies NULL or spaces values in a data field in the table or absence of one or more records in the table (Mahanti 2019). Sometimes, values such as "UNKNOWN", "Not Applicable" and "N/A" are also used to represent missing data. Below are a few examples of missing data values;
  - The customer last name is missing for 20% of the customers in the customer database of an organization.
  - The state code is missing for 15% of the customers in customer database of an organization.
  - The date of birth is missing for 18% of the patients in the patient database in a hospital.
  - The material data values are missing for 35% of the assets in the underground asset database in a utilities company.
  - The product description is missing for 24% of the products in the product database of a retail store.
- **Accuracy**—This data quality dimension relates to the extent to which data correctly describes the real-world object, entity, situation, phenomena or event, and their attributes. It is a measure of the correctness of the content of the data (which requires an authoritative source of reference to be identified and

available). Data inaccuracy results when the data does not match with reality. Correct information can come from different sources: a database of record, a similar corroborative set of data values from another table, dynamically computed values, or perhaps the result of a manual process (Loshin 2009) such as comparison with physical documents and machine generated data, which serves as the source of truth. Examples of data inaccuracy are as follows:

- The customer names and addresses have spelling mistakes.
- Telephone numbers with incorrect country codes and area codes.
- Incorrect date of birth values due to juxtaposing of month and day values.

- **Conformity/Validity**—This data quality dimension relates to the extent the data complies with a set of internal or external standards or guidelines or standard data definitions including metadata definitions like data type, size, and format. Comparison between the data items and metadata enables measuring the degree of conformity. Most common example of data values complying with formats are date data elements, like date of birth, account opening date, account closing date, issue resolution date, order date, shipping date, and delivery date. Date data values should comply with a specific format. The USA follows the MM/DD/YYYY format where MM stands for month, DD stands for the day of the month, and YYYY stands for the year. Other countries, like India, UK, and Australia follow the DD/MM/YYYY format.
  Other examples of Conformity/Validity are:
  - Gender having values M and F for male and female respectively.
  - Time should be recorded in a specific standard format, say in a 24 h, HH:MM:SS format, where HH stands for hours, MM stands for minutes, and SS stands for seconds. For example: 23:08:56.

- **Uniqueness**—This data quality dimension relates to the fact that there should be no duplicate records captured for the same entity or event in the same data set or table (Mahanti 2019). Uniqueness is also known as non-duplication and is the inverse of the assessment of the data quality dimension, duplication.
  Common examples of duplication are:
  - Same product being stored multiple times in the same table because their descriptions are outlined slightly differently.
  - Same customer being stored multiple times in the same table because of names being represented differently as illustrated in the Table 2.3. “Monalisa Zing”, “Mona Zing”, and “Lisa Zing” are the same customer with one record having the office address and the other two having the same home address and all three records having the same email ids.

- **Consistency**—This data quality dimension relates to the extent to which the data values are identical for all instances of an application. The data across the enterprise should be in sync with each other. The format and presentation of data should be consistent across whole dataset relating to the data entity (Mahanti 2019). If the data values are inconsistent, then at least one of the values is inaccurate.

Table 2.3 Duplicate record for customer—"Monalisa King"

Customer ID	First name	Last name	Street address	Postal code	State code	Country code	Email ID
CID0000278	Monalisa	Zing	10, DF Ave, SOP	2127	NSW	AU	King12@gmail.com
CID0000290	Mona	Zing	12, O'Neill Ave, Newington	2127	NSW	AU	King12@gmail.com
CID0000357	Lisa	Zing	12, O'Neill Ave, Newington	2127	NSW	AU	King12@gmail.com

Common examples of data inconsistency are:

- Date values captured inconsistently across different sets. For example, date of birth is captured in the 'MM/DD/YYYY' format in one dataset, 'DD/MM/YYYY' format in another dataset, and 'YY/MM/DD" format in yet another dataset (see Fig. 2.11).
- Names are captured inconsistently across different datasets. For example, dataset one has the full first name and last name, the second dataset has a shortened first name and full last name while the third dataset has the first name shortened differently and full last name (see Fig. 2.11).
- Gender values are represented differently in different datasets (see Fig. 2.11).

- **Integrity**—This data quality dimension refers to the validity of relationships between data entities or objects, and ensures that there are no missing linkages. Example of data integrity issue is payroll record for a new employee that exists in the employee payroll table but the employee record is missing from the employee master table.

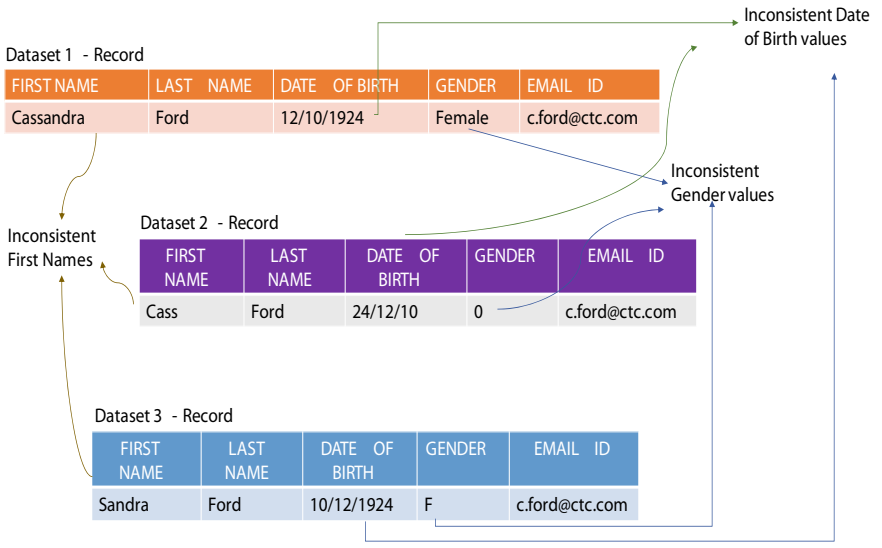

Fig. 2.11 Data inconsistency example

- **Timeliness**—This data quality dimension refers to whether the data is available in a timely manner as per expectations usually defined in service level agreements (SLAs). The timeliness dimension is driven by the fact that it is possible to have current data that do not serve any purpose, because they are late for a specific usage (Batini et al. 2009).
- **Currency**—This data quality dimension refers to the degree to which the data is sufficiently up-to-date for the specific context and usage.
- **Volatility**—This data quality dimension refers to the frequency with which data changes with time. Volatility has an inverse relationship with the currency dimension and vice versa, that is, more volatile the data, it remains current for lesser amount of time. Different data have different levels of volatility, some data like hourly temperatures and stock market data are more volatile than the others like passport expirations, motor vehicle license expiration, while some others like date of birth and place of birth, do not change at all.
- **Data Coverage**—This data quality dimension refers to the extent of the availability and comprehensiveness of the data when compared to total data universe or population of interest (McGilvray 2008). More the data coverage, the greater is the ability of the data to suit multiple applications, business processes, and functions. Reference data and master data which are usually widely shared have high coverage compared to transactional data which record an event and are usually point in time.
- **Data Security**—This data quality dimension refers to the extent to which access to data is managed aptly to prevent illicit access, and the data is protected from unauthorized modifications, deletes, insertions, and corruption throughout the data life cycle. More sensitive the data, greater are the risks and implications from compromise of the data, and greater security measures and controls need to be incorporated and enforced to ensure adequate protection of data.
- **Data Accessibility**—This data quality dimension refers to the ease with which data and/or metadata (data about data) can be located, as well as the appropriateness of the form, medium, environment, and access mechanisms, through which users can quickly and easily access and obtain the data to fulfil their business needs.
- **Reliability**—This data quality dimension refers to the extent data is complete, relevant, accurate, free of duplicates, consistent, and traceable to a trustworthy source.
- **Granularity**—This data quality dimension refers to the extent to which the data in data elements can be sub-divided, so that they can be aggregated or rolled up at different levels. Storing data at the finest grain is becoming increasing important for audit, risk management, and compliance purposes. This is because while it is possible to roll up data at different levels, it is not possible to determine the finer grain of data from the higher aggregated level. For example, account level data for a customer is at finer grain than the total customer balance. All the account balances can be aggregated to arrive at a customer's total balance; however, it is not possible to determine each account balance for a customer, given the total customer balance.

- **Traceability**—The extent to which data can be traced to its source system and information related to its origin, history, first inserted date and time, updated date, and time are recorded. Traceability is extremely important from a reliability perspective, and is required for audit and compliance purposes.
- **Credibility**—This data quality dimension refers to the extent to which the good faith of a provider of data or source of data can be relied upon to ensure that the data actually represents what the data is supposed to represent, and that there is no intent to misrepresent what the data is supposed to represent (Chisholm 2014).
- **Trustworthiness**—This data quality dimension refers to the extent to which the data is sourced from data sources that are dependable. Several factors such as ability to trace data back to authoritative sources, number of data issues reported, number of requests to use the data, and availability of data quality statistics have an impact on trustworthiness.
- **Reputation**—This data quality dimension refers to the extent the data is trusted and highly regarded in terms of their source and content. Reputation of data is built over time, and both data and data sources can build reputation (Wang and Strong 1996).

Data quality dimensions can be divided by the way they can be measured (Mahanti 2019):

1. **Objective or Quantitative**—for which a concrete value can be calculated; completeness, accuracy, conformity, consistency, integrity, timeliness, currency, and volatility are examples of objective data quality dimensions.
2. **Subjective or Qualitative**—which cannot be quantified but depend on the users' perspective of the respective dimension and measured indirectly via surveys; credibility, trustworthiness, and reputation are examples of subjective data quality dimensions.

2.8 Need for Good Data Governance

Any organization that captures, stores, and manages data needs to have good data governance. This is because strategic enterprise assets need to be governed, and data is a strategic enterprise asset.

Data governance empowers and facilitates good behavior with respect to data, and restricts behavior that creates risks with regard to data. Data is your biggest strength and power if it is of good quality. However, data requires love and care, and will stay healthy, if governed and managed properly, but will become your greatest weakness in the absence of adequate management and governance.

With digitization, large volumes of data are being created, captured, managed, and stored in large number of digital repositories in organizations, and being used by almost everyone in the organization. Data that is captured, flows through

multiple applications, systems, departments, and processes in large organizations, and there is likely to be multiple versions of truth.

Organizations are likely to have different owners for different systems, and in the absence of data governance, the owners or their delegates make data related decisions with respect to their systems in isolation, resulting in inconsistencies in the data capture, data format, data definitions, availability, usability, data consistency and integrity issues. For example, consider a utilities organization that captures its asset information (e.g. boilers, pipes) in two different systems. Each of these systems are managed by different departments, and there is no overarching governance. Since there is no overarching governance, there are no uniform standards for storing the data resulting in the data being stored in different formats in the two systems causing inconsistencies.

Some companies have reported up to 30% of their customer data being duplicated and replicated across their enterprise. This business problem occurs whenever critical data (for example, customer and product) standards, and governance practices are not consistently and universally applied across all points of data acquisition, creation, and existence (Duemig 2011).

In the absence of effective enterprise wide data governance, there is inconsistent application of data validation rules, and business rules used to transform the data are also inconsistent across systems, resulting in data quality issues.

Consistent production of bad quality data results in data consumers to doubt the credibility of the data source and lose trust in the data, culminating in users not using the system anymore. Therefore, ineffective data governance paves a pathway for poor data quality.

Lack of effective governance and proliferation of bad data can increase the risk quotient of your organization. The sensitive data might be exposed leading to thefts, data breaches, data corruption, legal, and compliance issues. Also, in the absence of data governance, presence of data, and its percolation to the other parts of the organization is not properly managed resulting in a scenario where business users need the information, but are not aware whether it is available or where is it available, and who in the organization to approach for it. This can also lead to compliance issues. For example, with GDPR, which came into effect in early 2018, EU residents have the right to request that all of their data be deleted from an organization database. With data not being governed at all or not being governed properly, there would be inconsistent data for the same person sitting in multiple systems that an organization is not aware of, making it extremely difficult, if not impossible to guarantee that all data relating to that particular individual is deleted on request. This kind of situation exposes your organization to huge risk and possibly hefty fines. Ineffective or lack of data governance can have significantly destructive effects on success, for example, losing control of the use of data, lawsuit by disgruntled users, low quality data (Hagiu 2014), bad company image and reputation, and privacy and security issues. For example, the article titled, "7 Controversial Ways Facebook Has Used Your Data" (Luckerson 2014) which disclosed surprising things Facebook can do with user data: e.g. tracking user movements or using user data in ads without consent (Lee et al. 2017). According to this article, Facebook paid more than $20

million for lawsuit settlement by disgruntled users (Lee et al. 2017). Figure 2.12 summarizes the issues that rise from ineffective data governance or lack of data governance that equates to anarchy at the worst.

In the absence of governance (anarchy) or with ineffective data governance, people usually do not know where the data resides in the organization, and who to contact in relation to the data in question as there is no defined ownership from business for the data within the organization. It is difficult to obtain consensus on data related issues as data users are spread across the organization, spanning different departments, and in the absence of a data governance framework there is lack of clear data ownership, absence of or lack of clear definition of roles and responsibilities in relation to data, and no conflict resolution process with respect to data issues. Even if a conflict resolution process exists, there is no governance structure to enforce the process of conflict resolution.

In the absence of governing principles, processes and practices around the data, chaos would reign. Hence, data governance is not an option but a necessity. When asked about the industries that had been lax in the implementation of data governance, this is what Phil Watt, Director, Elait Australia and John Zachman, Author

Fig. 2.12 Issues arising from ineffective data governance or absence of data governance

of "The Framework for Enterprise Architecture" (The "Zachman Framework"), Zachman International had to say (Mahanti 2021).

Phil Watt—"The public sector has relied on manual processes for data governance, resulting in large gaps in knowledge of systems and where the data is being used. They appear to have adopted data governance recommendations some years ago but due to financial constraints (budget and long buying cycles), they do not have the infrastructure in place to deal with data governance today. This is being slowly acknowledged but remediation programs are taking some time to be rolled out."

John Zachman—"….there are likely patterns that would be helpful in identifying lax industries and rationalizing their laxity state. I will suggest a metaphor for illustration. Back in the '70's timeframe, the oil industry was thriving such that they could "spill more oil than they sold and still make a profit." Any industry where they are making so much money or getting so many tax dollars (which would include Public Sector Departments) that they did not need efficiency, has no incentive to invest in infrastructure, that is, anything that is "indirect" in an accounting sense or not required in a regulatory sense."

The rigor for data governance as in terms of the time, effort, and cost spent for data governance, would depend on the risks associated with the data, that the organization manages.

If the current management structures in place satisfactorily govern the data without having a formal data governance program or governing body, then you may not need to establish formal data governance in the organization. An informal system of data governance may work for small businesses which do not have complex data landscape, have no overlapping data, and are not driven by compliance or regulatory requirements that need formal governance around data. However, highly regulated industries, like financial institutions and healthcare industries, need to have more extensive and rigorous data governance functions in place to prevent financial fraud and adequately protect sensitive information. When asked about the industry sectors that have embraced data governance the most, it was no surprise when financial services (Phil Watt) and healthcare (John Zachman) were answers (Mahanti 2021). In his interview, Phil Watt, Director, Elait Australia, states that (Mahanti 2021):

> Financial services companies have been investing in this space fairly heavily over the last 5 + years, seeking both technology and process solutions.

John Zachman, Author of "The Framework for Enterprise Architecture" (The "Zachman Framework"), Zachman International states (Mahanti 2021),

> Any industry (or Public Sector Department), that is struggling or under heavy public scrutiny is a candidate for investing in infrastructure, particularly infrastructure that enhances data quality. Maybe, health care is a reasonable current example.

2.9 Informal Versus Formal Data Governance

Generally, any organization that captures, manages, and uses data has some form of data governance in place. While all organizations have some form of governance for individual applications or business units, the data practices lack sufficient breadth, depth, and alignment of a formal governance program, instead allowing individuals, departments or business units to make their own rules and standards (Tucci 2010).

Informal data governance is characterized by loosely or poorly established data policies and practices, lack of written procedures, and/or failure to clearly assign data responsibilities. The informal structures lack established guidelines and authority for conducting data governance. The associated risks are higher, data problems occur more frequently and are not always addressed effectively, offering no assurance of data integrity, quality, security, or confidentiality of personally identified information.

Conversely, formal data governance structures are characterized by established policies, guidelines, practices, and defined data roles for specific staff. Formal data governance does the following (Mauzy et al. 2016):

- establishes policies and procedures that are reviewed and adjusted periodically to ensure high-quality and trustworthy data;
- creates checks and balances to minimize risks, address risks if they emerge, and maintain data integrity;
- provides direction and increases transparency of data to those who handle it (that is, collect, enter, process, transfer, and report), and to those who use the data for continuous improvement;
- maintains data confidentiality and contributes to confidence in the data with internal staff members, clients, and those externally connected to the organization; and
- outlines processes and responsibilities to ensure that all data issues are addressed.

Table 2.4 outlines the differences between informal and formal data governance.

Organizations need to move from informal governance to formal data governance when one of the below four situations occur:

1. The organization becomes so large that the traditional management is not able to address data-related cross-functional activities (TDGI).
2. Organizations acquire other companies through mergers and acquisitions. Prior to the merger or acquisition, the companies were separate entities in their own right, and each of the companies had their own unique data landscape. Post-merger or acquisition, these data landscapes need careful governance to ensure that the data mergers also happen smoothly, and roles and responsibilities are clearly defined for post data merger activities too.

Table 2.4 Difference between informal and formal data governance (Adapted from Mauzy et al. 2016)

Informal data governance	Formal data governance
Data policies and data standards are developed or changed in a haphazard fashion or there is a development and change process in place that is not widely enforced or followed	There is a formal process for development and change of data policies and data standards, that is enforced and followed across the organization
Data collection schedule, data definitions, and/or data processes change in a haphazard fashion with minimal regard to alignment with other agency data efforts. These changes make consistent reporting and data use difficult	Data collection schedule, data definitions, and/or data processes change only after considering impact on internal and external constituents, only in ways consistent with the purpose and vision of the broader data system
Data roles and responsibilities of internal and external agency staff are fluid, frequently changed, and/or known to only a select few staff	Data roles and responsibilities are clearly defined, written in policy, and adherence is monitored. New staff are trained on data policies and their particular role
Unregulated child-level data sharing is allowed among internal and external agency staff and across external agencies	Regulated child-level data sharing is allowed among internal and external agency staff and across external agencies, with data classifications and data access processes and approval mechanisms in place to ensure that only authorized users are allowed to access data
Stakeholders have little or no opportunity to provide input on data systems and data issues	Active stakeholder input
Limited or less rigor around data security and data accessibility	Data are accessible only by staff with assigned permissions based on position and training, and sensitive data are encrypted
Lack of effective controls to monitor risks or ensure high quality data	Effective controls in place to flag and mitigate risks, and ensure high quality data
Unwanted data tends to accumulate	Ensure processes are in place for regular and consistent disposal of unwanted data
Data issue resolution process is adhoc; decision rights, roles and responsibilities are not defined or not well defined	Data issue resolution processes are defined and adhered to, and roles and responsibilities are in place to prioritize, investigate, and resolve data issues in a timely manner

3. The organization's data architects, SOA teams, risk management, legal, audit or other horizontally-focused groups need the support of a cross-functional program that takes an enterprise (rather than siloed and departmental) view of data issues (TDGI).
4. There are several data intensive initiatives being implemented in the organization, and a formal framework is needed to ensure that these initiatives work with

each other without inhibiting one another, and the use of data fits the requirements of all these initiatives and future initiatives too (TDGI).
5. Regulatory, compliance, legal or contractual requirements call for formal data governance.

In this book, the term "data governance" will be used interchangeably with the term "formal data governance." Also, the terms "data governance" and "information governance" are used interchangeably.

2.9.1 *Warning Signs that Indicate, You Need Formal Data Governance*

Below are some of the indicators that tell you that your organization needs formal data governance:

- You are making bad decisions because the data quality is poor.
- You do not know where you can find a certain piece of data in the organization.
- You do not know whom to contact regarding a piece of data.
- You do not know whether the information that you are looking for is captured or stored in the organization.
- You do not know the pockets in the organization, where the sensitive data is stored.
- You do not know the purpose behind the access and collection of specific data.
- You do not know who are using specific data, and what they are doing with it, or for what purpose they are using the data.
- You find it extremely difficult to get access to critical data.
- Interdepartmental finger pointing is common, when you have problems with data.
- You are struggling to obtain stakeholder consensus on data related issues, where the related data spans across different departments.
- You experience difficulty in understanding and using the data, as metadata and supplementary information is not available.
- While there is a conflict resolution process, it is not enforced.
- You miss opportunities because it takes a considerable amount of time to get the information you need (Pixentia 2016).
- Your competitors are beating you because you can't react fast enough (Pixentia 2016).
- You have no or little confidence in your data and cannot trust your data.
- You are spending more time fixing the data than using it to derive actionable insights.
- You are spending more money and time in managing data than you can afford.
- You are encountering difficulties and spending a lot of effort finalizing and delivering financial and regulatory reports.
- There are inadequate security and access controls with respect to sensitive data that puts your organization at risk.

- While there are a number of reports, there is a never-ending backlog for request for new reports very similar to the ones that already exist.
- You are encountering delays in implementation of new capabilities and tools.

The 2018 State of Data Governance Report indicates that 98% of organizations consider data governance important. Furthermore, 66% of respondents say that understanding and governing enterprise assets has become more or very important for their executives (Erwin and UBM 2018).

Not all data require the same level or rigor of governance as not all data are created equal, and do not have the same business value. Some data need to be more tightly governed as opposed to others, that can be lightly governed.

As a general rule, the more the data is shared across and beyond the organization, the more formal, governance needs to be (Turner 2018).

The data value chain on a broad level has data producers, data publishers, and data consumers. Data producers are man or machine generating the data. Data publishers are individuals that capture and provision data. In situations where data producers capture and provision data too, the data producers play the role of a data publisher too. Data consumers are individuals/groups who use or consume the data.

If there is just one data consumer, then there can be one to one relationship between the data publisher and data consumer, and with data sharing only restricted to one consumer, data governance and data management practices do not need to be as rigorous as when one or more of the following scenarios are in play (Fig. 2.13):

- Data is consolidated from multiple data sources maintained by different data producers/publishers,
- Data has multiple consumers, and
- Data traverses through multiple intermediate systems before reaching the data consumers.

This is because more the number of systems and people involved, greater are chances of inconsistencies, security breaches, and greater is the need for standardization, protection, and agreement between all the different stakeholders relating to data capture, usage, retention, and quality. Hence, more rigorous governance practices with well-defined policies, processes, controls, standards, and metrics need to be in place.

Master data and reference data are widely shared and used or manipulated by multiple business units. They need to be more heavily governed while data that is shared sparingly, or used for one–off analysis need to be relatively lightly governed (Turner 2018). Also restricted data and confidential data need to be governed with more rigor than private or public data. Public data is also widely shared but since public data is not sensitive, it is lightly governed. Also, data needed for regulatory purposes need to be tightly governed. Transactional data are less rigorously governed but if required for compliance purposes, will need to be more rigorously governed.

Metadata also needs to be governed, and the level of governance would be determined by the governance of the data it is associated with. For example, if the data is heavily governed, the associated metadata also needs to be heavily governed.

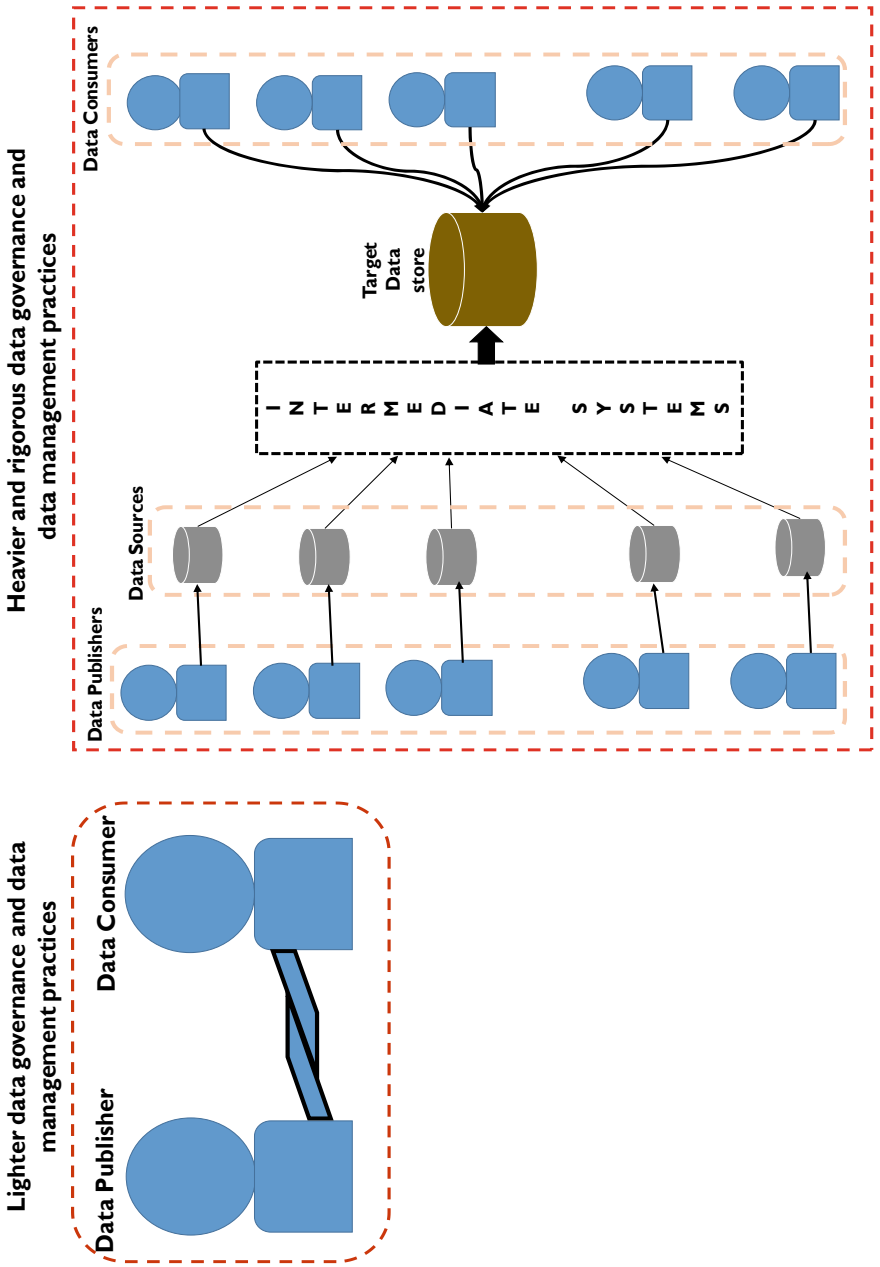

Fig. 2.13 Light versus heavy governance

2.10 Data Governance is not the Same as Data Management or Data Quality

There is a lot of confusion regarding the terms data governance, data quality, and data management. Data governance is used interchangeably with the terms data quality or data management, owing to the fact that they are closely related.

2.10.1 Data Governance and Data Management

While data management is the action of actually managing the data, and involves data related activities such as data architecture, data modeling, data storage, and data quality management, data governance is concerned with establishing and managing the authorities and accountabilities, rules, processes, practices for managing data, and guiding the data related activities and execution of policies and practices set forth in the data governance framework.

Data management takes the decisions, policies, and rules made by data governance and puts these decisions, policies, and rules into action by implementing them within the processes, architecture, and technology, so that the data governance objectives are accomplished. In other words, data governance provides the base to develop appropriate data management processes and procedures. While the driver of data management is to ensure an organization gets value out of its data, effective data management (which includes data quality management, data security and data privacy) is one of the drivers of data governance.

Data governance is a strategic business initiative that determines and prioritizes the financial benefit the data brings to organizations as well as mitigates the business risk of poor data practices and quality (Goetz 2015). Data management is usually an IT initiative, which is defined by a set of technologies that executes on business defined rules and standards to ensure data supports the information requirements of various stakeholders like customers, employees, suppliers, vendors, partners, and shareholders.

Data governance is not defined or driven by technology, though technology supports the program objectives through augmentation, automation, and scale which cannot achieved via manual effort only. Technology is an enabler and facilitator of data governance rather than the driver.

Data management is the action of managing the data. Data management and data governance though closely related are different and should not be used interchangeably.

Data governance and data management can be compared to the horse-cart scenario where data governance is the horse and data management is the cart as data governance guides the data related activities like the horse guides the cart.

Once the governance direction and structure is put into place, we are then expected to manage our data activities accordingly. The Data Management Capability Assessment Model (DCAM) reinforces this concept with its opening definition statement in the data management program component; "Data governance is the backbone of a successful data management program" (Gorball 2016).

Data management can be considered as an overarching umbrella which encompasses the disciplines or functions included in the data governance framework (Norris-Montanari 2016). DAMA has identified 11 major functions of data management which it calls as knowledge areas, as given below (DAMA International 2014):

- Data Governance—exercise and enforcement of rules, processes, policies, practices, standards, controls, decision rights, and people accountabilities to manage data as a strategic enterprise asset.
- Data Architecture Management—technology and infrastructure design with respect to data collection, storage, transformation, distribution, and consumption as an essential part of the enterprise architecture.
- Data Modeling and Design—defining, analyzing, and scoping data requirements to support business processes and then designing a data model (which may include a conceptual, logical, and physical model) to represent and communicate these data requirements.
- Data Quality Management—defining, measuring, analyzing, improving, and sustaining data quality.
- Data Security Management—classification, privacy, confidentiality, appropriate access, authentication, and auditing of data.
- Data Warehousing and Business Intelligence (BI) Management—collection, integration, and presentation of data to knowledge workers for the purpose of business analysis and decision-making.
- Data Integration and Interoperability (DII)—relates to the acquisition/capture, movement, consolidation, and delivery of data within and between data stores, IT systems/applications, and enterprises.
- Document and Content Management (DCM)—management of the capture, storage, protection, access, and version control of unstructured data found outside relational databases and data lakes, and usually stored in digital files in an organization.
- Metadata Management—capturing, categorizing, storing, maintaining, integrating, controlling, managing, and publishing metadata.
- Reference and Master Data Management—managing shared data to reduce redundancy and ensure better data quality through standardized definition and use of data values (DAMA International 2014).
- Data Storage and Operations—storage deployment and management of structured physical data assets including backup, archival, purging, and performance management.

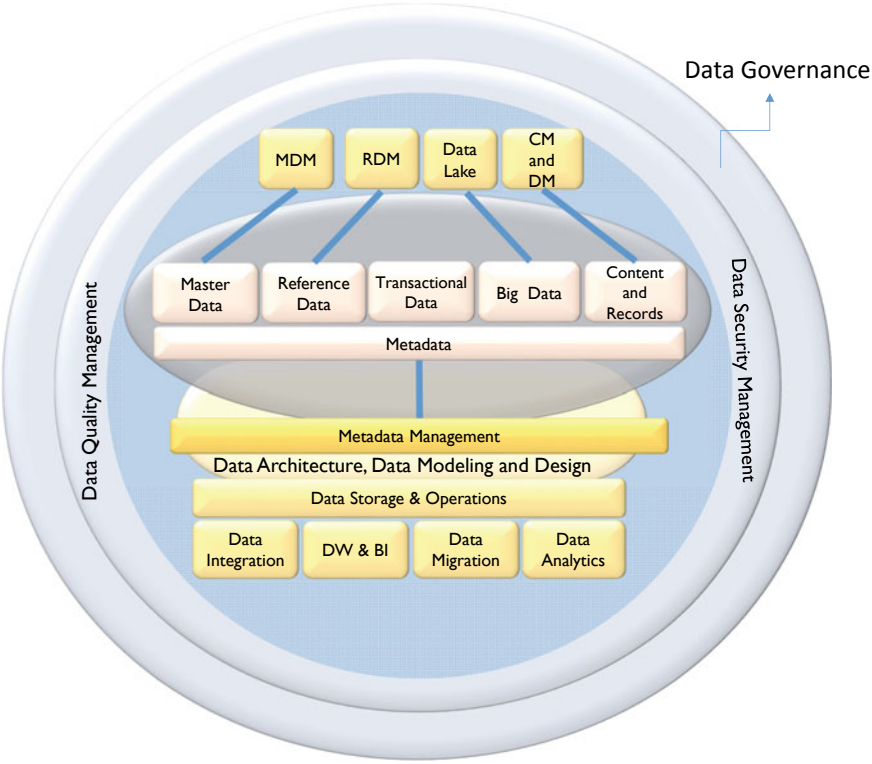

Fig. 2.14 Data governance tying together the data management functions

Data governance is identified as one of the core components of data management tying together the other data management functions as summarized in the Fig. 2.14. We will discuss the alignment of data governance with each of these data management functions in detail in the Chapter—Data Governance and Data Management Functions and Initiatives.

The data governance function guides all the other data management functions by defining, reviewing, communicating, enforcing policies, processes, rules, and standards in each of the other data management functions, identifying, managing, prioritizing, escalating, resolving issues, and overseeing data management projects.

2.10.2 Data Governance and Data Quality

While both data governance and data quality endeavor to maximize the value of data to meet business expectations and requirements, and involve people, processes, and technology to achieve success, the terminologies—data quality and data

governance are not synonymous but are often used interchangeably under the misconception that they are the same.

The underlying reason for this confusion lies in the fact that they are very closely related and interdependent data management disciplines and doing one without the other is of little worth. Data quality and data governance are complementary functions that have primarily different responsibilities. Data governance has to do with creating the framework, policies, processes, and rules by which organizations will collect, maintain, access, and use data. The purpose of data quality is to ensure that the data owned by the organization is of high quality.

Data quality and data governance have a symbiotic relationship, with data quality being a crucial driver and enabler for data governance and data governance being a critical factor in the success of data quality initiatives. Data governance asks questions like:

"What should we do?",
"What should we not do?", and
"Who is accountable?".

Data governance establishes the dos, don'ts, roles, responsibilities, and authorities around data. Data quality answers, "How will we do?" and deals with the nuts and bolts to achieve high quality while adhering to the do's and don'ts established by data governance.

Data governance usually provides context, standpoint, and assigns priority to data quality issues based on the business need and criticality.

From a business perspective, data quality is all about whether the data meet the needs of the information consumer (Scarisbrick-Hauser and Rouse 2007).

Data quality can be defined as the degree to which data meets business stakeholders' requirements and are fit for use for the context along line of the various characteristics like accuracy, completeness, timeliness, currency, validity, non-duplication, consistency, and integrity. These characteristics are called data quality dimensions and we have discussed them in brief in the section—*Data Quality Dimensions* in this chapter. Data quality management is a discipline to achieve high quality data and there are different data quality solutions to monitor, fix, and sustain data quality.

Data governance is the exercise and enforcement of policies, processes, guidelines, rules, standards, metrics, controls, decision rights, roles, responsibilities, and accountabilities to manage data as a strategic enterprise asset. Data governance provides guidelines, standards, and rules regarding data quality and data quality management. Without data governance, it is extremely difficult to achieve and sustain data quality throughout the enterprise. This is because a lack of data governance is characterized by the absence of a governing framework which is rigorous in its definition, enforcement of data standards and policies, clear accountabilities, and responsibilities around data.

While data governance deals with the definition and responsibilities of data policies, processes, and standards; data quality deals with the actual implementation and monitoring of these policies, processes, and standards.

I will use the example of constructing and maintaining a building to illustrate the differences and relationships between data governance, data quality and data management.

If building components (like bricks, nails, boards, windows, and doors) were data, then the overall building construction and maintenance is data management. The quality of these components and the building is data quality; any issues with the building construction like leakages, cracks, and material defects are data quality issues; the management of the quality of these components as well as the building is data quality management; and the storage of the raw materials is data storage.

The building architecture patterns, plans, overall structure, and the way the different components are linked together is the IT, data architecture and data modeling; the strength of the building and the doors, windows, locking, and bolting system to prevent thieves from entering reflects the data security management aspect.

The governance aspects are different roles, responsibilities, accountabilities, processes, standards, policies, rules, controls, and decision rights involved in the building construction and maintenance. For example, the architect would be designing the model, a site manger would be overseeing the project, and laborers would be responsible for the construction of the building.

The building development would need to follow standard processes and policies. Certain rules need to be followed when constructing the building—for example, the foundation should have certain depth, and each floor should be of a certain height. Also, governance would ensure that the different functions involved in building construction work in sync with each other.

2.11 Data Governance Goals

Some universal goals of data governance programs are as below (drawn and adapted from The Data Governance Institute):

- Facilitate better decision-making,
- Reduce operational friction and increase operational effectiveness,
- Protect needs of business and data stakeholders,
- Promote and adopt a common approach to handling data issues,
- Ensure creation, use of standard, and repeatable processes,
- Ensure transparency of your approach and process, and
- Reduce cost, increase productivity, and effectiveness through coordination of efforts.

In addition to the above goals, depending on the focus of the data governance program you will have other goals too. Some of these goals address general

infrastructure and culture of the organization, such as identifying data stakeholders and how they will fit in relation to solving data-related issues (The Data Governance Institute).

2.12 Data Governance—The Key Elements

Data governance is a function that provides policies, processes, standards, rules, roles, responsibilities, accountabilities, decision rights, and ensures that appropriate controls and metrics are there in place to oversee the effective capture and management of data across the enterprise, and encourage appropriate behaviors in the usage of data. Data governance execution has several interacting components which can be grouped under three broad areas:

- People,
- Processes, and
- Tools and technology.

2.12.1 People

Data governance is an enterprise wide program and involves a large number of stakeholders. It is an organized structure consisting of decision-making bodies and individuals, with clearly defined accountabilities, roles, and responsibilities in association with data.

The organizational structure and operating model for data governance depends on a number of factors such as the organization size, culture, existing hierarchies, organizational maturity, business drivers of data governance, whether operations are distributed or centralized, and hence would differ from one organization to another.

Typically, data governance implementation would involve a multitude of roles at different levels from both IT and business organization and a cross-functional group collaborating to create formal, standard, consistent policies, processes, rules, and standards across the organization.

Some of the common stakeholders are data producers, data publishers, data consumers, data owners, data stewards, and data custodians as shown in Fig. 2.15. The common groups/bodies involved in data governance at different levels are Executive Steering Committee, Data Governance Council, Data Stewardship Council, Data Governance Office (DGO), and Information Technology Partners as shown in Fig. 2.15.

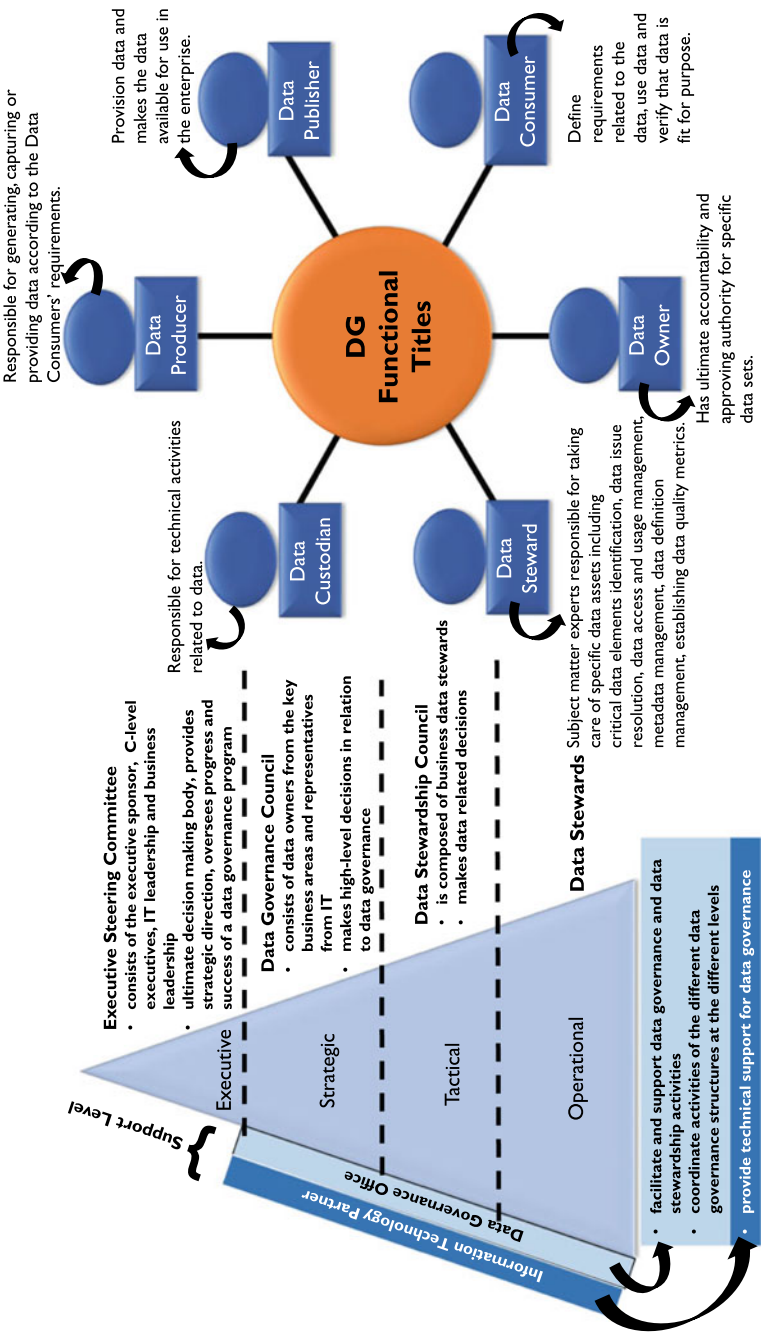

Fig. 2.15 People aspect of data governance

2.12.2 Processes

Data governance process component includes the following:

- Principles—high level statements of what an organization wants to achieve by implementing data governance.
- Policies—collection of statements of an organization's intent or rules controlling the several data related activities and operations - creation, capture and/or acquisition, retrieval/requests, transformation, management, quality, protection, sharing, and usage of data throughout its lifecycle.
- Guidelines—recommendations and best practices designed to achieve the policy's objectives by providing a background to design standards and implement processes.
- Processes—methods or set of steps to support the various activities needed to govern data as well as implement the data policies in the organization. While policies define what to do, processes establish how to do.
- Rules and Standards—encompass business rules, data quality rules, and data standards to ensure consistent results from people and processes using them.

2.12.3 Tools and Technology

Tools and technology facilitate and enable data governance through automation, scaling, augmentation, and auditability. A data governance program involves several components, activities, and aspects. Creating and managing all these components and aspects manually requires considerable time and effort, which gets unwieldy as the program evolves.

Many organizations start with using Microsoft Excel, Microsoft Word, Wikis, SharePoint lists, and existing document and content management (DCM) systems to manage the data governance activities and artifacts.

However, owing to the complexity of the organization's data landscape, multiple players at different levels, and multiple departments across the organization, as the data governance program grows, the set of standards, policies, procedures, definitions, and workflows become unwieldy to the extent that they cannot be easily managed in spreadsheets and documents. This is where tools and technology can help by enabling automation of some of these elements and improving efficiency.

2.13 Key Data Governance Business Drivers and Uses Cases

In response to a question regarding the biggest business driver of data governance, Andres Perez, Data Architect, IRM Consulting, Ltd. Co., made the following statements (Mahanti 2021):

Fig. 2.16 Data governance business drivers and use cases

> Organizations are using more data and increasingly sharing data across lines of businesses that traditionally were kept separate and, in doing so, realizing that the data definition and content is not meeting their business needs. Also, organizations are beginning to understand that data is a valuable resource that must be managed well to get better outcomes.

As the applications for data have grown, so too have the data governance business drivers. The business drivers of data governance are summarized in Fig. 2.16.

2.13.1 Compliance

Compliance and changing regulatory requirements are one of the biggest drivers of data governance. Compliance generally refers to actions that ensure behavior that complies with established rules as well as the provision of tools to verify that compliance. It encompasses compliance with laws, regulations, contractual requirements as well as the enterprise's own policies, which in turn can be based on best practices (Salido and Voon 2010).

In many industry sectors, the risks and impacts of non-compliance can be severe, with hard costs taking the shape of hefty fines, penalties, legal costs, and soft costs that can take the form of reputation damage and brand image that can be difficult to repair.

As per "*The 2018 State of Data Governance Report*" conducted by erwin, 60% of the respondents indicated the need to comply with regulatory mandates as the topmost driver for data governance (Erwin and UBM 2018).

Organizations in different industry sectors are required to comply with a growing number of external regulations (national and international), as well as with internal corporate governance policies designed to increase transparency and prevent corporate fraud.

Some of these regulations are:

- Health Insurance Portability and Accountability Act (HIPAA),
- Basel Committee on Banking Supervision's standard number 239 (BCBS 239),
- Sarbanes–Oxley Act (SOX),
- Dodd Frank Act,
- Foreign Account Tax Compliance Act (FATCA),
- Anti-Money laundering (AML),
- Know Your Customer (KYC),
- Common Reporting Standard (CoREP) and Financial Reporting Standard (FinREP),
- Payment Card Industry Data Security Standard (PCI-DSS), and
- European Union's General Data Protection Regulation (GDPR).

The European Union's General Data Protection Regulation (GDPR) came into effect in May 2018, requiring all organizations handling personal data belonging to EU residents to implement specific privacy and security controls or face stringent penalties. Data governance is mandatory under the new law, and failure to comply will leave organizations liable for huge fines—up to €20 million or 4% of the company's global annual turnover (Erwin and UBM 2018).

According to the case study—"Data Governance and Compliance for Financial Services" conducted by IBM, "85% to 95% of all regulatory evidence is now electronically stored information" (Gow 2006). 82% of companies face external regulatory requirements on electronic data (Rand 2013; Oracle 2014).

It is necessary to streamline the collection of reporting data to meet regulatory reporting requirements. Many regulations require documentation of the sources of

the data being reported and certification of their accuracy, and implementation of specific governance policies. Regulators demand the ability to actively monitor, control, store, search, retrieve, and analyze critical information. Complying with these regulations and policies can burden a company when it comes to how it handles its data (Panian 2010).

Zeljko Panian in his article—"Some Practical Experiences in Data Governance", highlights some common data issues that companies face when faced with compliance and regulatory requirements:

- It's hard to pull together the financial data from the dozens of different sources, from mainframe to spreadsheets, so there's always a lot of IT involvement to pull the data, which slows the whole process down.
- Our Chief Financial Officer is demanding access to all the information on a daily basis on his computer. We can't do that right now.
- The data we report to the auditors have to be clean and accurate, and right now it's nowhere close.
- Our different business units use different charts of accounts, so it takes days for our analysts to reconcile them.
- A lot of our compliance reporting is done via spreadsheet, which is not going to hold up with the auditors.
- These are sensitive data. We have to carefully control access to them, or we could face huge fines.

Without a robust enterprise-wide data governance, it is not possible to meet compliance and regulatory requirements. This is because, in order to be compliant, adequate controls are required over the data asset life cycle starting from the data creation or acquisition, capture, storage, maintenance, access, and retrieval to the retirement of the data as well as data security and data quality. Also, the data should align with predefined business rules and the data should be fit for purpose.

Data governance means ensuring that the organization has policies, processes, and controls in place to facilitate adequate management of data and ensure that all regulations are met in an organization's data practices. Data governance establishes standards, rules, policies, and processes that are required by regulations and corporate governance policies. It establishes and assigns roles and responsibilities that help to ensure accountability for the data. Data governance also helps to automate compliance-related tasks, lowering costs (Panian 2010).

Regulations do not remain static and change over time which might need existing processes, standards, and data controls to be modified and/or new processes, standards, and data controls to be implemented and enforced to meet changing regulatory requirements. Without a rock solid and sustained data governance program in place, an enterprise cannot keep up with the changing compliance needs. Good data governance enables compliance in a changing regulatory landscape, and takes into account changes in the organization's own business goals and objectives (Salido and Voon 2010). In short, data governance is the key to effective compliance (Mahanti 2021).

Data governance provides a platform for meeting the demands of regulations and helps achieve compliance goals. By adopting a principal of good data governance, an organization should be able to comply with anything asked for by a regulator and compliance will essentially become a by-product of an ongoing process (Cooper 2016).

The role of data governance in achieving compliance is discussed in detail in the Chapter—Data Governance and Compliance in the first book of the series—Data Governance and Compliance: Evolving to our Current High Stakes Environment.

2.13.2 Improving Customer Satisfaction

Improving customer satisfaction is another important data governance business driver. As per "*The 2018 State of Data Governance Report*" conducted by erwin, 49% of the survey respondents cited improving customer satisfaction as one the key business drivers for data governance (Erwin and UBM 2018).

Having visibility into your data also facilitates customer facing transparency, which means businesses have a single view of their customers, so they can quickly resolve issues, answer customer queries, and help align the products and services most relevant to customer needs, thus, improving customer satisfaction.

2.13.3 Reputation Management

As per "*The 2018 State of Data Governance Report*", conducted by erwin, 30% of respondents cited reputation management as a driver for implementing data governance. Reputation is not easy to measure but a good reputation is crucial for the success of a business.

Reputation management has always been imperative to businesses as bad reputation has a potential to reverse years of hard-won trust and brand value that a business has earned. Bad reputation also leads to a loss of competitive advantage.

However, with the growing number of data breaches every year, hackers are becoming more and more sophisticated, and with information flow being so smooth and fast in today's digital world, reputations are more fragile than ever.

In today's internet age and social media being active, bad news travels extremely fast; often considerably faster than businesses can respond. It is also extremely difficult to drive away bad news even long after the occurrence of the adverse incident. Earlier, old news would die a natural death after a certain period of time. However, in today's internet age, social media and search engines make it easy for information to circulate, even long after incidents have happened. One of the fastest ways to see your organization's reputation suffer today is to lose or expose sensitive data (Pastore 2018b). A study in the U.K. found that 86% of customers would not do business with a company that failed to protect its customers' credit card data

(Pastore 2018b). A big part of building and maintaining a good reputation today means circumventing slip-ups like those suffered by numerous companies like TalkTalk, Facebook, Equifax, Uber, Yahoo, and Wells Fargo.

In order to maintain good reputation, organizations need to govern their organization's data assets. This is because successful reputation management reflects in the data the organization collects and maintains. A strong data governance practice gives businesses the needed visibility into their data:

- what data are they collecting?,
- why are they collecting the data?,
- who can access the data?,
- where is the data stored?,
- what is the data used for?,
- who is accountable for the data?,
- how sensitive is the data?, and
- how is the data used?

This visibility can help protect reputations because knowing what data you have, how the data is used, what the data is supposed to be used for, and where the data resides helps improving controls around data security, access, and data protection. This understanding enables organizations to focus security spending on the areas of highest risk. Thus, they can take a more cost-effective but thorough approach to risk management (Pastore 2018b).

Having visibility into your data also facilitates internal transparency as well as customer facing transparency. Internal transparency is the ability to quickly and accurately answer questions posed by executives, auditors or regulators which is possible when you have data to support your answers. Both types of transparency help manage an organization's reputation. Businesses with a well-developed strategy for data governance are less likely to be caught off guard by a data breach months after the fact (Pastore 2018b), and are better positioned to provide a better customer experience resulting in customer satisfaction which in turn has a positive impact on the reputation.

2.13.4 Better Decision Making

As per "*The 2018 State of Data Governance Report*", conducted by erwin, 45% of respondents identify better decision making as the third key data governance driver (Erwin and UBM 2018).

Data governance success manifests itself as well-defined data that is consistent throughout the business, understood across departments, and used to pull the business in the desired direction. It also improves the quality of the data.

By moving data governance out of its IT silo, the employees responsible for business outcomes are part of its governance. This collaboration makes data more discoverable, more insightful, and more contextual.

The decision-making process becomes more efficient, as the velocity at which the data can be interpreted increases. The organization can also better interpret and trust the information it is using to determine course.

2.13.5 Data Security and Privacy

Data security and privacy are business drivers for data governance. Yahoo, Facebook, Equifax, Delta, Westpac, Excellus BlueCross BlueShield, Marriott International, Sonic, Whole Foods, and many other organizations in different industry sectors have been the victims of data breaches.

The media coverage of cracks in security that lead to the cyber theft of corporate intellectual property is harder to find. Organizations tend to keep cyber security incidents from being exposed in public. While media coverage of security loopholes that result in theft or manipulation of corporate intellectual property is not easy to find, data hacking resulting in privacy and security breaches can severely undermine an organization's value and damage its reputation. Such incidents may be the result of not having a true data governance foundation that makes it possible to understand the context of data—what data assets exist and where, the relationship between them and enterprise systems and processes, which data is sensitive and needs more protection, who is authorized to access specific data and how and by what authorized parties is the data used. This knowledge is critical to supporting efforts to keep relevant data secure and private. Creating policies and processes for data management and accountability and driving culture change so people understand how to properly work with data are important components of a data governance initiative, as is the technology for proactively managing and securing data assets (Erwin 2018).

Data governance results in reliable, up-to-date audit information, processes, and policies in place to classify data, control access to data, and a reduction in the overall risk of unauthorized access to data and the inappropriate use of that data.

Concrete data governance programs will have policies for archival, retirement, and purging of data, and will mandate that any non-essential or out-of-date data is destroyed in a secure manner. Data governance also makes sensitive and confidential data more readily discoverable and comprehensible, resulting in making well informed decisions regarding the security investments.

2.13.6 Improving Data Quality

Organizations are victims of poor data quality issues, such as missing data values or records, duplicated data records, data integrity issues, data inconsistencies, and data inaccuracies.

In the words of John A. Zachman, Author of "The Framework for Enterprise Architecture" (The "Zachman Framework"), Zachman International (Mahanti 2021),

> The big problem with data is semantic coherence, typically referred to as 'data quality.' If the data does not accurately reflect reality or is being conceived or understood inconsistently, then there WILL BE errors, defects, mis-understandings, conflicts, war, and/or untold costs of entropy of reconciliation.

Organizations generally start focusing on data quality because of one or both of the below reasons:

- Data intensive projects such as business intelligence, analytics, CRM, ERP, etc. are suffering because of bad data, and
- Laws and regulations are pushing the organizations to get a grip on the correctness, completeness, and accuracy of their data.

While technology is needed to improve data quality, you cannot achieve and sustain data quality with technology alone. You need data governance which facilitates to provide the necessary level of data quality control.

Data governance creates policies, processes, standards, roles, responsibilities, decision rights, controls, and metrics that results in better quality data including data accuracy, completeness, conformity, completeness, integrity, consistency, reduced duplication, and greater traceability.

Standard naming conventions and data definitions decreases inconsistencies and duplications, and increases conformity. Roles, responsibilities, and decision rights increase accountability for data and help resolve data issues. Controls and data quality metrics help ensuring that the data is of adequate quality.

2.13.7 Analytics

Analytics is another key driver for data governance. As per "*The 2018 State of Data Governance Report*" conducted by erwin, 27% of the survey respondents cited analytics as one the key business drivers for data governance (Erwin and UBM 2018).

Emily Washington, senior vice president of product management and Infogix, a company that makes data analytics and data management software, states (Woodie 2018):

> When organizations start collecting data, it is often on a small scale that is manageable. But as decisions are made based on analytical results, organizations start to see the real value in data collection and, as you guessed it, start collecting more data. But as they amass more data, organizations often lose control of their data's quality, origin, ownership — all key components to a successful data governance program (Woodie 2018).

Analytics is often referred to as fact-based decision making (Harris 2015), with data being the common denominator of all analytics be it *descriptive analytics* that assists organizations in understanding what has happened in the past and what is happening in the present, or *diagnostic analytics* that helps answer, why something happened, or *predictive analytics* that determines the probability of what will happen in future, or *prescriptive analytics* that focuses on finding the best course of action for predicted future scenarios.

Descriptive analytics, which includes a lot of traditional business intelligence reporting, often draws from rigidly governed sources such as master data management and data warehousing which have better data quality and better issue resolution processes, whereas predictive and prescriptive analytics often draw from more loosely governed sources, such as social media, open data, and data streaming from Internet-connected sensors (Harris 2015) which may not be of high quality.

Since your analytics results will be as good as your underlying data, and data is often required from different sources, a significant amount of effort goes into understanding the data, cleaning, and stitching the data from different sources before they can be used for analytics purposes. This is known as data preparation.

A sound data governance framework creates and maintains policies and processes that define, document, and communicate the data preparation process for analytics (Harris 2015) and brings together the different stakeholders to resolve data issues by establishing roles and responsibilities around data in the different systems.

When data governance helps your organization develop high-quality data with demonstrated value, your IT organizations can build better analytics platforms for the business (Pastore 2018a). Sound data governance is a must for analytics, as poor data governance will result in poor data quality, which in turn can compromise any strategies which have been constructed on the basis of data analytics, which in turn can lead to making poor decisions, that can eventually lead to financial and reputational risk for the organization. A well-functioning data governance program creates a single version of the truth by helping IT organizations identify and present the right data to users and eliminate confusion about the source or quality of the data (Pastore 2018a). Data governance also enables a system of best practices, subject matter experts in the form data stewards, accountabilities in the form data owners, and increased alignment and collaboration that are the hallmarks of today's analytics-driven businesses (Pastore 2018a).

2.13.8 Big Data

Big data is another key driver for data governance. Put in simple terms, big data is the data that contains a greater variety of data arriving in increasing volumes and with ever-higher velocity. As per Gartner's definition, big data is high-volume, high-velocity, and/or high-variety information assets that demand cost-effective and innovative forms of information processing that enable enhanced insight, decision making, and process automation (Gartner 2018). Research indicates that 90% of the

world's data has been created just in the last two years. Globally, we generate 2.5 quintillion bytes a day (Erwin and UBM 2018).

As per "*The 2018 State of Data Governance Report*" conducted by erwin, 21% of the survey respondents cited big data as one of the key business drivers for data governance (Erwin and UBM 2018).

The need for data governance in these cases is largely driven by the volume, variety, and velocity of the data, also known as the "3 V's of data" (volume, velocity, and variety), which tend to be positively correlated.

It is important to discover and understand the data before it is used. Big data results in the capture of huge amounts of data, but not all data is valuable or fit for purpose.

Data governance is needed so that the unneeded data is disposed off in a legally defensible way (Smallwood 2014). The big data environments are potential goldmines when looked at from an insight's perspective but without proper governance, accountability, organizational collaboration, and support, they can be black holes of unused data (Washington 2018).

2.13.9 Revenue Growth

One of the most important objectives of any business is to generate and grow revenue. One of the most effective ways to grow revenue is to increase cross-sell/up-sell rates, improve retention among existing customers, and attract prospective customers. To do so, organizations need a broad and deep understanding of their existing customers as well as products.

The customer, prospect, and product data are often scattered across a multitude of different business systems in an organization resulting in data quality issues like duplication and inconsistencies. To resolve these issues, companies must address the underlying organizational, architecture, process, and technical issues related to the data.

Data governance provides the framework for addressing complex issues such as improving data quality or developing a single view of the data at an enterprise level (Panian 2010), thus, enabling superior service, ability to better target campaigns and offers based on a specific customer's needs which in turn results in increased revenue.

Good data governance results in the utilization of data assets to make new sales and to achieve new business capabilities, creation of "sellable" information products, and a better understanding of customers, prospects, products, and (and other) hierarchies (Thomas).

2.13.10 Improving Operational Efficiency

Organizations are under pressure to increase operational efficiency to improve productivity and lower costs. One of the important ways' organizations can reduce costs and increase operational efficiency is to automate business processes (Jeston 2008). For example, financial organizations in Australia have automated their conveyancing and property settlement processes by using an online platform PEXA developed by Property Exchange Australia, to lower operational and administration costs, and to decrease fraudulent transactions.

While automating business processes increases efficiency, enterprise data issues prevent organizations from getting the utmost benefit out of the operational efficiency initiatives. This is because there are a multitude of systems and data across these systems do not have uniform standards and definitions.

Streamlining business processes across financial, human resource, sales, inventory, marketing, and other business systems requires that the structure and meaning of data be reconciled across those systems so that they have an enterprise wide standard definition and representation. However, this activity has often been an afterthought in operational efficiency initiatives (Panian 2010; Schmidt and Lyle 2010).

Usually data quality and master data management initiatives are set up to improve the quality of data. Data governance plays a critical role in the success of such projects, providing a structure consisting of roles, responsibilities, policies, processes, rules, and standards for addressing the organizational and process issues around the data.

2.13.11 Mergers and Acquisitions

With mergers and acquisitions, organizations are faced with the need to rationalize and reconcile the IT and data environments from merged or acquired entities. This is because different organizations have very different IT and data systems, data models, as well as business processes.

While there are substantial IT costs involved in mergers and acquisitions, there is a significant data aspect involved too. The goal is to accelerate the promised synergies from the merger, both in the form of cost reductions by eliminating redundancies, as well as revenue growth from increased cross-selling. However, there are significant differences around data in the entities involved in a merger or acquisition that need to be taken into consideration when consolidating data. The differences are as follows:

- Differences in data capture, processing, management, and maintenance.
- Differences in data models, data standards, and data definitions.
- Difference in data policies.
- Differences in data processes, procedures, and methods (Joss 2016).
- Differences in data management practices.

- Differences in data quality (Joss 2016).
- Strategies related to data may be different.
- Differences in business rules used to transform, derive or summarize data.
- Technologies used to manage and store data may be different.
- Differences in data culture.
- Differences in sponsorship in relation to data related activities.

All the above differences need to be taken into consideration when migrating and integrating data. The process of migrating and consolidating the data after a merger or acquisition is a huge task (Panian 2010), and involves much more than just technical integration.

Integration is usually bound by extremely tight timelines and is often underestimated initially, as the above differences in data are not taken into consideration.

IT groups must deal with unknown systems, resolve quality issues, and provide detailed documentation on how the information has been merged. IT organizations must reconcile different data definitions and models, and processes must also be put in place to ensure alignment of the various entities.

Also, what is often underestimated is the resistance to change as when two organizations merge; each organization has its unique way of doing things, and active engagement and collaboration is needed between the different groups to resolve differences in data and find a way to work together.

Data governance which brings together the different data groups, and establishes processes, roles, responsibilities, and decision rights around the different data domains, facilitates collaboration and alignment across both the business and IT groups, to resolve data conflicts. A data governance framework provides significant value in managing the organizational and technical complexities of merger or acquisition consolidation and accelerating positive business results (Panian 2010).

2.13.12 Partnering and Outsourcing

Organizations are increasingly using partners and outsourcers to manage parts of the value chain. They are focusing on core competencies and handing off non-core functions and processes to partners and outsourcing providers. For example, IT departments outsource application development, maintenance, and network management.

As business processes and IT systems are outsourced to external vendors and providers, the data associated with those processes and systems also relocates outside the boundaries of the organization, sometimes in a different country.

Organizations must ensure that the data are correctly migrated to the external provider. The data must be complete and accurate, and they have to be restructured to work in the third-party system and data should continue to integrate with the other systems still within the organization boundaries. It is important to note that although they have moved to a third party and reside outside the firewall, these data

still remain a core asset of the organization and should be adequately secured, and adequate controls over the data should be in place. A robust data governance framework is critical to managing data that are distributed in different locations, especially in defining the policies, standards, processes for interaction, and collaboration with external partners and outsourcers (Panian 2010).

2.14 Key Benefits of Data Governance

It is important to be aware of the benefits of data governance. Anee Buff, SAS Best Practices thought leader, in a recent Information Management column stated (Lawson 2014; Buff 2014):

> Data governance is a powerful pill as it not only knocks out the causes of the common data headaches, it helps prevent them. A finely tuned data governance program can reduce duplicate data throughout the organization, reduce errors in reporting and coding and reduce costs associated with poor data quality.

It can save money during data migrations, too, she writes, by improving accuracy and efficiency while reducing the time it takes to complete the migration (Lawson 2014).

As stated by Shannon Fuller, Director of Governance Advisory Services, Gray Matter Analytics in his interview, an effective data governance program enables trust in information and risk reduction. Trusted information aids information-based decision making and improves operational efficiency (Mahanti 2021).

Below are some of the key benefits of data governance (summarized in Fig. 2.17).

2.14.1 Common Understanding of Data

If existing data objects (for example, data elements and table names) are not named and defined meaningfully, consistently, and uniquely, they cannot be located, and therefore cannot be accessed, used or shared efficiently.

For example, if supplier name is referred to as provider name and vendor name with different definitions in different business units and their respective systems, then it would cause confusion. Supplier name, provider name and vendor name would imply three different entities when they are actually referring to the same entity. In the absence of data governance, naming and definition standards would not be in place and hence, inconsistent naming of data objects as well as inconsistent data definitions would be prevalent.

On the other hand, if two data elements in different systems have the same name but different definitions, it would again cause confusion as they would actually refer to two different entities but could be mistaken for the same entity.

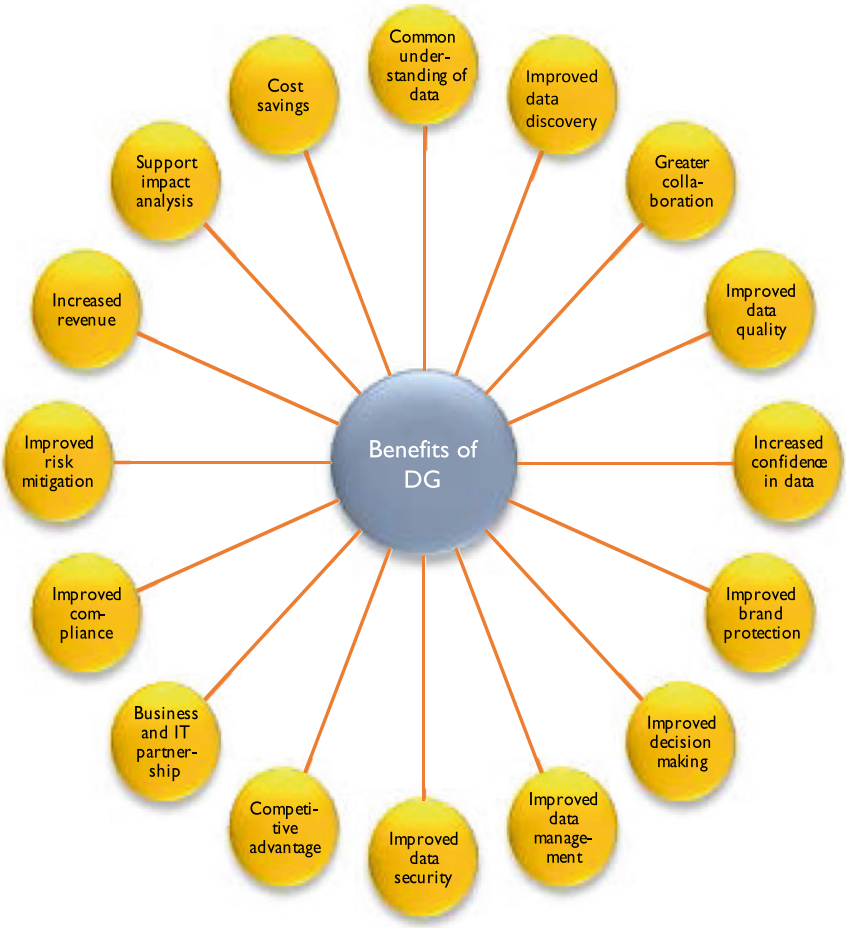

Fig. 2.17 Key benefits of data governance

Data governance provides a consistent view of, and common terminology for, data objects by providing standards for naming and definition of data objects across the enterprise, while individual business units retain appropriate flexibility.

Common names based on standard naming conventions with common definitions across the enterprise speeds investigation when trying to determine whether a data object already exists. Common names imply that there is a readily understandable business name and an abbreviated short physical name and prefix, based in part on a standard abbreviation list (Talend; Dennedy et al. 2014) and naming convention list.

2.14.2 Greater Collaboration

An effective data governance program breaks down barriers to collaboration by initiating conversations between data stakeholders from different business functions about how to best use and manage data. This cross-functional communication helps eliminate the misinterpretations that occur when different business units make assumptions about data that different groups capture, contribute, and share with the enterprise (CU*Answers 2015) and also paves way to acceptance of common enterprise wide data standards and definitions.

2.14.3 Improved Data Discovery

Data governance provides an advanced ability to understand the location of all data related to key entities by assigning data owners and data stewards (subject matter experts) in different business units and systems who are accountable and responsible for specific data respectively, as well as by ensuring the capture of information about different data objects in the corporate data dictionary. Data owners and data stewards can serve as the go-to people when people have questions related to a capture of a data object.

In case, the data object cannot be found in the data dictionary and data owners and data stewards also cannot be found, there is still the governance team who can direct to the appropriate person or bring it up in a common forum/working group that has data stakeholders from different business units/departments to help locate the data. Like a GPS that can represent a physical landscape and help people find their way in unknown landscapes, data governance makes data assets useable and easier to connect with business outcomes (Talend).

2.14.4 Increased Confidence in Data

Data governance builds trust in an organization's data assets by establishing a shared understanding through standard data definitions, terminologies, data availability, quality, and usability. Data governance results in an increased confidence in data-related decisions, a better ability to make timely data-related decisions, increased confidence in data appearing in financial and management reports and increased confidence in data strategy by providing a cross-functional team to weigh in on key decisions (Thomas).

2.14.5 Improved Brand Protection

McDonald's states, "Our brands need predictability in the market place—our consumers expect predictability." (Young and Ilieva; Broekhof and Stomp 2015). Data governance results in creation and enforcement of policies, processes, roles, and responsibilities that result in improved data quality and data security, which in turn reduces security incidents, which in turn has a positive impact on brand reputation.

2.14.6 Improved Decision Making

The best decisions are founded on both experience and data which represent facts. These decisions can be reactive to fix a problem after an event has occurred (for example, in case of data breach or recall) or proactive (for example in case of a strategy decision or forecasting). Bad data quality will therefore result in wrong or low-quality decision. As data governance results in better data quality, it indirectly leads to better decision making.

2.14.7 Competitive Advantage

With a solid data governance program, decision-makers have accurate, current, timely, and consistent data to react to market conditions, develop real time dashboards, see what is happening across their organizations, perform comparisons, respond to client and customer needs, deliver relevant products and services, and drive success.

2.14.8 Improved Data Management

Data governance brings the human and process dimension into a highly automated and data-driven world (Talend). It establishes policies, processes, standards, codes of conduct, and best practices as well as establishes enterprise wide roles and responsibilities around data management. By doing so, it ensures that the concerns and needs beyond traditional data and technology areas—including areas such as legal, risk, operations, security, and compliance as well as cross-department data issues are addressed consistently (Talend).

2.14.9 Improved Risk Mitigation

Data governance helps in risk mitigation by facilitating understanding of where the data is kept and what it is used for, and establishing standards and processes to improve the data quality.

The misuse of data results from a lack of understanding of the meaning and provenance of the data.

Data governance establishes roles, responsibilities, processes, and standards that ensure documentation, maintenance, availability of critical context, and meaning throughout the lifecycle of critical data which results in better understanding of the data and lessens the risks of data misuse. Knowing where all in the organization, critical, sensitive data resides, what for and how the different data are used, helps organizations better manage and mitigate risk, by working towards appropriate controls to secure and protect data from unauthorized access and manipulation, and prevent misuse of data.

2.14.10 Cost Savings

A solid data governance program ensures improved data quality that reduces the likelihood of data issues; which in turn results in reduced costs and hence savings.

Data governance results in reduced operational costs achieved by eliminating duplicate processes, eliminating manual steps related to data input, processing, analysis, and distribution, getting rid of data assets that are not used (and duplicate data), decreasing costs associated with poor data quality, and reducing administrative costs by defining clear roles and responsibilities for data management.

Data governance also results in improved compliance which in turn results in cost savings from avoidance of cost of penalties associated with non-compliance, not having to pay higher audit fees due to lack of confidence in "authoritative data", reduction of management attestation/certification costs and reduction of costs of pre-audit testing (Thomas).

2.14.11 Support Impact Analysis

Data governance increases the ability to do useful impact analysis (by providing authoritative business rules, system of record information, and data lineage metadata) and provides a capability to assess cross-functional impacts of data-related decisions (Thomas).

2.14.12 Business and IT Partnership

As per Dr. John R. Talburt, data governance helps to bring business and IT together in a more synergistic and comfortable partnership. In Dr Talburt's words, "Often, the relationship between business and IT are very strained. They are frequently at odds with each other, each blaming the other over a lack of responsiveness and issues of control." The implementation of the DG program is a great opportunity to bring business and IT into a more harmonious and productive partnership with increased collaboration and better communication channels between them (Mahanti 2021).

Other benefits of data governance are improved data quality, improved data security, increased revenue and improved compliance and these have been covered in the "*Key data governance business drivers and uses cases*" section in this chapter.

2.15 Concluding Thoughts

If your organization collects, stores and uses data, then you need data governance. This is because data is an important asset and an effective data governance program is essential in managing this asset (Mahanti 2021). In George Firican's words (Mahanti 2021):

> … the engine for economic growth is no longer fueled by gasoline, but more and more by data.

While data governance is one of the 'pillars' of data management (Advisiondigital 2016), it is often viewed as "nice to have". However, with the importance of data increasing exponentially, and compliance and analytics calling for better quality data and improved data protection, data governance is a "must have".

As stated by Andres Perez in his interview, effective data governance provides a beacon for data value realization and contributes to a data driven culture (Mahanti 2021). Data governance enables you to effectively and proactively manage data assets throughout the enterprise by providing guidance as in how to do it, in the form of policies, standards, processes, and rules, and defines role and responsibilities with respect to who will do what, with respect to data.

While data needs to be governed, not all data needs to be governed with the same rigor. Critical data that are sensitive data, data that need to be of high quality, and data that is widely shared needs to be governed rigorously with the exclusion of public data.

As data, industry, technologies, and compliance are evolving, data governance is becoming even more crucial in the corporate world. While the adoption of data governance is faster in regulated industries with compliance being the driver, even for non-regulated industries, the benefits surpasses the cost of establishing data

governance in the organization. This is because data is the life blood of an organization and hence, needs to be managed with care, and effective data management needs sound data governance.

As highlighted by Christopher Butler, Chief Data Officer, HSBC, UK in his interview (Mahanti 2021):

> The one thing for certain is that data will become more and more important as a differentiator for businesses going forward. Data Governance is at the heart of whether an organization will be in a position be benefit from its data.

References

Advisiondigital (22 November, 2016) Why do you need data governance? WorleyParson Group. http://digital.advisian.com/curious/why-do-you-need-data-governance/. Last accessed 9 Nov 2018

Ballou DP, Pazer HL (1985) Modeling data and process quality in multi-input, multi-output information systems. Manage Sci 31(2):150–162

Batini C, Cappiello C et al (2009) Methodologies for data quality assessment and improvement. ACM Comput Surv 41(3):1–52. https://doi.org/10.1145/1541880.1541883

Birchler U, Bütler M (2007) Information economics. In: Routledge advanced texts in economics and finance. Routledge, London, 462 pp

Broekhof M, Stomp R (30 September, 2015) Why do companies need data governance?. View points on innovation. http://viewpoints.io/entry/why-do-companies-need-data-governance. Last accessed 9 Nov 2018

Buff A (16 May, 2014) Headache number two: remedies for data migration pain. Information management. https://www.information-management.com/news/headache-number-two-remedies-for-data-migration-pain

Cappiello C, Francalanci C, Pernici B (2003) Time-related factors of data quality in multi-channel information systems. J Manage Inf Syst 20(3):71–91

Cebr (12 June, 2013) Data on the balance sheet. https://cebr.com/reports/data-on-the-balance-sheet/

Chisholm M (06 May, 2014) Data credibility: a new dimension of data quality? Information Management. https://www.information-management.com/news/data-credibility-a-new-dimension-of-data-quality. Last accessed 08 Aug 2017

Chisholm MD (2010) Definitions in information management: a guide to the fundamental semantic metadata. Design Media, Canada

Cooper D (15 August, 2016) The importance of data governance in meeting regulatory compliance. dataIQTM Blog. https://www.dataiq.co.uk/blog/importance-data-governance-meeting-regulatory-compliance. Last accessed 9 Nov 2018

CU*Answers (February, 2015) Data governance research—a business development exercise. https://www.cuanswers.com/wp-content/uploads/DataGovernanceResearch.pdf. Last accessed 9 Nov 2018

D'Onofrio D (Nov, 2017) The future is now: data as an asset in the corporate world. Snaplogic blog. https://www.snaplogic.com/blog/the-future-is-now-when-data-becomes-a-tangible-corporate-asset. Last accessed 22 Sept 2018

DAMA International (2014) DAMA-DMBOK2 Framework, Last accessed on June 30 2020 from https://dama.org/sites/default/files/download/DAMA-DMBOK2-Framework-V2-20140317-FINAL.pdf

DAMA UK Working Group (2013) The six primary dimensions for data quality assessment defining data quality dimensions, pp. 1–17. http://www.damauk.org/rw/CatViewLeafPublic.php?&cat=403. Last accessed 8 Aug 2017

Dennedy M, Fox J, Finneran T (2014) The privacy engineer's manifesto: getting from policy to code to QA to value. A-Press

Duemig K (2011) Treating data as an enterprise asset to achieve business value. IBM Corporation. ftp://ftp.software.ibm.com/software/data/sw-library/information-governance/Data_a_an_Enterprise_Asset.pdf

Eckerson W (June, 2011) Creating an enterprise data strategy: managing data as a corporate asset. http://docs.media.bitpipe.com/io_10x/io_100166/item_417254/Creating%20an%20Enterprise%20Data%20Strategy_final.pdf

English LP (1999) Improving data warehouse and business information quality. John Wiley and Sons, Indianapolis, IN

Erwin (2018) Examining the data trinity: governance, security and privacy. Erwin White Paper

Erwin, UBM (2018) The 2018 State of data governance report

EverEdge (26 June, 2019), Data valuation: the Holy Grail, Last accessed on 10 Feb 2020, from https://www.everedgeglobal.com/news/datavaluation/

Experian Data Quality (2014) Dawn of the CDO. www.edq.com/cdo

Experian Data Quality (2015) Rise of the data force. Experian data quality in conjunction with Spencer Stuart. https://www.spencerstuart.com/-/media/pdf%20files/research%20and%20insight%20pdfs/experian_rise_of_the_data_force_sept15.pdf

Financial Accounting Standards Board [FASB] (1980) Statement of financial accounting concepts no. 3: elements of financial statements of business enterprises, Stamford, Conn., Dec 1980

Firican G (22 August, 2018) 4 main roles of metadata. Lights on data blog. https://www.lightsondata.com/the-4-main-roles-of-metadata/

Foote KD (2 July, 2020) A brief history of data lakes. Dataversity. https://www.dataversity.net/brief-history-data-lakes/. Last accessed 28 July 2020

Fred J (2017) Data monetization—how an organization can generate revenue with data? Tampere University of Technology

Gallaugher J (2009) The data asset: databases, business intelligence, and competitive advantage. http://gallaugher.com/The%20Data%20Asset.pdf. Last accessed 22 Sept 2018

Gartner (2016) Gartner glossary—master data management. https://www.gartner.com/en/information-technology/glossary/master-data-management-mdm

Gartner (2018) Big data, Gartner IT glossary. https://www.gartner.com/it-glossary/big-data/. Last accessed 14 Oct 2018

Gidley S, Castanedo F (2017) Understanding metadata, 1st edn. O'Reilly Media Inc.

Glue Reply. The valuation of data as an asset: a consumption based approach. https://www.reply.com/Documents/13903_img_The-valuation-of-data-as-an-asset.pdf. Last accessed 22 Sept 2018

Goetz M (September 1, 2015) Data governance and data management are not interchangeable, Forrester blog. https://go.forrester.com/blogs/15-09-11-data_governance_and_data_management_are_not_interchangeable/. Last accessed 7 July 2018

Gorball J (24 March, 2016) Data governance versus data management. Kingland blog. https://blog.kingland.com/data-governance-vs.-data-management

Gow B (2006) Case study: data governance and compliance for financial services. IBM Corporation. http://www.sourcemediaconferences.com/CDISP07/pdf/Gow_Brett.pdf. Last accessed 13 Oct 2018

Hagiu A (2014) Strategic decisions for multisided platforms. MIT Sloan Manage Rev 55(2):71

Harris J (20 August, 2015) Data governance and analytics. In: The data roundtable, SAS blog. https://blogs.sas.com/content/datamanagement/2015/08/20/data-governance-analytics/. Last accessed 14 Oct 2018

Hodgson R (25 August, 2017) The "governance of data governance". TopQuadrant company blog. https://www.topquadrant.com/2017/08/25/how-to-establish-governance-of-data-governance/. Last accessed 22 Sept 2018

Inmon WH (2005) Building the data warehouse, 4th edn, Wiley. ISBN: 978-0-7645-9944-6
i-scoop (2016) Data is a business asset beyond imagination—here is why (and where). https://www.i-scoop.eu/big-data-action-value-context/data-business-asset/
ISO 11179. https://www.iso.org/obp/ui/#iso:std:iso-iec:11179:-4:ed-2:v1:en
Jeston J (2008) Business process management: practical guidelines to successful implementations, 2nd edn. Butterworth Heinemann, Burlington, MA, pp 44–45
Joss A (16 December, 2016) The role of data in mergers and acquisitions. Informatica Blog. https://blogs.informatica.com/2016/12/16/role-data-merger-acquisitions/#fbid=Qb7RoMLIVRp. Last accessed 14 Oct 2018
Knight M (12 December, 2017) Data management versus data governance: improving organizational data strategy, Dataversity. https://www.dataversity.net/data-management-vs-data-governance-improving-organizational-data-strategy/
Koutroumpis P, Leiponen A (2013) Understanding the value of (big) data. In: 2013 IEEE international conference on big data
Lawson L (22 May, 2014) How data governance can cut costs up- and down-stream. IT Business Edge. https://www.itbusinessedge.com/blogs/integration/how-data-governance-can-cut-costs-up-and-down-stream.html
Lederman R, Shanks G, Gibbs MR (2003) Meeting privacy obligations: the implications for information systems development. In: Proceedings of the 11th European conference on information systems (ECIS), Naples, Italy, 16–21 June. http://is2.lse.ac.uk/asp/aspecis/20030081.pdf. Accessed 29 June 2009
Lee SU, Zhu L, Jeffery R (2017) Data governance for platform ecosystems: critical factors and the state of practice. In: Twenty first Pacific Asia conference on information systems, Langkawi 2017. https://arxiv.org/ftp/arxiv/papers/1705/1705.03509.pdf
Loshin D (2009) Master data management. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA
Loshin D (2011) The practitioner's guide to data quality improvement. Morgan Kaufmann, Boston, MA
Luckerson V (2014) Controversial ways Facebook has used your data. Time. http://time.com/4695/7-controversial-ways-facebook-has-used-your-data/. Retrieved 03 Oct 2016
Mahanti R (2019) Data quality: dimensions, measurement, strategy, management and governance. ASQ Quality Press, Milwaukee WI, p 526, ISBN: 10:0873899776
Mahanti R (2021) Data governance and compliance, Springer Books, Springer, number 978-981-33-6877-4
Mauzy D, Bull B, Gould T (2016) Avoid the pitfalls: Benefits of formal Part C data system governance. SRI International, Menlo Park, CA
McGilvray D (2008) Executing data quality projects: ten steps to quality data and trusted information, 1st edn. Morgan Kaufmann, Morgan Kaufmann. Paperback ISBN: 9780123743695
MIT Technology Review Custom + Oracle [MIT+Oracle] (2016) The rise of data capital, White Paper, MIT technology review, Produced in partnership with Oracle
Myers D (2017) Conformed dimensions of data quality-open standard. http://dimensionsofdataquality.com/thestandard
Norris-Montanari J (17 October, 2016) What's the difference between data governance and data management? (Part 2). The data roundtable. SAS blogs. https://blogs.sas.com/content/datamanagement/2016/10/17/difference-in-data-gov-data-man2/
NTNU, Introduction to Big Data, Opphavsrett: Forfatter og Stiftelsen TISIP, Learning material, https://www.ntnu.no/iie/fag/big/lessons/lesson2.pdf
Oracle (2014) Take control of data governance and data quality. Enterprise data quality ipaper. http://www.oracle.com/us/products/middleware/data-integration/enterprise-data-quality/enterprise-data-quality-ipaper-2401613.pdf. Last accessed 13 Oct 2018
Orr K (1998) Data quality and systems theory. Commun ACM 41(2):66–71
Panian Z (2010) Some practical experiences in data governance. World academy of science, engineering and technology, vol 62. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.190.6948&rep=rep1&type=pdf

Pastore M (9 August, 2018a) Data governance helps build a solid foundation for analytics. Erwin expert blog. https://erwin.com/blog/data-governance-analytics/. Last accessed 14 Oct 2018
Pastore M (2 August, 2018b) Data plays huge role in reputation management. Erwin expert blog. https://erwin.com/blog/data-governance-reputation-management/. Last accessed 13 Oct 2018
Patrizio A (24 September, 2018) What is a data lake? Flexible big data management explained, InfoWorld. https://www.infoworld.com/article/3305843/what-is-a-data-lake-flexible-big-data-management-explained.html. Last accessed 28 July 2020
Pipino LL, Lee YW, Wang RY (2002) Data quality assessment. Commun ACM 45(4):211–218
Pixentia (16 September, 2016) 6 reasons you need data governance. http://blog.pixentia.com/6-reasons-you-need-data-governance. Last accessed 15 Sept 2018
Polikoff I (15 June, 2018) Data governance as a lifecycle-centric asset management activity. TopQuadrant company blog. https://www.topquadrant.com/2018/06/15/data-governance-as-a-lifecycle-centric-asset-management-activity/. Last accessed 6th Oct 2018
Pwc (2014) Deciding with data. https://www.pwc.com.au/consulting/assets/publications/data-drive-innovation-sep14.pdf
Rand (2013) Data governance simplified. www.randsecuredata.com/getattachment/resources/whitepapersreports/Survey-Results/RandSD_WP_DataGovSurveyResults_Final.pdf.aspx?ext=.pdf
Redman TC (1996) Data quality for the information age. Artech House Computer Science Library
Redman TC (1997) Data quality for the information age, 1st edn. ACM Digital Library, Artech House, Inc., Norwood, MA, USA. ISBN: 0890068836
Riley J (01 January, 2017) Understanding metadata: what is metadata, and what is it for?: a primer. NISO. ISBN: 978-1-937522-72-8
Rouse M. Metadata, WhatIs.com, TechTarget. https://whatis.techtarget.com/definition/metadata. Last accessed 6th Oct 2018
Salido J, Voon P (January, 2010) A guide to data governance for privacy, confidentiality, and compliance. Microsoft White Paper
Saxena A (18 March, 2019) What is data value and should it be viewed as a corporate asset?, Dataversity, Last accessed on 10 Feb 2020, from https://www.dataversity.net/what-is-data-value-and-should-it-be-viewed-as-a-corporate-asset/
Scarisbrick-Hauser A, Rouse C (2007) The whole truth and nothing but the truth? The role of data quality today. Direct Mark Int J 1(3):161–171. https://doi.org/10.1108/17505930710779333
Schmidt JG, Lyle D (18 May, 2010) Lean Integration: an integration factory approach to business agility. Pearson Education
Sebastian-Coleman L (31 December, 2012) Measuring data quality for ongoing improvement. Morgan Kaufmann. Print ISBN-13: 978-0-12-397033-6
Secure UD (2018) Understanding data criticality. University of Delaware. http://www1.udel.edu/security/data/criticality.html
Short J, Todd S (03 March, 2017) What's your data worth? Magazine Spring 2017, MIT Sloan Manage Rev, Last accessed on 10 Feb 2020, from https://sloanreview.mit.edu/article/whats-your-data-worth/
Smallwood RF (2014) Information governance: concepts, strategies, and best practices. John Wiley & Sons
Southekal PH (15 May, 2018) Why is data missing from the balance sheet? Data science central blog. https://www.datasciencecentral.com/profiles/blogs/why-is-data-missing-from-the-balance-sheet. Last accessed 22 Sept 2018
Talend. What is data governance (and why do you need it)?. Last accessed 8 Nov 2018
Tayi GK, Ballou DP (1998) Examining data quality. Commun ACM 41(2):54–57
The Apex (2 January, 2019) MIT on how to craft a successful data strategy. https://www.apexofinnovation.com/how-to-craft-a-data-strategy/
The Data Governance Institute. Goals and principles for data governance. http://www.datagovernance.com/adg_data_governance_goals/. Last accessed 9 Nov 2018
The Data Governance Institute [TDGI]. Data governance—the basic information. The data governance institute. http://www.datagovernance.com/adg_data_governance_basics/. Last accessed 9 Nov 2018

Thomas G. Demonstrating values, The data governance institute. http://www.datagovernance.com/demonstrating-value/. Last accessed 9 Nov 2018

Tucci L (22 July, 2010) Why you need a formal data governance program, and how to get started TechTarget. http://searchcio.techtarget.com/news/1517117/Why-you-need-a-formal-data-governance-program-and-how-to-get-started. Last accessed 31 March 2018

Turner N (24 January, 2018) Data governance and data architecture: alignment and accountability. Global data strategy. https://globaldatastrategy.files.wordpress.com/2018/02/edgo2017_data-governance-data-architecture.pdf. Last accessed 1 Jan 2019

Turnbull M (2014) AIIA speech: navigating analytics summit. navigating analytics summit 20 March 2014 Canberra. Australian Information Industry Association

Vincent DR (1990) The information-based corporation: stakeholder economics and the technology investment. Dow-Jones-Irwin, Homewood, Illinois

Wang RY, Strong DM (1996) Beyond accuracy: what data quality means to data consumers. J Manage Inf Syst 12(4):5–33

Washington E (16 April, 2018) Why data governance is crucial for big data environments. ReadITQuik. https://www.readitquik.com/articles/data/why-data-governance-is-crucial-for-big-data-environments/. Last accessed 4 Oct 2018

Watts S, Shankaranarayanan G, Even A (2009) Data quality assessment in context: a cognitive perspective. Decis Supp Syst 48(1):202–211

Wiki-Database model, https://en.wikipedia.org/wiki/Database_model#:~:text=A%20database%20model%20is%20a,uses%20a%20table%2Dbased%20format

Woodie A (2016) How do you value information?. Datanami https://www.datanami.com/2016/09/15/how-do-you-value-information/. Last accessed 22 Sept 2018

Woodie A (1 February, 2018) Five tips for winning at data governance. Datanami. https://www.datanami.com/2018/02/01/five-tips-winning-data-governance/. Last accessed 14 Oct 2018

Young G, Ilieva D. Product lifecycle management @McDonald's: an essential risk management tool. View points on innovation. http://viewpoints.io/entry/product-lifecycle-management-mcdonalds-an-essential-risk-management-tool/. Last accessed 9 Nov 2018

Zeithaml V (1988) Consumer perceptions of price, quality, and value: a means-end model and synthesis of evidence. J Market 52:22

Zentut, Dimensional modeling, http://www.zentut.com/data-warehouse/dimensional-modeling/

Chapter 3
Data Governance and Data Management Functions and Initiatives

Data that is loved tends to survive.
—Kurt Bollacker

Abstract The data management discipline has several functions or components with data governance being identified as one of the core components of data management tying together the other data management functions and data initiatives—for example, data architecture management, data modelling and design, master data management, reference data management, data warehousing and business intelligence, data quality management, metadata management, data security management, data storage and operations, document and content management, data integration and interoperability, data migration, big data, and analytics. In this chapter, we discuss the alignment and interaction of data governance in detail with each of these data management functions. The concept of big data, how big data differs from traditional data, and the application of data governance in big data, analytics, and data lakes has also been discussed in a separate section.

3.1 Data Governance and Data Management

Data is an enterprise asset at par with other critical enterprise assets such as finance, building, and machinery. However, data is not handled with the same rigor as other assets.

Data governance brings the same level of discipline and control to data management as is usual with governance of other critical enterprise assets (for example, finance, building, and machinery). Thus, it is imperative to institute and integrate data governance early on and in every phase of data initiatives, especially the ones that require large cultural and operational changes in organizations.

Data governance defines who has access to data and how, sets practices, processes, policies, rules, and standards for data management, and imposes roles and responsibilities for data stewardship and data ownership; this results in data management becoming a completely auditable and accountable exercise.

Tying data governance with data and data management initiatives ensure the business backing and active participation in initiatives that are often times perceived

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2021

R. Mahanti, *Data Governance and Data Management*,
https://doi.org/10.1007/978-981-16-3583-0_3

by the business as an absolute technology exercise owned and operated by IT (Shortis 2017).

Data governance represents the principal function of data management as it gives direction as to what decisions need to be made in data management, who makes these decisions, and their roles and responsibilities in relation to the decisions. Data management ensures these decisions are made, and appropriate action takes place in terms of implementation (Otto 2011).

As discussed in Chap. 2—Data and Its Governance, data management can be defined in two ways:

- Data management (the thing), and
- Data management (the discipline).

Data Management (the Thing):
Data management is the action of actually managing the data.

Data Management (the Discipline):
Data management can be considered as an overarching umbrella which encompasses different disciplines or functions. DAMA has identified 11 major functions of data management which it calls as knowledge areas (DAMA International 2014).

Data management is a multifaceted discipline comprising of several closely interacting subdisciplines or functions with data governance being one of the core functions tying together all the other data management functions—data governance, data architecture, data management, data modeling, data design, data quality management, data security management, data warehousing, business intelligence (BI) management, data integration and interoperability, documentation and content management, metadata management, reference data management, master data management, and data storage and operations.

Data governance is responsible for formalizing policies, processes, standards, rules, practices, roles and responsibilities within the different data management functions and ensuring that appropriate controls and metrics are in place, and data management is concerned with providing a set of tools and techniques to ensure that organizations actually have data to govern. Data governance is the data management function which provides all the other data management functions (for example, data quality, master data management, reference data management, and metadata management) with the necessary foundation and structure needed to ensure that data is managed as an asset.

Data governance is identified as one of the core components of data management tying together the other data management functions. We will discuss the alignment and interaction of data governance with each of these data management functions in detail in this chapter.

Data governance guides all the other data management functions by defining, reviewing, communicating, and enforcing policies, guidelines, processes, metrics, rules, and standards in each of the other data management functions and formalizing roles, responsibilities, accountabilities, decision rights and ensuring necessary controls are in place as shown in Fig. 3.1.

It also involves identifying, managing, prioritizing, escalating, resolving issues, and overseeing data management projects.

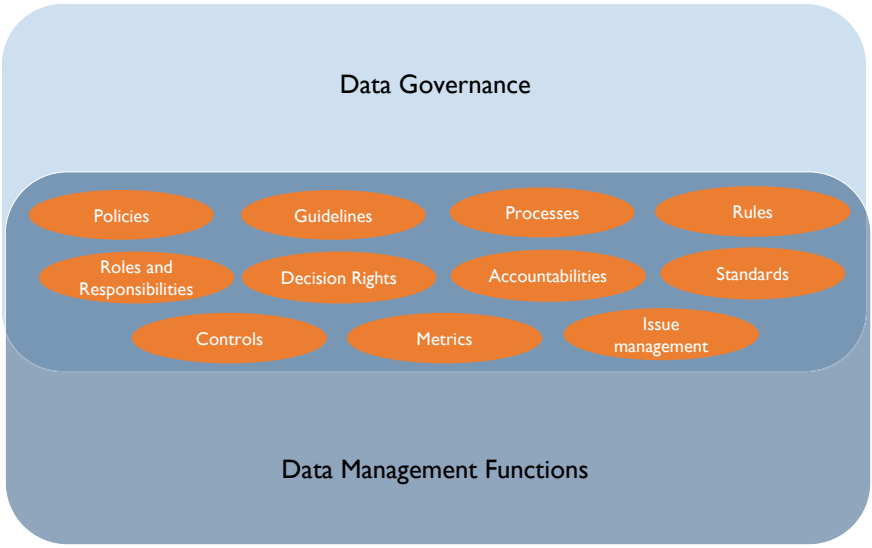

Fig. 3.1 Data governance and data management functions

3.2 Data Management Functions and Initiatives

As discussed in Chap. 2—"Data and Its Governance," data management can be considered as an overarching umbrella which encompasses the disciplines or functions included in the data governance framework (Norris-Montanari 2016). DAMA has identified 11 major functions of data management which it calls as knowledge areas as given below and summarized in Fig. 3.2 (DAMA International 2014):

- Data Governance—The exercise and enforcement of rules, processes, policies, practices, standards, metrics, controls, decision rights, and people accountabilities to manage data as a strategic enterprise asset.
- Data Architecture Management—technology and infrastructure design with respect to data collection, storage, transformation, distribution, and consumption as an essential part of the enterprise architecture.
- Data Modeling and Design—defining, analyzing, and scoping data requirements to support business processes, and then designing a data model (which may include a conceptual, logical, and physical model) to represent and communicate these data requirements.
- Data Quality Management—defining, measuring, analyzing, improving, and sustaining data quality.
- Data Security Management—classification, privacy, confidentiality, appropriate access, authorization, authentication, and auditing of data.

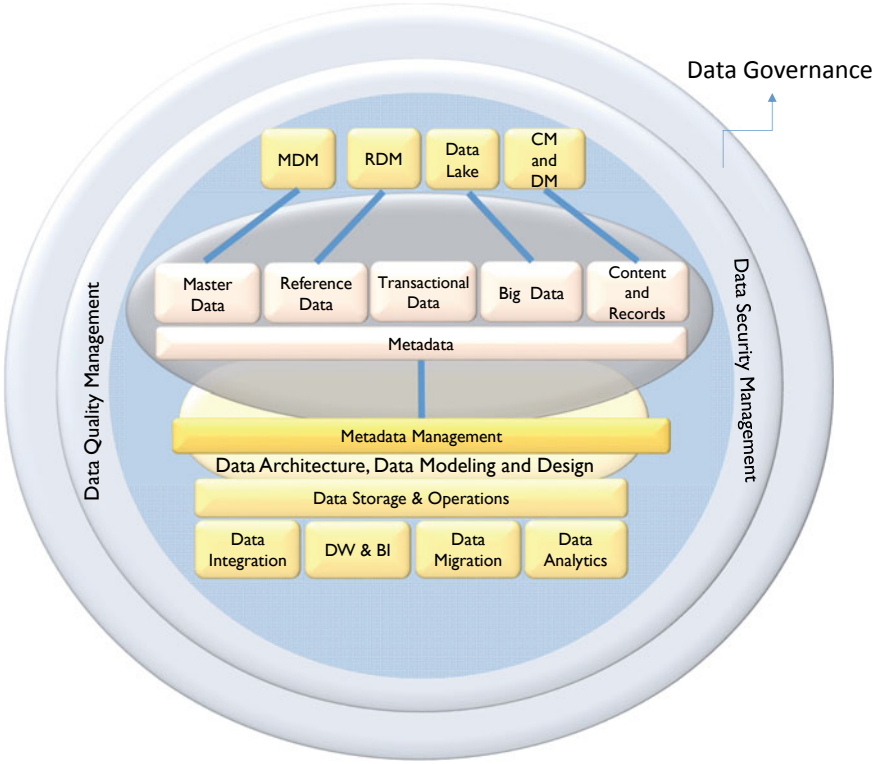

Fig. 3.2 Data governance tying together the other data management functions

- Data Warehousing and Business Intelligence (BI) Management—collection, integration, and presentation of data to knowledge workers for the purpose of business analysis and decision-making.
- Data Integration and Interoperability (DII)—relates to the movement and consolidation of data within and between data stores, IT systems, and enterprises.
- Document and Content Management—capturing, storing, categorizing, controlling the access and use of data and information stored outside relational databases.
- Metadata Management—capturing, categorizing, storing, maintaining, integrating, controlling, managing, and publishing metadata.
- Reference and Master Data Management—managing master data and reference data to improve data quality through standardized definitions and data values.
- Data Storage and Operations—storage, deployment, and management of structured physical data assets including backup, archival, purging, and performance management.

3.3 Data Architecture, Data Modeling, Design, and Data Governance

Both data governance and data architecture are data strategy components. Data architecture is a core component of data management, and an enterprise architecture domain. Organizations capture and store large amounts of data, which is growing exponentially and hence, needs to be managed adequately.

John Parkinson, Global Head of Data Architecture for HSBC RBWM, in his article titled "What is Data Architecture?" in LinkedIn, split and defined data architecture in two ways (Parkinson 2017).

- Data Architecture (the thing), and
- Data Architecture (the discipline).

Data Architecture (the Thing):
Data is quite a traveler. The way in which data travels through the organization through different systems can be thought of as plumbing of pipes from one point to another (Parkinson 2017). Data architecture can be thought of as the set of containers (systems) that holds the data throughout its lifecycle, from its capture, through to integration, to storage, and the pipes (transport system) through which the data flows as they are transported from one container to another container.

Data Architecture (the Discipline):
There is a need to represent and store enterprise data at different granularity levels depending on the business need. Data architecture as a discipline relates to controlling the design, the models, policies, rules, and standards required to store and manage the data in line with the business architecture. It is analogous to design of the pipe work so as to get the contents (the data) to the right place at the right time (Parkinson 2017).

Data governance and data architecture support each other in a number of different ways, and both have a shared goal of creating standards and guidelines to support the enterprise.

Data architecture provides an understanding of where and how the data is stored in the organization as well as the systems and applications through which the data traverses in the organization.

Data architecture also explains how the data is transformed as it is transported from one system to another. The flow of data between systems is documented in data flow diagram, which in turn drives governance in terms of processes, data ownership, and data stewardship. For example, data entity with industry classification codes is captured in systems A, B, and C, and the integrated data is stored in the data warehouse from where it is provisioned to the downstream systems. The industry classification codes which are themselves external reference data are stored in reference data tables in the data warehouse as per the architectural decision. There should be processes to ensure that the systems A, B, and C are providing the latest industry classification codes when they send entity information, the industry

classification code reference data in the data warehouse is reviewed at regular intervals, the review cycle SLAs have been agreed and approved by all stakeholders, is current, and controls are in place to flag incorrect industry classification codes, and roles and responsibilities for ensuring good quality data and resolution of data issues for industry classification codes has been established.

Data modeling is a significant component of data architecture. Data models—conceptual, logical, and physical models document the data design at different levels of abstraction.

Conceptual data models present the entities that are represented in the database and identify the highest-level relationships between the different entities.

Logical data models present the entities and relationships within and among them, with details about all attributes for each entity key structure (the attributes needed to define a unique instance of an entity and attributes to identify the relationship between different entities).

Physical data models represent the way the model will be built in the database and how data are physically stored in a database. They define the physical characteristics of the individual data elements that are required to set up and store actual data about the entities represented (Sebastian-Coleman 2012). A physical database model displays all table structures, including column name, column data type, column length, column constraints, key structures (for example, primary key and foreign key), and relationships between tables (Mahanti 2019).

Data architecture identifies what data should be the focus of data governance with data models assisting in defining the critical data elements and data governance helps in prioritizing the data for improvement (Turner 2018). Data architects have knowledge of the different data sources and an understanding of their structure, content and importance. They often act as business contacts for data governance activities. Data architects and data stewards should be assigned to each data domain and should work together to create a data asset register containing critical data elements.

Data modeling standards and guidelines defined by data architecture ensure consistency of data structures across the organization and governance ensures adherence to these standards and guidelines. Metrics can be used to assess the degree of adherence to these standards. The number of data elements not conforming to the naming standards is an example of such a metric. Some other examples to assess the quality of data models are completeness, correctness, structural soundness, correctness of definitions, and extent of alignment of the data stored in the fields with the corresponding metadata outlined for them.

Data architecture provides inputs to decisions about how the data should be governed. In addition to data models and data flow diagrams, other data architecture artifacts include data definitions and data standards, data dictionaries, data lineage, master data sets, reference data sets, reference data register, critical data elements information, information asset register, metadata, data store, data source inventories, and data policies related to the usage and management of data (Turner 2018).

These artifacts help the data governance team in data discovery, to make correct decisions about data policies and standards as well as assist them in performing root

cause analysis when data issues are raised by the business users, as well as establish responsibilities and processes around the different artifacts. For example, artifacts such data definitions, data lineage, and reference data register need to be up-to-date, and data governance helps in establishing processes, roles, and responsibilities for the same.

Data inventory and data flow diagrams help in identifying the possible business impacts associated with improving the data quality in the systems by understanding who uses the systems, for what purpose, and assist in the creation of metrics for tracking adherence and improvements. Additionally, these diagrams can also help to determine how to measure adherence to standards based on who creates and updates the data and in which systems.

Data governance helps establish accountability for key data architecture artifacts (example, ownership, decision rights, and business rules) and helps build the business case for data architecture. Data inventory and data flow diagrams overlaid with data accountability and ownership are key in identifying any gaps in accountability and ownership (O'Neal 2017).

Data governance ensures that metrics are in place to measure data architecture capabilities including data modeling, in line with guiding principles. For example, one of the guiding principles of data architecture is to reduce data replication. Some examples of data architecture capabilities metric along this guideline are number of data replications retired, reduced data storage cost, number of technical debts removed, and the number of data elements definitions standardized (O'Neal 2018).

The concepts—data governance and data architecture are inter-related, so processes that may appear to be wholly related to data architecture can play a key part in data governance. Data driven organizations need a strong foundation of data governance and data architecture. In addition, there should be good alignment between the enterprise data architecture and the data governance organization. Figure 3.3 shows the alignment between data governance and data architecture.

3.4 Data Governance, Data Integration, and Data Interoperability

Data Integration and Interoperability (DII) relates to the acquisition/capture, movement, consolidation and delivery of data within and between data stores, IT systems/applications, and enterprises.

Data integration is the process of combining data from multiple heterogeneous data sources into consistent forms, so as to provide users with an integrated and complete view of all the data. These heterogeneous data sources are autonomous and usually have different formats. Data integration is a complicated process. The complications arise from the need to consolidate data from the different systems that have different formats, standards, and quality levels into a single consistent format in the target system. Data interoperability is the capability for several systems to interconnect.

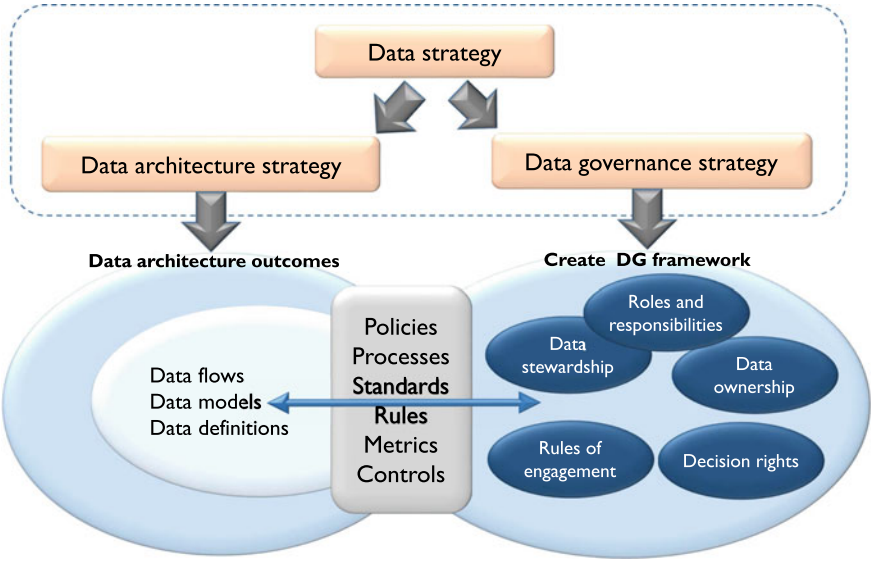

Fig. 3.3 Alignment between data governance and data architecture

Data governance provides oversight to DII solutions in the following areas (summarized in Fig. 3.4).

3.4.1 Stakeholder Engagement and Management

Data integration involves sourcing and consolidating similar data from disparate autonomous sources into a consistent format for storage in the target system which have multiple consumers. Hence, multiple teams, groups, and departments are involved. Sometimes data needs to be obtained from outside the organization too. A collaborative, trusting relationship, and ground rules for operation needs to be established.

It is important to identify the business and IT stakeholders, including the business subject matter experts and technology experts for all these systems. It is also essential to establish decision rights, roles, responsibilities, ownership, and conflict resolution processes. Data governance ensures that stakeholders are identified, and their roles and responsibilities, and accountabilities are established and enforced in relation to source data quality, target data quality, data model, data mapping and business rules for transformation of data. Business data stewards, technical data stewards, and data architects need to work together. However, the decisions relating to data quality and business rules should be driven by the business who are the consumers of the data. Data integration solutions often have multiple consumers of data, and there needs to be a consensus with respect to the decisions among these stakeholders and decision rights need to be defined.

Fig. 3.4 Role of data governance in DII solutions

3.4.2 Establish Governance Policies, Processes, and Best Practices

Business and IT should work together to create governance policies, processes, and best practices related to data integration in line with the organization's strategy, structure, enterprise architecture, environment, and regulatory mandates.

The policies and processes need to be established to ensure that the organization benefits from an enterprise approach to DII. For example, policies around the data quality of the source systems, processes, and artifacts to raise and resolve data quality issues need to be in place to have effective DII.

The policies, processes, and best practices should be documented and reviewed periodically and revised when necessary. The policies, processes, and best practices documents should be used to guide current and future data integration efforts.

3.4.3 Metadata Management and Data Lineage

Data lineage is the capture of data flow/movement and data transformations from the data source through intermediary systems to the target systems. Metadata management is the vital input to capturing enterprise data flow through various systems in an organization and presenting data lineage. It consists of metadata collection, integration, usage, and repository maintenance (Jain and Thomson 2013).

Development of DII solutions involves identifying the potential data sources, understanding, and profiling data elements in these sources. This in turn requires studying the metadata as a part of DII solutioning as well as capturing existing metadata and new metadata created as a part of DII solutions and storing it in a metadata repository or documents.

Metadata should be managed and maintained so that it is up-to date. Data governance ensures that metadata is up-to-date. We have discussed data governance and metadata management in the "*Data Governance and Metadata Management*" section of this chapter.

Data integration involves consolidating data from disparate systems in the organization and sometimes even from systems from different organizations. Data generally flows through the intermediate systems before being stored in the target system and hence good data lineage is important in DII solutions.

Data consumers need to be able to see the flow of data and know of the data origins in order to be able trust the data and have confidence in the data they use. Having good data lineage provides an end-to-end view and a means for the users to trace the data back to the sources, validate, and confirm that the data in the target systems are from trusted, authoritative sources, the data has been transformed as per business rules and there are adequate controls in place to govern the hand-offs between systems (Urso 2018).

Data governance ensures that data lineage information is documented and is up-to-date, that is, the data lineage is updated whenever changes are made to the data flow.

3.4.4 Security and Privacy

Different data sources involved in data integration have their separate security access controls and data protection mechanisms in place. When integrating data from these different data sources, it is imperative that an appropriate level of protection is provided to the consolidated data in the target system. The source system stakeholders (business and IT), target system stakeholders (business and IT) and the organization's IT security need to work together to ensure that adequate data protection measures are in place.

Data protection measures may include encrypting data and using appropriate authentication tools. Some data integration platforms have capabilities like role-based authentication and data encryption built in (Harvey 2018). Role-and

policy-based access controls are essential to govern, protect, and audit data and associated entitlements (Allen). If these controls are not effectively managed, there is a risk of unauthorized access and/or manipulation, loss or theft of data.

3.4.5 Data Sharing Agreements

DII solutions require multiple systems to communicate with one another. Common standards need to be established for sharing information. The common terminology used in identifying data sets and protocols should be reviewed by all stakeholders to avoid misunderstandings or problems in data transfer. On the other hand, different naming terminologies used for common data sets should be reviewed and standardized to avoid data transfer issues. Data transfer between systems requires data sharing agreement and a consistent process for sharing data, prior to the development of the interfaces. Data sharing agreement contains the responsibilities in relation to data sharing, purpose of data sharing and data usage or outputs prepared (for example, reports), data protection and security requirements, process to address data breaches, who should have access to the data, data availability times and SLAs. Data sharing agreement should be approved by the business and technical data stewards of the data in question. Additional approvals might be needed depending on the data in question. For example, sensitive data might need approval from risk, legal and/or compliance too. Data governance ensures that roles, responsibilities, and processes are in place for creation of data sharing agreements, and standards are in place to ensure consistency of protocols used in data transfer.

3.4.6 Data Integration Metrics

Effective data governance ensures that data integration metrics are measured, reported, and roles and responsibilities for the metric measurement and reporting have been established. Metrics are important to measure the ROI of data integration implementation. Data integration results in better data quality as it consolidates data sitting in silos in different sources. Therefore, data quality metrics such as accuracy, completeness, currency, and timeliness can help measure the benefits of data integration. The number of manual processes eliminated when automating integration, and the number of users using the data integration solutions can also be used to track success.

3.5 Data Governance and Reference Data Management

Reference data is widely used and shared throughout the organization, and is present in virtually every database and hence, need to managed and governed effectively.

Reference Data Management: It is the management of reference data and consists of a number of activities such as definition of reference data; maintenance of the permissible values/codes, descriptions, and definitions; standardization of codes; and sharing of reference data in a timely and consistent fashion across the enterprise to ensure high quality reference data is associated with master data as well as transaction data.

3.5.1 What is Reference Data?

Reference data are sets of permissible values or codes and their corresponding descriptions that are referenced, shared, and used by multiple systems, applications, data repositories, business processes, and reports, as well as other data like transactional and master data records (Mahanti 2019).

Reference data is any kind of data that is used for the sole purpose of categorizing other data found in a database, or solely for relating data in a database to information beyond the boundaries of the enterprise (Chisholm 2008).

3.5.2 Reference Data Categories

Reference data can be broken into two categories—

- Internal reference data, and
- External reference data.

3.5.2.1 Internal Reference Data

Internal reference data is reference data that are specific to a particular organization, and are created to describe or standardize their own internal business data for business concepts (for example status codes), to provide consistency across the organization. Internal reference data is managed entirely within an organization (Mahanti 2019).

3.5.2.2 External Reference Data

External reference data is reference data that is created and maintained by authorities outside the enterprise in order to provide and mandate standard values or terms to be used in transactions by specific or multiple industry sectors. This helps to reduce failure of transactions between different organizations and improve compliance by eliminating ambiguity of the terms. For example, the International

Organization for Standardization (ISO) defines and maintains currency codes (that are defined in ISO-4127) and country codes (that are defined in ISO 3166-1) (ISO/IEC 2008). Other examples of external reference data are SIC (Standard Industry Classification) codes and NACE (Nomenclature des Activités Économiques dans la Communauté Européenne) codes.

Reference data is extremely important, widely used, and needs to be effectively managed. However, the perceived structural simplicity of reference data, its relatively low volume, and slow rate of change, results in reference data being often ignored. However, in reality, reference data tables constitute around 20%–50% of the tables in a database, and tables containing same or similar reference data are used by multiple applications. Hence, they get widely duplicated across many applications in the organization, resulting in inconsistencies. Since reference data are used by master data as well as transaction data, data quality issues in reference data can have widespread impacts, such as errors in reporting and data integration (Chisholm).

3.5.3 Reference Data Governance

Governance around internal and external reference data involves ensuring that:

- Polices are in place for the management of internal and external reference data, and ensuring that these policies are enforced.
- Reference data inconsistencies can result in significant quality issues and have a considerable cost impact. Reference data management involves standardization of reference data, and data governance provides a layer of oversight and ensures that these standards are adhered to.
- Each business concept represented in reference data has formal assigned accountabilities and at the same time ensuring that these accountabilities are known across the enterprise (Chisholm) and has business data stewards assigned who are responsible for its quality. Since reference data is used by multiple business processes and systems in the organization, accountabilities and decision rights need to be established as to whose word will be final in case of a conflict. The responsibilities need to be established to answer questions—who can make changes to the reference data, definitions, values, and who has the approving authority. A detailed RACI (responsible, accountable, consulted, and informed) matrix should be in place to map various reference data management activities to the roles and responsibilities.
- Processes are in place for managing reference data so accountable parties have a standardized approach when creating reference data and making changes to them. Also, processes are in place for routing changes required in the internal reference data sets to the accountable parties so that these can be updated in a proactive and timely manner. This is especially important for reference data that are shared across multiple business process and systems.

- Ensuring each internal reference dataset is periodically reviewed for relevance, and kept up to date with business changes, and internal reference datasets that have relationships or dependencies are harmonized, especially in terms of update cycles (Chisholm).
- External reference data is reviewed at periodic intervals in line with review schedules set by the external organization that is the owner of the data. A process for review and update for the same needs to be set up. For example, ISO-3166-1 country codes and ISO-4217 currency codes are reviewed by ISO at regular intervals and organizations need to have a process to review and update these codes once ISO has published a revised version.
- Reference datasets updates are socialized and effectively distributed across the enterprise (Chisholm).
- Accountable parties for internal reference datasets provide adequate support for these datasets (Chisholm).
- Metrics are in place to monitor the quality of reference datasets.

External reference data is provided by authorities outside the organization. Hence, the external organization who provides the data, needs to accountable for its quality. However, it still needs to be discovered, selected, understood, ingested, and maintained in the organization to ensure that its integrity is preserved; hence policies, processes, controls, and roles and responsibilities should be there to ensure that the data is adequately governed. It is best to have a centralized model for external reference data management and governance, as ongoing maintenance is required after the initial ingest of an external reference dataset and in the absence of centralized body for governance, maintenance is often neglected. However, a more federated model is suitable for internal reference management because it is created and managed by many different subject matter experts (SMEs), with a centralized governance body creating policies, practices and accountabilities and ensuring the individual groups headed by the subject matter experts follow the standardized approach. The reference data governance body should have a good understanding of reference data management and have adequate domain knowledge of the related reference data (Chisholm).

3.6 Data Governance and Master Data Management

Master data is high value, key business information that defines attributes of instances of core business entities (for example—customers, products, and employees) in organizations. Master data are non-transactional data. Master data supports transactions (Mahanti 2019).

Master data management (MDM) is the sum total of processes, policies, standards, tools, and technologies used to create, store, consolidate, and enrich an organization's master data to ensure their quality, stewardship, and accountability.

Master data management enables creation of a single authoritative view of data or single version of truth also known as the golden record of data.

There is a common misunderstanding that data governance is the same as master data management. While successful implementation of MDM requires integration of data governance throughout the initiative, it is not the same as master data management. MDM creates one single source of truth for your master data. Data governance ensures agreement in identifying the master data elements associated with common business terminology, examining their authoritative sources, agreeing on their definitions and data standards, setting and enforcing policies, data rules, and accountabilities.

Successful implementation of MDM requires integration of data governance throughout the initiative. As stated by Aaron Zornes (2012)—

> MDM Without Governance…is Just Data Integration!

Lack of data governance integration in MDM projects to manage the people, processes, and data is one of the main reasons behind unsuccessful or underperforming MDM projects.

> The data governance, prioritization, people and process aspects of implementing an MDM solution will likely derail the project before the technology fails—Wang and Karel (Thomas 2008).

Since master data spans across application and business unit boundaries, master data management programs are often implemented in parallel with business data governance initiatives. Data governance plays a major role in master data management by assisting in the maintenance of the single version of truth and preventing data silos. It provides the basis required to address the business and data issues relating to master data.

Data governance has a critical role to play in master data management as follows and summarized in Fig. 3.5.

3.6.1 Agreement and Management of Critical Master Data Elements

Organizations have a lot of master data elements but not all of them are critical. Master data elements that are critical are candidates for both data governance and master data management. Generally, data governance helps ensuring agreement identifying the master data elements associated with common business terminology, scrutinizing their sources, determining the authoritative sources, getting consensus on which of the data should be mastered (that is critical master data) taking into consideration, the business unit functions and usages of the mastered data, and getting consensus on their definitions, data standards, and the associated metadata (for example, data length and data type), and managing them within the master data repository as the enterprise source of truth.

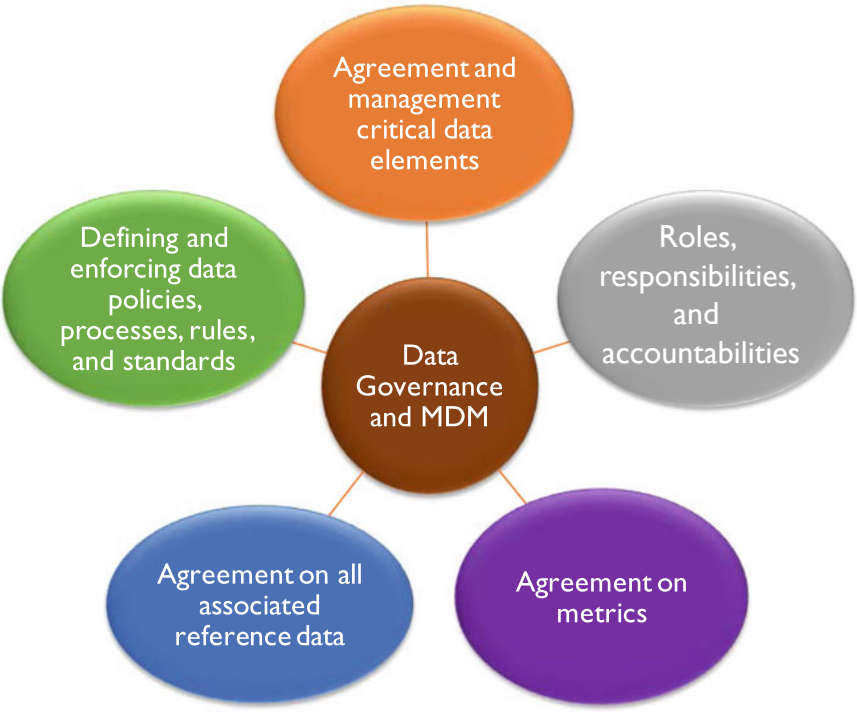

Fig. 3.5 Role of data governance in master data management

There can be multiple sources for the same data, and there needs to be an agreement on the sources from which data will be sourced from for the building the master data repository (Loshin 2013; Fryman 2017).

3.6.2 Defining and Enforcing Data Policies, Processes, Rules, and Standards

This involves determining the critical business policies that relate to data, and formulation of information policies that comprise of the specification of management objectives associated with data governance and master data management. Business rules need to be established for various activities involved in master data management (for example, master data profiling, data quality rules, and data consolidation rules) and governance provides a layer of oversight.

The business rules that govern the acquisition/creation, management, archive, and purging of the data that is being mastered, may vary by the type of data, applications and business unit functions, and individual departments/business units. These may

also be governed by applicable laws and regulations. Processes and standards need to be created for managing master data issues and for conflict resolution.

There needs to be an agreement on the business rules and policies that will be enforced for the management of the "golden record data" (Fryman 2017).

In addition, there needs to be consensus amongst all stakeholders for data quality rules and data integration rules for the golden record. These rules and policies need to address the needs of all the business units. All stakeholders should agree on the processes and standards that will be used to manage MDM data issues.

3.6.3 Roles, Responsibilities, and Accountabilities

Since master data management involves several stakeholders, the roles, responsibilities, accountabilities, and decision rights need to be established with a detailed RACI (responsible, accountable, consulted, and informed) matrices for each master data domain (Aykose 2019).

The right individuals in the organization should be authorized and empowered to administer well-defined governance policies and to institute the underlying organizational structure to make it possible by defining a management structure to manage the execution of the governance framework (Loshin 2013). Data ownership and data stewardship needs to be established for different master data elements. There should be an agreement on the business and technical resources that will be accountable for the mastered data.

Some master data, for example, customer data that are captured and used by several business units, need business data stewards and technical data stewards from multiple business units and applications to manage the master data. Decision rights need to be established for conflict resolution.

3.6.4 Agreement on Metrics

There needs to be an agreement on metrics such as metrics for gauging progress of the MDM implementation, metrics for master data quality, metrics for data issues and metrics for the governance of the MDM data (Fryman 2017). Governance teams need to work with data stewards for establishing data quality metrics and ensuring that the performance of MDM is tracked.

3.6.5 Agreement on All Associated Reference Data

Master data references reference data and hence, there needs to be an agreement on all the associated reference data that will be used with the mastered data. For

example, customer is a master data but the customer statuses such as active, inactive, deceased, and insolvent are usually maintained as reference data. Similarly, for address master data, state codes, and country codes are usually stored as reference data and address master data reference these data elements.

Reference data requirements should be addressed early on by the data governance team when the critical master data elements are being determined.

3.7 Data Governance, Data Warehousing, and Business Intelligence

Just like a warehouse stores goods transported from different sources in one place, a data warehouse integrates and stores data from different data sources (operational data stores, business applications like customer relationship management and enterprise resource planning systems, or from systems external to the organizations) in one place in the organization. A data warehouse is a specific case of data integration. Figure 3.6 shows the high-level view of the data warehouse.

While the data stored in the data warehouse is also available in the source systems, the source system data is not in a form suitable for business intelligence, analytical processing, and reporting purposes. Data is populated in a data warehouse through a process called Extract, Transform, and Load (ETL) as shown in Fig. 3.6.

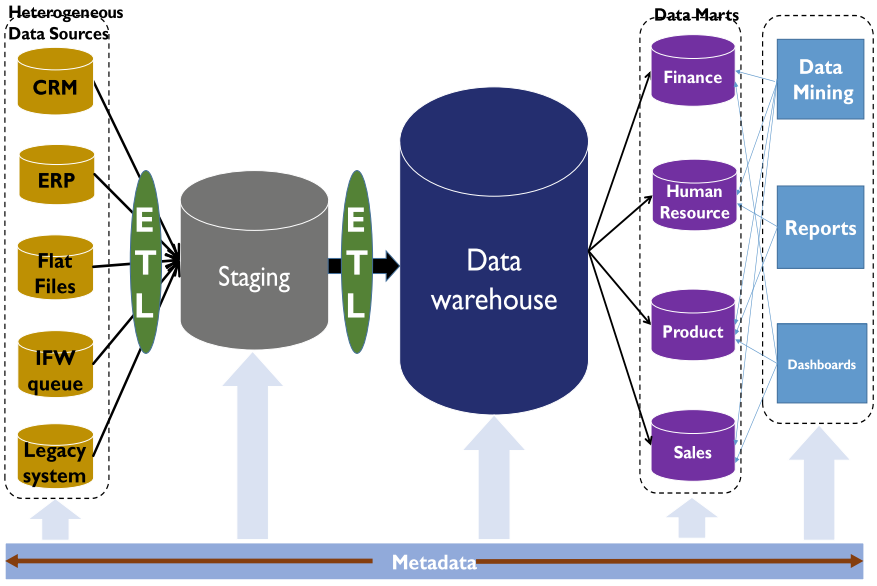

Fig. 3.6 Data warehouse—a high level view

"Extract" involves extraction of data from sources internal and/or external to the organization. Extraction of data is followed by "Transform" which involves transforming data into a standard format as per business rules, which includes validation rules so that the data in the data warehouse is understandable to and usable by the organization's analytical community and management. Some examples of transforming data are transforming dates to the same format, splitting full name into title, first name, middle name and last name, and designating fixed values for gender. "Load" involves the transformed data being loaded into the tables in the data warehouse.

Business intelligence (BI) is a blanket term that consists of the applications, infrastructure, tools, and best practices that enable access to, analysis, conversion of data, and visualization (such as reports and dashboards) to drive, improve, and optimize decisions and performance (Gartner 2019).

The results of a global survey of more than 400 senior IT and business professionals carried out by Forbes in association with Qlik reveals that executives realize the value of governance in BI, as more than three quarters of the executives (78%), say that data governance is either vital or important to their BI operations (Forbes 2016).

Many organizations equate data governance with data storage in a central repository, and have a perception that data governance is about data storage in a central repository like a data warehouse. However, it is the roles and responsibilities, decision rights, processes, policies, standards, rules and controls around storing, processing, and accessing critical data from the repository to ensure that it is secured effectively and that the data and the associated metadata stored is of high quality and meets business needs, that forms a part of data governance.

A data warehouse initiative is cross functional as a data warehouse sources data from numerous heterogeneous sources that belong to different functional areas or data domains and multiple stakeholders are involved in data warehousing.

Data governance involves activities around identifying these stakeholders, establishing decision rights, and clarifying accountabilities. A detailed RACI (responsible, accountable, consulted, and informed) matrix should be in place to map various activities to the roles and responsibilities. With multiple stakeholders, there is bound to be conflict, and data governance establishes a conflict resolution process and establishes who would have the final say, in case of a deadlock.

In order for a data warehousing initiative to be successful, the data in data warehouse should meet quality thresholds defined by the consumers. In order for this to happen, the source systems need to meet quality thresholds and data from the different source systems needs to be integrated, transformed, and stored in the data warehouse such that the data meets all the stakeholders' requirements.

Data governance ensures that these quality thresholds are defined, agreed upon, and accountabilities are in place that data meets the quality thresholds and controls are in place to monitor the same. Processes should be defined by data governance team to log data issues. Working groups consisting of business data stewards and technical data stewards who are responsible for data in each of the different source systems and data consumers should be held at regular intervals (usually weekly) to

resolve data issues in case quality thresholds are not met. A data governance program can help ensure good quality data in the data warehouse.

Before implementing a data warehouse, organizations should have a data governance framework that define the accountabilities, the rules, the policies, and the procedures organizations use for deciding which data should be kept, for how long, in what format, the security requirements around data, who should and shouldn't have access to what data, the data backup strategy, and the data standards to be adhered to.

Data governance should oversee the data warehouse lifecycle and assess the impact of the new data policies and processes on the existing business processes. Data governance in conjunction with the data warehouse architecture ensures creation of policies, rules, and standards related to data warehouse model, naming conventions, warehouse data creation/modifications, deletion related rules, and ensures the enforcement of these policies, rules, and standards.

When new data elements are being introduced into the data warehouse, data governance ensures that standards are being adhered to. Data warehouses results in a lot of metadata discovery and creation, and data governance plays an important role in overseeing the creation and management of metadata repositories and documentation of data lineage.

An effective data governance program ensures that controls and data quality metrics are agreed upon with the stakeholders, and are used to monitor the quality of critical warehouse data along the established data quality dimensions. It also assigns accountabilities for the ongoing monitoring of the quality of critical data being loaded into the warehouse and outlines a process for issue management (reporting, tracking, and resolution). Some of the data quality metrics will be discussed in the third book of the series—Data Governance Success.

A data warehouse which contains historical data for different data domains in an organization, serves as the foundation for analytical and decision making, querying and reporting, and, therefore, needs to take into consideration the data requirements across the enterprise, and requires consensus from all stakeholders that a robust data governance program would provide.

Data warehouse initiatives require organizations to make many decisions that involve data from several heterogeneous sources, to enable cross-application analysis and rules for consolidating and transforming data that needs to be defined and recorded. In addition, a data governance program for a data warehouse can also provide analysis for external data that is brought into the warehouse, can offer the oversight to enforce standards and rules after the decision support system becomes operational, and the process to resolve data quality issues (Smith 2009).

Data mart is a subset of the data warehouse and contains data related to a particular business unit (like sales, human resources, and finance). Data marts also need the same level of data governance as data warehouses.

An organizational approach to data governance is required to provide the oversight and guidance for decision-oriented data across business units in an organization, to enable data mart data to be useful across those divisions. A decision-support program may start small (a small data warehouse with data

related to a few critical subject areas or a small set of integrated marts) and grow in a staggered fashion to an enterprise approach as data governance is accepted and sustained across the organization (Smith 2009).

Data governance program as part of a data warehouse initiative has a number of benefits including more informed decisions driven by good quality data, reduction of data duplication, inconsistencies and conflicting definitions/standards/calculations, statutory and regulatory reporting using high quality data reducing risk of non-compliance, security mechanisms and role based access to prevent unauthorized access and manipulation, and an integrated approach to data management and usage throughout the enterprise (Smith 2009).

3.8 Data Governance and Data Migration

Data migration projects deal with the migration of data from one data structure to another data structure or transformed from one platform to another platform with a modified data structure (Mahanti 2019).

The data definitions are different in the source system and target system. If the source system involved in data migration is a legacy system, since often the legacy systems are department based, the concept of formal data definitions and standards, and the processes needed to manage them across functional boundaries simply don't exist (Standen 2009). Data governance will be needed to facilitate the creation of these processes.

Data migration projects are perceived as technical centric exercise and are often outsourced and not given sufficient oversight, which in turn greatly slows down an organization's digital transformation progress.

While data migration projects are one off projects, they should be checked against the data policies and standards and should be approved by the governing body. Working groups should be set up to understand data definition differences between the systems involved, and ownership and accountabilities for different activities involved in data migration need to be established. The data migration should meet data conversion standards and data quality standards. With policies containing data quality requirements and standards for compliance, communicating and enforcing the same will ensure that data that is migrated to the target system meets the quality threshold. Also, with effective data governance, there is clarity on roles and responsibilities with respect to rights to create, approve, edit or remove data from the source system.

While data migration is a one-off exercise, it is an opportunity to put in place a true data governance strategy (if one does not exist) including aspects such as data management processes, organization responsible for the data quality and data quality measures. Putting in place such a strategy implies (Clément et al. 2010):

- Defining the role of each teams/individuals in relation to the data: data creators, data custodians, data consumers (Burgess 2007).
- Documenting and enforcing data management processes related to data entry, data quality, data maintenance, and data administration.
- Publishing data quality measures at regular intervals, and raising a flag if data quality does not meet the optimum quality level defined during the users' interviews. Data quality is a continuous improvement cycle (Weber et al. 2009); it is therefore beneficial to capitalize on the learning and the experience gathered from the data migration, and to put in place a robust well-defined data governance strategy (Clément et al. 2010).

3.9 Data Governance and Metadata Management

There is a common misconception that data governance is metadata management. While metadata management is a fundamental component of data governance initiatives, metadata management is not the same as data governance.

Data governance uses metadata management to enable policies related to data definition, data usage, data security, data lineage and heritage. However, a great deal of metadata management activities is manual that can be time taking.

As stressed by Dr. John Talburt, Acxiom Chair of Information Quality at the University of Arkansas at Little Rock, and Lead Consultant for Data Governance and Data Integration with Noetic Partners Inc., metadata automation can enable data governance, and one of the greatest problems faced in data governance is the lack of metadata automation. In his words, "*...Even though we now understand how important it is to curate and track all data across the organization, actually doing it is still largely a manual operation. While many software companies are beginning to address these problems and produce DG specific software tools, we still have a long way to go. Large enterprises need an army of data stewards to describe, classify, and manage their data assets. Because so much of traditional computing is only focused on the process and not on the data, tracking where data inputs come from, how the data are transformed, and where the outputs go has largely been left to be recorded and managed by people.*"

Systems should generate metadata with data and use the metadata with metadata automation acting as a data governance enabler by preventing the need to track the data manually. In a nutshell, data governance uses metadata management to force management discipline on the collection, discovery, control of data, and mitigate risks associated with the data.

Christian Bremeau, CEO and President of Meta Integration Technology states (Dennis 2017) that,

> a solution to both Metadata Management and Data Governance should be integrated in order to succeed.

Metadata management and data governance work hand in hand to implement the right controls on enterprise data (Ghosh 2018). Metadata is a type of data and as such needs to be governed like other data types like master data and reference data.

As highlighted by Dr. John Talburt, Acxiom Chair of Information Quality at the University of Arkansas at Little Rock, and Lead Consultant for Data Governance and Data Integration with Noetic Partners Inc., in his interview, "*we need to think in terms of thousands of bytes of metadata for each byte of data. Just think of the business and technical descriptions and requirements, data movement and provenance, data classification, data access control, data quality, and the many other metadata relevant to just one item of operational data.*"

While organizations have a large volume of metadata (each data object and element has associated business and technical metadata), managing and governing all of it would be overwhelming and costly exercise, and one that is not required.

Not all metadata need to be managed and governed with the same rigor. As a general rule, the more critical the data is and the more widely the data is shared across and beyond organization, the associated metadata also becomes critical and the more formal metadata management and governance needs to be in place for the associated metadata.

Metadata helps break data silos—when the same data entity or element sits in different locations in an organization, what ties or links them together is the metadata.

For example, data element, "customer date of birth" which is an attribute of the data entity—customer and which represent the real world entity—customer can be known as date of birth in one system, consumer DOB in the second system, DOB in the third system, and customer DOB in the fourth system. Without metadata to indicate that all these data elements allude to the data of birth of the customer, it is very difficult to determine what they exactly mean and that these data elements store the same information—the data of birth of a customer. Without metadata, data would not have context and data governance would not have a base to work with.

Amnon Drori, CEO of Octopai highlights the importance of metadata management in data when he states that data governance success depends on the effectiveness of metadata management as governance requires understanding the context, journey, or data lineage which is not possible without metadata management (Ziton 2018).

Metadata comes in many forms and is stored in various locations in the organization. Some sources of metadata are as below and shown in Fig. 3.7:

- Word documents,
- Excel spreadsheets,
- Software tools (for example, data modeling and data integration tools),
- Applications,
- E-commerce, and
- Websites.

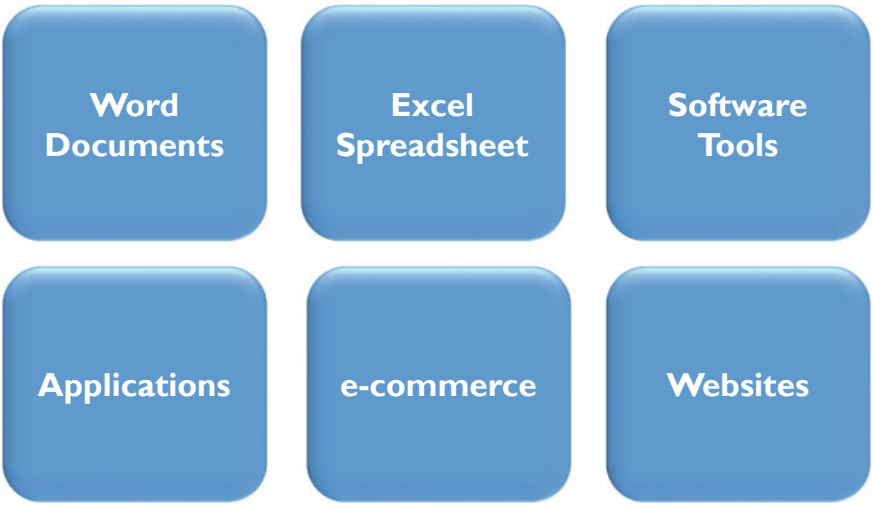

Fig. 3.7 Different sources of metadata

Metadata can be formal, structured, and documented; they can also be informal and unstructured; they can also not be documented and only reside within the heads of selected individuals (generally, data subject matter experts) in the organization.

An organization that implements a data governance program without addressing metadata management will be unsuccessful because many of the activities and tasks carried out by data stewards are focused on metadata and revolve around the process of managing it (Smith 2015). On the contrary, effective metadata management empowers the data governance and data stewardship teams to monitor conformance to corporate data expectations as well as enforce alignment with corporate data policies (Loshin).

Data governance uses metadata to enforce management discipline on the collection and control of data (InfoLibrarian Corporation). Data governance helps establish and enforce policies, standards, formal roles, and responsibilities around critical metadata and holds owners accountable for the quality of metadata. For example, data (for example, customer data—customer names, and addresses) are stored in different databases in an organization and are used and shared by multiple applications, business units, and departments. Without standards and consistency around metadata associated with the data, there would be inconsistencies in the data element definitions and conflicting sources of customer metadata resulting in metadata that is not-reliable. This would result in problems around data usage and decision making. For example,

- In Database 1, customer name is represented by the data element CUSTOMER NAME and is a string field which can hold 255 characters, except for blanks, ".", and ""; no special characters are allowed.

- In Database 2, customer name is represented by the data element C_NAME and is a string field which can hold 100 characters, except for blanks and "."; no special characters are allowed.
- In Database 3, customer name is represented by the data element CLIENT_NAME and is a string field which can hold 100 characters, but no special characters other than blanks are allowed.

The above would result in inconsistent data across systems and since different terminologies were used for the same business term, it would be difficult to determine whether they all represent the same thing.

Standardizing data and metadata and their upkeep is accomplished through the creation and enforcement of data policies as a part of data governance and metadata management. Standardization effort not only promotes consistency of data across the organization and thereby improves data quality but also boosts business users' confidence that the data are accurate and reduces risk of data misuse.

There are a number of established international standards for metadata structure, and additional guidance on strategy and implementation has been provided by standards groups, such as the International Organization for Standardization (ISO), the American National Standards Institute/National Information Standards Organization (ANSI/NISO), and other bodies, such as the Dublin Core Metadata Initiative (DCMI) (Smallwood 2014).

Effective metadata management involves architectural components, people, and processes for collecting, storing, maintaining, distributing, and managing access to metadata in a methodical fashion, across the organization.

This involves creation and maintenance of a metadata repository, which is a database storing the integrated metadata sourced from different metadata sources (such as spreadsheets, documents, software tools, applications, websites, and e-commerce). The metadata stored in the metadata repository is typically used by multiple users, tools, and applications and controlled to eradicate inconsistencies and redundancies.

The metadata model also known as a meta model is the physical model designed to store the metadata in the metadata repository. It has metadata elements, tables, and relationships. A metadata repository should have the following characteristics (Marco 2017) as summarized in Fig. 3.8.

- **Generic**: Generic means that the physical meta model looks to store metadata by metadata subject area as opposed to application-specific subject area.
- **Integrated**: The metadata repository also provides an integrated view of an organization's major metadata subject areas. The repository should allow the user to view all entities within the organization, and not give a limited view of just entities loaded in a particular database or entities pertaining to an application.
- **Current**: The metadata repository contains current metadata, as in the metadata that is not outdated. In other words, the metadata should be periodically updated to reflect the current technical and business environment.

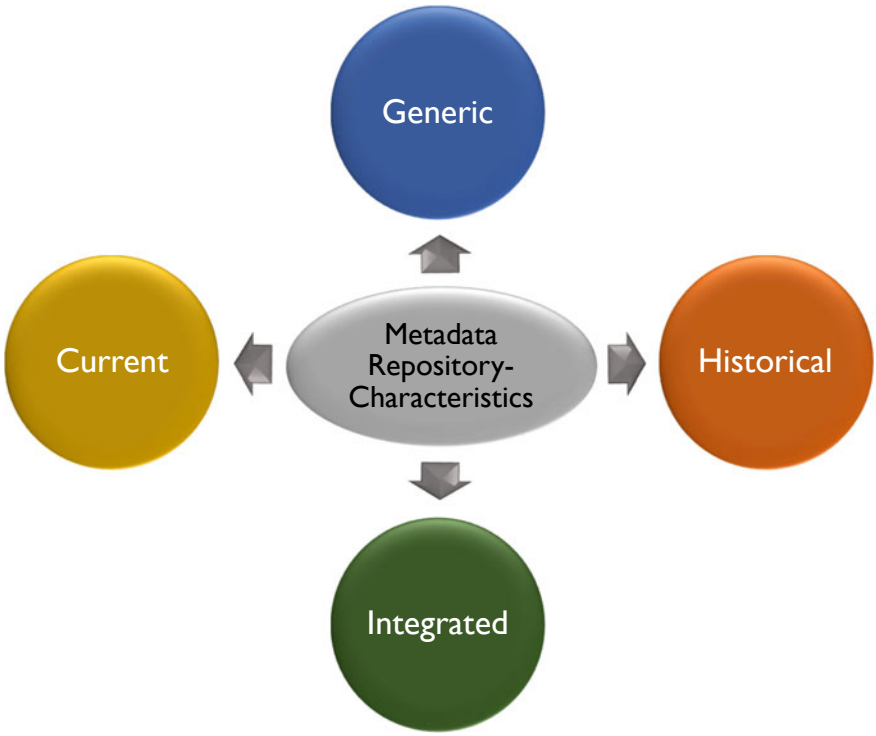

Fig. 3.8 Metadata repository characteristics (adapted from Marco 2017)

- **Historical**: Metadata repositories should contain the historical metadata too. A good repository will also hold historical views of the metadata, even as the new view of the metadata is in place to reflect the current technical and business environment. The different views of the metadata at different points in time help the organization to understand how their business and technical environment has transformed at these points in time and what those changes were. This is critical when the repository supports a data warehouse (as data warehouses store historical data) (Marco and Jennings 2004).

The effective management of a metadata repository is one of the essential activities of a data steward within a governance practice enabling enforcement of data management policy as well as facilitating locating data easily (InfoLibrarian Corporation).

Data stewards and other data management professionals should ensure that there are processes in place to maintain metadata related to critical data elements and metadata should be reviewed periodically to ensure that it is up to date and if not, then it should be revised accordingly. Data owners should be held accountable for capture of metadata as in approving definitions and data quality rules and implementation of changes to the data. Stewards are also responsible for classification

and definition of the data entities and the data elements. The technical administration is the responsibility of a metadata repository administrator. Some of the roles such as data consumers and data modelers who are also consumers of metadata are actively involved in defining the metadata.

While data modelers are involved from a technical perspective, data consumers who are typically business users understand the data, the context of usage, different uses of data, and are best placed to provide definitions and domain values.

In absence of a metadata repository, the data stewards would be limited to working with manual processes, documents, and spreadsheets to accomplish their critical tasks.

Without effective management and governance of metadata, the organization would have siloes of metadata that will provide conflicting information. Data governance provides standards, policies, and controls to ensure integrity, quality and stability of both data and metadata.

Metadata governance involves looking at metadata roles, responsibilities, standards, lifecycles, statistics, and processes around how terminologies and definitions are categorized and updated, and ensuring that the metadata is up to data (as in changes in definition are reflected and truly represents the data), in addition to how operational activities and related data management projects integrate metadata (Knight 2017). Metadata is data about data so if there are changes in the data stored in a particular data field, processes should be there in place to ensure that the metadata is updated if needed. For example, say a data field—"Region" used to store country names but now stores continent names too, the definition of the "Region" field should be modified to reflect the same. Also, if a new data field has been created, then the metadata for the new field should be captured correctly and should be available when the new field is available in production.

Metadata is also data and the common data quality metrics such as accuracy, completeness, consistency, currency, volatility, and timeliness also apply to metadata. Governance ensure that metrics (for example, metadata quality statistics) are defined for critical metadata (metadata associated with critical data) to measure, monitor, help control the effectiveness of metadata management, and the degree of conformance and enforcement of respective data policies.

3.10 Data Governance, Document, and Content Management

Documents and contents not stored in databases need to be managed to justify the collection of sensitive data, to support and improve processes (business, IT, and data), and to ensure compliance with regulations. In addition to the large volume of data records stored in databases, organizations have a large amount of information stored electronically as well as non-electronically (that is in physical documents).

Most of the records and content created today are electronic; for example, email content, attachments, messaging, images, content on a website, digital documents,

and so on. A lot of the content is unstructured, and reside outside relational databases or data lakes. Document and content management encompasses the activities of the capture, storage, protection, access, and version control of unstructured data found outside relational databases and data lakes, and usually stored in digital files in an organization.

Document and content management includes two sub-functions:

- Document management, and
- Content management

3.10.1 Document Management

Document management is the storage, inventory, and control of structured documents. Some examples are contracts, tax documents, emails, and government/legal notices. Document management includes records management. A distinction between documents and records is that, documents are any documents created within an organization or brought into it during the day to day business activity, and that can be altered. Records are documents that are completed, and will not be altered (Docsvault 2020; Paperalt 2018).

Data governance is often confused with the term records management and both are thought to be synonymous. However, this is a misconception. The fact is that records management is an important component of data governance and a strong data governance program requires a unified records management strategy which gives records managers the ability to apply standard policies and classification schemes to the content stored in different applications as well as locations (Mahanti 2019; Alfresco 2018).

ARMA International published the Generally Accepted Recordkeeping Principles® (GAAP) in 2009 and revised these principles in 2017. GAAP constitute a generally accepted global standard that provides high-level framework of good practices for records management and information governance. The principles are as follows (ARMA International):

- **Principle of Accountability**: A senior executive (or a person of comparable authority) shall oversee the information governance program and delegate responsibility for information management to appropriate individuals.
- **Principle of Transparency**: An organization's business processes and activities, including its information governance program, shall be documented in an open and verifiable manner, and that documentation shall be available to all personnel and appropriate, interested parties.
- **Principle of Integrity**: An information governance program shall be constructed so the information assets generated by or managed for the organization have a reasonable guarantee of authenticity and reliability.
- **Principle of Protection**: An information governance program shall be constructed to ensure an appropriate level of protection to information assets that

are private, confidential, privileged, secret, classified, essential to business continuity, or that otherwise require protection.

- **Principle of Compliance**: An information governance program shall be constructed to comply with applicable laws, other binding authorities, and the organization's policies.
- **Principle of Availability**: An organization shall maintain its information assets in a manner that ensures their timely, efficient, and accurate retrieval.
- **Principle of Retention**: An organization shall maintain its information assets for an appropriate time, taking into account its legal, regulatory, fiscal, operational, and historical requirements.
- **Principle of Disposition**: An organization shall provide secure and appropriate disposition for information assets no longer required to be maintained, in compliance with applicable laws and the organization's policies.

Document management activities revolve around the identification of existing documents and records, creation of document and record policies, classification of documentation and records, and storing records and documents.

Document and record access, retrieval, circulation, review, revisions, retention, backup, recovery, destruction, disposal, audit, and monitoring should be in accordance with policies and processes. Appropriate controls should be there in place to prevent unauthorized access and modifications.

Document management systems are used to manage documents and records. It must be noted that document management systems do not create these documents and records.

3.10.2 Content Management

Content management refers to the processes, techniques, and technologies for organizing, categorizing, and structuring access to information content, resulting in effective retrieval and reuse.

Sometimes, content management is referred to as enterprise content management (ECM), implying the scope of content management is across the entire enterprise.

The activities involved in content management are classifying content, connecting, and relating content through tags and taxonomies, maintaining tags and taxonomies, documenting and indexing content metadata, publishing content, managing content, version control, review, maintenance, content archival, and destruction. These activities should be governed and should adhere to policies and processes. Appropriate access controls should be there in place to ensure that right individuals and/or teams have the right level of access to the content. Some might have access to view content, which some might have access to view as well as modify content.

Content management systems are used to manage content. It must be noted that content management systems do not create content.

3.10.3 Document Management System (DMS) Versus Content Management System (CMS)

Both DMS and CMS have centralized storage and administration, automated workflows, robust security features, review trails, advanced searching options, facilitate easy retrieval, and effective distribution of information (Business-Software.com 2020). Because of these similarities, they are often thought to be the same.

CMS are different from DMS in one important area—the type of information they manage. Document management solutions are designed specifically for data contained in structured documents and traditional files like Word, PowerPoint, Excel spreadsheets, PDF, and other popular formats (Business-Software.com 2020; Paperalt 2018).

Content management systems, on the other hand, are more about the logical organization and improved accessibility of various types of structured and unstructured electronic information. Content management systems can not only manage structured documents managed by document management applications, but a broader range of digital assets (Business-Software.com 2020; Paperalt 2018). Some examples are text, audio, video, Flash, infographics, and multimedia files, as well as raw data collected from various third-party Internet sources.

For effective document and content management, it is important to first identify the departments or business units who will be accountable for managing content, documents, and records. Given that there is a lot of content, document, and records, only critical content, documents, and records should be targeted when planning for managing content, documents, and records. Criticality of content, documents, and records can be ascertained by the following:

- Business value, and
- Associated risk level.

Level of control depends on the business value of the content and associated risk level. If the content is of high business value and the associated risk is high, rigorous controls will be needed to ensure protection against theft, unauthorized access and modification. On other hand, if the content is of low business value and the associated risk is low, light controls will be needed. Governance ensures the appropriate level of controls are there for the content, and audits are conducted at regular intervals to review controls, and modify controls if the criticality of the documents and content have changed.

Governance also establishes policies and processes that need to adhered to in relation to document and content management, with document and content management systems providing a framework for managing content, and creation, communication and enforcement of polices and processes. Governance also ensures the data content management policies are periodically reviewed and revised so that they are up to date.

Organizations need to have documented data content management policies that justify the collection of sensitive data, and should have processes and standards, rules and controls for managing that data. There should be polices and processes around record retention, access, privacy, protection, distribution, record preservation/archiving, and destruction in accordance with organizations needs, and legal and regulatory requirements.

The roles and responsibilities involved in the management of documents and contents should be established, and staff should be trained so that they understand the activities and processes and their specific responsibilities in relation to document and content management. Without having a shared understanding as to who is in charge and who is responsible for a specific content, record, document, and their management, content and document management will be inconsistent and chaotic.

There are several aspects of accountability involved—strategic authority to design content strategies and make decision on content management project projects so that they are in line with the overall organization strategic goals, implementation accountability for document and content management system and specialist/subject matter expert input to strengthen the content strategy. Some of the basic roles and responsibilities are:

- **Creator**—responsible for creating and editing content or documents.
- **Editor**—responsible for editing the content message and the style of delivery, including translation and localization (Wiki).
- **Publisher**—responsible for releasing the content or document for use.
- **Administrator**—responsible for managing level of access to folders and files, usually achieved by assigning access rights (such as view only and/or edit/write permissions) to appropriate user groups or roles. Administrators may also support users in other ways as in sorting environment issues (Wiki).
- **Consumer**—the person who reads or uses content or documents after it is published or shared.

While content and document management systems provide core control functions such as organization and search, permissions, state and workflow management, versioning, and dependency management, ownership and roles should be established, and accountable business units and departments should manage documents, contents, and records in accordance with the defined policies and processes.

3.11 Data Governance and Data Security Management

Data security is the extent to which access to data is restricted and regulated appropriately using suitable security mechanisms to prevent unauthorized access, modification, deletion and theft. Data security is a high priority issue, as the risk of not securing critical data is simply too high. Below points highlight the problem of not securing data:

- Each cyber incident costs U.S. companies a reported $7.1 million on average, or $221 per record (Source: IBM; Allen).
- 468 major breaches were recorded in 2011 and the number went up to 1,175 in 2012. In 2013, the number went up further and stood at 1,731, showing a steep rise in the number of breaches (Details—Veris Community Database; Allen).
- 63% of organizations do not have appropriate data security measures prior to new information technology installation (Source: Thales Security (2017); Allen).
- Hotel and hospitality powerhouse Marriott recently revealed a massive data breach that led to the theft of personal data for an astonishing 500 million customers of its Starwood hotels (Sandwell 2018).
- 51% of organizations have global data breach in the past five years, with 56% of these have had multiple data breaches (Akamai 2018; Ponemon Institute 2017).
- 1.9 billion data records were leaked or stolen during just the first half of 2017, well surpassing a total of 1.37 billion in all of 2016 (Gemalto; Akamai 2018).
- Average cost of an APT data breach is $18 million; 50% is damage to brand reputation (Akamai 2018; Ponemon Institute).
- The average cost of a lost or stolen record is $141 (Ponemon Institute 2017; Alfresco 2018).
- In 2017, the average number of breached records by country was 24,089. India had most breaches annually with over 33k files; the US had 28.5k (Ponemon Institute's 2017 Cost of Data Breach Study; Sobers 2019).
- In 2016, 3 billion Yahoo accounts were hacked and this was one of the biggest breaches of all time (Oath.com 2017; Sobers 2019).
- In 2016, Uber reported over 57 million riders' and drivers' information being stolen by hackers (Uber Newsroom 2016; Sobers 2019).
- In 2017, 412 million user accounts were stolen from Friendfinder's sites according to breach notification site LeakedSource (Sobers 2019).
- The 2017 Cyber Security Breaches Survey found 46% of all British businesses had suffered a data breach or cyber-attack in the last 12 months, rising to 67% for larger organizations (Fraser and Price 2018).

Data governance plays an important role in the security arena. It ensures that the right individuals have the right access rights, and the level of data exposure is controlled so that only individuals who have adequate permissions can view the data.

While data security management ensures that enterprise data is adequately protected, a data governance program ensures that this data is accessible across the organization in a controlled manner, and appropriate controls, roles, and responsibilities are in place to ensure that the data is adequately protected.

Anu Tirupathi in her article, "Information Security to Justify Data Governance" uses the analogy of a home to describe the relationship between data governance and data security. According to Tirupathi, it is important to keep your home secure from intruders trying to gain entry or hackers trying to break into your computers to steal your personal information. To secure data, precautions are taken by having locks for the doors, installing security alarms, setting up anti-virus software on computers, setting up strong passwords for logging into the computer and

applications, and password protection of files and folders. Only authorized individuals living in the home have the keys to unlock and enter the house. Access to individual rooms might need additional keys. Only individuals who have the associated passwords can access the computer and respective applications, files, and folders. Not all individuals might have access to all applications, files, and folders, that is, applications and data are locked down and secure.

However, while a secure home is better than an insecure home, a secure home automatically does not become a convenient or good home, that people enjoy living in. A good home allows individuals in the home to move about, use the various rooms and appliances, without conflict and be free from internal chaos. To ensure smooth working of the household, roles are assigned to individuals to manage the household (Tirupathi 2017).

Data security management is the data management function which deals with protection of data throughout its life cycle. Data security management comprises the practices, policies, standards, rules, processes, technologies, and tools for protecting data in alignment with business needs by providing adequate authentication, authorization, access, and auditing of data.

Policies need to be defined around data security—classification, access, authorization, auditing of data, and treatment of sensitive data. Governance ensures that these policies are enforced. Organizations have vast amounts of data and to protect and secure all data is neither possible nor is desirable, nor do all data need to be protected with the same rigor. Greater the sensitivity of the data and risk associated with the data in case of unauthorized data access and disclosure, data loss or data corruption, more rigorous processes and controls are needed for protecting the data and mitigating the risks associated with the data. It is important to discern what data need to be protected, in order to set priorities and plan appropriately so as to minimize data security management costs. For example, row-level and column-level access control can be implemented to restrict access to sensitive data that require additional security. Compliance might also require certain data to be adequately protected. One example is the credit card numbers should not be fully displayed on the application screen; only the last 4 digits should be displayed and rest of the digits should be masked.

Data classification is the process of organizing data into categories for its most effective and efficient use and protection. For new data that is created or acquired and stored in the organization, organizations can establish processes that enable the users to classify the data they create, send, modify or otherwise touch. Since data classification is a resource intensive exercise, in case there is no business need or associated risk, the older data can gradually be retired without being classified. Alternatively, organizations can classify their backlog of existing data, using data discovery (Sotnikov 2018).

While there is no one-size-fits-all approach that can be applied to data classification, the data classification process can be broken down into four key steps, which can be customized to meet your organization's specific data security requirement, as you plan and formulate your data security strategy. The key steps involved in the data classification process and summarized in Fig. 3.9 are as follows.

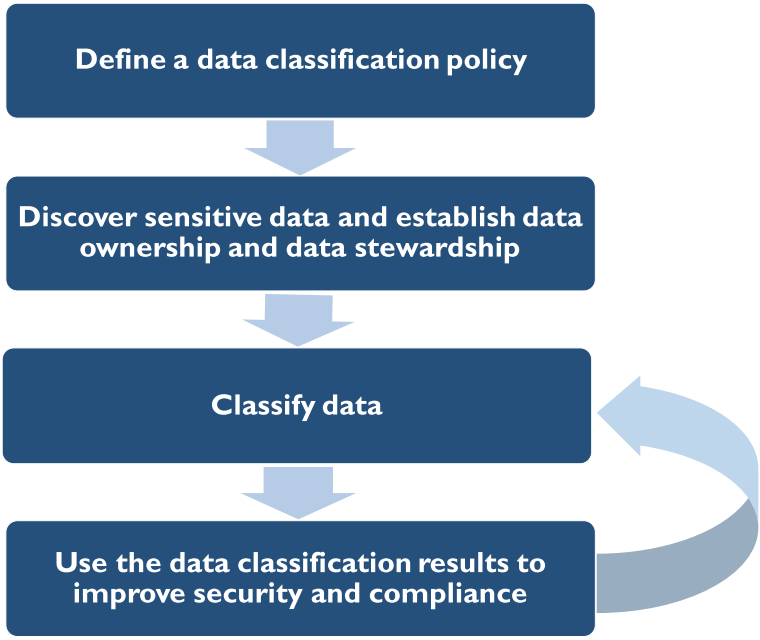

Fig. 3.9 Data classification process

3.11.1 Define a Data Classification Policy

The first step is to define the data classification policy which is a document that includes a data classification framework, a list of roles and responsibilities for identifying sensitive data, and descriptions of the various data classification levels. While a data classification policy differs from organization to organization, it should be simple and brief and include the following sections. Figure 3.10 summarizes the different components of a data classification policy document:

- **Purpose**: The purpose of the data classification policy document and organization objectives that it is set to achieve.
- **Scope**: The scope addresses what data is targeted for data classification and who in the organization it applies to.
- **Data Classification Scheme**: The data classification scheme is the classification of data from a data security and protection perspective. Each organization can have different data classifications schemes from a data security and protection perspective, depending on the organization specific business requirements and legal, contractual and regulatory mandates that affect the organization, and this should be defined in the data classification policy document. In this section we discuss two data classification schemes—

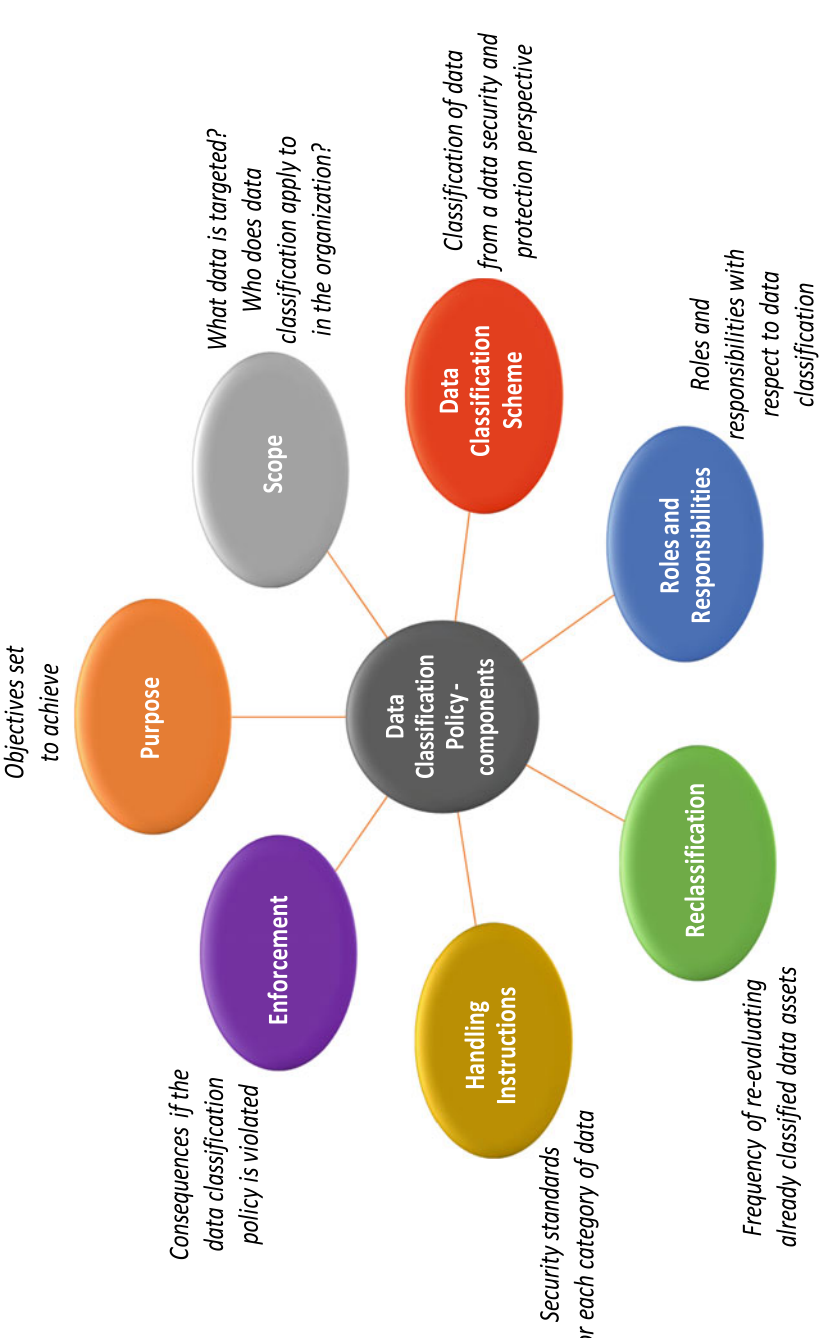

Fig. 3.10 Data classification policy components

- **Data Classification Scheme 1**—Data classified based on data sensitivity and risks associated with data.
- **Data Classification Scheme 2**—Data classified by the level of impact to the organization, if confidentiality, integrity, or availability is compromised.

Figure 3.11 shows the two data classification schemes at a high level.

As discussed in Chap. 2—"Data and Its Governance," from a perspective of data security, data can be classified into the following categories (see Fig. 3.12):

- Restricted data,
- Confidential data,
- Private or Internal data, and
- Public data.

3.11.1.1 Restricted Data

Data should be classified as restricted when the unauthorized disclosure, alteration or destruction of that data could cause an extremely high level of risk—compliance risk, reputation risks or other risks to the organization. Loss of confidentiality or integrity of this category of data can lead to harm to an individual including identity theft, bad press coverage/publicity, lawsuits, and significant reputational damage and costs in terms of fines and penalties.

Restricted data is super-sensitive and highly confidential data. It requires the highest level of privacy and security protection for storage, access, and transmission. Hence, the highest level of security, access controls, and protection should be applied to restricted data. Special authorization should be required for collection, storage, processing, use, and transmission of restricted data. Records containing restricted data elements should be stored securely.

Data records containing restricted data are generally not available for public inspection, and can be only accessed by, and shared with authorized individuals who have legitimate purpose of accessing that data via secure authentication mechanisms and secure mode of transport respectively. Unauthorized access or disclosure of restricted data should be notified. Data protected by global, local, national privacy regulations, and confidentiality agreements are restricted data.

3.11.1.2 Confidential Data

Confidential data is moderately sensitive information that needs to be protected from unauthorized access, and is only intended for limited dissemination.

The difference between restricted data and confidential data is the prospect, duration, and the level of harm incurred. The likelihood, duration, and level of harm are lesser in case of confidential data than restricted data. Unauthorized access or disclosure of confidential data can have adverse impacts though such impacts are

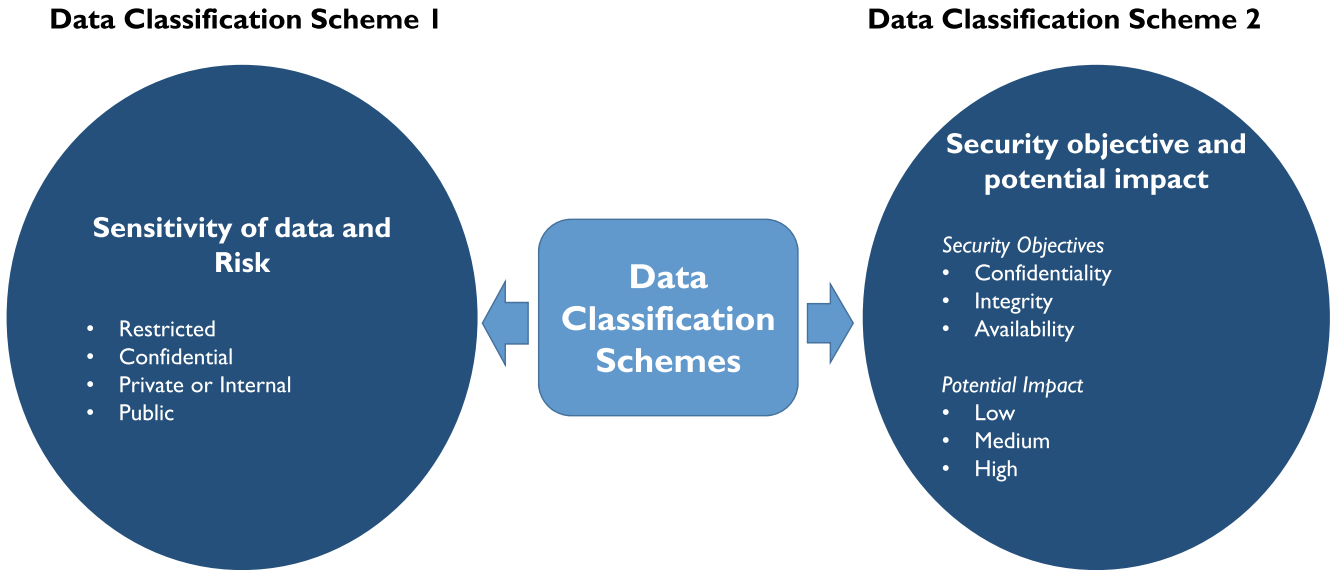

Fig. 3.11 Data classification schemes

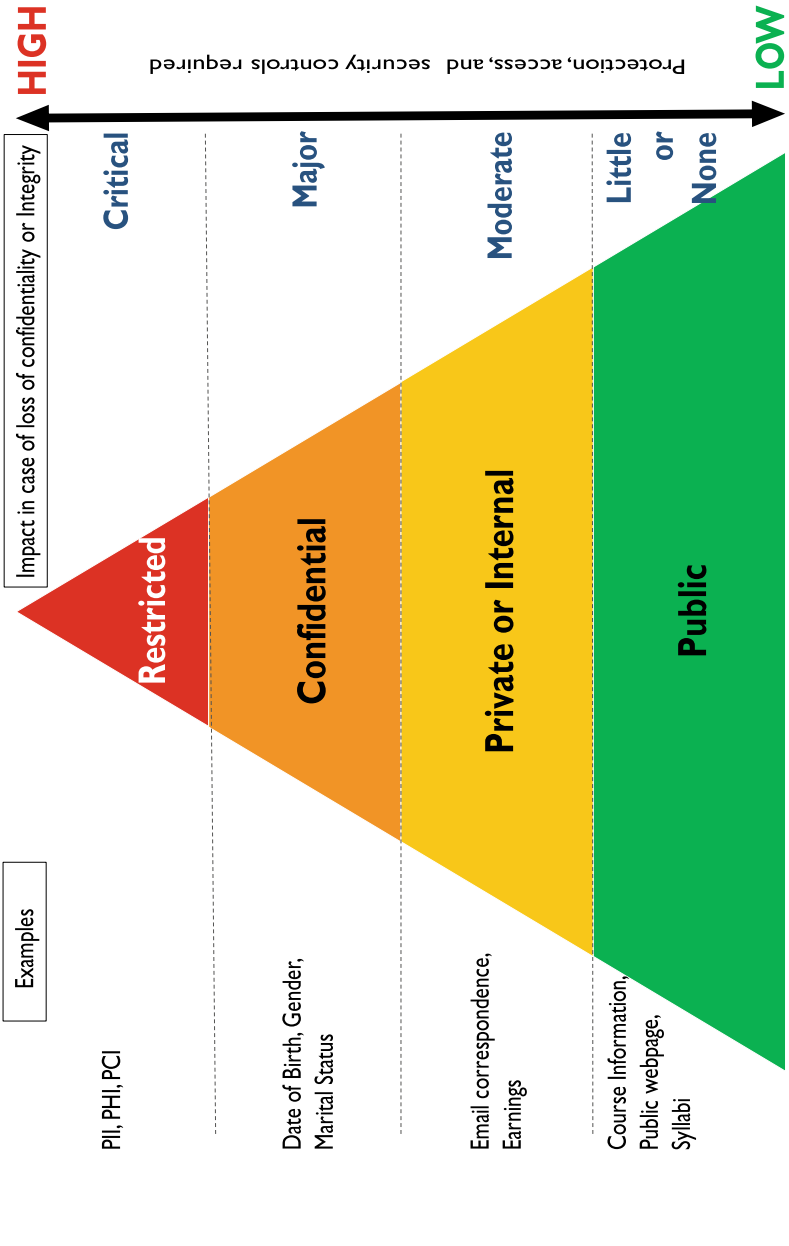

Fig. 3.12 Data classification by data sensitivity and protection (Mahanti 2021)

less severe than in case of restricted data. To address this risk, some level of protection or access restriction may be required, though this is less than the level required for restricted data.

Data classified as confidential could potentially become classified as restricted, if, in aggregation, data could be reconstructed to reveal personally identifiable information (PII).

3.11.1.3 Private or Internal Data

Data should be classified as private or internal when the unauthorized disclosure, alteration or destruction of that data could result in a moderate level of risk to the organization. By default, organization that is not explicitly classified as restricted, confidential or public data should be treated as private or internal data.

As the name suggests, the data is meant for internal use in the organization and is not meant for public release. While confidentiality of private data is preferred, the data might be subject to open records disclosure. Some level of protection or access restriction may be required though this is less than the level required for confidential data.

3.11.1.4 Public Data

Data should be classified as public when the unauthorized disclosure, amendment or destruction of that data would result in little or no risk to the organization. Public data is least sensitive data and has the lowest security requirements and can be released without restriction. Hence, little or no controls are required to safeguard the privacy of public data. However, control is required to prevent its accidental/untimely release and unauthorized modification or destruction.

Data Classification Scheme 2—Data is classified by the level of impact to the organization, if confidentiality, integrity, or availability is compromised. Table 3.1 shows the mapping between confidentiality, integrity and availability and the potential impact (Evans et al. 2004). Table 3.1 also summarizes the potential impact definitions for each security objective—confidentiality, integrity, and availability.

- **Roles and Responsibilities**: The description of the roles and responsibilities with respect to data classification.
- **Reclassification**: Frequency of re-evaluating already classified data assets and description of scenarios, which would require reassessment of classified data assets. The frequency of re-evaluating classified data assets is generally once per year. Examples of scenarios that might need reassessment of a classified data asset is when the data asset in question has been modified.
- **Handling Instructions**: Security standards that state suitable practices for handling each category of data (such as how data in each category must be stored, accessed, shared, and transported, what permissions/access rights should

Table 3.1 Potential impact definitions for each security objective—confidentiality, integrity, and availability (Evans et al. 2004)

	Potential impact		
Security objective	Low	Moderate	High
Confidentiality Preserving authorized restrictions on information access and disclosure, including means for protecting personal privacy and proprietary information	The unauthorized disclosure of information could be expected to have a **limited** adverse effect on organizational operations, organizational assets, or individuals	The unauthorized disclosure of information could be expected to have a **serious** adverse effect on organizational operations, organizational assets, or individuals	The unauthorized disclosure of information could be expected to have a **severe or catastrophic** adverse effect on organizational operations, organizational assets, or individuals
Integrity Guarding against improper information modification or destruction, and includes ensuring information non-repudiation and authenticity	The unauthorized modification or destruction of information could be expected to have a **limited** adverse effect on organizational operations, organizational assets, or individuals	The unauthorized modification or destruction of information could be expected to have a **serious** adverse effect on organizational operations, organizational assets, or individuals	The unauthorized modification or destruction of information could be expected to have a **severe or catastrophic** adverse effect on organizational operations, organizational assets, or individuals
Availability Ensuring timely and reliable access to and use of information	The disruption of access to or use of information or an information system could be expected to have a **limited** adverse effect on organizational operations, organizational assets, or individuals	The disruption of access to or use of information or an information system could be expected to have a **serious** adverse effect on organizational operations, organizational assets, or individuals	The disruption of access to or use of information or an information system could be expected to have a **severe or catastrophic** adverse effect on organizational operations, organizational assets, or individuals

be granted to whom, when and what data must be encrypted or masked, and retention, archival, and purging terms and processes). Since these guidelines may change, it is best to maintain them as a separate document (Sotnikov 2018).

- **Enforcement**: The enforcement section outlines the consequences if the data classification policy is violated by the users.

3.11.2 Discover Sensitive Data, Establish Data Ownership, and Data Stewardship

Once the data classification policy is established and circulated to employees in the organization, a decision needs to be made whether only new data need to be classified or all the existing data in the organization needs to be classified.

If the existing data in the organization contains sensitive data and risks associated with leaving the data unprotected is high, then such data would need to be located in order to be able to protect the data. If a decision is made on classification of existing organizational data then a data discovery process is required to discover the data and then apply data classification policies to the data.

Data discovery is an expensive as well as time consuming exercise. Therefore, the benefits and the risks should be taken into consideration and weighed before making a decision to embark on a data discovery exercise.

Another important part of a robust data governance program is data ownership and data stewardship. Once data owners and data stewards are identified for the different data sources where sensitive data reside, they need to make sure they understand the data flow, and determine where sensitive data may exist within their area of responsibility.

3.11.3 Classify Data

Classify the data as per the organization's data classification policy. Data classification is also one of the most of critical elements of an effective data governance program. The data classification document should contain the following details (captured in a tabular format in Table 3.2):

- **Data Domain**—Data domain is used to identify a particular group of data assets aligned with an operational business unit or line of business. For example, customer and product.
- **Data Entity**—While data domain specifies the group classification of business data entities at their highest level of data object abstraction, data entity specifies the group of data elements under data domain that you want to classify.

Table 3.2 Data classification document template

Data item id	Data domain	Data entity	Data description	Data perspective	Data location	Data security classification	Business risk category	Business risk description

- **Data Description**—Data description includes addition details about the data domain and the data entity.
- **Data Perspective**—For example, master data, transaction data or reference data, email, etc.
- **Data Security Classification**—Data security classification can be one of the following values depending on the data sensitivity and level of protection required.
  - Restricted.
  - Confidential.
  - Private or Internal.
  - Public.

Individual data elements in a data entity may have different levels of sensitivity. In this case, the most restrictive classification of any of the individual data elements should be used.

- **Data Location**—Location of the data. This can be a folder name, document name, database or table where the data is stored.
- **Business Risk Category**—Business risk category defines the type of risk associated with the data. For example, compliance risk, strategic risk, operations risk, reporting risk, and so on.
- **Business Risk Description**—Business risk description describes the risk associated with the data and the adverse business impact if the data is compromised.

Quite a few applications enable tagging the data security classification to the data object itself. For example, Documentum, a document management system has a parameter for tagging the data security classification and needs to be populated before you can import and store a document in Documentum.

It is important to reassess the classification of organizational data, on a periodic basis to ensure that the assigned classification is still appropriate based on changes to legal, regulatory, contractual mandates as well as changes in the business value or use of the data in the organization. This assessment should be conducted by the appropriate data steward. The data steward should determine the frequency of conducting this assessment, based on available resources.

If a data steward concludes that the classification of a specific data collection or data set has changed, an analysis of security controls should be conducted to determine whether existing security and access controls are consistent with the new classification and determine the gaps in the controls. If gaps are found in the prevailing controls, they should be amended in a timely fashion, so that they are appropriate with the level of risk presented by the gaps. For example, if data earlier classified as internal data is now classified as confidential, more rigorous security and access controls would be needed to protect the data. On the other hand, if the data earlier classified as confidential is now classified as private, then security and access controls need to be more relaxed than earlier.

3.11.4 Use the Data Classification Results to Improve Security and Compliance

Once you know the risks associated with data that you have and its storage locations, you can review your security policies and procedures to assess whether all data is protected by risk-appropriate measures (Sotnikov 2018).

By categorizing all your sensitive data, you can prioritize your efforts, control costs, and improve data security management processes. Ensuring the security and protection of sensitive data and mitigating the risks of unauthorized access, modification, loss and disclosure of these data is a top priority for an effective data governance program.

Knowing the data stores and schemas that may contain sensitive data will allow the organization to properly document the administrative controls to protect sensitive data via a formal policy as well as define the scope of the technical controls including access controls to be implemented to ensure the data remains private and secure.

Technical policies that enforce data protection for sensitive data and ensure that only the authorized processes and users can decrypt sensitive data should be established.

Creation of rules for segregation of duties and role-based access and role-based security rules can help secure data by allowing users to access and view only that data that they should be viewing.

It is imperative to ensure that all controls are actively monitored and enforced for the protection of that sensitive data (Martinez 2017). It is also important to conduct periodic review of user access lists, and revoke access if their roles and responsibilities no longer warrant access to the data.

Data classification is an ongoing process as new data are created, and existing data are moved and stored in multiple locations in organizations. Proper administration of the data classification process will help ensure that all sensitive data is protected (Sotnikov 2018).

The typical governance roles in data security management are as follows:

- Sponsor: This is the individual is responsible for managing the organization's data security program. Many organizations have a Chief Security Officer (CSO), whose responsibilities in the data security management area comprise of the following:
  - Overseeing the organization's data security program.
  - Documenting and disseminating data security policies and procedures.
  - Coordinating the development and implementation of an organization-wide data security training and awareness program.
  - Coordinating a response to actual or alleged breaches in the confidentiality, integrity, or availability of organizational data.

- Data Owners: Data owners are the accountable for a collection of data sets or data domain, and have approving authority. Data owners are involved in creation of data security policies around data classification, access, authentication, authorization, and audit, which is a collaborative effort. Data owners review and approve the processes, standards, and rules regarding data security defined by data stewards for operationalizing the data security policies.
- Data Stewards: Data stewards are subject matter experts (SMEs) for a collection of datasets. From a data security perspective, data stewards ensure that the data in the datasets are classified appropriately as per an organization's data classification policy. They carry out risk assessment with the help of the data custodian and technical data stewards, and define the risks associated with the data. The data stewards also define a set of rules that determine who is eligible for accessing data based on business function, job description, support role, and so on, and ensures that data custodians implement appropriate security controls to protect the confidentiality, integrity, and availability of enterprise data. The data stewards should have an understanding of whether or not any organizational policies govern the data sets that they are responsible for. For example, policies exist to help govern financial information, health information, and personal identifiable information, and so on, and data stewards should have an understanding of these policies to classify data, assess the risk tolerance and design rules for access correctly. However, policies are high level guidelines. Data stewards need to define processes, standards, and rules in alignment with the policies to help enforce them. For example, a policy requires passwords to follow guidelines for strong passwords; data stewards in conjunction with the technical data stewards or data custodians, would need to define detailed standards for what would be the constitution a strong password. For example, standards for strong password to access a database might look like the below:

  1. The password should be a minimum of 12 characters and a maximum of 15 characters in length.
  2. The password should start with an alphabet followed by a combination of digits or characters.
  3. The password should have at least 3 alphabets and 3 numbers.
  4. The password should have a mix of uppercase, lower case letters, and special characters.

- Data Custodian: From a data security perspective, data custodians have operational responsibility for the organizational data, and they are responsible for implementation of adequate security and access controls (that is physical and technical safety mechanisms) to guard the privacy, integrity, and availability of enterprise data. Implementing appropriate data security controls, validating, and monitoring them are important aspects of data security management and governance. They are also responsible for documenting the administrative and technical procedures for implementing the security controls. For example, as per guidelines, passwords need to be changed every month. Data custodians would

have to document the technical and administrative procedures to make sure this happens.
- Data Consumer: Data consumers are individuals who access and use data. From a data security perspective, data consumer adheres to the policies, procedures and guidelines for usage and protection of data, and reports actual or suspected vulnerabilities and/breaches in the privacy, integrity or availability of the data to the organization's the local security office or local security point.

While the role titles may vary from organization to organization, the various activities discussed above should be assigned to individuals, who should be accountable for them.

Metrics and reporting are the areas of information security that fall behind in most organizations where proper governance is not in place (Pironti 2006). The following metrics can be used to assess the progress and effectiveness of your data security program:

- Number of orphaned data assets without a data owner.
- Number of data security policies, standards, procedures, and metrics with committed owners.
- Reduction in data security incidents and data breaches.
- Number of data security incidents including data breaches that become public knowledge per year.
- Percentage of data assets not correctly classified.
- Coverage of data security controls.
- Access anomalies.
- Data security and privacy awareness levels.

3.12 Data Governance, Data Storage, and Operations

Data storage and operations is the data management function that deals with the design, implementation, and support of data stored in an organization's data systems, with an intent to maximize its value throughout the lifecycle of data, starting from data creation/acquisition to data disposal. Any data initiative (for example, data integration, MDM, data warehousing, and data migration) would need data to be stored and maintained, and hence data storage and operations is one of the core functions of data management.

Data storage and operations involves activities such as acquisition and implementation of a database environment, restoring data state, data retention, backup, and purging, database performance tuning and monitoring. It also involves defining technical requirements around data storage that will meet organizational needs, designing architecture and model to store the data, dealing with licensing and technology support, installing, administering technology, and resolving issues related to technology.

Effective data governance ensures that policies, processes, rules for acquisition, migration, retention, archival, expiration, and destruction of data are established, enforced, reviewed, and revised (if needed). Data governance also ensures that roles, responsibilities, and accountabilities are in place to accomplish the activities. Business data stewards, technical data stewards, data architects, network administrators, data analysts, and security analysts help in planning for volumes, performance, internal/external retention, archival needs, backup, recovery, and purging while data base administrators (DBAs) do the actual implementation. It is prudent to prepare checklists to ensure that policies are enforced, and all tasks are performed at a high level of quality.

With cloud computing, which in simple terms is the delivery of on-demand computing resources—from applications to storage and processing power, organizations can rent data storage from a cloud service provider, and store data in data centres outside the organization.

Data governance in relation to data stored in cloud needs to take into consideration the security and privacy aspects of the cloud computing solution, data accessibility, data quality, usability, compliance aspects (for example, whether the data centres used by the cloud providers are ISO 27001 compliant), and SLAs that the cloud providers have. Depending on these factors and the sensitivity of the data, risks involved, and business uses of data, decisions need to made as in whether the data assets should reside on-premise or can be migrated to a cloud. Data governance makes sure that processes that ensure these factors are taken into consideration are in place, when making a decision about storing data in the cloud. There should be effective processes for ownership, stewardship, control for access, retention, archival, purging, and oversight, over all data assets residing in data stores located inside the organization and cloud storage.

3.13 Data Governance and Data Quality Management (DQM)

Generally speaking, DQM can be defined as the "quality-oriented management of data as an asset" (Weber et al. 2009, p. 4:4). Data quality management (DQM) is the data management function that involves management of combination of right people, processes, and technologies to achieve high quality data to achieve desired business outcomes (Mahanti 2019; Knowledgent 2017).

All data management functions contribute to the quality of data. Speaking ahead of the October Gartner Data and Analytics Summit 2018 in Frankfurt, Ted Friedman, vice president and distinguished analyst at Gartner, stresses on the impact of poor data quality in his statement,

> As organizations accelerate their digital business efforts, poor data quality is a major contributor to a crisis in information trust and business value, negatively impacting financial performance (Moore 2018).

A recent Gartner research has found that organizations believe poor data quality to be responsible for an average of $15 million per year in losses (Moore 2018). This figure indicates that data quality management should be given high priority in organizations. Lack of governance is one of the factors contributing to poor-quality data.

Data quality management goes beyond the management of the technical data quality management of data throughout its life cycle. It also includes defining the data quality requirements correctly and ensuring that processes and standards to capture and transform and store data are adequate. It requires business and IT teams to work together to define data quality thresholds for respective data quality dimensions relevant to the context of usage of the data, measuring and assessing data quality dimensions, performing root cause analysis to analyze the gaps, layout, and implement the action plan for improvement (Mahanti 2019). These activities help ensure that data is fit for use or fixed for use.

Data quality management includes establishing data quality policies, processes, data quality assessment (including data profiling to enable measurement of the current state of data quality and analysis of the gaps to reach the future state of improved data quality), data cleansing, data quality process improvement, data validation, data monitoring, data quality awareness, and education. Data governance helps bringing together cross-functional team as well provides support for facilitation of these activities.

Data quality management includes a mix of reactive approach to fix data issues that can surface at any point of the data life cycle and proactive approach to prevent data quality issues from occurring in the first place by improving processes to capture high quality data and sustain data quality.

Effective data governance ensures that processes are in place to record and resolve data quality issues. It also ensures that the data quality issue log is maintained, and periodic meetings and working groups are organized with the different stakeholders to prioritize, investigate, and resolve the issues in a timely manner.

Data quality issue log is an artifact that can be used to record and prioritize the data quality issues based on their frequency of occurrence, severity, and business impact. Typically, a data quality issue log records the following information with respect to data quality issues:

- Issue Id—A unique number or combination of character and number to reference an issue and help distinguish one issue from another.
- Issue Description—The description of the data quality issue. For example, in a utilities company, material data is missing for underground pipes.
- Data Domain/Data Subject Area—The domain the data issue belongs to. For example, in a utilities company, the data issue—material data missing for underground pipes, would belong to the asset data domain.
- Issue Classification—It describes the type of issue, whether it is a process issue or related to a data quality dimension (for example, completeness, consistency, and accuracy), or compliance.

- Business Impact—It describes the impact that the data quality issue has on business. There could be one or more impacts on following lines—adverse financial impacts, incorrect decision making and/or reporting, decrease in operational efficiency, customer dissatisfaction and so on.
- Occurrence Frequency/Magnitude—This describes the frequency of occurrence of the issue and/or the number of records impacted.
- Rating—The business impact, occurrence/frequency, and magnitude can be used to arrive at a rating for the issue that can be used for prioritization purposes.
- Reporter—This is the name of the person who reported the issue.
- Assigned to—This is the name of the person the issue is assigned to. There may be different individuals, teams or groups that the issue might get assigned to during the life cycle of the issue from it's discovery/reporting to its resolution. For example, the issue might get assigned to a team for investigation/analysis/solution proposal and post analysis it might get assigned to a different team to implement the solution.
- Issue Status—This field captures the current state of the issue, whether it is new, assigned, under investigation, under implementation, under testing/review or resolved.

The DQ issue log should be up-to-date and reflect the latest status of DQ issues.

DQM requires breaking the stovepipes separating data across business units and creating collaboration between business and IT functions, in order to address both organizational and technical perspectives. This requires a profound cultural change across the spectrum of leadership, authority, control, and allocation of resources, which means governance, specifically data governance (Lucas 2010).

With data governance, organizations are able to implement corporate-wide accountabilities for data quality management, encompassing professionals from both business and IT units (Wende 2007).

Data governance and data quality are separate but interdependent disciplines with an effective data governance program supporting the development of high quality data. As discussed in the previous chapter, data quality is one of the drivers of data governance, and effective data governance results in improved data quality. Effective data governance creates a collaborative structure with accountabilities, roles, and responsibilities established for managing and defining policies, business rules, standards, processes, controls, and metrics to provide the necessary level of data quality control. It also ensures that the data issues are tracked and resolved in a timely manner.

Some data quality tool capabilities can help automation of data governance processes. In order to effectively govern data, it is important to be aware of the current state of data and its usage.

Data profiling which is a technique used to understand the current state of data assets and is an integral part of data intensive projects and data quality discipline can play a similar role in data governance too.

Data discovery and assessing health of the enterprise data through profiling is a basis for deciding which data needs governance, especially with regards to

establishing and enforcing data standards for data's quality, models, architecture, metadata, interfaces, lineage, data security, access and usage rules.

Data stewards and others involved in a data quality program regularly capture metrics about the state of data's quality, then analyze the metrics report to assess whether data quality is up to the mark. If data quality is not up to the mark, then they need to define action plan and solutions to improve data quality.

Many data governance programs enforce policies about data standards and data quality metrics make a good starting point for such policies in where lies the overlap between data quality and data governance. This is because business rules and standards defined for data, "govern" how data is to be validated and verified; and the policies that drive these standards and rules are mandated by a governance committee (Russom 2011).

A governance organization can enable and accelerate a data quality program by-

- providing direction for data quality,
- prioritizing data quality initiative based on criticality of data in alignment with the corporate and data strategy,
- identifying and engaging data quality stakeholders and fostering collaboration by establishing committees and working groups to improve data quality,
- establishing decision rights, clarifying accountabilities and responsibilities with respect to data quality,
- developing, maintaining, enforcing consistent standards, policies, and procedures for data quality and ensuring these are adhered to,
- ensuring appropriate metrics are in place for tracking effectiveness of the data quality program,
- establishing conflict resolution processes to resolve data issues,
- establishing effective data assessment and monitoring processes to ensure the quality of existing data, and
- building data quality awareness through communication, education, training, and success stories.

Metrics would include measuring the data quality along dimensions like accuracy, completeness, consistency, and duplication and converting them into dollar values, for example, cost savings. Data quality metrics with examples will be discussed in the third book of the series—Data Governance Success.

Roles and responsibilities with respect to data quality involves establishing data ownership and data stewardship around the data sources and data sets from a perspective of resolving data quality issues and data standardization. The data owners and the data stewards are responsible for establishing data quality thresholds for critical data elements and datasets, prioritizing data quality issues, and defining data definitions and standards.

Businesses own data with IT in a support role in charge of the IT processes and architecture to deliver the data that meets business requirements. Data is often a shared entity and decision rights needs to be established in case of conflicts.

3.14 Big Data and Data Analytics

While big data is one the most hyped terms of the twenty-first century, with almost everybody in industry having heard of it, different people tend to interpret the term differently and have a different understanding of the term.

The term big data is often thought as large volumes of data, as implied by the term itself and also because of the fact that the term was used to refer to the volumes of data requiring new, large-scale systems/software to process the data in a reasonable time frame as the volume had increased to the point that it could no longer be processed on traditional system platforms (Talburt and Zhou 2015). When talking about big data, the two foremost questions that need to be answered are—

What is big data?

How is big data different from data or traditional data?

3.14.1 What is Big Data?

To answer the first question, Gartner defines big data as follows—

Big data is high-volume, high-velocity, and/or high-variety information assets that demand cost-effective, innovative forms of information processing that enable enhanced insight, decision making, and process automation (Gartner-Information Technology Glossary).

The above definition disabuses the notion that big data is only about large volumes, and illustrates the fact that large volume is one of the aspects of big data; that is the "big" in big data goes beyond volume.

For most organizations, "big data" arises from the multitude of human and/or machine interactions occurring every second, minute or hour of every day that can have an impact on your business. Examples of big data are website page browsing, phone calls, text messages, social media posts such as posts on LinkedIn, Twitter, and Facebook, and data streaming in from field equipment like smart meters, sensors, and satellites.

Today's big data may be tomorrow's normal data as technologies evolve, as proved by history. Hence, big data is a relative rather than an absolute concept.

3.14.2 How is Big Data Different from Data or Traditional Data?

The differences between big data and data or traditional data are along the lines of volume, velocity, and variety. These aspects are popularly known as the 3 V's of big data.

3.14.2.1 Volume

If the volume of the data is so large- in the order of petabytes (1 million gigabytes) or exabytes (1 billion gigabytes), that it cannot be stored in one location or an ordinary database, or needs specialized tools and technologies to organize or analyze the data, then the data in question is big data. On the other hand, traditional data have comprehensible proportions and can be captured or processed by traditional data processing software.

3.14.2.2 Velocity

If data is generated at exceptionally high rates or changes rapidly, such that traditional software cannot capture, store or organize it, then such data falls under the realm of big data.

While traditional data has batch velocity, big data has real time velocity. For example, before the advent of smart metering, meter data was captured once per month and stored as a record in relational databases. However, with the advent of smart metering, meter data is read at a much greater frequency, once every 30 min and qualifies as big data.

3.14.2.3 Variety

Variety refers to the assortment of data types and data sources. Big data is multi-structured and has a wide variety of formats, data structure types, and semantics, sourced from a wide variety of data sources. Big data can be structured, semi-structured, and unstructured. Traditional data are structured data that can be stored in relational databases. Most traditional data sources are in the structured realm (NTNU). We have discussed the varieties of data, namely structured data, unstructured data, and semi-structured data in Chap. 2—Data and Its Governance.

A common pattern that we see when we analyze the above characteristics of big data and compare with traditional data is amplification—that is big data is more in terms of volume, variety, and speed. These differences necessitate different architectures, platforms, tools, and technologies that can be used to store and process big data versus traditional data. Relational database systems which are adequate to store traditional data are not suitable for storing big data. The Hadoop ecosystem is used to store and process big data.

Veracity refers to the uncertainty and noise in big data or conversely the quality and trustworthiness of big data. Big data originating from sources outside the organization might not have rigor around data validation, and hence there is likely to be a quality issue.

The publicity around big data has mislead individuals to think that big data is more valuable than just data (that is traditional data), the notion being “bigger is

better", though this is not necessarily true. The truth is that neither big data nor traditional data by itself have any inherent business value. While the value of traditional data is directly proportional to the criticality of the business purpose they serve, the power of big data is in the analysis one does using it, and the insights, actions, outcomes or decisions that are driven by the results of big data analysis.

3.14.3 Data Analytics

Data analytics (DA) or Analytics is the process of examining data with the intent to draw conclusions about the information they contain, with the aid of specialized applications, systems, tools, technology, and software. Data can be used for descriptive, diagnostic, predictive, and prescriptive analytics (see Fig. 3.13).

3.14.3.1 Descriptive Analytics

Descriptive analytics is the simplest class of analytics reporting on data that helps answer what has happened in the past using traditional business intelligence, data

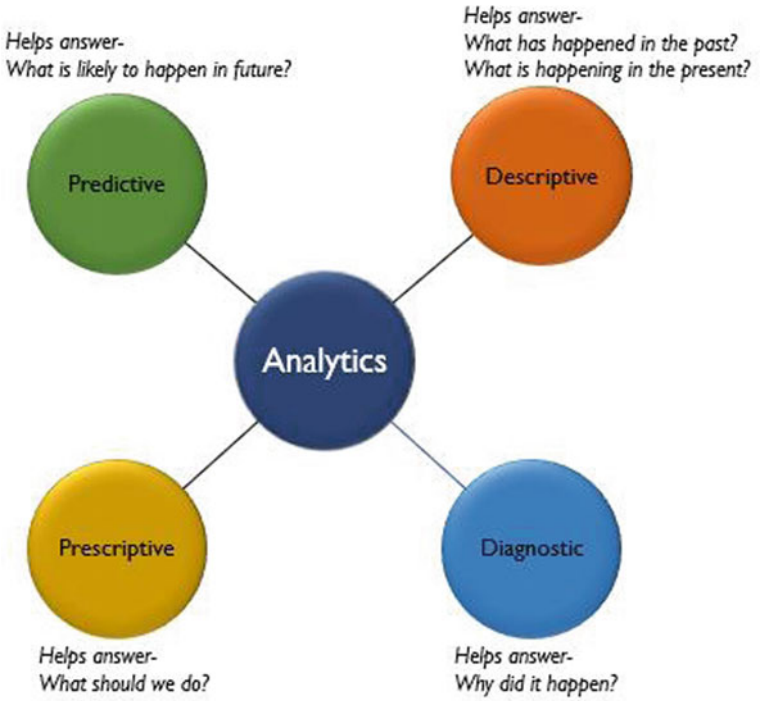

Fig. 3.13 Different types of analytics

mining, and visualization. Traditional data and big data (if stored) can be used for descriptive analytics.

3.14.3.2 Diagnostic Analytics

Diagnostic analytics uses historical data to help answer, why something happened, using drill-down, data discovery, data mining and correlations techniques.

3.14.3.3 Predictive Analytics

Predictive analytics uses data to help answer what is likely to happen in future, using statistical and machine learning algorithms that are probabilistic in nature.

3.14.3.4 Prescriptive Analytics

Prescriptive analytics is the next step from predictive analytics. Prescriptive analytics uses data to predict what is going to occur next and provide guidance on how to react to the prediction using optimization and simulation algorithms. In short it helps answer—"what should we do?"

3.15 Big Data, Analytics, Data Lake, and Data Governance

When it comes to data governance, the same components like policies, processes, decision rights, roles, and responsibilities apply to big data too. However, the data governance components that exist for structured data repositories (for e.g. relational database management systems and data warehouses), need to be extended to incorporate the effects of increased volume, velocity, variety, and veracity that uniquely characterize big data.

Ballard et al. 2014 in their book, "Information Governance: Principles and Practices for a Big Data Landscape" give an example of evolution of policies to address requirements related to increasing velocity that is one the characteristics of big data.

With batch processing, the time when the information would be available to the end users was typically measured in hours or days, which allowed for enough time for individuals to act in response. However, with big data, the time to respond is often in the order of milliseconds, which is beyond human capability to respond, and hence needs policies that need to drive and implement controls to define management of these high velocities. For example, for systems making an

automated response to data flowing into the environment, a policy is needed that requires incorporation of controls that prevent runaway processes and reduce the operational risk that is posed by the automated decision process. In real life, flash crashes that were observed in various stock exchange markets triggered the need for this type of control. These crashes were a result of automated selling that resulted in shares prices to fall down, which again triggered automated trades driving share prices even lower resulting in a cycle until the bottom was reached. The automated decision and market feedback set a downward spiral in motion. Today, these crashes are inhibited by automated responses to either slow trading or stop all trading completely which are enforced through controls as a part of the policy (Ballard et al. 2014).

As explained by Dr. John Talburt in his interview, while data governance is important to efficient data management in relational management systems, it is absolutely essential in Hadoop File system computing required in case of Big data (Mahanti 2021).

Dr. Talburt continues to explain that, big data not only bring large volumes of records per dataset, it brings large volumes of different datasets. Keeping track of these datasets and understanding their content is difficult using traditional RDMS client/server, an approach which starts with a pre-defined data model (Mahanti 2021). This problem is overcome by storing data in a data lake. Data lake is one of the uses cases of big data. Data lakes are data repositories, which do not need a pre-defined data model, and can store any data (structured, unstructured, and/or semistructured data) in their native format.

In case of big data, organizations need to ingest new data quickly without ETL into a model. At the same time, they still need to track what's in the data and know where it's located, but for more datasets and more quickly. Big data computing puts even more pressure on DG. Many have observed, big data computing using the raw data ingestion ("data lake") approach without good DG results in a "data swamp" (Mahanti 2021). This observation is reinforced by research from Gartner, who warn "the data lake will end up being a collection of disconnected data pools or information silos all in one place…Without descriptive metadata and a mechanism to maintain it, the data lake risks turning into a data swamp (Johnson 2019)."

The difference when applying data governance to big data in comparison to traditional data is the swiftness that organizations should have throughout the data life cycle as organizations need to refine big data efficiently and quickly in order to derive business value, which is a complex and burdensome task given the velocity, volume, veracity, and/or variety of data, and at the same time be able to answer questions about the origins of the data, what does the data mean, whether the data is trustworthy, what can and cannot be done with the data, and also have controls to adequately protect it in case it contains sensitive information, and detect anomalous behavior.

Effective data governance is very important to be able to assess the scale of impact, answer these questions and at the same time have policies and processes to ensure data quality, data privacy, data security, and data lineage. With new data

sources, business requirements may be in conflict with policy mandates and an escalation process needs to be in place sort out these conflicts.

When organizations create a big data environment with a lot of new data sources each having its own structure, the credibility, quality, and relevance of the data in new data sources and the associated risks need to be first determined before a decision can be made to source data into the organization and policies and processes need to ensure the same.

Big data often comes from unvetted sources and might contain sensitive data such as personal identifiable information, or viruses. The quality of the data might also need to be assessed before a decision is made to use the data and the initial analysis should be conducted within a secure environment and based on the assessments results, a decision would need to be made whether to ingest data.

The impact of volume, velocity, variety, and veracity of the data sources on different data quality dimensions like accuracy, completeness, duplication, security, and accessibility needs to be assessed and acceptable thresholds and controls need to be established. Since data creation is not controlled by your organization, there is no way to improve the original data quality. Depending upon the purposes served by the data, option would be one of the following:

- to not use it,
- use it as is, or,
- convert the data into a more acceptable form.

If the decision is made to "use the data as is" or "convert the data into a more acceptable form", the raw data will be ingested and stored in a data lake. A data lake is a repository for storing diverse data "as is" in a scalable manner.

While a governed data lake can provide a strong base for analytics, an ungoverned data lake will become an unmanageable data swamp and a liability.

PwC quote Sean Martin, CTO of Cambridge Semantics as saying "We see customers creating big data graveyards, dumping everything [into the data lake] and hoping to do something with it down the road. But then they lose track of what's there. The main challenge is not creating a data lake, but taking advantage of the opportunities, it presents (Johnson 2019)."

Data lakes acts like a reservoir and stores historical data that can be used later. There should be strategic thinking behind storing data in a data lake rather than "store data for now and we will think how to use it later". There should be control over who loads, what data is loaded, when the data is loaded, and also how the data is loaded. Anti-dumping data governance policies and processes should be defined for this purpose. These policies and processes should enforced by data stewards to prevent the data lakes from turning into data swamps (Russom 2017). Metadata for the data loaded into the data lake as well as the data lineage should be documented. This makes it easier for users to locate data and trust the data.

Data lakes store a lot of data, some of which will be sensitive data. Governance oversees the quality and security of the data in the data lake, and helps data

discovery and access by ensuring data ownership and data stewardship, data definitions, data lineage, data quality thresholds, and monitoring.

Without data governance, a data lake would become unmanageable, and consist of unusable heaps of data that cannot be trusted or easily located. However, with effective data governance, a data lake could serve as a flexible and usable repository of data. In short, effective data governance should be in place to oversee the landing, acquisition, processing, usage, storage, access, and security of big data.

Data governance policies, processes, controls, roles, and responsibilities including ownership and stewardship need to be in place to cater for the vetting, analysis, and decision of these new sources as well as data capture, data classification, data lineage, metadata management, data access, data retention, data security, and data privacy.

Data governance ensures that ownership, stewardship, accountabilities, and decision rights for these new data are established, and the data consumers who need the data exist before building big data solutions.

Data definitions and usage standards also take a new dimension when it comes to big data. The definition of the content, which starts with a concept in case of unstructured data, needs to defined, and then redefined as data is transformed and structure of the data needs to be redefined as it is processed and made available to different users.

Raw data from different sources (some of which such as social media platforms and open data loosely governed) in big data environments are complex and not in an appropriate format, and as such cannot be subjected to analytics. They need be cleansed, stitched together, and converted to an appropriate format (this process of data preparation is known as data wrangling using tools from vendors such as Trifacta or Alteryx) before they can be suitable for sophisticated analysis.

The complexity of the data preparation process depends on the complexity of the underlying data and data sources. Descriptive analytics is based on structured data in data warehouses which are governed. However, predictive and prescriptive analytics are based on variety of complex data (structured, semi structured, and unstructured) from different sources some of which may be loosely governed.

A single person in the organization cannot complete the data preparation task and the process requires a cross-functional team consisting of the data consumers (who will be using the results of analysis), data stewards or subject matter experts, data analysts, architects, and data scientists.

Governance needs to document policies, processes, roles, and responsibilities around data preparation Also, there should be ability to trace the finals results back to the sources, with adequate metadata to show context and representation and the business rules for transformation, which needs effective governance. Data governance helps establish trust in the data analytics results. Dr. John Talburt, Acxiom Chair of Information Quality at the University of Arkansas at Little Rock, and Lead Consultant for data governance and data integration with Noetic Partners Inc. in his interview states that though organizations understand the importance of data governance, they want to jump to end an immediately start using data analytics in hopes of achieving break-through insights without paying attention to data governance

and data quality. In his words—"…They are attracted to all of the buzz about using AI and machine learning, while skipping over the hard part of data governance, data integration, and data quality," (Mahanti 2021).

With big data, there is metadata associated with the data object in most cases, but there is not content-related metadata. To process metadata for the content and derive the appropriate contexts for the data, taxonomies, semantic libraries, and ontologies can be used (Krishnan 2012).

These data elements may be new to the enterprise and need to be treated as lookup data and reference data for processing, and maintained as such. Also, what needs to be understood is the retention requirements for the data, while some data will be useful for a short window and hence, can be disposed after that time window, some will be needed for historical analysis and might need to be stored in the data warehouse.

Typical metrics would include increase in revenue through new business opportunities and customer retention or expense reduction through risk mitigation or threat prevention.

3.16 Concluding Thoughts

Data governance is one of the most important disciplines of enterprise data management. It has interactions with all other data management functions and has a role to play in different data initiatives. For any data initiative to succeed, there need to be controls and data governance incorporates the right degree of control. In this chapter, we see the interactions between data governance and other data management functions and initiatives, namely, data architecture management, data modeling and design, data quality management (DQM), data security management, data warehousing and business intelligence (BI) management, data integration and interoperability (DII), document and content management, metadata management, reference data management, master data management, data storage and operations, and big data, data lake, and data analytics.

Tony Epler, Chief Data Strategist, PricewaterhouseCoopers states in his interview (Mahanti 2021)—

> For PwC to become a more data-driven organization, it was important to develop leading-edge capabilities in business intelligence, data mining, data analytics and knowledge management in order to monetize data, provide more value to customers and successfully grow the business.

In order to derive the best value from data, organizations would typically have a number of data initiatives (including data governance being implemented) and these initiatives need to be aligned with an overarching data governance framework with collaboration between the different business and data stakeholders would be needed to get the true value of data. In fact, as highlighted by Jill Dyché, Principal, Jill Dyché, LLC. in her interview (Mahanti 2021),

Many companies won't agree to formalize data governance until there's an authoritative platform for targeted data. Others require proofs of concept – often for data correction or data integration efforts – before being willing to launch a data governance team.

References

Akamai (January, 2018) Cybercrime by numbers. IntelligentCIO. Accessed 14 Jan 2018. http://www.intelligentcio.com/me/wp-content/uploads/sites/12/2018/03/Cybercrime-by-the-Numbers-Scale-Vulnerability-Infographic.pdf

Alfresco (2018) Build-in information governance e-book

Allen M. Top data security concerns around data integration. MarkLogic Blogs. https://www.marklogic.com/blog/top-data-security-issues-integrating-data/. Accessed 14 Jan 2018

ARMA International. ARMA generally accepted recordkeeping. Principles®. https://www.arma.org/page/principles. Last accessed 10 Jan 2019

Aykose M (17 December, 2019) All you need to know about master data governance. Digitalist Magazine. https://www.digitalistmag.com/cio-knowledge/2019/12/17/all-you-need-to-know-about-master-data-governance-06201884

Ballard C, Compert C, Jesionowski T, Milman I, Plants B, Rosen B, Smith H (2014) Information governance: principles and practices for a big data landscape. IBM Redbook

Burgess M (2007) Data quality what is it, and do we agree? http://www.dmsg.bcs.org/web/images/stories/PDFs/where%20did%20it%20all%20go%20wrong%20-%20mikhaila%20burgess.pdf. Accessed 14 Jan 2018

Business-Software.com (2020) Content management vs. Document management: What's the Difference? Accessed 06 Aug 2018 https://www.business-software.com/article/content-management-vs-document-management-whats-the-difference/

Chisholm M. The foundations of successful reference data management. Top Quadrant White Paper. https://www.topquadrant.com/docs/whitepapers/TopBraid_ReferenceDataManagement Whitepaper-3-18-15.pdf. Accessed 14 Jan 2018

Chisholm M (February, 2008) What is master data? B-eye-network. http://www.b-eye-network.com/view/6758

Clément D, Soumaya BH-G, Brigitte L (2010) Data quality as a key success factor for migration projects. ICIQ. http://mitiq.mit.edu/ICIQ/Documents/IQ%20Conference%202010/Papers/4A1_DQinDataMigration.pdf. Accessed 14 Jan 2018

DAMA International (2014) DAMA-DMBOK2 framework. https://dama.org/sites/default/files/download/DAMA-DMBOK2-Framework-V2-20140317-FINAL.pdf. Last accessed 30 June 2020

Dennis AL (2017) Metadata management and data governance: the essentials of enterprise architecture. DataVersity. https://www.dataversity.net/metadata-management-data-governance-essentials-enterprise-architecture/. Accessed 14 Jan 2018

Docsvault (2020) 5 major differences between document management and records management. https://www.docsvault.com/5-major-differences-document-management-records-management/. Last accessed 10 Jan 2020

Evans DL, Bond PJ, Bement AL Jr (February, 2004) Standards for security categorization of federal information and information systems, FIPS PUB 199. https://csrc.nist.gov/csrc/media/publications/fips/199/final/documents/fips-pub-199-final.pdf. Accessed 14 Jan 2018

Forbes (24 October, 2016) Strong data governance enables business intelligence success. https://www.forbes.com/sites/forbespr/2016/10/24/strong-data-governance-enables-business-intelligence-success-says-new-forbes-insights-study/#4eee105c582d. Accessed 14 Jan 2018

Fraser N, Price G (2018) Risk, opportunity and what it means for security professionals. Royal Holloway University of London, ISG MSc Information Security Thesis Series 2018 Special Report from Computer Weekly

Fryman L (7 September, 2017) MDM needs data governance (and here's why). https://www.collibra.com/blog/mdm-needs-data-governance-heres/. Last accessed 4 Jan 2018

Gartner, Information Technology Glossary. https://www.gartner.com/en/information-technology/glossary/big-data. Accessed 14 Jan 2018

Gartner (2019) Business intelligence (BI), gartner it glossary. https://www.gartner.com/it-glossary/business-intelligence-bi/. Accessed 14 Jan 2020

Gemalto Breach Level Impact Report. http://breachlevelindex.com/assets/Breach-Level-Index-Report-H1-2017-Gemalto.pdf. Accessed 14 Jan 2018

Ghosh PG (22 May, 2018) Metadata management and analytics: what is the intersection? Dataversity. https://www.dataversity.net/metadata-management-analytics-intersection/. Last accessed 18 July 2019

Harvey C (2 October 2018) How to create a data integration strategy, Datamation. https://www.datamation.com/big-data/data-integration-strategy.html. Accessed 14 Jan 2019

InfoLibrarian Corporation. Data governance and stewardship metadata. http://www.infolibcorp.com/metadata-management/data-governance. Accessed 14 Jan 2018

ISO/IEC (2008) ISO/IEC 38500: corporate governance of information technology. ISO (The International Organization for Standardization) and IEC (The International Electrotechnical Commission)

Jain S, Thomson B (1 August, 2013) Data lineage: an important first step for data governance, B-eye network. http://www.b-eye-network.com/view/17023. Accessed 14 Jan 2018

Johnson P (18 March, 2019) Overcome data lake challenges with 4 pillars of governance, Matillion Blog, Last accessed on 28 July 2020 from https://www.matillion.com/resources/blog/overcome-data-lake-challenges-with-4-pillars-of-governance/

Knight M (25 May, 2017) Fundamentals of metadata management. Dataversity. https://www.dataversity.net/fundamentals-metadata-management/. Last accessed 10 Jan 2018

Knowledgent (2017) Building a successful data quality management program. Knowledgent Group Inc. https://knowledgent.com/whitepaper/building-successful-data-quality-management-program/. Last accessed 2 Jan 2017

Krishnan K (2012) Data governance for big data. SourceMedia Inc. and Information Management, CBIG Consulting. https://www.cbigconsulting.com/wp-content/uploads/2014/03/data-governance-for-big-data.pdf

Loshin D. Role of metadata in a data governance strategy, white paper, sponsored by Erwin

Loshin D (2013) Data governance for master data management and beyond. SAS Institute. https://www.sas.com/content/dam/SAS/en_us/doc/whitepaper1/data-governance-for-MDM-and-beyond-105979.pdf. Last accessed 20 Jan 2018

Lucas A (2010) Towards corporate data quality management. Portugese J Manag Stud XV(2). https://www.ejmanager.com/mnstemps/127/2010%20MIS-Towards%20corporate%20data%20quality%20management.pdf

Mahanti R (2019) Data Quality: Dimensions, measurement, strategy, management and governance. ASQ Quality Press,Milwaukee WI. 526 pp., ISBN: 9780873899772

Mahanti R (2021) Data governance and compliance, Springer Books, Springer, number 978-981-33-6877-4

Marco PD (05 April, 2017) Metadata management fundamentals. EWSolutions. https://www.ewsolutions.com/metadata-management-fundamentals/. Accessed 14 Jan 2018

Marco D, Jennings M (2004) Universal meta data models. Wiley

Martinez E (2 October, 2017) Data governance: what are the components to keep data secure? Rackspace blog. https://blog.rackspace.com/data-governance-components-keep-data-secure. Accessed 14 Jan 2018

Moore S (19 June, 2018) How to create a business case for data quality improvement, smarter with gartner. Gartner. https://www.gartner.com/smarterwithgartner/how-to-create-a-business-case-for-data-quality-improvement/. Last accessed 1 Jan 2019

Norris-Montanari J (17 October, 2016) What's the difference between data governance and data management? (Part 2) The DataRoundtable, SAS Blogs, https://blogs.sas.com/content/datamanagement/2016/10/17/difference-in-data-gov-data-man2/. Last accessed 1 Jan 2019

NTNU, Introduction to big data. Opphavsrett: Forfatter og Stiftelsen TISIP, Learning Material. https://www.ntnu.no/iie/fag/big/lessons/lesson2.pdf. Last accessed 1 Jan 2019

O'Neal K (24 August, 2017) Data architecture and data governance: what's the relationship? First San Francisco partners blog. https://www.firstsanfranciscopartners.com/blog/data-architecture-data-governance-relationship/. Last accessed 1 Jan 2019

O'Neal K (23 October, 2018) Metrics for data architecture effectiveness. First San Francisco Partners Blog. https://www.firstsanfranciscopartners.com/blog/data-architecture-metrics/. Last accessed 1 Jan 2019

Oath.com (October, 2017) Yahoo provides notice to additional users affected by previously. https://www.oath.com/press/yahoo-provides-notice-to-additional-users-affected-by-previously/. Accessed 14 Jan 2018

Otto B (2011) Data governance, business & information systems engineering. Gabler Verlag, pp 241–244. https://core.ac.uk/download/pdf/159155213.pdf. Accessed 10 June 2020

Paperalt M (29 August, 2018) Document management vs records management vs content management, Paper alternative blog. https://info.paperalternative.com/blog/document-mangement-vs-records-management-vs-content-management. Last accessed 10 June 2020

Parkinson J (September, 2017) What is data architecture? LinkedIn. https://www.linkedin.com/pulse/what-data-architecture-john-parkinson. Last accessed 1 Jan 2019

Pironti JP (2006) Information security governance: motivations, benefits and outcomes. Inf Syst Control J 4. http://www.isaca.org/Journal/Past-Issues/2006/Volume-4/Pages/default.aspx. Last accessed 1 Jan 2019

Ponemon Institute. The economic impact of advanced persistent threats

Ponemon Institute (2017) Data protection risks & regulations in the global economy study. http://www.experian.com/assets/data-breach/white-papers/2017-experian-global-risks-and-regulations-study.PDF. Last accessed 1 Jan 2019

Russom P (July, 2011) TDWI checklist report: using data quality to start and sustain data governance. TDWI Research, Sponsored by Informatica. https://tdwi.org/research/2011/07/tdwi-checklist-report-using-data-quality-to-start-and-sustain-data-governance/asset.aspx?tc=assetpg. Accessed 03 Feb 2019

Russom P (16 October, 2017) The Data Lake Manifesto: 10 Best Practices, TDWI https://tdwi.org/Articles/2017/10/16/ARCH-ALL-Data-Lake-Manifesto-10-Best-Practices.aspx?Page=2

Sandwell D (5 December 2018) Massive marriott data breach: data governance for data security, Erwin Expert Blog. https://erwin.com/blog/marriott-breach-data-governance-for-data-security/. Accessed 14 Jan 2018

Sebastian-Coleman L (31 December, 2012) Measuring Data Quality for Ongoing Improvement, Morgan Kaufmann Print ISBN-13: 978-0-12-397033-6

Shortis J (6 March, 2017) Why your master data management needs data governance. https://www.infogix.com/master-data-management-needs-data-governance/. Last accessed 1 Jan 2019

Smallwood RF (2014) Information governance: concepts, strategies, and best practices. Wiley

Smith AM (April 2009) Data governance as part of a data warehouse initiative, vol 3, issue 4. EIMI Archives, Enterprise Information Management Institute, EIMInstitute.org. http://www.eiminstitute.org/library/eimi-archives/volume-3-issue-4-april-2009-edition/data-governance-as-part-of-a-data-warehouse-initiative

Smith AM (2015) How metadata management relates to data governance and MDM. SearchDataManagement TechTarget. https://searchdatamanagement.techtarget.com/answer/How-metadata-management-relates-to-data-governance-and-MDM. Accessed 14 Jan 2018

Sobers R (January, 2019) 60 must-know cybersecurity statistics for 2019, varonis data security blog. https://www.varonis.com/blog/cybersecurity-statistics/. Accessed 14 Jan 2020

Sotnikov I (20 April, 2018) Data classification: what it is, why you should care and how to perform it, information and data manager (idm). https://idm.net.au/article/0011947-data-classification-what-it-why-you-should-care-and-how-perform-it. Accessed 14 Jan 2020

Standen J (2009) Data migration Part 4—creating a data dictionary how to tackle master data management. Datamartist. http://www.datamartist.com/data-migration-creating-a-data-dictionary-how-to-tackle-master-data-management

Talburt JR, Zhou Y (2015) Entity information life cycle for big data: Master data management and information integration (1st. ed.). Morgan Kaufmann Publishers Inc., San Francisco, CA, USA

Thales Security (2017) Thales data threat report. Adavanced Technology Edition, Thales Security and 451 Research. http://enterprise-encryption.vormetric.com/rs/480-LWA-970/images/2017-Thales-Data-Threat-Report-Advanced-Technology-Edition.pdf. Accessed 14 Jan 2018

Thomas W (2008) Master data management: companies struggle to find the truth in massive data flows. CIO

Tirupathi A (19 July, 2017) Information security to justify data governance. The Data Administration Newsletter. http://tdan.com/information-security-to-justify-data-governance/21862. Accessed 14 Jan 2018

Turner N (24 January, 2018) Data governance & data architecture: alignment & accountability. Global Data Strategy. https://globaldatastrategy.files.wordpress.com/2018/02/edgo2017_data-governance-data-architecture.pdf. Last accessed 1 Jan 2019

Uber Newsroom (2016) Data security incident. https://www.uber.com/newsroom/2016-data-incident/. Accessed 14 Jan 2018

Urso V (2018) The importance of data lineage in data governance. Perficient blogs. https://blogs.perficient.com/2018/10/18/importance-data-lineage-data-governance/

Weber K, Otto B, Osterle H (2009) One size does not fit all—a contigency approach to data governance. ACM J Data Inf Qual 1(1) (Article 4)

Wende K (2007) A model for data governance—organizing accountabilities for data quality management. In: Proceedings of 18th Australasian conference on information systems

Wiki. https://en.wikipedia.org/wiki/Content_management

Ziton G (1 August 2018) The marriage of metadata and data governance. http://tdan.com/the-marriage-of-metadata-and-data-governance/23523. Accessed 14 Dec 2018

Zornes A (July, 2012) MDM software vendors struggling with data governance. Tech Target

Chapter 4
Data Governance Technology and Tools

Technology alone is not enough.
—Steve Jobs
The advance of technology is based on making it fit in so that you don't really even notice it, so it's part of everyday life.
—Bill Gates

Abstract Data governance tools and technology is one of the critical success factors in the successful implementation of data governance. Data governance tools and technologies can form an important part of an overall data governance strategy and implementation as they can automate repetitive activities and processes, enhance productivity, and reduce operational costs. In this chapter, we discuss the various components and aspects of data governance that can be facilitated by technology and tools. We also describe the distinction between data management tools and data governance tools. It is important to do a data governance tool readiness check as in whether your organization is ready for acquiring data governance software. We discuss the various points that you need to consider when assessing your readiness for purchasing a data governance tool as well aspects you need to consider when assessing the data governance tools and technology in the marketplace. There are number of vendors in the market who offer data governance tools with different functionalities and capabilities. The different players in the market that provide tools for supporting data governance will also be discussed.

4.1 Data Governance and Technology

Data governance is neither defined nor driven by technology. It is not a technology in itself.

A successful data governance program involves a combination of people, processes, as well as tools and technology. While tools and technologies can help assist with certain aspects of data governance, it is important to note that data governance is not a technical initiative but a business initiative that revolves around organization behavior.

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2021

R. Mahanti, *Data Governance and Data Management*,
https://doi.org/10.1007/978-981-16-3583-0_4

There is no tool or technology that can achieve all aspects of data governance for you. Tools and technology provide support, and enable the people and process aspects of data governance through computerization, scaling, and augmentation. However, the human element and the organizational structure comprising of a clear definition and institution of the roles, accountabilities, and responsibilities, the behavioral and cultural changes required to govern the data, and implementing and sustaining data governance cannot be performed by data governance tools or technologies.

In other words, tools and technology play a supportive, facilitating, and supplementary role in data governance with the human element being the driving force behind it. People establish and follow data governance processes and leverage tools and technologies to enable and automate these processes. Figure 4.1 shows the relationship between the people, process, tools, and technology components in the data governance program.

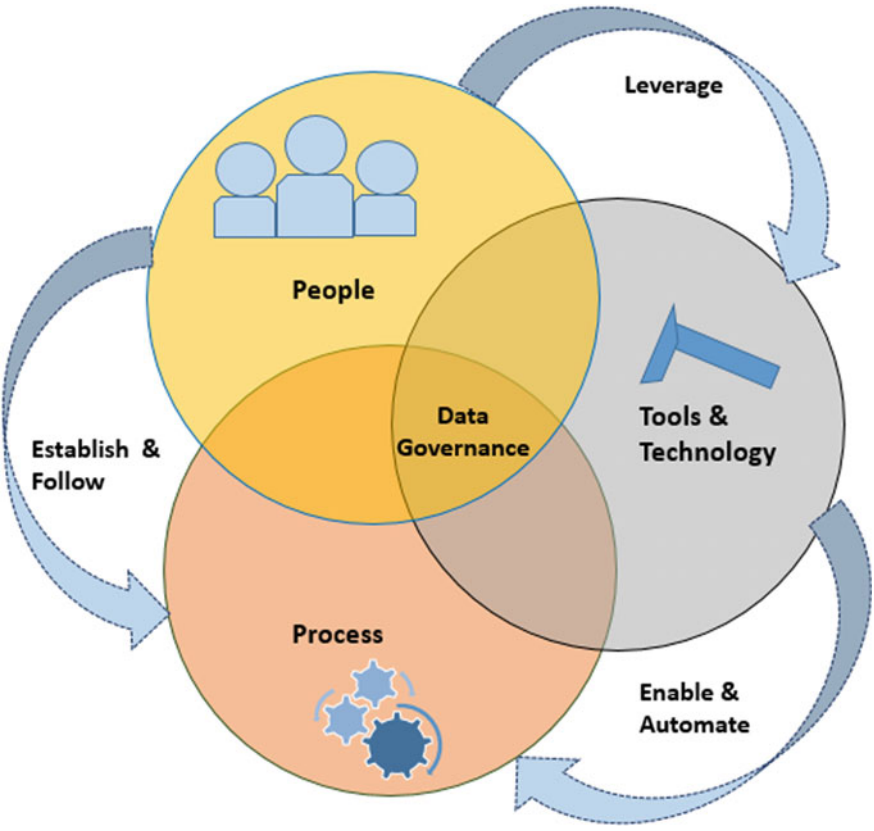

Fig. 4.1 Data governance—combination of people, process, and technology

In this chapter, we discuss the various components and aspects of data governance that can be facilitated by technology and tools. We also describe the distinction between data management tools and data governance tools.

Organizations often make the mistake of purchasing data governance tools and technology before they have understood the goals and strategy of the data governance program and have defined the roles and responsibilities as in "who" is going to do "what", assuming the tool will achieve data governance for them.

However, data governance tools are not plug and play tools and you need to invest time and effort to learn to use them to be able to derive value.

Consider the analogy of the hammer used to drive a nail through a wall. The hammer by itself will not know at which point in the wall it should drive the nail into or (for that matter) by itself drive the nail into the wall. You will have to decide what you want to hang on the wall so that you can determine the dimensions of the nail required to hold the item in place, where on the wall would you like to put it so that you know the exact point on the wall where the nail needs to be driven at, and who is going to do the activity (if it is not you). You must already have answers for these questions before you can get a suitable nail and hold it to the desired point in the wall with one hand and hold the hammer in the other, and strike the hammer against the nail till it gets driven into the wall.

The data governance tool is analogous to the hammer here and certain readiness checks need to be completed before you can start researching the market and start seeking the different data governance tool vendors.

We will discuss about the readiness checks that need to be carried out before exploring the market to purchase a data governance tool in the "*Data Governance Tool Readiness, Selection, and Acquisition*" section in this chapter. Also discussed in this section are the different aspects that you need to keep in mind for comparing and selecting the appropriate data governance technologies and tools from a large number of options available in the marketplace.

There are a large number of vendors who offer different categories of data governance tools with varying features and capabilities. The different players in the market that provide tools for supporting data governance will be discussed in the section—'*Data Governance Tool Vendors*' of this chapter.

4.2 Data Governance Tools Versus Data Management Tools

Data management tools provide capabilities that apply and execute on business defined standards and rules to data. These tools also ensure that the data supports the information requirements of customers, employees, partners, and shareholders.

As per Forrester, data governance tools provide capabilities that support the administrative tasks and processes of data stewardship. These tools support the creation of data policies, manage workflow, provide monitoring, help in measurement of policy compliance, and data use (Peyret and Goetz 2014).

There is a common misconception that data governance tools are repurposed data management tools (Ladley 2018) and that data management tools provide data governance functionalities.

Data management tools generally provide some functionality required by data governance. For example, data governance needs understanding of the current state of the data assets and data profiling tools can be used for the same. However, the output of all data management tools in their original form may not be easy to implement or even be understood by a business user.

Data management tools can also be adapted to assist data governance. For example, a data dictionary tool can be adapted to handle a data glossary that manages business data elements definition. Some data management tools can also be upgraded to include data governance capabilities. However, while there is some functionality overlaps between the data governance tools and data management tools, they are not the same. For example, data governance tools do not create executable (.exe) files (Ladley 2016).

4.3 Data Governance Elements That Can Be Supported By Tools

A data governance program involves several components, activities, and aspects. Creating and managing all these components, activities, and aspects manually requires considerable time and effort, and the exercise may get unwieldy as the program evolves. Many organizations start with using Microsoft Excel, Microsoft Word, Wikis, SharePoint lists, and existing document and content management systems to manage the data governance activities and artifacts which have been summarized in Fig. 4.2.

Robert Seiner, Author of Non-Invasive Data Governance has provided a data governance toolkit consisting of simple templates—Common Data Matrix, Activity Matrix, and Communications Matrix to kickoff the documentation efforts that are a helpful start when embarking upon a data governance initiative.

However, owing to the complexity of the organization's data landscape, multiple players at different levels, and multiple departments across the organization, as the data governance program grows, the set of standards, policies, procedures, definitions, and workflows becomes unwieldy to the extent that they cannot be easily managed in spreadsheets and documents. This is where tools and technology can help by enabling automation of some of these elements and improving efficiency. This has been reinforced by Dr. John Talburt, Acxiom Chair of Information Quality at the University of Arkansas at Little Rock, and Lead Consultant for Data Governance and Data Integration with Noetic Partners Inc. in his interview statement (Mahanti 2021)—

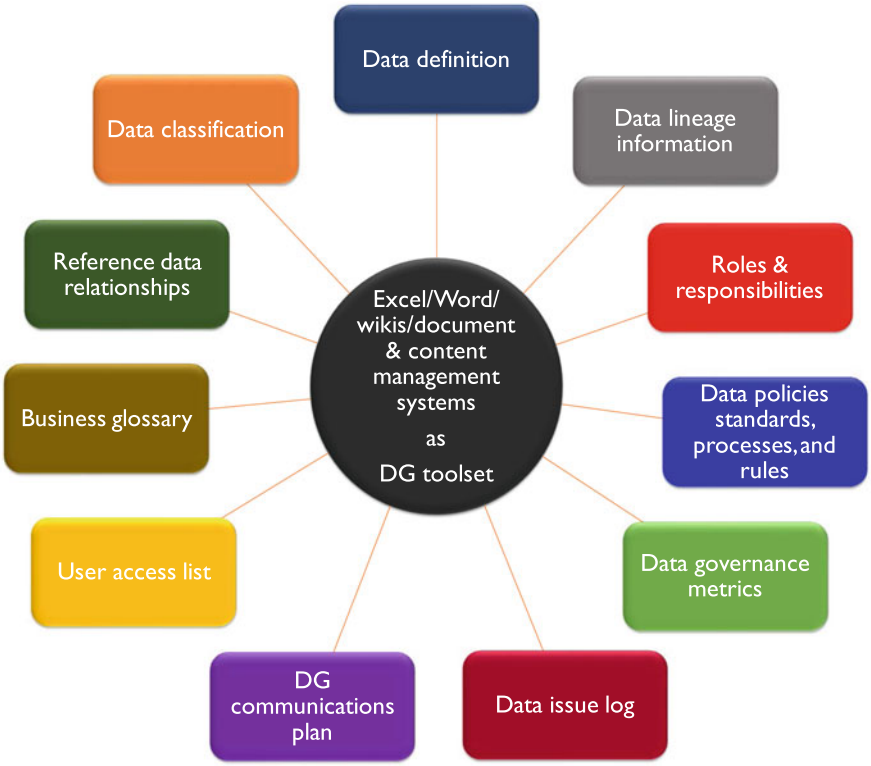

Fig. 4.2 Data governance activities and artifacts supported by common toolkits like Excel, Word, Wikis, SharePoint and document and content management systems

> …it is well-known in the DG practitioner community that the most commonly used DG tool is Microsoft Excel. While this approach might work to get DG started, or even suffice for a small organization, it does not have enough capability to sustain a robust DG program in larger enterprises.

The data governance elements that can be supported by tools are as follows and summarized in Fig. 4.3.

4.3.1 Managing Data Artifacts

One of the key objectives of data governance is to oversee the management of data and data artifacts like data definitions, data classifications, data models, business glossaries, and metadata to achieve consistency in data and data definitions across the organization.

Fig. 4.3 Data governance activities that can be supported by vendor tools

There are thousands of data elements that are stored across the organization and to achieve data consistency as well as consistency in their definitions is not an easy task. Generally, organizations start with using excel spreadsheets for capturing and maintaining data artifacts, like data definitions, data classifications, business glossary, and data lineage but the complexity of the data landscape and the number of data elements and terms call for more sophisticated tools for managing the same.

A data governance tool that has data discovery capabilities will facilitate scanning and identifying the data elements, as well as their values, patterns, outliers, and metadata values.

Data discovery is the first step to an organized management of data. A data discovery tool ensures that all critical data repositories are located and catalogued. It also helps detect suspect data sources that might be outdated or not credible or inappropriate. Data discovery is also a prerequisite for constructing data lineages.

The capability to account for data lineage is another functionality that is important for successful data governance. Data lineage is the documentation and visualization of the life cycle of data. It also includes the data's origin and its travel through different systems to the final destination and helps identify sources, uses, and interdependencies of data.

If the data cannot be traced back to its source or origin, the credibility of the data is called into question. Such data cannot be trusted and is in essence, of no value to a business. Enterprise data travels through and across multiple systems in an organization and the lack of data lineage can cause major issues in compliance and regulatory reporting if the business cannot show the movement of such data across the enterprise.

While organizations have data elements and systems in the order of thousands, and these data elements travel through multiple, complex, and interconnected systems before reaching the destination, tracing the data elements from destination back to the source, recording, and tracking lineage manually is an impossible exercise.

Data governance tools that cater for tracking data lineage can help, and tools from different vendors have different capabilities related to data lineage such as lineage discovery, recording, extraction, demonstration of the end-to-end lineage, and drill-up and drill-down capabilities when displaying lineage information, visual representations of data lineage, and the ability to expose lineage to be used by third party applications.

Business glossary which contains the definition of business data and relationships between them, and links business terms with the operational data entities and data elements is a building block of data governance as it provides clarity of the terminologies, and minimizes inconsistencies and misunderstandings.

While organizations use excel spreadsheets as a tool to start document, manage, and maintain their business data definitions, the large number of business data elements makes this exercise infeasible.

For example, consider the business term—Customer Name in an organization. The business term, Customer Name corresponds to a data element Cust_Name in one database, Customer_Name in a second database, and Buyer_Name in a third database. An organization would have thousands of such terms and relationships. Maintaining the same in an excel spreadsheet is inefficient and time consuming. This is where business glossary tools can assist. Different business glossary tools have different areas of focus ranging from business data definition discovery and collection to managing semantic complexities to industry specific offerings.

4.3.2 Metadata Management

Metadata management is a key component of a data governance program and is critical for effective governance of data, as it is a vital input to capturing enterprise data flow, discovering and presenting data lineage, collecting data requirements, and sharing business rules related to the data elements, which can provide the groundwork for the directions for ensuring compliance with data policies.

Managing metadata in an organization consists of designing the metadata repository, metadata collection, integration, and populating the repository with the same and using the information stored in the repository, and repository maintenance.

Organizations implementing ETL tools sometimes already have a metadata repository which is generated by the tool already in use in the organization. This is because, some ETL vendors offer metadata management applications that have capabilities to record and manage the ETL metadata. In these cases, these repositories can be extended to cater for the broader metadata management requirements. In some cases, the metadata management applications have capabilities to record and manage metadata associated with source and target systems. Data governance can leverage these metadata repositories.

The metadata repository should integrate with a business glossary so that the technical metadata can be linked to the business metadata created in the business glossary to add meaning and context to data assets.

Having a data governance tool that can capture metadata from all data sources in your organization and stored in a central repository gives visibility to all parts of the organization. It also enables easier and faster collaboration to maintain accurate and consistent data. These tools should be able to connect to a wide variety of data sources, have capabilities to link technical metadata and business definition, have data lineage capabilities, have collaboration capabilities, have good search capabilities and filters to facilitate fast location of relevant data assets, have change tracking and change notification capabilities, and be able to monitor updates.

4.3.3 Governance Organizational Structure

Achieving success in data governance requires documentation and management of the different data governance roles and responsibilities. A spreadsheet can be used to store and maintain the mapping between the roles and responsibilities and the data assets domains that they are responsible for. This is important in the initial stages of your data governance journey when you are designing and implementing data governance in one data domain or subject area. For example, who are the data stewards, data owners, and what are the data domains that they are responsible and accountable for respectively? However, as you scale up and expand your data governance throughout the organization, the spreadsheet approach of managing role and responsibilities can become unwieldy and seeking out a marketplace vendor might be a better option.

A data governance tool can help assign, track, and manage roles and responsibilities, as well as capture governing structures like committees, forum working groups, and data governance councils needed to support the data governance program. It can also help in tracking the activities of the members, and manage user permission levels. Every organization names its data governance roles and responsibilities slightly differently and a tool should be able to support custom roles and responsibilities, for example, “Subject Matter Expert” and “Informed”.

The governance organization structure will be discussed in detail in the third book of the series—Data Governance Success.

4.3.4 Data Security and Privacy

To protect data from unauthorized access including data breaches, to ensure protection from data tampering, and ensuring the completeness, consistency, and integrity of data, the ability to classify data, ensuring robust data security, privacy practices, and limiting access control is extremely important.

A data governance tool enables you to classify data based on its use (Ladley 2016) or data sensitivity—for example, whether it's regulated, confidential, public, life-limited or used as a reference item or a fact is also important. Appropriate security controls need to be in place to protect sensitive data. Vendor tools can be used to manage and monitor data centric security.

User lists in spreadsheets that contain user ids, departments, roles, and permissions can be used to manage and monitor access controls. However, given the strict regulatory and compliance requirements of the day, for critical applications, access control would need to be automated.

Access control can be automated using tools which can impose role-based access and constraints as per business requirements, manage user permission levels, monitor user base and their permissions at regular intervals to verify their suitability of access, identify users with unnecessary access and revoke their access. For example, a staff who has moved to a different department, or left the organization or gone on an extended leave, should no longer have access to certain applications. Automated monitoring of access can reveal such users and can also facilitate revoking of these permissions. While in most cases, you still need a person of appropriate authority, the application owner or data owner or the user's line manager to make a decision as to whether the user's access should be retained or revoked, a software that automates the process for monitoring access is more foolproof, less time consuming, and more efficient than the manual monitoring of access control.

Inappropriate or suspicious data access can also be tracked through file and data repositories' activity, and audit trails that are automatically generated and record changes to data repositories using data governance software.

4.3.5 Program Management and Workflow Management

Data governance program management and workflow management which is performed through documentation and emails can be simplified through the use of a workflow and program management software. Workflow and program management software provides organization-wide collaboration, assists with data governance processes by allowing definition, management, authorization, and approval of workflows to improve collaboration among stakeholders, facilitates effective decision making, tracks progress of data governance activities, and facilitates more effective distribution of policies.

Generally, these types of tools support separate approval processes and workflows for custom data attributes, policy annotations, business definitions and comments (Ladley 2016). Workflows can be also used to speed up data management process, and for managing and escalating data governance issues.

4.3.6 Data Stewardship Activities

Data stewardship is "the set of activities that ensure data-related work is performed according to the policies and practices as established through governance" [The Data Governance Institute [b]]. Data stewardship is the operational facet of a data governance program that involves the actual routine work of governing the enterprise's data (Plotkin 2013).

Data stewardship is the process of managing the lifecycle of data right from the capture and creation of data to its final retirement. Data stewardship involves defining and maintaining data models, documenting the data, and defining the rules and policies to oversee the management of data to ensure that the data is of high quality.

Data stewardship lessens data ambiguity and inconsistency through metadata and semantics. It also helps to ensure that the usage of data is consistent across the organization.

In other words, data stewardship results in better quality data which results in improved decision making. Data stewardship tools that can help automate stewardship activities, define common data models, define semantics and rules needed to cleanse and validate data, define user roles, workflows, and priorities, and delegate tasks to the appropriate individuals can help attain better governance.

These tools should help data stewards manage data policies, data standards, data issues, business rules for data quality and master data, business terms, categories, codes lists, and their values. Such tools should also have administrative facilities, which would allow administrators to add and delete data stewards.

4.3.7 Business Alignment

Proper business alignment and connecting data governance processes to other aspects of business is an extremely important element in data governance implementation. This involves documentation of business processes hierarchies, business strategies and plans, monitoring of business strategies and plans, assessing the business value of data, and tracking data governance metrics to evaluate data governance program progress and effectiveness.

4.3.8 Communication and Collaboration

Data governance requires a strong partnership between the business and the IT team. In other words, neither business nor IT can achieve data governance by working in isolation.

While "Communication and Collaboration" is not a data governance component or functionality, it is a critical success factor and hence, tools that incorporate collaboration and communication mechanisms such as task management and messaging can speed up the interactions and enhance understanding between the different players in business and IT, resulting in faster project delivery and fewer iterations. For example, IT users tend to understand technical metadata whereas business users are more comfortable in dealing with business metadata. A metadata management tool that has collaboration capabilities that enables the technical metadata users to message business glossary users can collaborate easily to define and maintain a common taxonomy of metadata definitions, along with their business names and definitions.

4.3.9 Data Management Activities and Data Quality

Data governance ties together all data management activities. Data governance makes decisions, policies, standards, and rules while data management implements those decisions, policies, standards, and rules. Hence, an effective data governance system requires tracking of the data management activities, to ensure that the data is being managed adequately and the data fulfils the purpose and is of good quality.

Improved data quality is usually one of the drivers of data governance implementation. Data quality and data governance have a symbiotic relation. Hence, tools and technology that support and optimize the following data quality functions can certainly help.

4.3.9.1 Data Profiling

Data profiling is a process to capture statistics that provides a picture of the current state of its data assets. Data profiling also provides some useful characteristics of the underlying data like completeness, null values, blanks, outliers, and patterns, that helps to assess the health of the underlying data in terms of the content, structure, and quality issues like duplication, inconsistencies, and missing data.

Data profiling helps provide an insight into the data's' relative strengths and weaknesses. Data profiling can be done manually by scanning through the data values, if the sample size is small and consists of a few data elements. Data profiling can be done using SQL queries if you understand the data and all the business scenarios for the data very well. However, for large and complex data sets, a data

profiling tool is the best option for assessing the health of such data sets to find out the hidden unknown anomalies. Once you have found the data quality issues, you would need to clean the existing data and profile the data post cleansing to assess the improvements in data quality.

4.3.9.2 Data Cleansing

Once you have profiled your data and found out the data quality issues and their potential impacts, you would need to devise an action plan to prevent bad quality data from being created in the first place as well as clean the existing data, which already has issues, which would involve a combination of the following:

Data Parsing and Standardization: Parsing is a process that involves the disintegration, conversion, and segregation of text fields and content into constituent atomic parts and subsequently, organizing the values into consistent formats, based on industry standards or local standards and user-defined business rules which involves comparison with regular expression, reference table entries, patterns or token sets using an assortment of techniques, including machine learning. Data standardization involves converting data elements to forms that are uniform and standard.

Data Matching: Matching is a process to identify similar data within and across the data sources and determining whether two records represent the same thing.

Data Linkage: Data linkage is also known as record linkage or linking, which is necessary when joining data sets based on entities that may or may not share a common identifier (for example, database key). The absence of a common identifier may be due to differences in record sizes, data types, data models, formatting conventions, abbreviations, and storage locations.

Matching and linking establish connectivity between records that are alike, and may even merge a pair of records into a solitary surviving record in preparation for data cleansing (Mahanti 2019).

Data Merging or Data Consolidation: Data merging or data consolidation is the consolidation of the data from the matched set into a single record based on a set of business rules.

Data Augmentation: Data augmentation, also known as data enrichment involves enhancing the value of internally held data by incorporating additional related data attributes from an organization's internal and external data sources that are not directly related to the base data (for example, consumer demographic attributes or geographic descriptors).

A data quality technology that supports the above cleansing functions would be required to detect and correct bad data to improve the quality of the existing data, as

performing any of the above functions manually is an impossible or an extremely cumbersome and time consuming exercise.

4.3.9.3 Data Monitoring

The critical data elements need to be monitored periodically and proactively to determine that the data quality meets the required data quality threshold, that is, the minimum degree of conformance or acceptable level of data quality deviation.

The technologies used for data profiling can be used for data monitoring too. Data quality dashboards can be used to display data quality trends. However, data quality dashboards will not show the individual data records or data elements values.

4.3.10 Master Data Management (MDM) and Reference Data Management

Data governance programs are often driven by master data management which aims at achieving a single golden record. Effective MDM programs embrace data governance from the very beginning.

In the absence of good reference data management, the data quality deteriorates and changes to regulatory requirements cannot be properly addressed. Since the reference data changes slowly, companies use spreadsheets to maintain and manage them manually but there is high risk of error in this approach.

Vendor tools that blend data governance, reference data management and MDM can be useful in efficiently consolidating and managing the reference data and master data throughout the enterprise, while at the same time ensuring that the rest of the enterprise data can leverage these data and achieve accurate and consistent master data and reference data faster than those solely focused on MDM or reference data management.

4.3.11 Data Governance Metrics

Data governance metrics are necessary to measure progress and success of the data governance program. Once an organization has established the data governance metrics based on its context and goals, data governance can be fully implemented in the organization.

It is necessary to capture, publish, and monitor metrics at regular intervals to track progress. Technology can help automate the capture and calculation of quantitative

metrics. It is also important to be able to see the trends over a period of time to see how the program is moving from the inception of data governance implementation. Dashboards or scorecards are tools that can be used to publish and look at the data governance metrics figures and trends. While an organization can start with excel dashboards to publish data governance metrics and trends, as the data governance program matures, organizations should consider acquiring vendor tools which have more sophisticated presentation features. Data governance metrics will be discussed in detail in the third book of the series—Data Governance Success.

4.3.12 Data Policy Management

Data policy management is the creation and management of the documented set of guidelines for the upkeep of proper management and usage of data. It is one of the most important aspects of data governance implementation. Hence, capability for enabling the establishment of governance principles, policies, standards, controls, and rules that formalize the declarations of the policy and regulations are good initial candidates for automation.

Microsoft Word or wiki can be used to start with documenting the data policies and document management systems like SharePoint or Documentum can be used to hold different versions of the data policies in a central repository. However, as your data governance program matures, you might want to consider vendor data governance tools to facilitate creation and management of data policies.

While the ownership, responsibilities, and accountabilities around data policies need to be established, data governance tools support the life cycle associated with defining, approving, communicating, and fully integrating data policy compliance. Data governance tools can also help in supporting data policy creation and management through data policy definition, metadata management, a centralized repository for sharing information about data policies, documentation of role definitions and associated procedures, documentation of the terms of data quality SLAs, guidance for operational roles based on defined policies, and services for measurement and monitoring of compliance to data quality rules (Loshin 2010).

4.3.13 Data Issue Resolution

Data issue resolution involves documentation, prioritization, assignment, reassignment, tracking, reporting, and closing data issues. It is an important aspect of data governance.

Generally, organizations with data governance tools that have these capabilities can replace the manual activity of tracking using a spreadsheet. In addition, tools should also have the capability to send an email notification to the assignee so that the concerned individual is aware that an action is required on his end.

4.3.14 Managing Other Artifacts

Unstructured or semi-structured information stored in Microsoft Excel, Microsoft Word, Microsoft PowerPoint, (for example—manuals, charters, product specifications, work products related to documenting roles, policies, standards and processes, and emails) that need to be stored and reviewed on an ongoing basis as part of a data governance program, need to secured, tracked, and managed. The hierarchy of roles, policies, standards, and processes also need to be maintained.

Document and content management tools that provide adequate document protection and security, maintain the document integrity, have version control mechanism, enable document classification, and provide access can be useful to manage these artifacts. These tools generally provide a number of features like drill-down, view, download/check-in and out, and abilities to add, change, and delete documents in the folders. As your data governance program evolves, you can assess and evaluate vendor data governance tools to manage these artifacts.

4.4 Data Governance Tool Readiness, Selection, and Acquisition

When it comes to purchasing and acquiring data governance tools and technologies, there is a readiness aspect that you need to consider before even looking at the options available in the market.

You need to have clarity as to why you need the tool, what you want the data governance tool or technology to accomplish for you, what manual governance tasks and activities are you looking forward to automating, what roles are responsible for completing those tasks, whether those roles sit in the business or IT, the capabilities you need, the requirements the tool should fulfill, and the business benefit derived from fulfillment of these requirements.

You need to understand all the scenarios which you wish to automate and the functionalities required for the same. It is a good practice to start with the existing toolkits in your organization like Microsoft Excel spreadsheets, Microsoft Word, content and document management systems to help manage your data governance program initially, and when you have absolute clarity about requirements that would benefit from automation, start your research on data governance tool vendors.

As of today, there is no single data governance tool that has all the capabilities and features to meet all data governance requirements, in terms of data governance management coverage and collaboration.

Hence, depending on the specific data governance business requirements and the corresponding functionalities required, a single or multiple category of tools might be needed for implementing these functionalities. There are a number of data governance tools available in the market by different vendors. However, these tools

are expensive and in some cases, multiple tool categories may be required for fulfilling your requirements but you may not have sufficient budget for the same. Hence, you would need to prioritize your requirements to target the tool required for the highest priority requirement.

If your organization has already heavily invested with one vendor, it makes sense to lookup the data governance capabilities offered by that vendor as tools from a single vendor will have better interoperability. Also, you should perform a cost benefit analysis, to be able to justify what value the data governance tool brings to the business before looking for suitable data governance tools in the marketplace.

Based on your requirements, you should have all the use cases and test the data governance tools with your organization's data and use cases, to see that it is meeting your expectations, and has the features and capabilities to fulfill the requirements. The tool may work well for some use cases but might fail for others. Figure 4.4 summarizes the readiness aspects that you should consider before looking up the market place for a DG technology or tool.

Both business and IT departments need to sit and work together when assessing, purchasing, and deploying data governance tools. This is because businesses do not understand the technology architecture and IT might not understand the business requirements and hence, if either party makes a decision to purchase a tool without

Fig. 4.4 Data governance technology and tool readiness—aspects to consider

consultation and involvement of the other, an organization might end up buying a tool that does not ultimately suit their data governance needs.

Data governance tools should help manage your data better and should fit into your business model, operating infrastructure and environment. It should also integrate well with the existing technologies in your organization landscape. Not all vendor tools have the same features and capabilities. You need to compare features and capabilities across different vendors to see which is best for your organization's data governance needs.

The users for data governance tools lie both in IT and business. However, tools that are to be used by the business side, need to be more user friendly, easy to learn, and use. It is important to understand the learning curve involved in using the tools effectively. Therefore, product documentation, product training, availability of vendor support and the quality, nature, and extent of the same are important aspects to evaluate when choosing a vendor. Vendors that provide good training and have good supporting material like user guides and a good support model would be a better bet in terms of long-term sustainability.

With advances in technology, organizations are collecting and storing a lot of data such as partner data, cloud data, social data, mobile data, big data, and data from the Internet of Things (IoT), in addition to data stored in a number of relational databases and file systems. The ability to connect to all data sources in the organization is another aspect that needs to be assessed when choosing a tool. Not all tools can connect to all data sources. Organizations generally have multiple heterogeneous data sources and the chosen tool should be able to connect to all these sources. Hence, looking at the connectivity options is important when selecting a vendor tool.

Another aspect that you need to consider is residence of the data governance tool in your infrastructure. Whether it is offered as a web service or distributed via on-premises servers or is it going to be cloud based? Many of the vendors offer cloud-based licensing which is a good option in that the tool can be used without significant internal technology disruption (Ladley 2018).

Scalability and performance should also be kept in mind when choosing a tool. The tool you select should be able to accommodate a significant amount of new data and users without compromising on speed or effectiveness. A cloud-based tool with flexible storage and processing power will find it easier to tackle an unexpected inflow of data and users.

As your organization's data governance program evolves and grows, you might find that various aspects of data governance cannot by implemented using tools from the same vendor and multiple vendor tools are required to meet your data governance needs. In such cases, you need to test the integration capabilities of these tools, and verify interconnectivity and interoperability to ensure that these different vendor tools integrate with each other as well as with the existing technologies before making your decision to purchase them.

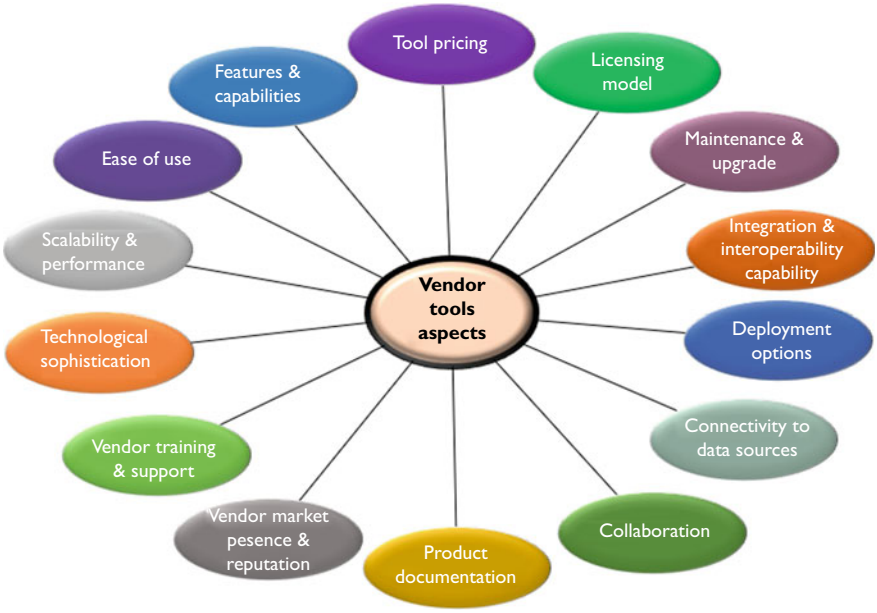

Fig. 4.5 Vendor tools—aspects to consider

Other aspects that you would need to consider when choosing the data governance tools are collaboration capabilities, technological sophistication, vendor's market presence, standing, and reputation in the market place, tool price, licensing model (per user or per organization), and maintenance/upgrade costs.

It is also important to understand the vendor's roadmap and strategy, and choose the ones that have concrete strategies and prototypes/early releases geared toward the strategy, process, and administrative aspects of governance and not just data management. Figure 4.5 summarizes the aspects that you need to consider when assessing vendors tools in the marketplace.

4.5 Data Governance Tool Vendors

With rising importance of data governance, the need for sophisticated tools to support data governance is also gaining more prominence. The initial tooling around data governance were designed for use by tech savvy professionals, and the outputs of the tools had to be converted by the technology professionals into a language that could be understood by the business stakeholders.

There are many sophisticated vendors tools in the marketplace for some aspects of data governance such as data management, data quality, master data management, and data modeling. These tools have been around for a considerable number

of years. However, the data governance tool market is still evolving and maturing. Forrester has grouped data governance vendors into several categories as follows (Peyret and Goetz 2014) (see summarization in Fig. 4.6):

- Data management platform vendors like IBM, Informatica, and SAP who provide a broad data governance offering pulled together from the numerous products they have acquired or themselves developed over years. These vendors have substantial integration capabilities.
- Business intelligence (BI) platform vendors such as Information Builders and SAS who have integrated some data governance capabilities in their products.
- Data governance specialists like Collibra and Global IDs who recognized a market need for data governance tools not served by the existing vendors a few years back and started working on building tools and capabilities to address these needs.

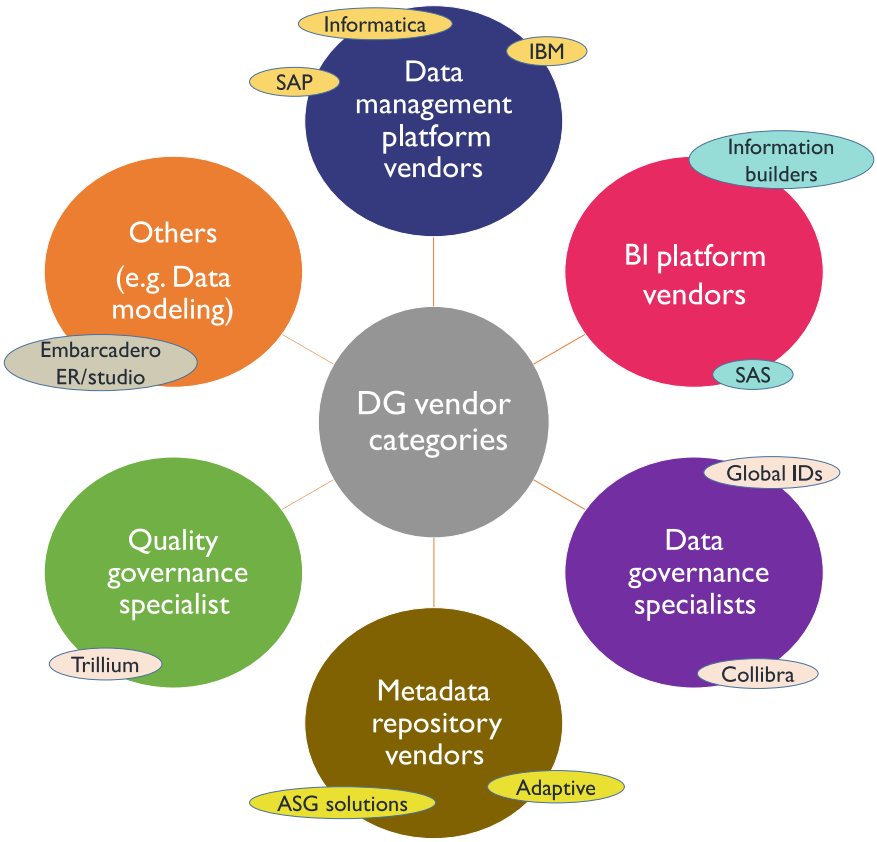

Fig. 4.6 Categories of data governance tool vendors as grouped by Forrester research

- Metadata repository vendors like Adaptive and ASG Software Solutions that have added collaboration and dashboarding capabilities to support data governance.
- Quality governance specialist like Trillium Software which makes an effort to address data governance management capabilities to ease quality remediation and profiling, including for business users.
- Other vendor categories include data modeling vendors, such as Embarcadero ER Studio. While they are not data governance management solutions strictly speaking, many customers are using such repository, modeling, and collaboration products to develop data governance management capabilities.

In addition to the vendors analyzed by Forrester research, there are other vendors like Datum, Oracle, Atacamma, Infogix, Alex Solutions, and Alation that offer tools with data governance capabilities.

Datum provides a cloud-based and on-premises data governance platform that assesses the impact of data on processes. It provides role-based access, and also has focus on enterprise collaboration.

However, none of the vendors solely provide entire coverage for data governance.

ALEX Solutions provides a cloud-based data governance solution designed to support business and technical users.

Infogix acquired Data3Sixty in early 2017 to enlarge its platform and offers support for policies, external and internal data stewards, collaboration, and external data source management (Violino 2018; Peyret et al. 2016).

Collibra Data Governance Center is an enterprise-wide data governance solution that provides tools that have capabilities such as data catalog, business glossary, reference data, policy management, and stewardship. In addition, it provides a Data Helpdesk tool that helps users to flag data issues and route issues to correct stakeholders for resolution (Collibra 2016). Collibra Connect also lets organizations merge their existing services with the data governance center, and allows them to continue to utilize and integrate their existing automated tasks (such as data quality report results or reports of metadata changes) with the platform.

Collibra Data Governance Center can be installed on premises on Windows or Linux systems or hosted on a private cloud via the Data Governance Center Cloud Edition, which can be accessed via a web browser (Houpes 2018a, b). Collibra On-the-Go is a mobile app that provides access to the Data Governance Center and its applications from any mobile or tablet device with a Windows or an iOS operating system.

IBM's InfoSphere Information Server (IIS) provides a wide-ranging set of tools to manage data governance efforts. These tools use the same database to store information. The Information Analyzer tool allows you to profile data to determine its current quality state. Information Governance Catalog tool, formerly Business Glossary, merges business and technical information, and displays both business and technical data lineage.

The Workflow functionality in the Information Governance Catalog tool allows managing the process of development and publishing business assets, including business terms, categories (business domains), policies, and business rules. Information Governance Catalog is available as an on-premises software with client–server architecture. It can be installed on AIX, Linux and Windows server systems, and users can access the software via the Microsoft Internet Explorer or Mozilla Firefox Web browsers (Houpes 2018a, b).

Diaku was acquired by Informatica in early 2017 to expand its data governance and compliance capabilities. Informatica provides tools with user interfaces that are designed for technical users and non-technical users (that is business users). Informatica's data governance platform consists of a collection of products that have mostly been built on top of the Informatica Intelligent Data Platform. These products—Axon, Secure@Source, and Enterprise Information Catalog, along with Informatica's well-known data profiling and cleansing capabilities address the core governance issues of data quality, data cataloging, compliance, privacy, and policy management (Howard and Howard 2017).

Global IDs focuses on data catalog features with some advanced capabilities for GDPR, and data analysts and scientists. Governance software, such as Collibra Data Lineage, Solidatus, Informatica Master Data Management, and SAS are designed to show the end-to-end lineage internally and to regulators. Oracle MDM and SAP's Master Data Governance Product are well-known solutions that blend data quality, governance, and policy management with master data management.

Alation Data Catalog automates updating the business glossary by using all data sources. It can also map data entries to underlying data objects, such as reports, and allows room to post usage policies and data quality warnings. Table 4.1 provides a mapping between data governance functionalities and corresponding vendors and tools (as per information available at the time of writing this book).

4.6 Conclusion and Final Thoughts

In this chapter, we discussed the distinction between data management tools and data governance tools. We also discussed the different data governance elements that can be automated by tools and technologies.

It is important to carry out a data governance tool readiness check to ensure whether your organization is ready for acquiring data governance software. We also discussed the various points that you need to consider when assessing your readiness for purchasing a data governance tool, as well as aspects you need to consider when assessing the data governance tools and technologies in the marketplace.

There are a number of vendors in the market who offer data governance tools with different functionalities and capabilities and we have discussed them in this chapter.

Table 4.1 Data governance functionalities and corresponding vendors and tools

Data governance functionality	Vendors/solution/tools
Business glossary	• Alation Data Catalog • Adaptive Business Glossary Manager • Collibra Data Lineage • Solidatus • Infogix (Data3Sixty) Data Collaboration Suite • Informatica (Daiku) Axon • IBM InfoSphere Information Governance Catalog • SAS Business Data Network • Information Builders Omni-Gen
Metadata management	• Adaptive Metadata Manager • ASG-metaGlossary • Informatica—Enterprise Information Catalog • Information Builders • Global IDs
Workflow and program management	• Adaptive workflow feature • Collibra Data Governance Center • SAP's Master Data Governance • IBM Information Governance Catalog • Profisee
Data lineage	• Adaptive • Collibra Data Lineage • Informatica (Diaku) • SAS
MDM	• Oracle MDM • SAP's Master Data Governance • Informatica Master Data Management
Reference data management	• Collibra Data Governance Center-Reference Data • Information Builders • Global IDs
Governance organizational structure	• Collibra Data Governance Center • SAP Master Data Governance
Policy management	• IBM InfoSphere Information Governance Catalog—Stewardship Center • Collibra Data Governance Center—Policy Manager • Infogix Data3Sixty—Govern • SAP- Information Steward • Datum • Global Data Excellence • Informatica—Enterprise Information Catalog and Axon • SAS • SAP's master data governance
Data stewardship	• Collibra Data Governance Center—Stewardship • BackOffice Associates Data Stewardship Platform • Datum—Information Value Management • Alex Solutions • Global Data Excellence • SAP Information Steward

(continued)

Table 4.1 (continued)

Data governance functionality	Vendors/solution/tools
Data issue resolution	• Collibra Data Governance Center—Data Helpdesk
Data quality	• IBM • Informatica • Trillium • BackOffice Associates • Alex Solutions • SAP • Global IDs • SAS
Data security and privacy	• Informatica Secure@Source

Data governance tools and technologies can form an important part of an overall data governance strategy and implementation as they can automate repetitive activities and processes, enhance productivity, and reduce operational costs. Dr. Stan Rifkin, Director of R&D, Master Systems Inc., in his interview statement lays emphasis on a mutual adaption model when he states "*... adapt the organization in light of the technology and adapt the technology in light of the particular organization* (Mahanti 2021)."

However, technology is an enabler and facilitator for data governance, as data governance is not a technical program. Governance is actually about the processes, people, related roles and responsibilities, and the organization. Tools are just implemented to make those tasks/jobs easier.

With the vast amount of data which organizations are capturing, storing, and managing, it is impossible to have an effective data governance program without tools for automation. However, data governance implementation requires significant changes in organizational behaviors and unless the people are willing to change, data governance tools cannot do much for you. This is reinforced by Andres Perez, in his interview statement as follows (Mahanti 2021):

> Most organizations fail to erect the proper organizational capabilities to facilitate managing information across its value chain; they tend to focus on technologies and are swayed by vendors into believing that the "silver bullet" technology they provide is all they need.

Christopher Butler, Chief Data Officer, HSBC, UK shares his view of the role of technology in the implementation of data governance as follows (Mahanti 2021)—

> As with all technology – these are tools that are used to measure, report and manage data as part of the governance process. However, what you measure and record is more important than just having a tool in place. The information, dashboards and processes must all fit within a holistic environment and be adapted to demonstrate value.

Vendor data governance tools, technologies and supporting software can only add value to a data governance program when there are well-defined data governance policies, processes, standards, and rules in place and a well-organized data

governance team with well-defined roles and clear demarcation of responsibilities with these data governance teams having the necessary expertise to effectively install as well as use the technologies to help support the adoption and enforcement of data governance processes and procedures. As stated by Dr. John Talburt (Mahanti 2021)—

> In general… "A new system is not the answer." Most obstacles in the adoption of new computing paradigms are "people problems," not technology issues. Viewing and managing information as an enterprise asset will require the same kind of cultural change that was needed for moving from procedural to object-oriented program, data warehousing, and now the shift to HDFS-based applications and cloud computing.

References

Collibra (2016) Collibra data governance center. https://www.collibra.com/wp-content/uploads/Collibra-DGC-DigitalBRO.pdf. Last accessed 19 Aug 2018

Houpes TJ (2018a) What you should know about Collibra Data Governance Center, TechTarget Search DataManagement. https://searchdatamanagement.techtarget.com/feature/What-you-should-know-about-Collibra-Data-Governance-Center. Last accessed 19 Aug 2019

Houpes TJ (2018b) What to know about the IBM Information Governance Catalog. TechTarget Search DataManagement. https://searchdatamanagement.techtarget.com/feature/What-to-know-about-the-IBM-Information-Governance-Catalog. Last accessed 19 Aug 2019

Howard D, Howard P (August, 2017) Informatica data governance, a bloor in-detail paper, bloor. http://www.integrationworx.net/sites/default/files/downloads/informaticadatagovernance.pdf

Ladley J (2016) How to narrow down your choices for buying a data governance tool. TechTarget Search DataManagement.https://searchdatamanagement.techtarget.com/feature/How-to-narrow-down-your-choices-for-buying-a-data-governance-tool. Last accessed 10 Aug 2018

Ladley J (2018) How data governance software helps ensure the integrity of your data. TechTargetSearch DataManagement. https://searchdatamanagement.techtarget.com/feature/How-data-governance-software-helps-ensure-the-integrity-of-your-data. Last accessed 10 Aug 2019

Loshin D (June, 2010) Operationalizing data governance through Data Policy Management, Knowledge Integrity, Inc. http://dataqualitybook.com/kii-content/OperationalizingDataGovernance.pdf. Last accessed 10 July 2018

Mahanti R (2019) Data quality: Dimensions, measurement, strategy, management and governance. Milwaukee: ASQ Quality Press. ISBN: 978-0-87389-977-2

Mahanti R (2021) Data governance and compliance: evolving to our current high stakes environment. Springer Books, Springer, number 978-981-33-6877-4

Peyret H, Goetz M (2014) The forrester wave™: data governance tools, Q2 2014, Forrester Research. http://vubtechtransfer.be/medialibrary/The_Forrester_Wave___Data.pdf. Last accessed 19 Aug 2018

Peyret H, Leganza G, Goetz M, Kramer A, Lynch D (2016) The forrester wave™: data governance Stewardship applications, Q1 2016, Forrester Research, Inc.

Plotkin D (2013) Data stewardship. Elsevier Science Publication

Violino B (01 February, 2018) 12 top products for data governance stewardship, information management. https://www.information-management.com/news/12-top-products-for-data-governance-stewardship?regconf=1. Last accessed 19 Aug 2018

Chapter 5
Data Governance and Data Management—Concluding Thoughts and Way Forward

Step by step, you make your way forward.
—Gretchen Rubin

Abstract This chapter discusses in brief—data and its governance, the data governance stakeholders, how data governance ties together the data governance functions and initiatives, and the data governance success factors.

5.1 Data and Its Governance

Data is an asset that needs to be governed effectively. Data has unique properties. Data is intangible—not physical in nature, non-rivalrous, non-fungible, contextual, accumulative, not consumed with use, easily transformed, easily shared, easily replicated and easily deleted. Data can be classified in a number of different ways such as data domains, uses of data, entities, acquisition or creation of data, data criticality, location of data, the sensitivity of data, and the level of protection it requires.

Organizations that have limited, ineffective or no data governance processes are most likely to have issues such as multiple versions of truth, lack of ownership or accountability of data, data quality issues, lack of understanding and use of data, loss of credibility of data source and lack of trust in data, difficulty in locating enterprise data, difficulty conforming to regulatory requirements, increasing risk of data breaches, legal, and compliance issues, and loss of reputation due to data breaches, legal, and compliance issues (Giordano 2010).

Data governance has several drivers, uses cases, and benefits, which we have discussed in detail in Chap. 2 of this book.

5.2 Data Governance Stakeholders

Some of the common stakeholders in a data governance program are data producers, data publishers, data consumers, data owners, data stewards, and data custodians. The common groups/bodies involved in data governance at different

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2021

R. Mahanti, *Data Governance and Data Management*,
https://doi.org/10.1007/978-981-16-3583-0_5

levels are Executive Steering Committee, Data Governance Council, Data Stewardship Council, Data Governance Office (DGO), and Information Technology Partners. Each of these stakeholders have responsibilities and/or accountabilities around data that need to be defined clearly.

5.3 Data Governance and Data Management

Data Governance represents the principal function of data management. It gives direction as to (Otto 2011):

- what decisions need to be made in data management, and
- who makes these decisions, and their roles and responsibilities in relation to the decisions.

Data governance ties together several data management disciplines and data initiatives—data architecture management, data modeling and design, data quality management (DQM), data security management, data warehousing and business intelligence (BI) management, data integration and interoperability (DII), document and content management, metadata management, reference data management, master data management, data storage and operations, big data, data lake, and data analytics.

Both data governance and data architecture are data strategy components. Data architecture provides inputs to decisions about how the data should be governed. Data governance in turn helps establish accountability for key data architecture artifacts (example, data models, data flows, and business rules) and helps build the business case for data architecture.

Data governance provides oversight to DII solutions in the areas of stakeholder engagement and management, establishment governance policies, processes, best practices, metadata management, data lineage, security, privacy, data sharing agreements, and data integration metrics.

Data governance assists metadata management by establishing metadata roles, responsibilities, standards, lifecycles, and statistics, in addition to how operational activities and related data management projects integrate metadata (Knight 2017). Metadata management ensures standardized metadata, which can help data governance in data discovery of same/similar data entities.

Data governance helps effective data security management by assisting in identification and management of stakeholders, creation of data security and classification policy, processes, and standards, discovery of sensitive data and assessment of risk, establishing data ownership and data stewardship, accountabilities and decision rights, data classification, ensuring appropriate processes and controls are in place to prevent data loss, theft, and corruption, ensuring that metrics are in place to assess the progress and effectiveness of your data security program, and that progress and effectiveness of the data security management program is tracked using the same.

Effective data governance in data storage and operations revolves around ensuring that policies, processes, and rules for acquisition, migration, retention, archival, expiration, and destruction of data are established and enforced, and roles, responsibilities, and accountabilities are in place to accomplish these activities.

Reference data governance ensures that policies and processes are there to manage internal and external reference data in a standardized fashion, and formal and well-defined accountabilities are there in place to manage reference data quality, and metrics are there to track the quality of reference data.

Data governance helps master data management in the identification and management of stakeholders, agreement regarding the critical data elements to master, assisting in defining policies, processes, rules, and standards for managing the golden record, establishing clear accountabilities and responsibilities for stakeholders involved in the master data initiative, establishing consensus on all associated reference data, and ensuring that metrics are agreed on and are in place to track effectiveness of master data management.

Many organizations equate data governance with data storage in a central repository and have a perception that data governance is about data storage in a central repository like a data warehouse. However, it is the roles and responsibilities, decision rights, processes, policies, standards, rules, and controls around storing, processing, and accessing critical data from the repository to ensure that it is secured effectively, and that the data and the associated metadata stored is of high quality and meets business needs, that forms a part of data governance. A data warehouse has a lot of stakeholders and data governance also assists in identification and management of these stakeholders.

Data mart which is a subset of the data warehouse and contains data related to a particular business unit (like sales, human resources, and finance) also need the same level of data governance as a data warehouse.

Organizations have lots of data in the form of documents and content that are not stored in databases but reside in document and content management systems. Governance is necessary to ensure that policies, processes, controls, roles, and responsibilities are there to ensure adequate protection, authorized access, and distribution.

The role of data governance in data quality management is supporting the development of high-quality data by ensuring the establishment and enforcement of policies, processes, roles, responsibilities and accountabilities, decision rights, standards, controls, metrics, and rules for creation and management of good data quality, and tracking and resolution of data issues.

The differences between big data and data or traditional data are along the lines of volume, velocity, and variety. The difference when applying to data governance to big data in comparison to traditional data is the agility that organizations should have throughout the big data life cycle.

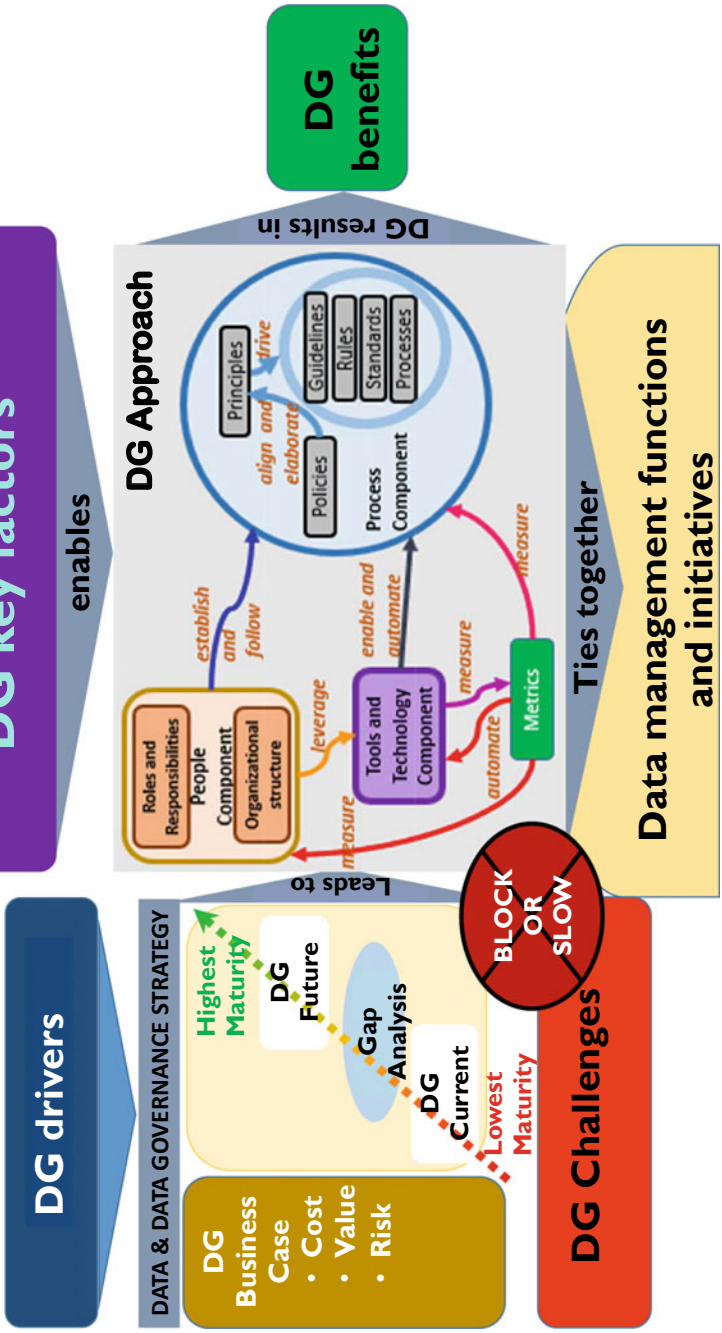

Fig. 5.1 Data governance in a page

5.4 Data Governance—The Way Forward

A successful data governance program involves a combination of people, processes, as well as tools and technology. The people and process components have been discussed briefly in this book but will be discussed in detail in the third book of the series—Data Governance Success.

Implementing data governance is not rocket science. However, it does involve a significant amount of effort, time, investment, and cultural change. Data governance is an ongoing endeavor. It has several dimensions to it, and implementing and sustaining the initiative is challenging and time consuming.

The various factors that need to considered when implementing the data governance program are strong executive sponsorship, strategy and business case, adequate training and education, organizational change management, communication and collaboration, skillsets, knowledge and abilities, DG framework, incremental approach to data governance, and DG tools and technology. We discussed the tools and technology components in detail in Chap. 4 of this book. The rest of the factors will be discussed in the third book in the series—Data Governance Success.

Figure 5.1 summarizes all aspects of DG that are discussed in this trilogy—Data Governance: The Way Forward.

References

Giordano AD (2010) Data integration blueprint and modeling: techniques for a scalable and sustainable architecture IBM Press, 2010

Otto B (2011) Data governance, business & information systems engineering, Gabler Verlag, pp 241–244, https://core.ac.uk/download/pdf/159155213.pdf

Knight M (2017) Fundamentals of Metadata Management, Dataversity, from https://www.dataversity.net/fundamentals-metadata-management/. Last accessed 10 Jan 2018

Appendix A
Restricted Data

Restricted data is sensitive data that is protected by law or policy. It requires the highest level of access control and security protections, whether in storage or in transit. Common examples of restricted data are as follows:

- Payment Card Industry (PCI) Information,
- Protected Health Information (PHI)/Electronic Protected Health Information (ePHI),
- Sensitive Personal Identifiable Information (PII), and
- Personally Identifiable Education Records.

A.1 Payment Card Industry (PCI) Information

Payment Card Industry Information is related to credit, debit, or other payment cards and is defined as a payment card number (generally a debit card number or credit card number; it is also referred to as a primary account number or PAN) in combination with one or more of the following data elements:

- Cardholder name,
- Service code,
- Expiration date,
- CVC2, CVV2 or CID value,
- PIN or PIN block, and
- Contents of a payment card's magnetic stripe.

Payment Card Industry Information is governed by the Payment Card Industry (PCI) Data Security Standards set by the PCI Security Standards Council.

© The Editor(s) (if applicable) and The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2021

R. Mahanti, *Data Governance and Data Management*,
https://doi.org/10.1007/978-981-16-3583-0

A.2 Protected Health Information (PHI)

Protected Health Information (PHI) refers to individually identifiable health information, that is:

(a) Transmitted by electronic media;
(b) Maintained in electronic media; or
(c) Transmitted or maintained in any other form or medium.

A.3 Individually Identifiable Health Information (IIHI)

Under HIPAA, individually identifiable health information (IIHI) is a subset of health information, and (TOC_HIT):

(a) Is created or received by a health care provider, health plan, employer, or health care clearinghouse;
(b) Relates to past, present, or future health or condition of an individual, provision or care, or payment; and
(c) Identifies the individual or there is a reasonable basis to believe that the information could be used to identify the individual.

PHI is considered individually identifiable if it contains one or more of the following identifiers:

- Name,
- Address (all geographic subdivisions smaller than state including street address, city, county, precinct or zip code),
- All elements of dates (except year) related to an individual including birth date, admissions date, discharge date, date of death and exact age if over 89,
- Telephone numbers,
- Fax numbers,
- Electronic mail addresses,
- Social security numbers,
- Medical record numbers,
- Health plan beneficiary numbers,
- Account numbers,
- Certificate/license numbers,
- Vehicle identifiers and serial numbers, including license plate number,
- Device identifiers and serial numbers,
- Universal Resource Locators (URLs),
- Internet protocol (IP) addresses,
- Biometric identifiers, including finger and voice prints,
- Full face photographic images and any comparable images, and

- Any other unique identifying number, characteristic or code that could identify an individual.

Not all IIHI is PHI. IIHI in the hands of a non-covered HIPAA entity is not PHI. IIHI in educational or employment records is not PHI, regardless if it is held by a covered entity (TOC_HIT).

A.4 Electronic Protected Health Information (e-PHI)

e-PHI is all PHI a covered entity creates, receives, processes, stores, and/or transmits in electronic form. A few examples include (UCS):

- Medical record number or social security number;
- Patient demographic data, for example—residence address, date of birth, date of death, sex, and email address;
- Digital photograph of the patient stored in the computer or secondary storage; and
- Medical records, reports, and test results, emailed to the patient.

A.5 Sensitive Personal Identifiable Information (PII)

To understand Sensitive Personal Identifiable Information, we first need to understand what Personal Identifiable Information is. The National Institute of Standards and Technology (NIST) defines PII as—

Any information about an individual maintained by an agency, including:

(1) *any information that can be used to distinguish or trace an individual's identity, such as name, social security number, date and place of birth, mother's maiden name, or biometric records; and.*
(2) *any other information that is linked or linkable to an individual, such as medical, educational, financial, and employment information.*

Personal identifiable information (PII) is information related or linked to an individual that can identify the individual. For example, an individual's first name is not sufficient to identify one individual from another individual; however, if both first name and social security number is present together, then the identity of the individual can be ascertained. Hence, an individual's name by itself is not PII but together with the social security number it is PII. On the other hand, attributes such is passport number or social security number alone are sufficient to identify an individual uniquely.

A sensitive PII is a PII which is not available to the public and if lost, compromised, or disclosed could result in substantial harm, embarrassment, inconvenience, or unfairness to an individual.

While sensitive PII is PII, not all PII is sensitive PII. For example, an individual's name, address, and phone number listed in the public telephone directory is PII, however, this combination is not sensitive PII as the information is already available to the public. However, if the phone number is unlisted, then a combination of the same elements becomes sensitive PII.

A.6 Personal Data from GDPR Perspective

The **EU Data Protection Directive (95/46/EC)** defines personal data as follows:

> "personal data" shall mean any information relating to an identified or identifiable natural person ('Data Subject'); an identifiable person is one who can be identified, directly or indirectly, in particular by reference to an identification number or to one or more factors specific to his physical, physiological, mental, economic, cultural or social identity.
>
> Personal data, in the context of GDPR, covers a much wider range of information when compared to PII and can include social media posts, photographs, lifestyle preferences, transaction histories, and even IP addresses (Seaman 2016). In other words, all PII is personal data but not all personal data is PII.

A.7 Personally Identifiable Education Records

Personally Identifiable Education Records are described in the Family Educational Rights and Privacy Act (FERPA) of 1974 (20 USC §1232g) and are defined as any education records that contain one or more of the following personal identifiers (CMU):

- Name of the student;
- Name of the student's parent(s) or another family member;
- Social security number;
- Student number;
- A list of personal characteristics that would make the student's identity easily traceable; and
- Any other information or identifier that would make the student's identity easily traceable.

Appendix B
Glossary of Terms

B.1 Asset

An asset can be described as any entity that has value, creates and maintains that value through its use, and has the ability to add value through its future use (Polikoff 2018).

B.2 Confidential Data

Confidential data is moderately sensitive information that needs to be protected from unauthorized access, and is only intended for limited dissemination.

B.3 Critical Data

Critical data are important for an organization to be able to operate efficiently. Loss of integrity or availability of critical data may result in considerable disruption, delay or degradation of vital services or functions, and would generally have a moderate short-term impact on business continuity or operational effectiveness.

B.4 Data

The International Organization for Standardization (ISO) defines data as "re-interpretable representation of information in a formalized manner suitable for communication, interpretation, or processing" (ISO 11179).

© The Editor(s) (if applicable) and The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2021

R. Mahanti, *Data Governance and Data Management*,
https://doi.org/10.1007/978-981-16-3583-0

B.5 Databases

Databases are a collection of related tables.

B.6 Data Classification

Data classification is the process of organizing data with similar characteristics into different groups for its most effective storage, protection, retrieval, distribution, and usage.

B.7 Data Criticality

Data criticality is a reflection of how important is the data to an organization's mission, functions, and processes in terms of integrity and availability.

B.8 Data Domain

Data domains also known as data subject areas are group classifications of business data entities at their highest level of data object abstraction and are not impacted by project related changes.

B.9 Data Governance

Data governance is the exercise and enforcement of policies, processes, guidelines, rules, standards, metrics, controls, decision rights, roles, responsibilities, and accountabilities to manage data as a strategic enterprise asset.

B.10 Data Lake

Data lake is a data repository that stores structured data, unstructured data and/or semi-structured data in their native or raw format.

B.11 Data Management (the Discipline)

The overarching umbrella which encompasses the different disciplines or functions of data management such as but not limited to data quality management, data security management, master data management, and metadata management.

B.12 Data Management (the Thing)

The action of actually managing the data.

B.13 Database Management System (DBMS)

Software or programs used to create and maintain database is called Database Management System (DBMS).

B.14 Data Profiling

Data profiling is a process to capture statistics that provides a picture of the current state of its data assets.

B.15 Data Quality

Data quality is the fitness of data for use in a given context or for a set of specific tasks in hand.

B.16 Data Quality Dimensions

The different dimensions used to characterize and measure the quality of data are called data quality dimensions. Each data quality dimension captures a particular aspect of data quality. Some examples of data quality dimensions are accuracy, completeness, and uniqueness.

B.17 Database Schema

Database schema is the structure in a database management system that holds the physical implementation of a data model.

B.18 Dataset

A dataset is a collection of data in the form of records (rows and columns) extracted from data files or database tables, for a specific purpose.

B.19 Data Stewardship

Data stewardship is the operational facet of a data governance program that involves the actual routine work of governing the enterprise's data (Plotkin 2013).

B.20 Data Warehouse

A data warehouse is a database repository that sources, integrates and consolidates data from multiple heterogeneous sources including OLTP databases, and covers a wide range of subject areas or data domains, provides a centralized view of data across the organization, and supports complex query processing, analytics and reporting.

B.21 Datamart

Data marts are essentially, specialized, sometimes local databases or custom built data warehouse offshoots that store data related to individual business units (e.g. sales, HR, and marketing) or specific subject areas (e.g. product, customer, asset, and event) or for addressing concerns of a particular business problem (e.g. increasing sales) (Mahanti 2019).

B.22 Dimension Modeling

Dimensional modeling is a data structure technique that aids business users to query data stored in the data warehouse.

B.23 Master Data

Master data is the consistent and uniform set of identifiers and extended attributes that describe the core entities of the enterprise, and are used across multiple business processes (Gartner 2016).

B.24 Metadata

Metadata is the data or additional information that describes other data, like master data, transactional data, and reference data.

B.25 Mission Critical Data

Mission critical data are vital for an organization to be able to operate efficiently. Loss of integrity or availability of mission critical data would result in extreme delay or degradation of vital services or function to the extent that they may not be able to deliver, and would have significant short-term impact and possible long-term impact on business continuity or operational effectiveness (Secure UD 2018).

B.26 Non-critical Data

Non-critical data are data whose loss of integrity or availability would result in a small delay or degradation of services and functions.

B.27 Normalization

Normalization, also known as data normalization is a refinement process and an efficient method of restructuring data in a database so as to eliminate or minimize the extent of unnecessary duplication and ensure that the data dependencies are logical without compromising the integrity of the stored data (Mahanti 2019).

B.28 Private or Internal Data

Private or internal data is data that has a moderate level of risk associated with its unauthorized disclosure, alteration or destruction of that data.

B.29 Public Data

Public data is data that has negligible or no risk and minimal adverse impacts associated with its unauthorized disclosure, distribution, modification, usage or destruction.

B.30 Reference Data

Reference data is a known set of permissible values that are referenced and shared by other data like master, transactional data, and systems with an aim to create a standard vocabulary, structure, and format across different systems.

B.31 Relational Database Management System (RDBMS)

Software system used to maintain relational databases is called relational database management system (RDBMS).

B.32 Restricted Data

Restricted data is sensitive data that is protected by law or policy. It requires the highest level of access control and security protections, whether in storage or in transit.

B.33 Semi-structured Data

Semi-structured data is data that may be erratic or incomplete, have a structure that may change rapidly or unpredictably, and organizational structures that makes it easier to analyze.

B.34 Structured Data

Structured data are data that have clearly defined data types, and are structured by predefined data models and schema.

B.35 Table

A table is collection of rows called records or tuples, and columns called fields to represent a set of characteristics of a particular entity. Tables generally store a huge amount of data in them.

B.36 Transactional Data

Transactional data is the data that describes the changes resulting from transactions or an event and always has a time dimension associated with it.

B.37 Unstructured Data

Unstructured data are data that have no identifiable structure and are not structured via pre-defined data models or schema, and hence are not stored in relational database tables.

Appendix C
Bibliography

Each source that I read, I would look through the bibliography and the footnotes, and use that as a map for the next thing I would read.

—Alexander Chee, American Fiction Writer

DAMA International (2017) DAMA-DMBOK: data management body of knowledge, 2nd edn. Technics Publication
DAMA International (2009) The DAMA guide to the data management body of knowledge (DAMA-DMBOK). Technics Publications
Seiner RS (2014) Non-invasive data governance. Technics Publications
Ladley J (2012) Data governance: how to design, deploy and sustain an effective data governance program (The Morgan Kaufmann series on business intelligence), 1st edn. Paperback. ISBN: 9780124158290
Plotkin D (2013) Data stewardship. Elsevier Science Publication
Sebastian-Coleman L (2018) Navigating the Labyrinth. Technics Publications
Mahanti R (2019) Data quality: dimensions, measurement, strategy, management and governance. Quality Press, ASQ, ISBN: 9780873899772
Mahanti R (2021) Data governance and compliance, Springer Books, Springer, number 978-981-33-6877-4
Talburt JR, Zhou Y (2015) Entity information life cycle for big data, Morgan Kauffmann, ISBN: 978-0-12-800537-8
Seaman J (10 December, 2016) GDPR: the difference between personally identifiable information (PII) and personal data, Linkedin. https://www.linkedin.com/pulse/gdprthe-difference-between-personally-identifiable-jim-seaman/
CMU (Carnegie Mellon University). Guidelines for data classification. https://www.cmu.edu/iso/governance/guidelines/data-classification.html

© The Editor(s) (if applicable) and The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2021

R. Mahanti, *Data Governance and Data Management*,
https://doi.org/10.1007/978-981-16-3583-0

UCS. https://its.ucsc.edu/security/restricted.html
Polikoff I (15 June, 2018) Data governance as a lifecycle-centric asset management activity, meta_connections. TopQuadrant company blog. https://www.topquadrant.com/2018/06/15/data-governance-as-a-lifecycle-centric-asset-management-activity/. Last accessed 6th Oct 2018
TOC_HIT. The office of the national coordinator for health information technology guide to privacy and security of health information, Version 1.0 022112, https://www.healthit.gov/sites/default/files/pdf/privacy/onc_privacy_and_security_chapter4_v1_022112.pdf
ISO (11179) https://www.iso.org/obp/ui/#iso:std:iso-iec:11179:-4:ed-2:v1:en
Gartner (2016) Gartner Glossary-Master Data Management, https://www.gartner.com/en/information-technology/glossary/master-datamanagement-mdm
Secure UD (2018) Understanding data criticality, University of Delaware, http://www1.udel.edu/security/data/criticality.html

Index

A
Aaron Zornes, 97
Absence of data governance, 46, *See* Lack of data governance
Access controls, 111, 115, 118, 124, 153
Amazon, ix, 6, 17, 19
American Institute of CPAs, 14
Amnon Drori, 105
Analytics. *See* Data analytics
Andres Perez, 60, 167
Andrew Lo. *See* Andrew W. Lo
Andrew W. Lo, 24 *See also* Andrew Lo
Anee Buff, 72
Anu Tirupathi, 114
ARMA International, 110
Assets, 14, 15, 17, 19, 20, 24, 25
  asset-definition, 14
  characteristics of assets, 14
  fixed assets, 15, 17
  intangible asset, 19
  tangible asset, 17, 19
  valuation, 20

B
Balance sheets, 16, 17, 19
  challenges of listing data in the balance sheet, 17–19
Ballard, 135
Batch processing, 135
Big data, 6, 68, 132, 133, 135–137
  big data-definition, 132
  Hadoop, 133
  variety, 133
  velocity, 133
  veracity, 133
  volume, 132
  3 V's, 68, 132
Big data environments, 69
Big data governance versus traditional data governance, 136
Big data versus traditional data, 132, 134
Business glossary, 151, 152
Business intelligence, 54, 86, 100, 101
Business risk category, 124

C
Cambridge Analytics, 19
Cambridge Semantics, 137
Charles Babbage, xiii
Charles E., 24
Chief Security Officer, 125
Christian Bremeau, 104
Christopher Butler, 78, 167
Classification, 26
Cloud, 6, 128
Cloud computing, 128
Collibra Data Governance Center, 164
Collibra On-the-Go, 164
Common understanding of data, 72
  common names, 73
Communication and collaboration, 155
Competitive advantage, 6, 7, 25, 64, 75
Compliance, 6, 47, 50, 62, 63, 77, 109, 115
Confidential data, 27, 37, 118, 179
  examples, 38
Conflict resolution process, 101
Content management, 110, 111 *See also* Enterprise content management (ECM)
  content management systems, 111
Controls, 38, 39, 67, 111, 112
Costing, 17
Country codes, 95

© The Editor(s) (if applicable) and The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2021

R. Mahanti, *Data Governance and Data Management*,
https://doi.org/10.1007/978-981-16-3583-0

Critical data, 35, 45, 77, 101, 102, 113, 179
Critical data assets, 21
Critical data elements, 88, 157
Critical master data elements, 97
Critical metadata, 106, 109
Critical success factor, 155
Cross-functional communication, 74
Cultural change, 130
Customer relationship management (CRM), 10
2017 Cyber Security Breaches Survey, 114

D
Dashboards, 158
Data, 1, 2, 5–8, 15, 19–21, 24, 39, 43, 44, 68, 78, 83, 87, 161
  asset to liability conversion , 19
  data-definition, 8
  data hierarchy, 11, 12
  data organization, 9
    columns. *See* Fields
    composite primary key, 9
    database management system (DBMS), 9
    database model, 9
    databases, 9, 10, 13
    database schema, 9
    database systems, 9
    data domains, 9
    data elements, 11
    dimensional model, 11
    fields, 9, 11
    file, 11
    foreign key, 9
    key, 9
    primary key, 9
    record, 9, 11 *See also* Row
    relational tables, 9
    relations, 9
    row, 11
    row, record, 9
    subject areas, data domains, 9
    tables, 9, 11
    tuple, 11
  data set, 12
  evolution of data, 1, 5, 24
  misconceptions, 25
  time dimension, 33
  unique properties of data, 16, 21
  varieties of, 31
Data aggregation, 15
Data analytics, 134
  data preparation, 68, 138
  data wrangling, 138
  descriptive analytics, 68
  diagnostic analytics, 68
  predictive analytics, 68
  prescriptive analytics, 67
Data architects, 88
Data architecture, 87–89
Data architecture management, 54, 85
  data architecture artifacts, 88
Data architecture (the discipline), 87
Data architecture (the thing), 87
Data as an asset, 24–26
Data asset register, 88
Data assets, 7, 14, 15, 17, 20–23, 33, 77
  life cycle, 21, 22, 63
  valuation, 20
Data asset versus data liability, 26
Data augmentation, 156
Database schema, 182
Data breaches, 66, 114
Data classification, 26, 34, 36, 39, 115, 120, 123, 125, 180
  data classification document, 123
    data classification document template, 123
  data classification policy, 116, 123
  data classification process, 115, 116
  data classification scheme, 116, 119
  data security classification, 124
  reclassification, 121
Data cleansing, 156
Data conflicts, 71
Data consolidation, 156
Data consumer, 51, 92, 127
Data criticality, 33, 180
Data custodian, 126
Data definitions, 25, 67, 138
Data dictionary, 74
Data discovery, 74, 88, 123, 130, 150
Data discovery tool, 150
Data domains, 33, 123, 152, 180
Data-driven, 25, 26, 75, 77, 89, 139
Data elements, 13, 40
Data explosion, 5
  factors leading to, 2
Data flow diagrams, 87, 89
Data governance, 3, 7, 14, 20–23, 26, 44, 45, 47, 50, 53–56, 58, 63–78, 83–85, 89, 90, 93, 101–103, 106, 128, 138, 149, 155, 180
  benefits, 71–77
  business drivers, 59, 61–71
  goals, 57
  level of governance, 51, 77
  light versus heavy governance, 52
  people aspect, 58, 59, 171

perception, 171
process component, 60
guidelines, 60
policies, 60
principles, 60
processes, 60
rules and standards, 60
stakeholders, 58, 60, 169
tools and technologies, 60, 145–148, 155, 159, 165
Data governance and cloud, 128
Data governance and data architecture, 87, 89, 90
Data governance and data management, 2
Data governance and data management functions, 85
Data governance and data migration, 103
Data governance and data quality, 55, 56
Data governance and reference data management, 93
Data governance and technology, 145
Data governance framework, 53, 68, 72
Data governance metrics, 157
Data governance strategy, 103, 104
Data governance tools, 147, 151–153, 158–161
Alation Data Catalog, 165
cloud-based tool, 161
Collibra Connect, 164
data governance functionalities, 166
data governance tool readiness check, 165
misconception, 148
plug and play tools, 147
readiness aspects, 159
Data governance tools versus data management tools, 147
Data governance tool vendors, 163
Data governance vendors
ALEX Solutions, 164
business intelligence (BI) platform vendors, 163
data governance specialists, 163
data management platform vendors, 163
datum, 164
Diaku, 165
Infogix, 164
Informatica, 163, 165–167
metadata repository vendors, 164
quality governance specialist, 164
Data initiatives, 127, 170
Data integration, 89, 90, 92
Data integration and interoperability, 54, 86, 89
Data integration metrics, 93
Data intensive initiatives, 49
Data interoperability, 89
Data issue resolution, 158
Data lakes, 13, 136–138
Data lake versus data mart, 13
Data landscape, 47, 48, 59, 148, 150
Data lineages, 21, 76, 88, 89, 92, 102, 105, 150, 151, 164
Data linkage, 156
Data management, 2, 7, 17, 53–55, 75, 77, 83–85
data management disciplines, 172
data management (the discipline), 84
data management (the thing), 84
Data Management Association (DAMA), 54, 84, 85
Data Management Capability Assessment Model (DCAM), 54
Data management tools, 147, 148
Data marts, 12, 13, 102, 182
Data matching, 156
Data mergers, 48
Data merging, 156
Data migration, 103
Data modeling, 88
Data modeling and design, 54, 85
Data modeling standards, 88
Data modeling vendors, 164
Data models, 88
conceptual data models, 88
data model quality metric, 88
logical data models, 88
physical data models, 88
relational model, 9
Data monitoring, 157
Data normalization. *See* Normalization
Data object, 74
Data owners, 74, 126, 131
Data ownership, 123
Data policy management, 158
Data processing, 1
Data producers, 51
Data profiling, 129, 130, 155, 157, 181
data profiling tool, 156
SQL queries, 155
Data protection, 92
Data publishers, 51
Data quality, 20, 39, 56, 67, 75, 76, 104, 137, 155, 181
data inaccuracy, 41
data quality-definition, 39
data quality issues, 66, 95
data quality metrics, 67, 131
data quality threshold, 21, 101–103, 129, 131, 137, 157
missing data, 40

Data quality and data governance, 56, 130, 155
Data quality dashboards, 157
Data quality dimensions, 20, 21, 40–44, 56, 181
accuracy, 40
completeness, 40
conformity, 40
conformity/validity, 41
consistency, 41
credibility, 44
currency, 43
data accessibility, 43
data coverage, 43
data inconsistency, 42
data integrity, 42
data security, 43
duplication, 41
granularity, 43
grain, 43
non-duplication. *See* Uniqueness
objective data quality dimensions, 44
reliability, 43
reputation, 44
subjective data quality dimensions, 44
timeliness, 43
traceability, 44
trustworthiness, 44
uniqueness, 41
volatility, 23, 43
Data quality management, 54, 56, 85, 128–130
data quality issue log, 129
issue classification, 129
issue status, 130
rating, 130
Data security, 35, 113, 115, 126
availability, 122
confidentiality, 122
integrity, 122
Data security and privacy, 66
Data security management, 54, 85, 114, 115
Data security strategy, 115
Data set, 182
Data sharing agreement, 93
Data sources, 44
Data stakeholders, 73, 74, 90, 128
Data standardization, 156
Data standards, 74
Data stewards, 74, 88, 124, 126, 131
Data stewardship, 123, 154, 182
Data stewardship tools, 154
Data storage and operations, 54, 86, 127
Data strategy, 74, 87
data strategy components, 170
Data subject areas, 33, 180
granularity, 33
Data swamp, 136, 137
Data value chain, 51
Data visibility, 64, 65
Data warehouse, 10–13, 100–102, 182
data marts, 10, 171
dimensional modeling, 11
dimension, 11
fact, 11
fact tables, 11
Data warehouse versus data lake, 12, 13
Data warehousing, 54, 86
Datum, 8
Davide Cervellin, 23
Delimited text files, 11
Depreciation, 17
depreciation cycles, 18
Digital age, 2, 5
Digital storage, 6
Digitization, 45
Dimensional modeling, 182
Document and content management, 54, 86, 110, 112
document and content management tools, 159
documents and contents, 109
Document management, 110, 111
data governance versus records management, 110
document management systems, 111, 158
documents versus records, 110
documentum, 124, 158
Doug Laney, 20
"Drowning in data" problem, 15

E

eBay, 6, 19, 23
Elait Australia, 47
Electronic Protected Health Information (e-PHI), 177
Embarcadero ER Studio, 164
Emily Washington, 67
Enterprise Content Management (ECM). *See* Content management
Enterprise Data Warehouse. *See* Data warehouse
Enterprise Resource Planning (ERP), 10
Entity, 8, 9
Entity - data categories, 30
Erik Brynjolfsson, 24
Erwin, 62, 64, 67, 69
ETL tools, 152
Excel dashboards, 158
Excel spreadsheets, 151

Extensible Markup Language (XML), 11
External data, 32
External reference data, 29, 87, 94, 96
External regulations, 62
Extract, Transform, and Load (ETL), 100
EY, 16

F
Facebook, ix, 6, 17, 19, 32, 46, 65, 132
Facts, 1, 8, 19, 75
Fair Market Value (FMV), 17
Financial Accounting Standards Board (FASB), 14, 17
Financial services, 24, 47
Fixed width text files, 11
Forbes, 101
Formal data governance, 48, 50
  need of formal data governance-indicators, 50
Formats, 41
Forrester, 23, 147, 163
Forrester research, 163, 164
Framework for Enterprise Architecture, The, 47, 67
Fred, 20
Friendfinder, 114

G
Gartner, 20, 68, 128, 129, 132, 136
General Data Protection Regulation (GDPR), 45, 62
Generally Accepted Recordkeeping Principles® (GAAP), 19, 110
George Firican, 77
Global IDs, 165
Gray Matter Analytics, 72
Greater collaboration, 74

H
Healthcare , 47
Hard skills, 20
Historical data, 33
HSBC, 78
HSBC, UK, 167

I
IBM, 62
IBM's InfoSphere Information Server, 164
IFRS, 19
Impact analysis, 76
Individually Identifiable Health Information (IIHI), 176
Industry classification codes, 87, 88
Ineffective data governance, 45, 46
Infogix, 67
Informal data governance, 48
Informal versus formal data governance, 48, 49
Information Analyzer tool, 164
Information governance, 50
Information Governance Catalog tool, 165
Integration, 71
Internal data, 27, 32
Internal reference data, 94
Internal transparency, 65
International Accounting Standards Board, 14
International Organization for Standardization (ISO). *See* ISO
International standards for metadata structure, 107
Internet age, 2, 64
Internet era, 6
Internet of Things (IoT), 6, 161
IRM Consulting, Ltd. Co., 60
ISO, 96 *See also* International Organization for Standardization (ISO)
IS0-3166-1, 19
  ISO-3166-1 country codes, 96
ISO-4217, 19
  ISO-4217 currency codes, 96
ISO data standards, 19

J
Jake Stein, 13
James Dixon, 13
JavaScript Object Notation (JSON), 11
Jenn Riley, 30
Jill Dyché, 139
Jill Dyché, LLC., 139
John A. Zachman, 67
John Parkinson, 87
John R. Talburt, 24, 77, 104, 105, 136, 138, 148, 168 *See also* John Talburt
John Talburt. *See* John R. Talburt
John Zachman, 46, 47 *See also* John A. Zachman
Jora Gill, 24

K
Koutroumpis and Leiponen, 20
Kroger Co, The, 7

L
Lack of data governance, 45, 56 *See also* Absence of data governance
Laura Sebastian-Coleman, 25
LinkedIn, 6, 7, 16, 32, 132

Location of the data, 35

M
Magnetic tapes, 7
Mark Weinberger, 16
Marriott, 114
Master data, 17, 18, 27, 33, 38, 39, 43, 51, 96, 183
  examples, 29
Master data elements, 97
Master data management, 54, 70, 86, 96–99, 157, 173
  golden record of data, 97
  single golden record, 157
Master Systems Inc., 167
Matching and linking, 156
McDonald, 75
Mergers and acquisitions, 48, 70
Metadata, 6, 8, 13, 30, 31, 43, 51, 92, 104, 105, 137, 183
  business metadata, 31
  level of governance, 51
  process metadata, 31
  sources of metadata, 105
  technical metadata, 31
Metadata automation, 104
Metadata discovery, 102
Metadata governance, 105, 109
Metadata management, 54, 86, 92, 106, 107, 151
Metadata management and data governance, 105
Metadata management tool, 155
Metadata model, 107
Metadata quality statistics, 109
Metadata repository, 107, 109, 152
  metadata repository characteristics, 108
Meta Integration Technology, 104
Meta model, 10
Metrics, 88, 96, 131
Microsoft, 7
Microsoft Excel, 149
Microsoft Word, 158
Mission critical data, 35, 183
Mobile computing, 6
Multiple versions of truth, 45

N
National Institute of Standards and Technology (NIST), 177
Noetic Partners Inc., 6, 104, 105, 138, 148
Nomenclature des Activités Économiques dans la Communauté Européenne (NACE), 95
Non-critical data, 35, 183
Normalization, 10 *See also* Data normalization
NOSQL databases, 9, 32

O
Octopai, 105
OLTP databases, 10
One-size-fits-all, 115
Online Analytical Processing (OLAP), 10
Online Transaction Processing (OLTP) systems, 10
Operational efficiency, 70, 72, 130

P
Paper files, 5
Parsing, 156
Partners and outsourcers, 71
Payment Card Industry Information, 175
Pentaho, 13
Performance, 128, 161
Personal Data from GDPR, 178
Personal Identifiable Information (PII), 177
Personal information, 39
Personally Identifiable Education Records, 178
Peter Aiken, 7, 24
PEXA, 70
Phil Watt, 46, 47
Poor data quality-impacts, 128
Prashant Southekal, 17
PricewaterhouseCoopers, 7, 26, 139 *See also* PwC
Private data, 27, 38
  examples, 38
Private or internal data, 38, 121
Protected Health Information (PHI), 176
Public data, 27, 38, 51, 121
  examples, 38
Punch cards, 7
PwC. *See* PricewaterhouseCoopers
PwC, 26

Q
Qlik, 101

R
Reference and master data management, 54, 86
Reference data, 29, 33, 43, 51, 93–95
Reference data governance, 95, 96, 173
Reference data management, 94, 95
Regulations, 19, 62, 63, 67, 109
Regulators, 6
Regulatory requirements, 62
Relational database management system (RDBMS), 9

Reputation management, 64
reputation, 64
Responsible, Accountable, Consulted, and Informed (RACI), 95, 99, 101
Restricted data, 27, 37–39, 118, 175
examples of restricted data, 37
Rigor, 47
Risk management, 65
data misuse, 76
Risk quotient, 45
Robert Seiner, 148
Role-and policy-based access controls, 93

S
Scalability, 161
Schema-on-read, 12
Schema-on-write, 12
Scorecards, 158
Sean Martin, 137
Security standards, 121
Segregation of duties, 125
Semi-structured data, 32
examples, 32
Sensitive Personal Identifiable Information, 177 *See also* Sensitive PII
Sensitive PII. *See* Sensitive Personal Identifiable Information
Sensitivity, 35
Service level agreement (SLA), 43, 88
Shannon Fuller, 72
SharePoint, 58, 148, 149, 158
Single version of the truth, 68
Smart metering, 133
Social media, 6, 64
Soft skills, 20
Source system, 44
Spreadsheets, 152, 153, 158
Standard data definitions, 41
Standard Industry Classification (SIC), 95
Standard naming conventions, 67
Stan Rifkin, 24, 167
2018 State of Data Governance Report, The, 51, 62, 64, 67, 69
Stitch, 13
Structured data, 31
examples, 31
Structured Query Language (SQL), 31
Susan T. Harris, 24

T
Technologies, 6, 25, 53, 84, 127, 157, 167
Ted Friedman, 128
Text files, 11
Tony Epler, 7, 26, 139
Tools and technologies, 6, 132, 145, 146
Traditional data, 31
Transactional data, 18, 27, 43, 51
examples, 27
Transaction data, 27

U
Uber, 114
UK Treasury, 17
University of Arkansas, 6, 104, 105, 138, 148
Unstructured data, 32
examples, 32
User lists, 153
Uses of data, 33

V
Value, 14
Value of data, 7, 15, 17, 19, 23, 24
perceptions, 23

W
Wall Street Journal, 17
Wang and Karel, 97
WeWork, 24
Wiki, 58, 148, 149, 158
Workflow management, 153
Workflows, 154

Y
Yahoo, 114

Z
Zachman Framework, 46, 47
Zachman International, 47, 67
Zeljko Panian, 63