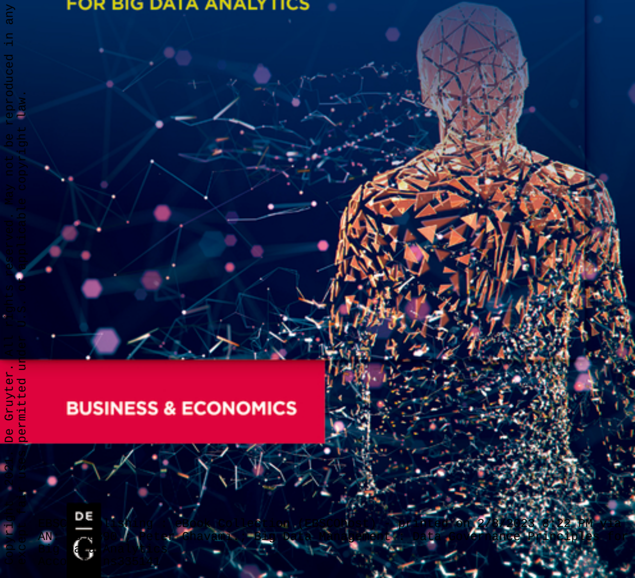
DE GRUYTER

Peter Ghavami

BIG DATA MANAGEMENT

DATA GOVERNANCE PRINCIPLES FOR BIG DATA ANALYTICS



BUSINESS & ECONOMICS

Peter Ghavami

Big Data Management

Peter Ghavami

Big Data Management

Data Governance Principles for Big Data Analytics

DE GRUYTER

The author and publisher have taken care in preparations of this book, but make no expressed or implied warranty of any kind and assume no responsibility for errors or omissions. No liability is assumed for the incidental or consequential damages in connection with or arising out of the use of the information or designs contained herein.

ISBN 978-3-11-066291-7
e-ISBN (PDF) 978-3-11-066406-5
e-ISBN (EPUB) 978-3-11-066432-4

Library of Congress Control Number: 2020943612

Bibliographic information published by the Deutsche Nationalbibliothek
The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data are available on the Internet at http://dnb.dnb.de.

Cover image: DKosig/iStock/Getty Images Plus
Typesetting: Integra Software Services Pvt. Ltd.
Printing and binding: CPI books GmbH, Leck

www.degruyter.com

To my beautiful wife Massi,
whose unwavering love and support make these accomplishments possible
and worth pursuing.

Acknowledgments

This book was only made possible as a result of my collaboration with many world-renowned data scientists, researchers, CIOs, and leading technology innovators who have taught me a tremendous amount about scientific research, innovation, and more importantly, about the value of collaboration. To all of them I owe a huge debt of gratitude.

Peter Ghavami
September 2020

https://doi.org/10.1515/9783110664065-202

About the Author

Peter K. Ghavami received his PhD in Systems and Industrial Engineering from the University of Washington in Seattle, specializing in big data analytics. He has served as head of data analytics at several financial institutions including CapitalOne Financial. He received his BA from Oregon University in Mathematics with an emphasis in Computer Science. He received his MS in Engineering Management from Portland State University. His career started as a software engineer, with progressive responsibilities at IBM as systems engineer. Later he became director of engineering, chief scientist, VP of engineering and product management at various high technology firms.

Before coming to CapitalOne Financial, he was director of informatics at UW Medicine leading numerous clinical system implementations and new product development projects. He has been a strategic advisor and VP of informatics at various analytics companies.

He has authored several papers, books, and book chapters on software process improvement, vector processing, distributed network architectures, and software quality. His first book, titled *Lean, Agile and Six Sigma Information Technology Management* was published in 2008. He has also published a book on data analytics titled *Big Data Analytics Methods: Analytics Techniques in Data Mining, Deep Learning and Natural Language Processing*, which has become a popular textbook on data science.

Peter is on the advisory board of several clinical analytics companies and is often invited as a lecturer and speaker on this topic. He is certified in ITIL and TOGAF9. He is a member of IEEE Reliability Society, IEEE Life Sciences Initiative, and HIMSS. He has been an active member of the HIMSS Data Analytics Task Force and advises Fortune 500 executives on data strategy.

https://doi.org/10.1515/9783110664065-203

Contents

Acknowledgments —— VII

About the Author —— IX

Introduction —— XVII

Part 1: **Big Data Overview**

Chapter 1
Introduction to Big Data —— 3
The Three Dimensions of Analytics —— **6**
The Distinction Between BI and Analytics —— **7**
Analytics Platform Framework —— **8**
Data Connection Layer —— **9**
Data Management Layer —— **10**
Analytics Layer —— **13**
Presentation Layer —— **16**
The Diverse Applications of Big Data —— **17**
The Future of Big Data —— **18**
Traditional Data Storage Models and Methods —— **20**
Why Data Governance Matters —— **22**

Chapter 2
Enterprise Data Governance Directive —— 24
Data Privacy Governance and GDPR —— **24**
The Impact of California Consumer Privacy Act —— **26**
The New York Shield Act —— **27**
Forming Data Subject Access Rights with Big Data —— **28**
Data Management Body of Knowledge (DMBOK) —— **29**
DMBOK2: What Is It? —— **29**
Nine Reasons for DMBOK2 —— **29**
Data Maturity Model (DMM) —— **32**

Part 2: **Big Data Governance Fundamentals**

Chapter 3
Data Risk Management —— 39
Top 10 Data Breaches —— **39**

What Is Data Risk Management? — 42
What Is Governance? — 43
Why Big Data Governance? — 44
Data Steward Responsibilities — 44
Corporate Governance — 45
Big Data Governance Certifications — 45
The Case for Big Data Governance — 46
Operational, Tactical, or Strategic Data Governance? — 46
TOGAF View of Data Governance — 47
Managing Risk: Data Lake versus Data Warehouse — 52
Data Governance in the Cloud — 54
Implementing Data Governance in the Cloud — 56
The Eight Best Practices for Big Data Governance — 56

Chapter 4
NoSQL Storage and Security Considerations — 58
Hadoop Overview — 59
HDFS Overview — 59
MapReduce — 61
Security Tools for Hadoop — 63
Snowflake Technology Overview — 63

Chapter 5
The Key Components of Big Data Governance — 64
Myths About Big Data and Hadoop Lake — 65
Data Governance Is More than Risk Management — 66
First Steps Toward Big Data Governance — 66
What Your Data Governance Model Should Address — 67
Data Governance Tools — 67
Big Data Governance Framework: A Lean and Effective Model — 69
Organization — 69
Data Quality Management — 70
Metadata Management — 71
Compliance, Security, and Privacy Policies — 72

Chapter 6
Big Data Governance Framework — 73
Introduction to Big Data Governance Rules — 74
Organization — 75
Data Governance Council — 75
Data Stewardship — 76

Authority: Policy, Decision Making, Governance — 77
Governance Activities: Monitoring, Support, and More . . . — 78
Users: Developers, Data Scientists, End Users — 78

Chapter 7
Master Data Management — 80
Metadata Management — 83
Big Data Classification — 84
Security, Privacy, and Compliance — 85
Apply Tiered Data Management Model — 87
Big Data Security Policy — 88
Big Data Security: Access Controls — 89
Big Data Security: Key Policies — 89
Data Usage Agreement Policy — 90
Security Operations Policies — 91
Information Lifecycle Management — 92
Quality Management — 92
Big Data Quality and Monitoring — 92
Data Classification Rules — 93
Data Quality Policies — 94
Metadata Best Practices — 95

Chapter 8
Big Data Governance Rules: Best Practices — 99
Sample Data Governance and Management Tools — 100
Data Governance in the GRC Context — 101
The Costs of Poor Data Governance — 102
Big Data Governance Budget Planning — 102
What Are Other Companies Doing? — 102

Chapter 9
Big Data Governance Best Practices — 104
Data Governance Best Practices — 104
Data Protection — 104
Sensitive Data Protection — 105
Data Sharing Considerations — 106
Adherence to Governance Policies — 107
Data Protection at the High Level — 107
Data Tokenization — 108
Low Level Data Access Protection — 108

Security Architecture for the Data Lake — 108
Big Data Classification — 111
Raw Data — 111
Keyed Data — 111
Validated Data — 112
Refined Data — 112
Hadoop Lake Data Classification and Governance Policies — 112
Metadata Rules — 115
Data Structure Design — 115
Sandbox Functionality Overview — 117
Detailed Access Policies — 117
Split Data Design — 118

Chapter 10
Big Data Governance Framework Program — 120
Overview — 120
Integrity of Information, Data Resources, and Assets — 121
Benefits of Compliance and Risks of Noncompliance — 122
Definitions: Terms that Your Program Must Define — 123
I Organization — 124
II Metadata Management — 125
III Data Classification — 127
IV Security, Privacy and Compliance Standards — 129
Authentication — 131
Authorization — 131
Visibility and Audit Reports — 133
V Data Usage Agreement (DUA) — 133
VI Security Operations and Policies — 134
VII Information Lifecycle Management — 135
VIII Data Quality Standards — 136
Data Quality Reporting — 138

Part 3: **Big Data and Model Risk Management**

Chapter 11
Why Data and Model Risk Management? — 143
The Future of Big Data Governance — 143
Examples of Data and Model Bias — 145
Model Risk Management Overview — 147

Stage 1: Establish the Foundation of MRM — 148
Stage 2: Operationalize and Execute — 148
Stage 3: Enhance and Optimize — 148

Summary — 151

Index — 153

Introduction

The future of business is big data. While the wealth of an organization may be displayed in balance sheets and electronic ledgers, the real wealth of the organization is in its information assets – in data and how well the organization harnesses value from it.

While open source storage systems for big data (such as Hadoop) promise to provide the ultimate flexibility and power in storing and analyzing data, because Hadoop was not designed with security and governance in mind, we face new and additional challenges in managing data to meet corporate and IT governance standards. I offer the best practices in data governance after sampling the best and most successful policies and processes from around the world and offer you a simplified, low cost, but highly effective handbook to big data governance. That's why this book is indispensable to implementing big data analytics.

Knowledge is information and information is derived from data. Without data governance and data quality, without adequate data integration and information lifecycle management, the chance of harnessing this value and leveraging from data will be very limited.

According to expert reports, data volumes in 2020 are about 50 zettabytes, compared to 2010 when there were just around 1.2 zettabytes[1] (1.2 billion gigabytes) of data worldwide. The volume of data is expanding rapidly, doubling every 12–18 months. It's expected that worldwide data volume will grow more than three times by 2025, reaching over 175 zettabytes.[2] The majority of this data is unstructured in the form of PDFs, spreadsheets, images, multimedia (audio, video), geolocation data (GPS), emails, social content, web pages, machine data, as well as GPS and sensor data.

The purpose of this book is to present a practical, effective, no frills, and yet low-cost data governance framework for big data. You'll find this book to be concise and to the point, highlighting the important and salient topics in big data that you can implement to achieve an effective data governance structure but at a low implementation cost. The premise of the policies and recommendations in this book are based on best practices from around the world in big data governance. I've included best practices from some of the most respected and leading-edge companies who have successfully implemented big data and governance.

To learn more about big data analytics, you can read two companion books. The first book is titled *Clinical Intelligence – The Big Data Analytics Revolution in Healthcare: A Framework for Clinical and Business Intelligence*. It can be found at:

[1] Each zettabyte is roughly 1000 exabytes and each exabyte is roughly 1000 petabytes. A petabyte is about 1000 terabytes.

[2] https://www.statista.com/statistics/871513/worldwide-data-created/

https://doi.org/10.1515/9783110664065-205

https://www.createspace.com/4772104. The second companion book is titled *Big Data Analytics Methods: Analytics Techniques in Data Mining, Deep Learning and Natural Language Processing 2nd Edition* (ISBN 9781547417957). It can be found on Amazon and at fine booksellers.

This book consists of four major parts. Part 1 offers an overview of big data and open source big data storage options like Hadoop. Part 2 is an overview of big data governance concepts, structure, architecture, policies, principles, and best practices. Part 3 presents the best practices in big data governance policies. Finally, Part 4 includes a ready-to-use template for governance structure written in a flexible format that you can easily adapt to your organization.

The contents of this book are presented in a lecture-like manner using a presentation slide deck style that is available from the publisher for academic courses or corporate training programs. The companion book, mentioned above, covers the data science aspects of big data for those who are interested in big data analytics.[3]

Now, let's start our journey through the book.

[3] Ghavami, P. (2019). *Big Data Analytics Methods: Analytics Techniques in Data Mining, Deep Learning and Natural Language Processing* (2nd ed.). Boston, MA: De Gruyter.

Part 1: **Big Data Overview**

Chapter 1
Introduction to Big Data

Data is the new *gold*. And *analytics* is the machinery that mines, molds, and mints it. Big data analytics is a set of computer-enabled analytics methods, processes, and discipline of extracting and transforming raw data into meaningful insight, new discovery, and knowledge that helps make more effective decision making. Another definition describes big data analytics as the discipline of extracting and analyzing data to deliver new insight about the past performance, current operations, and prediction of future events.

Before there was big data analytics, the study of large data sets was called *data mining*. But big data analytics has come a long way in a decade and is now gaining popularity thanks to the eruption of five new technologies: big data analytics, cloud computing, mobility, social networking, and smaller sensors. Each of these technologies is significant in its unique way to how business decisions and performance can be improved and how vast amounts of data can be generated.

Big data is known by its three key attributes known as the three Vs: *volume, velocity, and variety*. The world's storage volume is increasing at a rapid pace, estimated to double every year. The velocity at which this data is generated is rising, fueled by the advent of mobile devices and social networking. In medicine and healthcare, the cost and size of sensors has shrunk, making continuous patient monitoring and data acquisition from a multitude of human physiological systems an accepted practice.

With the advent of smaller, inexpensive sensors and the volume of data collected from people, internet, and machines we're challenged with making increasingly analytical decisions quickly from the large sets of data that are being collected. This trend is only increasing giving rise to what's known in the industry as the "big data problem": the rate of data accumulation is rising faster than people's cognitive capacity to analyze increasingly large data sets to make decisions. The big data problem offers an opportunity for improved predictive analytics and fact-based decisions.

Big data is now regarded as one of the leading strategic business imperatives. Research points to an increasing number of executives who believe that without big data they will face extinction.[1] Every day more than 4 petabytes of data is created on Facebook, 500 million tweets are sent, and more than 5 billion searches are made online. Several companies now offer big data storage solutions for on-premise (on-prem) and cloud-based solutions. Open source platforms like Hadoop data vaults

[1] Columbus, L. (2018, May 23). 10 Charts That Will Change Your Perspective of Big Data's Growth. *Forbes*. https://www.forbes.com/sites/louiscolumbus/2018/05/23/10-charts-that-will-change-your-perspective-of-big-datas-growth/#5a90e3a62926

https://doi.org/10.1515/9783110664065-001

and warehousing products including Cloudera, MapR, and HortonWorks have been implemented by most major corporations. The dominant cloud platforms such as Amazon AWS and Microsoft Azure offer some form of Hadoop, elastic storage, and other brands of parallel data storage solutions. Companies like Data Bricks, Snowflake, and Denodo are growing in the midst of the appetite to store and manage ever-expanding data sets.

The NoSQL (Not only SQL) and non-relational database movement has evolved beyond Hadoop to include Snowflake, Data Bricks, Elastic Search, and Spark. These unconventional solutions offer new data management tools that allow storage and management of large structured and non-structured data sets. But the biggest challenges of big data governance remain mostly uncharted territory at this time.

The variety of data is also increasing. For example, medical data was confined to paper for too long. As governments around the world, such as the United States, pushed medical institutions to transform their practice into electronic and digital format, patient data became digital and took on diverse forms. It's now common to think of electronic medical records (EMR) as including diverse forms of data such as audio recordings; MRI, ultrasound, computed tomography (CT), and other diagnostic images; videos captured during surgery or directly from patients; color images of burns and wounds; digital images of dental x-rays; waveforms of brain scans; electrocardiograms (EKG); and the list goes on.

How will we manage and govern this vast and complex sea of data? What will be the costs of poor or no data governance? The goal of this book is to show the components and tips on establishing both effective and lean data governance. This book will contemplate the costs of doing nothing but presents a framework to bring data governance to big data.

New types of data include structured and unstructured text. It will include server logs and other machine generated data. It will include data from sensors, web sites, machines, mobile and wearable devices. It will include streaming data and customer sentiment data about you. It includes social media data including blogs, LinkedIn, Twitter, Instagram, Facebook, and local RSS feeds about your organization, your people, and products. All these varieties of data types and many more can be harnessed to provide a more complete picture of what is happening in delivering value to your customers.

The traditional data warehouse strategies based on relational databases suffer from a latency of up to 24 hours. These data warehouses can't scale quickly with large data growth; and because they impose relational and data normalization constraints, their use is limited. In addition, they provide retrospective insight and not real-time or predictive analytics. Big data analytics will become a more real-time and "in-the-moment" decision support tool than the traditional business intelligence process of generating batch-oriented reports.

Semantics are critical to data analytics. As much as 60–80% of all data is unstructured data in the form of narrative text, emails, or audio recordings. Correctly extracting the pertinent terms from such data is a challenge. Tools such as *natural language processing* (NLP) methods combined with subject-specific libraries and ontologies are used to extract useful data from the vast amount of data stored in a Hadoop *Data Lake*. Hadoop Data Lake is a common term that describes the vast storage of data in the Hadoop file system. *Data lake* is a term increasingly used to refer to the new generation of big data warehouses where all data is going to be stored using open source and Hadoop technology. However, understanding the differences among sentence structure, context, and relationships between business terms is critical to detecting the semantics and meaning of data.

Given the rising volume of data and the demand for high-speed data access that can handle analytics, IT leaders are contemplating investing in a variety of tools and architectures. For about a decade, most solutions explored non-SQL solutions. SQL was a fundamental feature of data warehouses. It lost favor when NoSQL technologies like Hadoop and MongoDB emerged. NoSQL databases are ideal for storage and retrieval of unstructured data. But a new breed of data storage systems emerged that combine the best of both SQL and non-SQL data storage models.

Amazon offers RDS that comes with PostgreSQL by default, but gives you the option to choose Oracle or other databases. Microsoft Azure Datawarehouse has evolved to perform quite well on a variety of data schema and structures. Today SQL interfaces on top of Hadoop, Kafka, Spark, and many other database systems.

Transactional systems and reporting platforms were not designed to handle high-speed access to big data for analytics and thus are inadequate. As a result, specialized data analytics platforms are needed to handle high-volume data storage and high-speed access required for analytics.

Big data analytics requires very fast data load and data read functionality. In response to the faster database performance needs, dedicated analytics platforms like Hadoop, Snowflake, Denodo, Data Bricks, and other NoSQL databases and open source tools for data lakes have been adopted. Handling streaming data requires special database tools that can read and store data in real time. Some of the common open source tools include Kafka and NiFi. Other solutions include Amazon's AWS Kinesis.

While big data analytics promises phenomenal improvements in every industry, as with any technology acquisition, we want to take prudent steps toward adoption. We want to define criteria for success, gauge *return on investment* (ROI), and its data-centric metric that I define as *return on data* (ROD). A successful project must demonstrate palpable benefits and value derived from new insights. Implementing big data analytics is a necessary and standard procedure for most organizations as they strive to identify any remaining opportunities in improving efficiency, strategic and marketing/sales advantage, and cutting costs. Managing data and data governance are critical to the success of big data analytics initiatives as the volume of data increases.

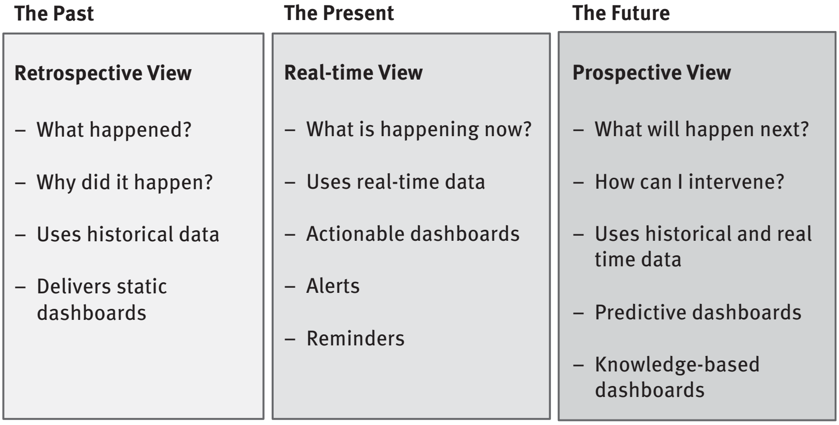
In addition to implementing a robust data governance structure as a key to big data analytics success, as I mentioned in my earlier book *Lean, Agile and Six Sigma*

IT Management, successful implementation of big data analytics also requires a team effort combined with lean and agile practices. A team effort is required because no single vendor, individual, or solution satisfies all analytics needs of the organization, and data governance is no different. Collaboration and partnerships among all user communities in the organization as well as among vendors is needed for successful implementation. A lean approach is required in order to avoid duplication of efforts, process waste and discordant systems.

The future of big data analytics is bright and will be so for many years to come. We're finally able to shed light on the data that have been locked up in the darkness of our electronic systems for years. When you consider other applications of analytics yet to be discovered, there are endless opportunities to improve business performance using data analytics.

The Three Dimensions of Analytics

Data analytics efforts may focus on any of the three temporal dimensions of analysis: retrospective analysis, real-time (current time) analysis, and predictive analysis. *Retrospective analytics* can explain and provide knowledge about the events of the past, show trends, and help find root causes for those events. *Real-time analysis* shows what is happening right now. It works to present situational awareness, send alarms when data reaches a certain threshold, or send reminders when a certain rule is satisfied. *Prospective analysis* presents a view into the future. It attempts to predict what will happen and determine the future values of certain variables. Figure 1.1 shows the taxonomy of the three analytics dimensions.



The Past

Retrospective View

- What happened?
- Why did it happen?
- Uses historical data
- Delivers static dashboards

The Present

Real-time View

- What is happening now?
- Uses real-time data
- Actionable dashboards
- Alerts
- Reminders

The Future

Prospective View

- What will happen next?
- How can I intervene?
- Uses historical and real time data
- Predictive dashboards
- Knowledge-based dashboards

Figure 1.1: The Three Temporal Dimensions of Data Analytics.

The Distinction Between BI and Analytics

The purpose of *business intelligence* (BI) is to transform raw data into information, insight, and meaning for business purposes. Analytics is for discovery, knowledge creating, assertion and communication of patterns, associations, classifications, and learning from data. While both approaches crunch data and use computers and software to do that, the similarities end there.

With BI, we're providing a snapshot of the information, using static dashboards. We're working with normalized and complete data typically arranged in rows and columns. The data is structured and assumed to be accurate. Data that is out of range or outliers are often removed before processing. Data processing uses simple, descriptive statistics such as mean, mode, and possibly trend lines and simple data projections to extrapolate about the future.

In contrast, analytics deals with all types of data both structured and unstructured. In medicine, about 80% of data is unstructured in the form of medical notes, charts, and reports. Analytics does not mandate data to be clean and normalized. In fact, it makes *no assumption* about data normalization. It can analyze many varieties of data to provide views into patterns and insights that are not humanly possible. Analytics methods are dynamic and provide dynamic and adaptive dashboards. They use advanced statistics, artificial intelligence techniques, machine learning, feedback, and natural language processing (NLP) to mine through the data. They detect patterns in data to provide new discoveries and knowledge. The patterns may have a geometric shape that data scientists believe may have mathematical representations that explain the relationships and associations among data elements.

Unlike BI dashboards that are static and show snapshots of data, analytics methods provide data visualization and adaptive models that are robust to changes in data and, in fact, learn from changes in data. While BI uses simple mathematical and descriptive statistics, data analytics is highly model-based. A data scientist builds models from data to show patterns and actionable insights. Feedback and machine learning are concepts found in analytics. Table 1.1 illustrates the distinctions between BI and analytics.

Table 1.1: Business Intelligence vs. Data Analytics.

Business Intelligence	Advanced Data Analytics
Information from processing raw data	Discovery, insight, patterns, learning from data
Structured data	Unstructured & structured data
Simple descriptive statistics	NLP, classifiers, machine learning, pattern recognition, predictive modeling, optimization, model-based

Table 1.1 (continued)

Business Intelligence	Advanced Data Analytics
Tabular, cleansed & complete data	Dirty data, missing & noisy data, non-normalized data
Normalized data	Non-normalized data, many types of data elements
Data snapshots, static queries	Streaming data, continuous updates of data & models, feedback & auto-learning
Dashboards snapshots & reports	Visualization, knowledge discovery

Analytics Platform Framework

When considering building analytics tools and applications, a data analytics strategy and governance is recommended. One strategy is to avoid implementing point-solutions that are stand-alone applications and do not integrate with other analytics applications. Consider implementing an analytics platform that supports many analytics applications and tools integrated in the platform. A 4-layer framework is proposed here as the foundation for entire enterprise analytics applications. The 4-layer framework consists of a data connection layer, a data management layer, an analytics engine layer, and a presentation layer as shown in Figure 1.2.

In practice, you'll make choices about what software and vendors to adopt for building this framework. The data management layer includes the distributed or centralized data repository. This framework assumes that modern enterprise data warehouses will consist of distributed and networked data warehouses using an open source environment like Hadoop.

The analytics layer may be implemented using SAS and/or the R statistical language or solutions from other vendors who provide the analytics engines in this layer.

The presentation layer may consist of various visualization tools such as Tableau, QlikView, Microsoft Power BI, or other applications. In a proper implementation of this framework, the visualization layer offers analytics-driven workflows and therefore tight integration between the presentation layer and the other two layers (data and analytics). This is critical to successful implementation.

The key management framework that spans the entire 4 layers is data management and governance. This framework must include the organizational structure, operational policies, and rules for security, privacy, and regulatory compliance.

Let's examine each layer a bit more closely.

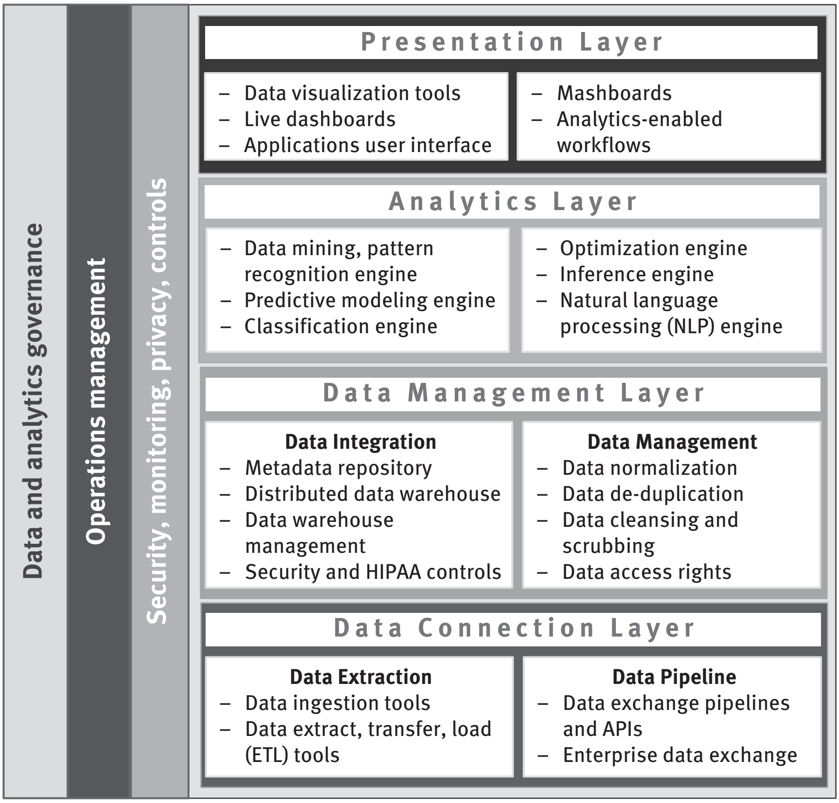

Data and analytics governance
Operations management
Security, monitoring, privacy, controls
Presentation Layer
– Data visualization tools
– Live dashboards
– Applications user interface
– Mashboards
– Analytics-enabled workflows
Analytics Layer
– Data mining, pattern recognition engine
– Predictive modeling engine
– Classification engine
– Optimization engine
– Inference engine
– Natural language processing (NLP) engine
Data Management Layer
Data Integration
– Metadata repository
– Distributed data warehouse
– Data warehouse management
– Security and HIPAA controls
Data Management
– Data normalization
– Data de-duplication
– Data cleansing and scrubbing
– Data access rights
Data Connection Layer
Data Extraction
– Data ingestion tools
– Data extract, transfer, load (ETL) tools
Data Pipeline
– Data exchange pipelines and APIs
– Enterprise data exchange

Figure 1.2: The 4-Layer Data Analytics Framework.

Data Connection Layer

In the data connection layer, data analysts set up data ingestion pipelines and data connectors to access data. They might apply methods to identify metadata in all source data repositories. This layer starts with making an inventory of where the data was created and is stored. The data analysts might implement *extract*, *transform*, followed by *load* (ETL) software tools to extract data from their source. Tools such as Talend and data exchange standards such as X.12 might be used to transfer data to the data management layer.

With the advent of Hadoop technologies and data lake storage models, data managers apply ELT (extract, load, transform) to gain faster access to data. Open source tools allow data engineers to extract data from a source, load it into Hadoop and then perform transformations. This approach makes data available faster.

Other data storage paradigms have been tried and some are still emerging. We'll review different data storage models in the upcoming chapters.

Data Management Layer

Once the data has been extracted, the data scientist must perform a number of functions that are grouped under the data management layer. The data may need to be normalized and stored in certain database architectures to improve data query and access by the analytics layer. We'll cover taxonomy of database tools including SQL, NoSQL, Hadoop, Shark, and other architectures in the upcoming sections.

In the data management layer, we pay attention to security and privacy standards. Standards vary by industry. Key standards in healthcare include HIPAA[2] and HITRUST[3]; in finance, standards include PCI-DSS[4]; in education FERPA[5]; for federal information systems the standards are FedRAMP[6] and FISMA; and the list goes on. The overarching privacy standards include the General Data Protection Regulation (GDPR) and the California Consumer Privacy Act (CCPA). Other governments and states are adopting similar guidelines and regulations. The data scientist will use the tools in this layer to apply security controls, such as those from HITRUST (Health Information Trust Alliance). HITRUST offers a common security framework (CSF) that aligns HIPAA security controls with other security standards.

The US government has established a government-specific set of security and operational standards for cloud operators. FedRAMP is the standard that government requires for all its cloud applications. Consequently, vendors who provide application services to the government must comply and test their SaaS applications to run in a FedRAMP cloud environment. In other words, such vendors must obtain FedRAMP certification for their cloud infrastructure. The US government has designated certain audit companies that conduct FedRAMP audits and provide certification to the vendors.

Data scientists may apply data cleansing programs in this layer. They might write tools to de-duplicate data (remove duplicate records) and resolve any data inconsistencies. Once the data has been ingested, it's ready to be analyzed by engines in the next layer.

Since big data requires fast retrieval, several organizations, in particular the Open Software Foundation, have developed alternate database architectures that allow parallel execution of queries, read, write, and data management.

[2] Health Insurance Portability and Accountability Act.

[3] Health Information Trust Alliance – HITRUST is a privately held company that has introduced a Common Security Framework (CSF) in collaboration with healthcare, technology, and information security organizations.

[4] The Payment Card Industry Data Security Standard.

[5] Family Educational Rights and Privacy Act of 1974.

[6] Federal Risk and Authorization Management Program.

Most analytics models require access to the entire data set because often they annotate the data with tags and additional information that are necessary for the models to perform.

The traditional data management approaches have adopted centralized data warehouse architectures – but there are pros and cons to this approach. The traditional data warehouse, namely a central repository or a collection of data from disparate applications, have not been as successful as expected. One reason is that data warehouses are expensive and time consuming to build. The other reason is that they are often limited to structured data types and difficult to use for data analytics, in particular when unstructured data is involved. Finally, traditional data warehouses insist on relational databases and tabular structures and normalized data. Such architectures are too slow to handle the large volume of data queries required by data analytics models. They require normalized relations between tables and data elements.

To overcome performance and implementation issues, data warehouse experts have come up with diverse data schemas.

When architecting data warehouses, there are two traditional schemas or approaches. A schema defines how data is organized in a data warehouse. The first approach is the *snowflake schema* and the other is the *star schema*. Star schema is more prevalent and has been more widely adopted in the past. Star schemas are denormalized data warehouses where normalization rules, typical of transactional relational databases, are relaxed during design and implementation. This approach offers simpler and faster queries and access to cube data. However, they share the same disadvantage with the non-SQL data bases – the rules of data integrity are not strongly enforced. The following is a brief overview.

Star Schema: More advanced data warehouses have adopted Kimball's Star schema or Snowflake schema to overcome normalization constraints. The Star schema splits the business process into *fact* tables and *dimension* tables. Fact tables describe measurable information about entities while dimension tables store attributes about entities. The Star schema contains one or more fact tables that reference any number of dimension tables. The logical model typically puts the fact table in the center and the dimension tables surrounding it, resembling a star (hence the name). A Snowflake schema is similar to Star schema but its tables are normalized.

Non-SQL database schema: In order to liberate data from relational constraints, several alternate architectures have been devised in the industry. These non-traditional architectures include methods that store data in a columnar fashion, or store data in distributed and parallel file systems, while others use simple but highly scalable tag-value data structures. The more modern big data storage architectures are known by names such as NoSQL, Hadoop, Cassandra, Lucene, SOLR, Spark, and others.

NoSQL database means "Not only SQL." It doesn't mean *no* SQL. It means that it includes SQL along with additional schema and storage techniques. A good example is Snowflake. It combines SQL referential data schema with additional techniques like parallel and columnar data storage that drastically improve database performance.

A NoSQL database provides storage mechanisms for data that is modeled in ways other than the tabular relations constraint of relational databases like SQL Server. The data structures are simple and designed to meet the specific types of data, so the data scientist has the choice of selecting the best fit architecture. The database is structured in a tree, columnar, graph, or key-value pair. However, a NoSQL database can support SQL-like queries. Hadoop, MongoDB, and Snowflake are examples of NoSQL data storage technologies.

Hadoop is an open source database framework for storing and processing large data sets on low-cost commodity hardware. Its key components are the Hadoop distributed file systems (HDFS) for storing data over multiple servers and MapReduce for processing the data. Written in Java and developed at Yahoo, Hadoop stores data with redundancy and speeds up searches over multiple servers. Commercial versions of Hadoop include HortonWorks and Cloudera. I'll cover Hadoop architecture and data management in more detail in the coming sections.

MongoDB is a NoSQL document database that can store data types and files of any format and content. Since it's schema-less, meaning that it does not relate to any schema compared to relational databases, any type of document can be stored as a data object. It's ideal for storage and rapid query of data files. Data is stored as a group of related documents with a shared common index as a collection. Behind the scenes, documents are stored in a key-value pair. MongoDB documents conform to the Binary JSON data format which makes it easy for developers to map the data used in applications to its associated objects in the database. MongoDB, a file-based database, is popular because it's easy to store data, easy to set up, and easy to scale. But there are ultimately limits both in terms of speed and data size when working with file-based databases, which has made other NoSQL data stores like Hadoop or AWS S3 more suitable choices for data lakes.

Snowflake is a new generation of SQL database architecture. The company offers a multi-cluster, shared data architecture built for cloud computing from the ground up. The technology offers a distributed data warehouse as a service. More specifically, the architecture is designed to perform extremely fast on both read and write functions. In addition, there is no need to install any hardware or software, and there is no maintenance of the data warehouse. Snowflake optimizes data storage of any data format, structured or unstructured datasets, in key-value pair or columnar fashion, optimized for fast query and analytical functions.

Cassandra is another open source distributed database management system designed to handle large data sets at higher performance. It provides redundancy over distributed server clusters with no single point of failure. Developed at Facebook to power the search function at higher speeds, Cassandra has a hybrid data structure that is a cross between a column-oriented structure and key-value pair. In the key-value pair structure, each row is uniquely identified by a row key. The equivalent of a RDBMS table is stored as rows in Cassandra where each row includes multiple columns. But unlike a table in an RDBMS, different rows may have different set of columns and a column can be added to a row at any time.

Lucene is an open source database and data retrieval system that is especially suited for unstructured data or textual information. It allows full text indexing and searching capability on large data sets. It's often used for indexing large volumes of text. Data from many file formats such as pdf, HTML, Microsoft Word, and OpenDocument can be indexed by Lucene as long as the textual content can be extracted.

SOLR is a high speed, full-text search platform available as an open source (Apache Lucene project) program. It's highly scalable offering faceted search and dynamic clustering. SOLR (pronounced "solar") is reportedly the most popular enterprise search engine. It uses Lucene search library as its core and often is used in conjunction with Lucene.

Hive is another open source Apache project designed as a data warehouse system on top of Hadoop to provide data query, analysis, and summarization. Developed initially at Facebook, it's now used by many large content organizations including Netflix and Amazon. It supports an SQL-like query language called HiveQL. A key feature of Hive is indexing to provide accelerated queries, working on compressed data stored in Hadoop database.

Spark is a modern data analytics platform; a modified version of Hadoop. It is built on the notion that distributed data collections can be cached in memory across multiple cluster nodes for faster processing. Spark fits into the Hadoop distributed file system offering 10 times (for in-disk queries) to 100 times (in-memory queries) faster processing for distributed queries. It offers tools for queries distributed over in-memory cluster computers that allow applications to run repeated in-memory queries rapidly. Spark is well suited to certain applications such as machine learning (which will be discussed in the next section).

Real-time vs. batch analytics: Much of the traditional business intelligence and analytics happen on batch data – a set of data is collected over time and the analytics is performed on the batch of data. In contrast, real-time analysis refers to techniques that update information and perform analysis at the same rate as they receive data. Real-time analysis enables timely decision making and control of systems. With real-time analysis, data and results are continuously refreshed and updated.

Analytics Layer

In this layer, a data scientist uses a number of engines to perform the analytical functions. Depending on the task at hand, a data scientist may use one or multiple engines to build an analytics application. A more complete layer would include engines for optimization, machine learning, natural language processing, predictive modeling, pattern recognition, classification, inferencing, and semantic analysis.

An *optimization engine* is used to find the best possible solution to a given problem. The optimization engine is used to identify the best combination of other variables to give an optimal result. Optimization is often used to find lowest cost, the highest utilization, or optimal level of care (in medical applications) among several possible decision options.

Machine learning is a branch of artificial intelligence that is concerned with construction and building of programs that learn from data. This is a basis for building adaptive models and algorithms that learn from data as well as adapt their performance to the data as it changes over time or when applied to one population versus another. For example, models based on machine learning can automatically classify patients into groups of having a disease or not-having a disease.

Natural language processing (NLP) is a field of computer science and artificial intelligence that builds computer understanding of spoken language and texts. NLP has many applications, but in the context of analyzing unstructured data, an NLP engine can extract relevant structured data from the unstructured text. When combined with other engines, the extracted text can be analyzed for a variety of applications.

A *predictive modeling* engine provides the algorithms used for making predictions. This engine would include several statistical and mathematical models that data scientists can use to make predictions. An example is making a prediction about patient readmissions after discharge. Typically, these engines ingest historical data and learn from the data to make predictions.

Pattern recognition engines, also known as data mining programs, provide tools for data scientists to discover associations and patterns in data, perform correlation analysis, and cluster the data in multiple dimensions. Some of these methods identify outliers and anomalies in data which help data scientists identify black-swan events in their data or identify suspicious or unlikely activity and behavior. Using pattern recognition algorithms, data scientists are able to identify inherent associate rules from the data associations. This is called *association rule learning*.

Another technique of this engine is building a regression model which works to define a mathematical relationship between data variables with minimum error. When the data includes discrete numbers, regression models work fine. But when data includes a mix of numbers and categorical data (textual labels), then logistic regression is used. There are linear and non-linear regression models and since many data associations in biological systems are inherently non-linear, the more complete engines provide non-linear logistic regression methods in addition to linear models.

Classification engines solve the problem of identifying to which set of categories a subject or data element belongs. There are two approaches, a supervised method and unsupervised method. The supervised methods use a historical set of data as the training set where prior category membership is known. The unsupervised methods use the data associations to define classification categories. The unsupervised classification is also referred to as clustering of data. Classification engines help data scientists to group patients or procedures, physicians, and other entities based on their data attributes.

Inference is the process of reasoning using statistics and artificial intelligence methods to draw a conclusion from data sets. Inference engines include tools and algorithms for data scientists to apply artificial intelligence reasoning methods to their data. Often the result of their inferencing analysis is to answer the question: "What should be done next?" where a decision is to be made from observations from data sets. Some inference engines use rule-based approaches that mimic an expert person's process of decision making collected into an expert system. Rules can be applied in a forward chain or backward chain process. In a forward chain process, inference engines start with the known facts and assert new facts until a conclusion is reached. In a backward chain process, inference engines start with a goal and work backward to find the facts needed to be asserted so the goal can be achieved.

Semantic analyzers are analytics engines that build structures and extract concepts from a large set of textual documents. These engines do not require prior semantic understanding of the documents. Semantic analysis is used to codify unstructured data or extract meaning from textual data. One application of semantic analytics is consumer sentiment analysis.

Statistical analysis: Statistical analysis tools include descriptive functions such as min, max, mode, median, plus ability to define distribution curve, scatter plot, z-Test, percentile calculations, and outlier identification. Additional statistical analysis methods include regression (trending) analysis, correlation, Chi-square, maxima and minima calculations, t-Test and F-test, and other methods. The open source R programming language has become a favorite for those working on statistics for research and business/industrial use. Others include MATLAB, SPSS, SAS, and Stata. For advanced tools, you can use an open source tool developed at Stanford called MADLIB, (www.madlib.net). Apache Mahout also includes a library of open source data analytics functions.

Forecasting and predictive analytics: Forecasting and predictive analytics are the new frontiers in medical data analytics. The simplest approach to forecasting is to apply regression analysis to calculate a regression line and parameters line, such as the slope and intercept value. Other forecasting methods use interpolation and extrapolation. Advanced forecasting tools offer other types of analyses such as multiple regression, non-linear regression, analysis of variance (ANOVA) and multivariable analysis of variance (MANOVA), mean square error (MSE) calculations, and residual calculations.

Predictive analytics are intended to provide insight into future events. Predictive analytics are model-driven and includes methods that produce predictions using supervised and unsupervised learning. Some of the methods include neural networks, PCA (principal component analysis), and Bayesian network algorithms. Predictions require the user to select the predictive variables and the dependent variables from the prediction screen.

A number of algorithms are available in this category that together provide a rich set of analytics functionalities. These algorithms include logistic regression, naive Bayes, decision trees and random forest, regression trees, linear and nonlinear regression, time series ARIMA,[7] ARTXp,[8] and Mahout analytics (collaborative filtering, clustering, categorization). Additional advanced statistical analysis tools are often used, such as multivariate logistic regression, Kalman filtering, association rules, LASSO[9] and Ridge regression, conditional random fields (CRF) methods, and Cox proportional hazard models.

Pattern analysis: Using machine learning, algorithm researchers can detect patterns in data, perform classification of patient populations, and cluster data by various attributes. The algorithms used in this analysis include various neural networks methods, principal component analysis (PCA), supervised and unsupervised learning methods such as k-means clustering, logistic regression, decision tree, and support vector machines.

Presentation Layer

This layer includes tools for building dashboards, applications, and user-facing applications that display the results of analytics engines. Data scientists often mash up several dashboards (called "Mashboards") on the screen to display the results using infographic graphs. These dashboards are active and display data dynamically as the underlying analytics models continuously update the results.

Infographic dashboards allow us to visualize data in a more relevant way with better illustrations. These dashboards may combine a variety of charts, graphs, and visuals together. Some examples of infographic components include heat maps, tree maps, bar graphs, a variety of pie charts and parallel charts, and many more visualization forms.

Advanced presentation layers include data visualization tools that allow data scientists to easily see the results of various analyses (classification, clustering, regression, anomaly detection, etc.) in an interactive and graphical user interface.

Several companies provide rapid data visualization programs including Tableau, QlikView, Panopticon, Pentaho, and Logi; and are revolutionizing how we view data.

[7] Auto regressive integrated moving average.

[8] Auto regressive tree model (ARTXp) is used for representing periodic time series data. This algorithm relates a variable number of past items to each current item that is being predicted.

[9] Least Absolute Shrinkage and Selection Operator (LASSO) is a regression analysis method for making predictions.

The Diverse Applications of Big Data

The big data storage technologies such as Hadoop have been designed to handle big data and distributed processing of massive amounts of data, but they offer other diverse use cases. In fact, there are six use cases for Hadoop as a distributed file system that you can utilize:

1. **Store all data.** Use Hadoop to store and combine all data of all types and formats. Store both structured and unstructured data. Store human generated data (application data, emails, transactional data, etc.) and machine generated data (system logs, network security logs, access logs, and so on). Since storage is inexpensive and Hadoop does not mandate any relational structure upfront, you are free to bring any data into a Hadoop data store.

   This approach is called the "data lake." With this freedom, of course, comes great responsibility: to maintain and track data in the data lake. You should be able to answer questions like: "What type of data do we have and where is it stored in the data lake?" "Do we have such and such data in the lake?" and "Where was the source of the data?"

   The ability and confidence to answer these questions lies in data governance.
2. **Stage data.** Use Hadoop as a staging platform to transform, re-transform, extract, and load data before loading it into the data warehouse. You can create a sandbox in Hadoop to specifically handle interim and intermediate data in a temporary storage area for processing. Tracking lineage of data and what type of processing was applied to it are important pieces of information to maintain as part of data governance. In this type of use case, Hadoop provides a powerful platform to handle ETL and ELT transformations. Hadoop or a cloud storage structure like S3 can provide a "sandbox" environment as a landing pad for incoming raw data and then transform the data into relational data formats that fit the data warehouse structure.
3. **Process unstructured and external data.** Use Hadoop to archive and update two types of data: 1) the unstructured data in your business, and 2) external data for analysis. Instead of using costly data warehouse resources to store these types of data, why not send them to Hadoop? This is now a common practice for many organizations. In particular, when you want to process social media data – which is typically semi-structured – Hadoop when combined with other open source tools like SOLR or Python libraries offers the ideal tools to parse and process textual information that deliver natural language processing, fast text search, social media analysis, and customer sentiment analysis.
4. **Process any data.** Use Hadoop to combine both internal and external data for processing. The Hadoop ecosystem offers a rich set of tools and processing options to join, normalize, and transform data of all types and sources. More generally, you can bring any data that's currently not in the data warehouse into

Hadoop and process it with internal data providing additional insights about your customers, products, and services.

5. **Store and process fast data.** Analyzing real-time data is a challenge that Hadoop can nicely accommodate. Consider a use case to analyze security access logs across all systems in your enterprise. One global company reports generating 2 petabytes (a petabyte is 1024 terabytes, or a million gigabytes) of access log data every day that needs to be analyzed in real time. There are millions of transactions per second that need to analyzed for a variety of use cases ranging from real-time fraud detection (in the case of credit card usage) or pattern detection (in the case of identifying hacker activity). These are known as fast data use cases. Use Hadoop components and tools to store fast data on the petabyte scale and analyze it on the fly as it gets stored in Hadoop in real time.
6. **Virtualize data.** The goal of data virtualization is to provide a seamless and unified view of all enterprise data no matter where the data resides, be it in data warehouses, in a big data lake, or even spread among many Excel spreadsheets. This vision is intended to address an endemic IT problem of not knowing where the information in the organization resides and how we can access that information to monetize it.

While Hadoop does not directly provide a data virtualization platform, it does contribute as a significant resource to the data virtualization architecture building block. There are open source tools and commercial solutions that over time provide robust solutions for data virtualization. Keeping pristine data management and data governance over Hadoop data will be a major factor for achieving successful data virtualization across the enterprise.

As we continue our discussion about big data management, there are two sources of standards and best practices that deserve to be referenced. One is the Data Management Association (DAMA) that has published a set of guidelines for data management in a library called the Data Management Body of Knowledge (DMBoK). The other source of information is the Data Maturity Model. In the next sections, I'll briefly review these two references to set the stage for later discussions.

The Future of Big Data

What holds for big data beyond 2020? Let's look at some predictions that our expert colleagues posit to come:

1. **Big data will be replaced by fast data.** Data volumes will continue to grow as the world volume of data doubles every 18 months. More data will be streaming and in real time. According to some experts, all companies are becoming data businesses. Interest in acquiring external data will increase (since plenty of such data will be collected and be available).

2. **More data types and higher abstractions of data will emerge.** An increase in data volume comes with an increase in the variety and types of data. Common data sources will include the internet of things (IoT), autonomous vehicles (all categories of autonomous robots in general), imaging devices, drones, smart advisers, virtual personal assistants, sensors, wearable devices, all connected people and intelligent appliances, and will form the future data origination points that generate diverse and unique data types. To handle all this data, new data formats and storage schema will be introduced to the market. Data will be categorized by the cognitive value that it will provide, not by the source that created it.
3. **New analytics methods will revolutionize data science.** New techniques, new tools, and algorithms will emerge that will drastically improve our ability to sift through this massive amount of data and analyze it. The new tools and algorithms will deliver incredible insights and AI solutions. But machine learning algorithms were developed decades ago – there are already cries from data science leaders for drastically new and revolutionary approaches to machine learning.
4. **More companies will hire a chief data officer.** Companies recognize the potential as well as the challenges associated with managing and maintaining good quality data at a strategic level. To fully leverage their data, companies will focus on how to achieve new forms of automation using artificial intelligence solutions. The role of chief data officer will continue to rise, but will lead, in many professions, to the role of chief AI officer.
5. **Big data will face bigger challenges around privacy.** New privacy regulations become more specific and personal. Future extensions to the privacy regulations set by the European Union (future revisions of GDPR for example) will aim to adopt privacy rules personalized to each individual's choices. Personalized control of data privacy will force more companies to accommodate personalized security and privacy procedures. Solutions to personalized privacy rules currently favor those organizations that can afford to create such solutions, though better open source and generic solutions will eventually level this particular playing field.
6. **Data and analytics will emerge as service business models.** Cloud providers are expanding their business models to include data as a service and analytics as a service. Examples include Microsoft's Cognitive Services and Amazon's SageMaker. The possibility that you may outsource your entire data services and data governance to third party companies, such as cloud providers, is becoming a reality. These types of cloud service arrangements are emerging, enabling businesses to focus on leveraging data and AI rather than fret the day-to -day data operations and data management routines.
7. **Data partnerships will drive business partnerships.** More businesses realize the synergy and strategic value of partnering around their data, sharing

algorithms, APIs, and building cross-industry data collectives. Mergers and acquisitions decisions will be driven not just by the business revenues, but by the value of data. Microsoft's acquisition of LinkedIn and GitHub are prime examples of this idea. The trend will accelerate in the future.

Traditional Data Storage Models and Methods

Many enterprises have invested heavily in centralized data storage platforms – with mixed success. These centralized data models include data warehouses and centralized data lakes. Collecting data in a central location has many advantages. It's easier to support and maintain a centralized platform, and it may be more effective to manage data in a consistent and uniform way. But the disadvantages of a centralized model take a big toll on achieving business objectives in a timely and cost-effective manner. Filling the data lakes often takes a long time to accomplish, partly due to poor data quality and difficulty getting data out of the old databases rather than a technical issue. Business leaders are impatiently looking to gain competitive advantage or drastically reduce operational costs via access to data on demand – as needed, when needed.

The centralized architectures are monolithic, domain agnostic, and costly to build. They often require duplicating and moving data from one storage platform to another. Data management has been through three generations of data storage architectures in its journey to maturity.

The first generation consisted of proprietary enterprise data warehouses that were extensions of enterprise databases. Companies created data marts in order to create domain-specific formats of the data for specific reporting called *views*. It was common to build different views (or data marts) of the same data for marketing, operations, finance, and other functional domains of the business. This approach created large amounts of technical debt in terms of hundreds if not thousands of unmanageable ETL jobs, tables, and reports. Only a small group of specialists could access or understand the data. It was easy to store data into these platforms, but very hard to get the data out. Relational and SQL-based databases were the hallmark of this generation.

The second generation, driven by the idea of a central and often open source data lake, delivered faster and easier access to data. The goal of this generation was to democratize data. Data democratization led to support for the self-service approach. The idea was that anyone should be able to access data with minimal or no coding. The data lakes required specialized and skilled developers called data engineers who developed automated data ingestion programs and supported data analytics functions. The highlight of this generation has been the Hadoop infrastructure which offered not only the ability to collect vast amounts of data, but also a parallel computing platform. Despite advances in this technology, many of the first generation issues continued: Data still had to be collected from a source into a centralized medium. The data lakes were not fast enough to keep up with the business demands of the organization. High data quality levels and use case context were often insufficient. This generation is best defined by using Not only SQL (NoSQL) architectures which removed many of the relational entity constraints and liberated data locked up in traditional data warehouses.

The third (current) generation builds upon lessons learned from the second generation. It embraces the cloud and containers, cloud managed data warehouses, data as a service, streaming and real-time data, data virtualization, and distributed data management. This generation of technologies creates high speed data storage and retrieval at a fraction of the cost that companies paid in previous generations. It works to resolve the technical debts accumulated by companies over the last two generations.

In the past, data was moved repeatedly to various data warehouses and then again moved to where computation was happening. This meant moving data multiple times to satisfy business intelligence and modeling needs. In the third generation, the idea is to leave data distributed but move the computations (the models) to the where data is located and avoid making duplications.

We need to shift the data storage paradigm from a central approach to a more modern distributed architecture. Some experts advocate creating a distributed data mesh architecture, while others promote data virtualization, and some are pursuing the data pipeline paradigms to overcome the limitations of a centralized architecture.

While the outcomes of data virtualization attempts have been less impressive than expected, the technology to make it practical is within reach. The current generation will produce data virtualization tools that will create a “one data” view of all enterprise data regardless of their physical structure, physical location, domain, or logical format.

Organizations are looking for rapid experimentation, model prototyping, and data transformations. Competitive pressures will place a higher demand for faster data cycle times to deliver data products on numerous use cases.

Data architectures in this generation will include creating a distributed data mesh or common data pipes. The data source is typically a transactional application or system. The consumer of data is commonly a data product such as a BI dashboard or a predictive model. In a distributed *data mesh*, all source systems and data consumption points are interconnected allowing each source and consumer of data to interact. In the common *data pipe* architecture, each source system offers a data stream where any privileged consumer of data can tap into the stream using simple APIs and consume the data.

The next generation of data storage will finally deliver on the promise of “data as a service.” In order to make data ubiquitous and accessible on demand, efforts to make data available will shift to making data smarter. Smart data objects will transform how we store and access data in the not so distant future.

Data objects will be smart enough to be self-aware and make themselves useful to other programs when needed. Smart data objects will consist of self-regulated, self-aware containers that advertise their function not by domain, but by the semantic purpose they serve. In a self-aware container, smart data objects are aware of their payload (data) quality and work to scrub or update themselves. The container includes data dictionaries, metadata, usage governance, and security policies.

Imagine smart data containers that are aware of their usage, frequency of use, and whether it's time to update their information to stay current. Each smart data container is programmed by specific logic and managed by a central catalog. Such catalogs offer a data marketplace across various functional silos of the organization and even beyond the boundaries of the company, providing access to external data in a consistent, containerized manner.

Why Data Governance Matters

In the digital economy where big data is a hot, strategic topic, data governance often takes a back seat to the lure of interesting data analytics models and ambitions to monetize data. But data governance is no longer a "nice to have" capability. It has become a key ingredient to every effective data strategy for a corporation.

In order to succeed in the gig economy where data is the oil and fuel, it's essential to master data quality, data management, privacy and ethical use of data, and data availability. Companies that get data governance right generate substantial benefits and financial results. These benefits include reduced operational costs and time savings, improved confidence and trust in data, faster adoption of new strategies, and accelerated return on investment.

Data governance provides important data quality controls. It enables companies to have consistent and reliable data sets. Reliable data, in turn, enables faster and better management decisions.

Data governance is more than just storing, cleansing, or consolidating data. It's about a framework of policies, practices, and business rules on how to collect, protect, and apply data to enable business. Data governance is an ongoing, repeatable process to properly manage data and mitigate potential risks. The right governance policies support and complement the overall data strategy. In fact, as we'll review later, data governance policies are critical to IT governance and, in turn, important to corporate governance. These days, the role of data governance has been more vital and visible in the organization where the data policy topics are often among boardroom agenda topics.

There are five important reasons to invest in data governance:

1. **Return on investment (ROI):** Data governance simply saves money and makes money. It helps the organization make better decisions, faster. Policies that help avoid data duplication, reduce likelihood of errors, and improve the

organization's access and understanding of data save the organization money. Better data quality means better understanding of markets, customers, and operational performance. Access to good, trustworthy data enables strategies that generate more revenues and value.

2. **Data consistency, trust, and reliability:** Data governance is crucial to creating data sets that are consistent, reliable, and repeatable. Without data governance, trust in data is diminished and management decisions will suffer. Operational performance metrics, KPIs (key performance indicators), and reports require dependable, consistent data. Data governance creates standards to ensure integrity of data, change control[10] and accountability for data to ensure business operations despite shifts in technology, environment, or personnel changes.
3. **Reduced risk:** Executives must be able to rely on their data when they report financial and operational results to stakeholders of their companies. Operating based on inconsistent data can lead to many risks, business liabilities, and internal conflicting decisions. Data breaches, exposure of data to the wrong parties, and non-compliance with regulatory agencies such as HIPAA or GDPR can bring high penalties and risks. Data governance defines the procedures to identify and mitigate risks associated with data ownership, data usage, and data-driven decisions.
4. **Problem solving:** Companies rely on data to help them solve problems. While they invest in technologies to improve their operations, they must align their technology with the reports and metrics they need to properly gauge their operational efficiency and effectiveness. Business intelligence dashboards, statistical models, machine learning, predictive modeling, and artificial intelligence (AI) applications are all important in solving problems and exploiting business opportunities. Governance ensures that data capture methods are consistent and that timely access to data is possible.
5. **Business alignment:** Data governance creates consistency with respect to business data, terms, and their meaning. This advantage of data management is often overlooked by executives. Data dictionary, metadata management, data lineage, and quality checks are critical to create high confidence in data. Thus, business can align on consistent definitions of terms, meanings of metrics, and reports. Data governance builds the clarity and confidence needed to align business around a set of common datasets, metrics, operational, and analytical measures.

[10] Change control is a data integrity management principle which ensures that changes to data are performed properly by people who have credentials to make changes. Data that is subject to Sarbanes Oxley must be controlled to ensure authenticity.

Chapter 2
Enterprise Data Governance Directive

To make a governance structure effective, it must contain all the right components and reference designs for a successful implementation. For the structure to be lean, it must be low cost to implement, low overhead to maintain, and free of excess baggage. In other words, you might ask: "What is the minimum viable framework that will get the biggest bang for the buck, so to speak, for my investment in big data governance?"

To reiterate, our goal is a lean and minimally viable governance framework that must include:

1. Organization
2. Metadata management
3. Security/privacy/compliance policies
4. Data quality management

I'll cover these components and more starting from a *tactical* data governance framework and build toward a *strategic* data governance structure. Governments around the globe including states within the US have adopted various data privacy initiatives and regulations. In this chapter, we'll cover three key policies to highlight their purpose, impact and their differences.

Data Privacy Governance and GDPR

Many organizations are impacted by the General Data Protection Regulation (GDPR), a set of regulations devised by the European Union to protect the privacy of individuals. Within the US, several states including California and Massachusetts have adopted similar regulations. GDPR regulations come with severe penalties and wide iron-fisted reach. These regulations apply hefty fines to non-compliant companies that can stretch to their parent companies, contractors, and affiliates.

GDPR was adopted in 2016 and enforcement began in May 2018. As a regulation, it's a binding legislative act that applies to the EU and EU citizens. Article 3 of the GDPR states that the regulation applies to all organizations who provide goods and services to the citizens of the EU, regardless of the organization's presence in the EU.

GDPR's intent is to protect individual privacy and private, personal information. It has a broad definition of what is often considered "personal data." Its definition includes any data that can be used to directly or indirectly identify an individual. *Any data* can include but is not limited to, names, addresses, email addresses, IP addresses, ID numbers, biometric identifiers (fingerprints, DNA,

https://doi.org/10.1515/9783110664065-002

iris patterns), occupation, location, medical/health data, physical or physiological characteristics, and even website cookies.

Article 5 of the GDPR specifies the basic regulations for processing personal data. These regulations are defined to ensure that personal data is lawfully collected, is accurate, is properly secured, and is limited in purpose, use, and duration of storage.

These regulations strive to give individuals more power over their personal information. One of the GDPR principles is to protect the individual's rights to know how their personal data is being used and the right to be forgotten (the right to have negative private information removed from internet searches and other directories). Under these regulations, the individual (also referred to as the "data subject") has a right to access their personal data and details regarding processing activities, as well as means to submit requests for data rectification, erasure of data, and to limit exporting the personal information (PII).[1]

Companies (also referred to as "controllers") must inform data subjects of their rights at the time of data collection. Data subjects have the right to data portability, meaning that the organization must provide a copy of the personal data to the data subject in a commonly used, machine readable format. The organization that processes the data (also referred to as the "processor") and controller must establish the process and technology to provide a method of enforcing GDPR regulations to protect the data subject's personal information.

Security is a critical element of the GDPR and of data governance. The regulations require companies who collect and process personal information to implement appropriate technical and organizational measures to ensure a level of security appropriate to the risk.

GDPR regulations demand much broader security and governance principles. They require the ability to ensure ongoing confidentiality, integrity, availability, and resilience of processing systems and services. These mandates specifically name techniques such as encryption of data and pseudonymization[2] of personal data as a measure to protect data privacy.

In addition to implementing these technical, organizational, and governance principles, the GDPR requires organizations to develop ongoing security, testing, and evaluation of effectiveness of governance and incidence response measures.

GDPR requires detailed governance documentation to be maintained for reporting. For example, an organization must be able to furnish records that show data was collected lawfully, consent was freely given, individuals' rights were properly managed, appropriate security measures were applied, required notifications were

[1] Personally Identifiable Information.

[2] *Pseudonymization* is the process of replacing a personally identifiable data field with an artificial data field or pseudonym. One data record can be consistently replaced throughout the data set with the same pseudonym. This technique is used to de-identify personally identifiable data.

sent, data protection impact assessments (DPIA) have been completed, and a data protection officer (DPO) has been designated (where necessary).

With the advent of the gig economy and widely adopted digital transformations, companies are likely to collect tremendous amount of personal data on their clients, potential customers, employees, and others. With such a significant business dependency on personal data comes a huge responsibility for data governance. We need a set of data governance principles that are founded on integrity, transparency, accountability, stewardship, standards, auditability, and change management.

Every organization must conduct at least four steps in order to prepare for GDPR compliance:

1. **Assess risk profile:** Conduct an internal audit and identify all personally identifiable data, data storage, and usage of such data. Determine your exposure, risk areas, and vulnerabilities. Generate a report of your risk assessment to your compliance executive.
2. **Categorize and label data:** Define and categorize your data by the risk level and impact. Label data sets into Very Low, Low, Medium, and High Risk categories. A client phone number, email address, and an online financial application are likely to rank as high risk. Start by implementing GDPR regulations on high and medium risk data sets.
3. **Become compliant:** Review GDPR policies against your data and identify specific action plans to implement policies. Identify the data access requirements for users, consent requirements for collecting data, and complaint usage for each category of data. As the sophistication and number of cyberattacks increases, it's urgent that you protect your most important data with preventive measures such as anonymizing or masking critical data.
4. **Prepare for audits and protection obligations:** How will you respond in the event of a breach? Prepare plans for mandatory breach reporting, increasing accountability, internal data protection audits, and remediation plans. Usage logging has become an important measure to auditability and tracking access to privacy information. Ensuring that such logs are available and conducting audits and drills will prepare you and your compliance organization to work closely in the case of a security or data breach event.

The Impact of California Consumer Privacy Act

The California Consumer Privacy Act (CCPA) is a law intended to enhance privacy and consumer protection for residents of California. The law was signed in 2018 and took effect in January 2020. It applies to all organizations that do business in California. The law effectively establishes the minimum requirements for nationwide privacy protection as most companies want to adopt and adhere to a single set of standards. The law applies to all businesses but not to non-profit organizations.

The CCPA gives consumers more control and information over how their data is being used. It also requires companies to be more transparent with handling consumer data. The definition of personal information is quite broad and CCPA gives a lengthy list of examples. These include social security number, driver's license and passport numbers, purchasing history, internet activity, geolocation data, employment-related information, and any information that can help draw a profile of a consumer (such as preferences or intelligence or abilities).

Upon request by a consumer, businesses must disclose information about the consumer's personal information including what is being collected, sold, or disclosed, the source of information, the business purpose for collecting or selling the information, and the categories of third parties with which the information is shared. The CCPA gives consumers a data portability right – namely, the right to access a copy of their personal information. In addition, companies must be able to honor a consumer's request to delete their personal information.

As a result, companies must develop data governance policies and capabilities to comply with CCPA regulations. Technical capabilities must give the user the ability to opt-out of the sale of their personal information. For example, this could require your website to provide a prominent link to a page titled: "Do Not Sell My Personal Information."

The CCPA enables consumers to submit a breach claim easily and empowers the California attorney general to impose penalties and fines for violations up to $2,500 per violation if not cured for 30 days.

There is some overlap and similarity between GDPR and CCPA, but there are also major differences. You might ask, can GDPR compliance also satisfy CCPA? Not really. The CCPA definition of personal information is more extensive than that of the GDPR. It also provides broader rights to request and delete data and offers different exceptions to the rule.

The CCPA is more empowering for the consumer to access and track the use of their personal information than the GDPR. In addition, CCPA's regulations for sharing personal information and commercial uses are far more stringent than the GDPR.

The New York Shield Act

Every state in the US has some type of consumer data security and privacy notification statute. The New York Shield Act expands on a number of consumer data protection areas. This law went into effect on March 21, 2020. In the past, personal identifier information (PII) included name, phone number, personal mark, or another identifier. The Shield Act broadens the list of PII data to include additional identifiers such as:

- Financial account numbers such as a credit card number
- A username and email address in combination with a password

- Biometric information used for access to systems
- Unsecured health information covered by HIPAA

The Shield Act requires companies and employers in possession of New York residents' information to adopt reasonable safeguards to protect and secure personally identifiable information. Even companies that do not have presence in New York may be required to comply since the law applies to any business that maintains the private information of New York residents.[3]

Forming Data Subject Access Rights with Big Data

Developing Data Subject Access Rights (DSARs) capability is crucial to becoming GDRP-CCPA compliant. Under GDPR regulation, individuals have the right to access their personal data. This is commonly referred to as subject access. The GDPR rules allows individuals to make subject access requests verbally or in writing. The company must respond within 30 days and it can't charge a fee.

A DSAR can be a written request made by an employee to their employer for information. All employees are allowed to request certain information from their employer about how their personal information is being used. Such a request is likely as part of an employee grievance, disciplinary, or employment related process.

Big data can help compliance in different ways, too. Organizations can use big data tools to classify all CCPA-impacted data across the enterprise, automate data flow monitoring, and fulfill data subject access requests (DSARs).

Big data can help support DSARs in three areas:

1. **Verify:** A data subject request can come in via email, web-link, or other communication means. The enterprise must confirm and verify the requestor's identity and store data about the entire fulfillment activity.
2. **Search:** The enterprise needs to search for all activities and transactions on the subject's personal information including the purpose and usage of collected data. The search must be extensive and include all source systems as well as data warehouses.
3. **Delete:** The company must be able to locate the personal information and delete related information across categories, attributes, and the company's purpose for collecting and processing the subject's information.

Companies who have invested in GDPR compliance are steps ahead since they can apply the lessons and capabilities from GDPR implementation toward becoming

[3] Philip Gordon and Jennifer Taiwo © Littler (2020, February 28). *The New York SHIELD act: What employers need to know*. SHRM. https://www.shrm.org/resourcesandtools/legal-and-compliance/state-and-local-updates/pages/new-york-shield-act.aspx

compliant with the CCPA. From big data governance readiness, we must consider developing CCPA readiness policies, checklists and strategies that include the following considerations:

- Inventory of consumer data
- Ability to automatically fulfill consumer data rights
- Update privacy policy and disclosure notifications
- Define breach response and notification work procedures
- Test and validate all aspects of CCPA policies and consumer rights

Data Management Body of Knowledge (DMBOK)

Data Management Association (now known as DAMA International) is an organization that devotes its efforts to advancing data management principles, guidelines, and best practices. One of the organization's goals is to be the go-to source for data management concepts, education, and collaboration on an international scale. It has published a set of guidelines that are publicly available under Data Management Body of Knowledge Version 2.0 (DMBOK2).[4]

DMBOK2: What Is It?

DMBOK2 is a collection of processes, principles, and knowledge areas about proper management of data. It includes many of the best practices and standards in data management. DMBOK may be used as a blueprint to develop strategies for data management for the entire enterprise or even at a division level. It was developed by DAMA, which has providing data management guidance since 1991. The data management DMBOK1 guide was released in 2009 and DMBOK2 in 2011 with an update in 2017. The guidelines were developed for IT professionals, consultants, data stewards, analysts, and data managers.

There are nine key reasons to consider DMBOK2 in your organization.[5] The key words are italicized:

Nine Reasons for DMBOK2

1. DMBOK2 brings *business* and *IT* together at the table with a common understanding of data management principles and vocabulary.

[4] https://www.dama.org/content/body-knowledge

[5] https://dama.org/sites/default/files/download/DAMA-DMBOK_Functional_Framework_v3_02_20080910.pdf

2. DMBOK2 connects *process* and *data* together so the organization can develop the appropriate processes for managing and governing its data.
3. DMBOK2 defines *responsibility* for data and accountability for the safeguard, protection, and integrity of data.
4. DMBOK2 establishes guidelines and removes ambiguity around change and *change* management.
5. DMBOK2 provides a *prioritization* framework to focus the business on what matters the most.
6. DMBOK2 considers regulatory and compliance requirements that enables intelligent *regulatory* response capabilities for the organization.
7. DMBOK2 helps identify *risks* and risk mitigation plans.
8. DMBOK2 considers and promotes the notion of *data as an asset* for the organization.
9. DMBOK2 links data management guidelines to process *maturity* and *capability*. The linkage allows the organization to identify a road map for improving data management processes.

The DMBOK2 covers nine domains of data management as shown in Figure 2.1. These nine domains are:

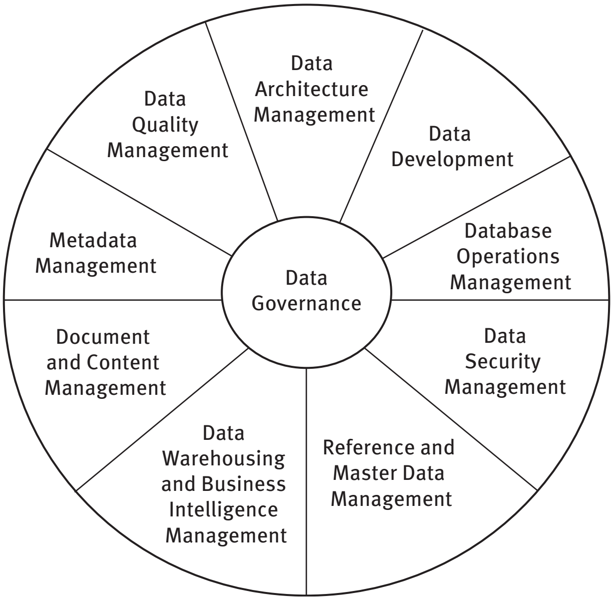

Data Architecture Management
Data Development
Database Operations Management
Data Security Management
Reference and Master Data Management
Data Warehousing and Business Intelligence Management
Document and Content Management
Metadata Management
Data Quality Management
Data Governance

Figure 2.1: DAMA International DMBOK2.

1. Data architecture management
2. Data development
3. Database operations management
4. Data security management
5. Reference and master data management
6. Data warehousing and business intelligence management
7. Document and content management
8. Metadata management
9. Data quality management

The contents of this book all relate to these nine domains from the perspective of big data, Hadoop, NoSQL, and big data analytics.

Data governance is at the center of the diagram. It is concerned with planning, oversight, and control over data management and use of data. The topics under data governance cover data strategy, organization and roles, policies and standards, issues management, and valuation.

Data architecture management is an integral part of the enterprise architecture. It includes activities such as enterprise data modeling, value chain analysis, and data architecture.

Data development is concerned with data modeling and design, data analysis, database design, building, testing, deployment, and maintenance of data stores.

Database operations management covers data acquisition, transformation, and movement; managing ETL, federation, or virtualization issues. It offers guidelines for data acquisition, data recovery, performance tuning, data retention, and purging.

Data security management ensures privacy, confidentiality, and appropriate access. It includes guidelines for security standards, classification, administration, authentication, logging, and audits.

Reference and master data management focuses on managing gold versions of data, replicas of data, and data lineage. It provides guidelines for standardized catalogs for external codes, internal codes, customer data, product data, and dimension management.

Data warehousing and business intelligence provides guidelines for managing analytical data processing and enabling access to decision support data for reporting and analysis. It contains policies regarding architecture, implementation, training and support, data store performance monitoring, and tuning.

Document and content management deals with data storage principles such as acquisition and storage of data, backup and recovery procedures, content management, retrieval, retention, and physical data asset storage management. These guidelines address storing, protecting, indexing, and enabling access to data found in unstructured sources (electronic files and physical records), and making this data available for integration and interoperability with structured (database) data.

Metadata management is about integrating, controlling, and delivering metadata. It also includes guidelines on metadata architecture, data integration, control, and delivery.

Data quality management is concerned with defining, monitoring, and improving data quality. It covers guidelines on quality specification, analysis, quality measurement, and quality improvement.

Data Maturity Model (DMM)

The CMMI Institute, a branch of Carnegie Mellon University, is funded by DoD and US government contracts, initially in response to the need for comparing process maturity among DoD contractors.

The *Capability Maturity Model Integration* (CMMI) was developed as a set of process improvement and appraisal programs by Carnegie Mellon University to assess an organization's *capability maturity*.

In 2014, the CMMI Institute released a model with a set of principles to enable organizations to improve data management practices across the full spectrum of their businesses. Known as the *Data Maturity Model* (DMM), it provides organizations with a standard set of best practices to build better data management structure and align data management with an organization's business goals.

In my opinion, the premise of process maturity is well grounded in the belief that better processes and higher process maturity are likely to produce higher quality and more predictable IT solutions.

The DMM model includes a *data management maturity portfolio* which includes DMM itself plus supporting services, training, partnerships, assessment methods, and professional certifications.

DMM was modeled after the CMMI maturity model to measure process maturity of an organization, improve efficiency and productivity, and reduce risks and costs associated with enterprise data.

DMM is an excellent framework to answer questions such as:

- How mature are your data management processes?
- What level of DMM maturity has your organization attained? How do you raise your maturity level?
- How do I capture the maximum value and benefits from data for the business?

The DMM model is comprised of five levels of maturity as shown in Table 2.1. The source of the table is CMMI Institute's standard for the Data Maturity Model. For more information, I recommend that you visit their website.[6]

[6] http://cmmiinstitute.com/data-management-maturity

Table 2.1: The DMM Model.

Level	Maturity Levels	Key Characteristics of Processes and Functions
Level 1	Initial	– Ad hoc, inconsistent, unstable, disorganized, not repeatable – Any success achieved through individual effort
Level 2	Managed	– Planned and managed – Sufficient resources assigned, training provided, responsibilities allocated – Limited performance evaluation and checking of adherence to standards
Level 3	Defined	– Standardized set of process descriptions and procedures used for creating individual processes – Activities are defined and documented in detail: roles, responsibilities, measures, process inputs, outputs, entry and exit criteria – Proactive process measurement and management – Process interrelationships defined
Level 4	Quantitatively Managed	– Quantitative objectives are defined for quality and process performance – Performance and quality practices are defined and measured throughout the life of the process – Process-specific measures are defined – Performance is controlled and predictable
Level 5	Optimized	– Emphasis on continuous improvement is based on understanding of organization business objectives and performance needs – Performance objectives are continually updated to align and reflect changing business objectives and organizational performance – Focus is on overall organizational performance – Defined feedback loop between measurement and process change

At the lowest level, Level 1, the organization’s activities around data management are ad hoc, unstable, and not repeatable.

At Level 2, the organization has data management processes, plans, some adherence to standards, and management objectives in place.

At Level 3, the organization has matured to establish and implement a standardized set of processes, roles and responsibilities, and a proactive process measurement system.

At Level 4, data management activities and performance are measured and quantitative objectives for quality and process management are defined and implemented.

At the highest level, Level 5, the organization optimizes its internal processes by continuous improvement, change management, and alignment with business objectives and changes in business needs.

The DMM model offers six domains (principle areas) for maturity assessment and process improvement. The six domains are: *data strategy, data governance, data quality, data operations, platform and architecture*, and *supporting processes*.

Table 2.2 represents the highlights of each domain and a brief description of each domain is provided.

Table 2.2: The Six Domains for Maturity Assessment and Process Improvement.

Data Strategy – Data management strategy – Data management function – Business case for data management – Funding, – Communications
Data Governance – Metadata management – Business glossary – Governance management
Data Quality – Data quality strategy – Data profiling – Data quality assessment – Data cleansing
Data Operations – Data requirements definition – Data lifecycle management – Provider management
Platform and Architecture – Overall data architectural approach – Data integration – Architecture standards – Data management platform – Historical data, archiving, and retention
Supporting Processes – Measurement and analysis, – Process management – Process quality assurance – Risk management – Configuration management

Data strategy is concerned with how the organization views the data management function, the business case for data management, funding, communications among data management roles, and adherence to data management strategy.

Data governance considers principles such as metadata management, implementation of the business glossary, and overall governance management.

Data quality is concerned with maturity of strategy and processes for data quality, data profiling, data quality assessment, data cleansing, and validation.

Data operations focuses on the operations aspect of DMM and evaluates the maturity of data lifecycle management, data requirements definitions, and provider management.

Platform and architecture are concerned with activities and procedures associated with data integration, overall data architecture creation, architecture standards, the data management platform, and data lifecycle practices such as data archiving and retention.

Supporting services is intended to empower the other five domains by implementing measurement and analysis, process management, process quality assurance, risk management, and configuration management practices.

Part 2: **Big Data Governance Fundamentals**

Chapter 3
Data Risk Management

Big data analytics is now a common strategy for many corporations to identify hidden patterns and insight from data. The three Vs of big data analytics (velocity, volume, and variety of data) are stretching the limits, if not breaking the traditional frameworks, around data governance and data management and data lifecycle management. The added volume of data comes with additional risk of data loss, data breach, and exposure.

In the current era of big data, organizations are collecting, analyzing, and differentiating themselves based on the analysis of massive amounts of information – data that comes in many formats, from various sources, and at a very fast pace.

Central to big data analytics is how data is stored in a distributed fashion that allows processing that data in parallel on many computing nodes that typically consist of one or more virtual machine configurations. The entire collection of parallel nodes is called a *cluster*.

Cloud computing, distributed computing, and data lakes have raised the risk profile and exposure to possible attacks and the high cost of data breaches. Some of the recent data breaches have cost companies and consumers alike. The intangible cost of data breaches may be almost impossible to overcome, in particular due to a tarnished brand image and loss of public trust.

Top 10 Data Breaches

Data governance is an important enterprise activity toward mitigating data issues and overall data risk. These days, companies will not survive without proper data governance. The list of security breaches is a constantly changing and dynamic list, a painful reminder of the vulnerability that may exist in our protection practices and our responsibility to safeguard data. These data breaches are not just demonstrative of poor IT security measures, but reflective of failures in data governance or poor data governance management. According to studies published by Health IT News (May 5, 2015), 42% of serious data breaches in 2014 occurred in the healthcare sector. The following are some of the most egregious breaches that led to substantial data loss.[1]

Sony Online Entertainment Services and Sony experienced repeated breaches. In 2011, more than 102 million records were compromised when hackers attacked

[1] For more up-to-date information about hacking incidents visit PrivacyRights.org at https://privacyrights.org/data-breaches?title=&page=177.

https://doi.org/10.1515/9783110664065-003

the PlayStation Network that links Sony's home gaming consoles and the servers that host the large multi-player online games. After the breach was discovered, the Sony PlayStation Network went dark worldwide for more than three weeks. Some 23 million accounts including credit cards, phone numbers, and addresses were compromised in Europe alone. Soon after, about 65 class action lawsuits were brought against the company at a cost of $171 million.

In 2011, Epsilon reported a breach that compromised 60 million to 250 million records. The Texas-based marketing firm that handles email communications for more than 2,500 clients worldwide reported that its databases related to 50 of its clients had been stolen. The email addresses of over 60 million clients were stolen, including those of customers from a dozen key banks, retailers, and hotels such as Best Buy, JPMorgan Chase, and Verizon.

Similar to Home Depot, Target reported a breach of its databases which caused tremendous customer complaints and damage to the company's image and trust from consumers.[2] The security breach at Target affected 40 million card accounts, potentially exposing credit and debit cards used at the store in 2013.

Evernote, the popular note-taking and archiving online company reported in 2013 that the email addresses, user names, and passwords of its clients had been exposed by a security breach. While no financial data was stolen, the notes of the company's clients and all the content had been compromised. The clients later became targets of a massive number of spam emails, phishing scams, and traps by hackers who even pretended to be from Evernote itself.

In 2013, Living Social, the daily-deals social site partly owned by Amazon reported that the names, email addresses, birth dates, and encrypted passwords of some 50 million customers had been stolen by hackers. In the same year, Yahoo! was breached, exposing more than three billion accounts which contained users' names, dates of birth, phone numbers, and passwords.

Blue Cross Blue Shield of New Jersey lost data affecting roughly 800,000 individuals – simply by losing a laptop. The data was not encrypted. This occurred in January 2014.

This is in addition to another breach of Sony Pictures Entertainment in 2014 when their computer systems were hijacked and the company's 6,800 employees plus an estimated 40,000 other individuals were impacted when their data, emails, and personal information were exposed.

The year 2014 was a relentless one of data breaches for many industries. JP Morgan Chase, the largest bank in terms of assets in the US revealed that personal information for some 76 million households and 7 million small businesses were exposed. Several other businesses including Neiman Marcus, P.F. Chang's, 11 casinos,

[2] https://www.usatoday.com/story/money/2017/05/23/target-pay-185m-2013-data-breach-affected-consumers/102063932/

and Michael's, a leading provider of art and hobby supplies, reported significant data breaches.[3]

In February of 2015, Anthem, formerly known as WellPoint, the second largest health insurer in the US reported 80 million records stolen by hackers. Stolen data included names, addresses, dates of birth, employment histories, and Social Security numbers.

Another incident affecting 56 million payment card customers of Home Depot occurred in 2014. The major hardware and building supplies retailer admitted to what it had suspected for weeks. Sometime in April or May of 2014 the company's point of sale systems in its US and Canada stores were infected with malware that pretended to be antivirus software, but instead stole customer credit and debit card information.

Verizon Communications, which acquired Yahoo! in 2017, admitted to the breach that all 3 billion accounts' information had been stolen.[4]

One of the worst security breaches occurred in 2007 at TJX Companies, the parent company of the retailer, TJMaxx and HomeGoods. At least 45 million credit and debit card numbers were stolen over an 18-month period, though some put the estimate at closer to 90 million. Some 450,000 customers had their personally identifiable information stolen including driver's license numbers. The breach eventually cost the company $256 million.

In 2018, some 500 million Marriott guests' information, including names, email addresses, phone numbers, passport numbers, birth dates, genders, and arrival and departure times were exposed after a breach.

For some time, Facebook stored its users' passwords in plain text – which exposed users' entire account information to Facebook's 20,000 employees – until the company fixed the issue.

In 2019, a hacker gained access to 100 million CapitalOne credit card and financial applications. The hacker was planning to sell this information online before she was arrested. The data was stored on the Amazon AWS cloud and the hacker exploited a misconfigured web application firewall to gain access to the information. The company expects the cost of remediation, including customer notifications, credit monitoring, technical and legal costs to reach $150 million.

With the advent of open source technologies like Hadoop and the proliferation of big data, the accumulation of vast amounts of data on customers, products, and social media, the risks associated with data breaches will only get more widespread and costly. Hence, investing in big data governance infrastructure, processes, and policies will prevent costly exposure to data breaches in the future.

[3] Hardekopf, B. (2015, January 13). *The Big Data Breaches of 2014*. Forbes. https://www.forbes.com/sites/moneybuilder/2015/01/13/the-big-data-breaches-of-2014/#54fd6245efe6.

[4] *All 3 billion Yahoo accounts were affected by 2013 attack*. (2017, October 3). The New York Times – Breaking News, World News & Multimedia. https://www.nytimes.com/2017/10/03/technology/yahoo-hack-3-billion-users.html.

The point to keep in mind is that data governance must pay close attention to security and privacy. It must enable the organization to implement security controls and measures that provide proper access for business needs with a clear and hard defense against hackers. For example, policies might state that data at rest and in transit should be encrypted and certain data (such as Social Security numbers or credit card numbers) are to be tokenized. Simply put, without the key to the tokens, even if the data was hacked, the hackers cannot obtain the data without the token keys.

Many of these data breaches could have been averted if the basic principles of security and data governance were applied by the organizations – studies of prior breaches have shown that if the affected organizations had applied these principles, they could have prevented 80% of their data loss events. I raise this point early in this chapter to highlight the much-needed focus and investment in data governance across organizations globally.

What Is Data Risk Management?

Given the growing volume and velocity of data, data management is a key responsibility for a company that collects, stores, and processes such data. Data risks, if not properly addressed, can be a real vulnerability and threat to a company's survival. *Data risk management* encompasses the processes that an enterprise uses for acquiring, storing, processing, and using its data, from inception to retirement, to mitigate data risk.

A proper data risk management structure minimizes the ability of data to be exposed, breached, wrongfully or unethically used, and accessed. Data risks can cause significant business losses due to poor data governance, data mismanagement, and inadequate data security.

When data risks and vulnerabilities are not assessed, managed and controlled properly, the financial and reputational damages can be devastating to a company. In case of a data incident, a company can be liable for expenses that include:

- Immediate and long-term fixes to its IT infrastructure, procedures, and staffing
- Payment of regulatory fines, costs of legal counsel, settlement of disputes, customer notifications, credit monitoring and public relations
- Decreased employee productivity and increased labor for time taken to contain an incident
- Damaged reputation and tarnished brand value
- Business and data center downtimes, and application downtimes that negatively impair business transactions

According to IBM, the average cost of a data breach in the US is $3.92 million, with the healthcare sector being the most costly industry to handle a breach (costing

$6.45 million on average).[5] The same report puts the average size of a data breach at over 25,000 records.

What Is Governance?

There are many definitions for data governance. One popular definition is: *Data governance* is the execution and enforcement of authority over the management of data and data-related assets. Data governance is and should be in alignment with the corporate governance policies as well as within the IT governance framework.

Along with data governance, you may have seen references to data stewardship – sometimes they're used synonymously. The two notions are related but different. *Data stewardship* is the formalization of accountability for the management of data and data-related assets.

A data governance framework relies on several other data management roles and will be defined later in the book. But the role of data stewards is critical to operational integrity and success of a big data governance framework.

There are two perspectives on who a data steward is. One definition views everyone in the organization who deals with data as being accountable for how data is treated. Another definition describes an actual job role of data steward as someone whose sole responsibility is to be accountable for their organization's treatment of data.

In the first definition, a data steward can be anyone in the organization whether business-minded or technical. The data steward is tasked with the responsibility and accountability for what they do with the data as they define, produce, and/or use data as part of their work.

In the second definition (and this is the definition that I use in this book and is typically implied in data governance policies) the data steward is one who defines, produces, or uses data as part of their job and has a defined level of accountability for assuring quality in the definition, production, or usage of that data.

The best practices in big data governance promote a partnership between IT and other functional areas, divisions, and lines of business when it comes to data governance. This is a departure from the heavy, iron-fisted approach that some IT organizations have employed in the past. Instead, the best practices bring clarity of purpose, accountability, and policies, and the philosophy that data stewardship is not solely an IT responsibility but a shared responsibility across the enterprise. Every employee is a data steward, responsible for proper use, storage, and protection of data.

Data governance best practices apply formal accountability and set the bar for the correct behaviors toward data governance in an organization. A data governance

[5] https://www.ibm.com/security/data-breach

framework must support existing and new processes to ensure the proper management of data, and the proper production and usage of data through its lifecycle.

Big data governance improves regulatory compliance, security, privacy, and protection of data across the enterprise. Best practices prescribe non-threatening data governance in a collaborative, federated approach to managing valuable data assets. The goal of best practices is to create a big data governance structure and framework that is 1) transparent, 2) supportive, and 3) collaborative.

Why Big Data Governance?

Another perspective of big data governance is to think of it as the convergence of policies and processes for managing *ALL* data in the enterprise – data quality, security, privacy, and accountability for data integrity. It's a set of processes that ensures important data assets are formally and consistently managed through the enterprise. It ensures data can be trusted and people are held accountable for low data quality.

Big data governance brings people and technology together to prevent costly data issues and ensure enterprise data is more efficiently managed and utilized. Finally, it enables the use of data as a strategic asset maximizing the return on investment by leveraging the power and value of big data.

Some of the reasons for data governance include:

- *Better compliance and fewer regulatory problems*. Without metadata compliance and data quality procedures, regulatory reports are not as reliable.
- *Increased assurance and dependability of knowledge assets*. With governance, you can trust your data and make decisions based on solid data.
- *Improved information security and privacy*. Data governance will reduce the risks and exposures associated with data loss or breaches.
- *Across the enterprise accountability*. Data governance enables more accountability and consistency of data handling across the enterprise.
- *Consistent data quality*. Improved quality reduces rework, waste, and delays. It improves quality of decision making.
- *Maximizing asset potential*. With data governance, the assets are known, discoverable, and usable across the enterprise, thus they raise the value and return on data (ROD).

Data Steward Responsibilities

Data stewards are formally accountable for defining rules around new data versus existing data. Their responsibilities include defining rules around logical and physical data modeling, metadata definition, the business glossary, and recognizing the system of record.

The responsibilities of data stewards extend to defining production rules for data produced internally, or data acquired from outside, data movement across the enterprise, and ETL (Extract, Transfer, Load) functions.

Other key responsibilities include defining and administering data usage rules and enforcing them – in particular, defining business rules, classifying data, protecting data, retention rules around data, compliance and regulatory requirements.

Corporate Governance

Corporate governance refers to the mechanisms, processes, policies, and rules by which corporations are controlled and directed. These mechanisms include monitoring the actions, policies, and decisions of corporations, their officers, and employees. Corporate governance and IT governance are the overarching structures with which big data governance must be aligned.

Big Data Governance Certifications

Many organizations opt to apply for and receive certification of their data governance structure. Certification can be granted by an outside organization. Certifying your data governance structure has many tangible and intangible benefits.

The value and benefits of data governance increase with the number of employees and amount of data that is maintained in the enterprise. Certification implies people in the organization understand the *value* of data governance; they understand data and are competent to protect it. Competency is created by leadership through setting direction, governance structure, and training.

The enterprise value of certification can be profound. Certification implies that people are educated in proper data management policies and follow the rules – the risk management rules, the business rules, and data change management rules. Certification also means that people are consistent in how they define data, product data, and use data across the enterprise.

With a big data governance certification, your organization can promote that they are: “Good with their data”! They can make declarations such as, “We protect our data better than our competition,” or “We improve the quality of our data,” “Our data is known, trustworthy and managed,” or “We centralize how we manage our data.”

Imagine an opposite and not so desirable situation: Let’s assume you’re presenting the results of a clinical random trial study to an FDA panel to receive approval for releasing a new drug to market. Imagine if your statement said: “We have the results but we don’t know much about the quality of our data, where it came from, and who has had access to it. Furthermore, we can’t even reproduce these results because we never safeguarded the source data nor took snapshots of our analysis." This would be a low point for a company that can't govern and manage its big data.

The Case for Big Data Governance

A number of incidents indicate the seriousness of security breaches that have caused substantial damage and cost to organizations.

A study released by Symantec found that on average each security breach costs an organization $5.4 million. Another study claims that the cost of cybercrime to the US economy is $140 billion per year.

Sony has experienced more than its share of breaches with hackers. In 2011, one of the largest breaches in recent history involved Sony's PlayStation network which costs as much as $24 billion according to some experts.

Some organizations like Sony have experienced multiple breaches of security at a very high cost to their reputation, financial impact, and loss of consumer trust. Data governance has become an important strategy both to enterprise governance and IT governance. It is the vehicle to ensure the data is properly managed, used, and protected.

Without proper security controls, big data can become *big breaches*. Put differently, the more data you have collected, the bigger the magnitude and cost of risk, and the more important it is to protect the data. This implies that we must have effective security controls on data in storage, inside our networks, and as it leaves our networks.

Depending on the sensitivity of data, we must control who has access to it, to see what analysis is performed, and who has access to the resulting analysis.

Hadoop has become a popular platform for big data processing. But, Hadoop was originally designed without security in mind. While Hadoop's security model has evolved over the years, security professionals have sounded the alarm around Hadoop's security vulnerabilities and potential risks. Increasingly, new third party security solutions come to market but it's not easy to craft an architecture integrating these solutions to create a comprehensive and complete security infrastructure.

Operational, Tactical, or Strategic Data Governance?

With big data and the proliferation of Hadoop, we're at the dawn of a new era: the age of data democracy or democratization of data. Data democratization delivers a lot of power and value but with it comes responsibility, and a bigger need for governance. There are three primary levels or perspectives to data governance: *operational*, *tactical*, and *strategic*.

Operational data governance focuses on daily operations and the safeguard of data, security, privacy, and implementing policies typically influenced by the IT organization with little involvement from business owners of data. Data is not treated as a strategic asset. There is no staff dedicated to the roles of data steward or data guardian in the organization.

Tactical data governance is concerned with the current and immediate issues associated with management of data and implementation of governance, and accountability of data governance. This approach spends its energy on governance measures, policies, roles, and responsibilities. It regulates the metadata and business glossary, standardization of terms, names, and data dictionaries.

Tactical data governance is often regarded as Data Governance 1.0. It concentrates on daily activities and maintenance of policies through data stewards and other members who support data governance decision making. There is typically little involvement or designation of an accountable executive or data risk officer(s) who provide ground cover and executive leadership to data governance, an organizational structure typically found in strategic data governance.

Sometimes referred to as Data Governance 2.0, *strategic data governance* is concerned with the propagation of data assets, the forthcoming objectives and challenges that will be encountered as data grows throughout the organization and the maintenance of this data; it is the long-term and future perspective of data governance. Strategic data management treats data as a corporate asset and works to optimize the value from this asset to the organization.

Strategic data governance brings a holistic view of standardization into the conversation, including evaluating the current organization's standard terms and ontology, identifying the gaps, and using advanced tools and techniques such as semantic ontologies and semantic concepts across the enterprise. In this approach, you'll find roles like the chief data officer (CDO) and accountable executive (AE), data risk officers (DRO), and a central *data council* (also referred to as data governance council) in the organization.

TOGAF View of Data Governance

The Open Group Architecture Framework (TOGAF) has published a set of architecture best practices and principles in its standard, TOGAF Version 9, released in 2011. The goal of TOGAF is to provide guidance and an open forum to enable interoperability and a boundary-less flow of information among vendor applications and systems. Since then, The Open Group Architecture Framework has released two more revisions with version 9.2 as the most recent edition.

TOGAF promotes architectural best practices through 10 distinct sets of guidelines and activities. These guidelines include:

– The preliminary phase
– Architecture vision
– Business architecture

- Information systems architectures
- Technology architecture
- Opportunities and solutions
- Migration planning
- Implementation governance
- Architecture change management
- Requirements Management

Specifically, TOGAF promotes adoption of key principles by the organization before drafting an architecture. Each principle defines a high-level aspiration, vision, and policy. The principle-driven approach to architecture is a powerful approach to bring common understanding and guidance through the architecture and design phases.

TOGAF version 9.1 provided a set of sample principles which include a wide range of information architecture topics, but for the sake of brevity, I outline a few guiding principles that pertain to data governance in the following section. These principles are excerpts from TOGAF training courseware.[6] For more details and a complete overview of TOGAF Version 9.1, I recommend visiting the site and reading the entire TOGAF documents.

Sample TOGAF Inspired Data Principles

Principle 1: Data is an Asset – Data is an asset that has value to the enterprise and is managed accordingly.

Rationale: Data is a valuable corporate resource; it has real, measurable value. In simple terms, the purpose of data is to aid decision-making. Accurate, timely data is critical to accurate, timely decisions. Most corporate assets are carefully managed, and data is no exception. Data is the foundation of our decision making, so we must also carefully manage data to ensure that we know where it is, can rely upon its accuracy, and can obtain it when and where we need it.

Implications: This is one of three closely-related principles regarding data: data is an asset; data is shared; and data is easily accessible. The implication is that there is an education task to ensure that all organizations within the enterprise understand the relationship between value of data, sharing of data, and accessibility to data.

Data stewards must have the authority and means to manage the data for which they are accountable.

We must make the cultural transition from "data ownership" thinking to "data stewardship" thinking.

The role of data steward is critical because obsolete, incorrect, or inconsistent data could be passed to enterprise personnel and adversely affect decisions across the enterprise.

Part of the role of data steward, who manages the data, is to ensure data quality. Procedures must be developed and used to prevent and correct errors in the information and to improve those

[6] *The TOGAF® standard, version 9.2*. (n.d.). The Open Group Publications Catalog. https://pubs.opengroup.org/architecture/togaf9-doc/arch/

processes that produce flawed information. Data quality will need to be measured and steps taken to improve data quality – it is probable that policy and procedures will need to be developed for this as well.

A forum with comprehensive enterprise-wide representation should decide on process changes suggested by the steward.

Since data is an asset of value to the entire enterprise, data stewards accountable for properly managing the data must be assigned at the enterprise level.

Principle 2: Data is Shared – Users have access to the data necessary to perform their duties; therefore, data is shared across enterprise functions and organizations.

Rationale: Timely access to accurate data is essential to improving the quality and efficiency of enterprise decision-making. It is less costly to maintain timely, accurate data in a single application, and then share it, than it is to maintain duplicative data in multiple applications. The enterprise holds a wealth of data, but it is stored in hundreds of incompatible stovepipe databases. The speed of data collection, creation, transfer, and assimilation is driven by the ability of the organization to efficiently share these islands of data across the organization.

Shared data will result in improved decisions since we will rely on fewer (ultimately one virtual) sources of more accurate and timely managed data for example Set of Architecture Principles Architecture Principles all of our decision-making. Electronically shared data will result in increased efficiency when existing data entities can be used, without re-keying, to create new entities.

Implications: There is an education requirement to ensure that all organizations within the enterprise understand the relationship between value of data, sharing of data, and accessibility to data.

To enable data sharing we must develop and abide by a common set of policies, procedures, and standards governing data management and access for both the short and the long term.

For the short term, to preserve our significant investment in legacy systems, we must invest in software capable of migrating legacy system data into a shared data environment.

We will also need to develop standard data models, data elements, and other metadata that defines this shared environment and develop a repository system for storing this metadata to make it accessible.

For the long term, as legacy systems are replaced, we must adopt and enforce common data access policies and guidelines for new application developers to ensure that data in new applications remains available to the shared environment and that data in the shared environment can continue to be used by the new applications.

For both the short term and the long term we must adopt common methods and tools for creating, maintaining, and accessing the data shared across the enterprise.

Data sharing will require a significant cultural change.

This principle of data sharing will continually ‘‘bump up against’’ the principle of data security. Under no circumstances will the data sharing principle cause confidential data to be compromised.

Data made available for sharing will have to be relied upon by all users to execute their respective tasks. This will ensure that only the most accurate and timely data is relied upon for decision-making. Shared data will become the enterprise-wide ‘‘virtual single source’’ of data.

Principle 3: Data is Accessible – Data is accessible for users to perform their functions.

Rationale: Wide access to data leads to efficiency and effectiveness in decision-making, and affords timely response to information requests and service delivery. Using information must be considered from an enterprise perspective to allow access by a wide variety of users. Staff time is saved and consistency of data is improved.

Implications: Accessibility involves the ease with which users obtain information.

The way information is accessed and displayed must be sufficiently adaptable to meet a wide range of enterprise users and their corresponding methods of access.

Access to data does not constitute understanding of the data. Personnel should take caution not to misinterpret information.

Access to data does not necessarily grant the user access rights to modify or disclose the data. This will require an education process and a change in the organizational culture, which currently supports a belief in ‘‘ownership’’ of data by functional units.

Principle 4: Data Trustee & Data Steward – Each data element has a trustee accountable for data quality and data steward accountable for proper management and usage of the data

Rationale: One of the benefits of an architected environment is the ability to share data (e.g., text, video, sound, etc.) across the enterprise. As the degree of data sharing grows and business units rely upon common information, it becomes essential that only the data trustee makes decisions about the content of data. Since data can lose its integrity when it is entered multiple times, the data trustee will have sole responsibility for data entry which eliminates redundant human effort and data storage resources.

A data trustee is different than a data steward – a trustee is responsible for accuracy and currency of the data, while responsibilities of a steward are broader and include data standardization, data usage policies, data access management and definition tasks.

Implications: Data Trustee and Data Steward roles resolve ambiguity around data ‘‘ownership’’ issues and allow the data to be available to meet all users’ needs. This implies that a cultural change from data ‘‘ownership’’ to data ‘‘trusteeship’’ and “stewardship” may be required.

The data trustee will be responsible for meeting quality requirements levied upon the data for which the trustee is accountable.

It is essential that the trustee has the ability to provide user confidence in the data based upon attributes such as ‘‘data source’’.

It is essential to identify the true source of the data in order that the data authority can be assigned this trustee responsibility. This does not mean that classified sources will be revealed nor does it mean the source will be the trustee.

Information should be captured electronically once and immediately validated as close to the source as possible. Quality control measures must be implemented to ensure the integrity of the data.

As a result of sharing data across the enterprise, the trustee is accountable and responsible for the accuracy and currency of their designated data element(s) and, subsequently, must then recognize the importance of this trusteeship responsibility.

Data Steward is responsible to ensure proper use of data according to access rules, regulatory and compliance rules, data sovereignty, data jurisdictional rules and contractual agreements with customers and/or third-party data vendors.

Sometimes the roles of data steward and data trustee are combined into one individual. You need to define the role as it fits your organization.

Principle 5: Common Vocabulary and Data Definitions – Data is defined consistently throughout the enterprise, and the definitions are understandable and available to all users.

Rationale: The data that will be used in the development of applications must have a common definition throughout the Headquarters to enable sharing of data. A common vocabulary will facilitate communications and enable dialog to be effective. In addition, it is required to interface systems and exchange data.

Implications: Data definitions, common vocabulary and business glossaries are key to the success of efforts to improve the information environment. This is separate from but related to the issue of data element definition, which is addressed by a broad community – this is more like a common vocabulary and definition.

The enterprise must establish the initial common vocabulary for the business. The definitions will be used uniformly throughout the enterprise.

Whenever a new data definition is required, the definition effort will be coordinated and reconciled with the corporate "glossary" of data descriptions. The enterprise data administrator will provide this coordination.

Ambiguities resulting from multiple parochial definitions of data must give way to accepted enterprise-wide definitions and understanding.

Multiple data standardization initiatives need to be coordinated.

Functional data administration responsibilities must be assigned.

Principle 6: Data Security – Data is protected from unauthorized use and disclosure. In addition to the traditional aspects of security classification (for example, Classified Secret, Confidential, Proprietary, Public) this includes, but is not limited to, protection of pre-decisional, sensitive, source selection-sensitive, and other classifications that fit your organization's line of business.

Rationale: Open sharing of information and the release of information via relevant legislation must be balanced against the need to restrict the availability of classified, proprietary, and sensitive information.

Existing laws and regulations require the safeguarding of national security and the privacy of data, while permitting free and open access. Pre-decisional (work-in-progress, not yet authorized for release) information must be protected to avoid unwarranted speculation, misinterpretation, and inappropriate use.

Implications: Aggregation of data both classified and not, will create a large target requiring review and de-classification procedures to maintain appropriate control. Data owners and/or functional users must determine whether the aggregation results in an increased classification level. We will need appropriate policy and procedures to handle this review and declassification.

Access to information based on a need-to-know policy will force regular reviews of the body of information.

The current practice of having separate systems to contain different classifications needs to be rethought. It is more expensive to manage unclassified data on a classified system. Up until now the only way to combine the two was to place the unclassified data on the classified system, where it remained. However, as we shall see in the coming sections, we can use Hadoop to create multiple data sandboxes that offer fit for purpose configurations. We can create different sandboxes to keep classified and unclassified data separate but on the same Hadoop distributed file system. In general, it's recommended that all data in the Hadoop lake be classified to maintain a manageable data lake.

In order to adequately provide access to open information while maintaining secure information, security needs must be identified and developed at the data level, not just as the application level.

Data security safeguards can be put in place to restrict access to "view only", or "never see". Sensitivity labeling for access to pre-decisional, decisional, classified, sensitive, or proprietary information must be determined.

Security must be designed into data elements from the beginning; it cannot be added later. Systems, data, and technologies must be protected from unauthorized access and manipulation. Headquarters information must be safeguarded against inadvertent or unauthorized alteration, sabotage, disaster, or disclosure.

Need new policies on managing duration of protection for pre-decisional information and other works-in-progress, in consideration of content freshness.

Managing Risk: Data Lake versus Data Warehouse

We often hear that the data lake and data warehouse are synonymous and a data lake is just another incarnation of a data warehouse. The truth is that there are major differences between them. A data lake is not a data warehouse. They are optimized for different purposes and to provide the best tool for that purpose.

The data lake can be thought of as a large storage space that can house any kind of data, data structures, and formats at a very low cost. A data lake is a storage repository that holds a vast amount of raw and processed (refined) data in its native format, including structured, unstructured, and semi-structured data. The data structure and requirements are not defined until the data is needed.

Tamara Dull, in a series of blogs, provided a good overview of the differences between a data warehouse and a data lake.[7] The summary of those differences is shown in Table 3.1.

Table 3.1: The Difference between a Data Warehouse and a Data Lake.

Data Warehouse	**vs.**	**Data Lake**
Structured, processed	DATA	Structured, semi-structured, unstructured, raw
Schema-on-write	PROCESSING	Schema-on-read
Expensive for large data volumes	STORAGE	Designed for low-cost storage
Less agile, fixed configuration	AGILITY	Highly agile, configure and reconfigure as needed
Mature	SECURITY	Maturing
Business professionals	USERS	Data scientists, data engineers, et al.

So how do these differences affect whether we choose data lake over a data warehouse?

- *Data*: A data warehouse only stores data that has been modeled and structured with clear entity relationships defined. A data lake stores any type of data – raw and unstructured as well as structured data.

[7] *Marketers ask: What can Hadoop do that my data warehouse can't?* (n.d.). LinkedIn. https://www.linkedin.com/pulse/marketers-ask-what-can-hadoop-do-my-data-warehouse-cant-tamara-dull/?src=aff-ref&veh=jobs_aff_ir_pid_27795_plc_Viglink%20Primary_adid_637360&trk=jobs_aff_ir_pid_27795_plc_Viglink%20Primary_adid_637360&clickid=Qa4Ud2RpYxyOU020TWXZ0S3wUkiwHFTONw9eQk0&irgwc=1 and *Data lake vs data warehouse: Key differences*. (n.d.). KDnuggets. https://www.kdnuggets.com/2015/09/data-lake-vs-data-warehouse-key-differences.html..

– *Processing*: Before loading data into a data warehouse, we must give it some structure and shape – i.e., we need to model it and set it up in tables and identify primary and foreign keys. This is called *schema-on-write*. It can take a long time and lots of resources to understand the schema of data upfront. With a data lake, you are able to load the data as is, raw and without any need to know or make any assumptions about its structure. You can enforce a structure with your programs when you read the data for analysis. That's called *schema-on-read*. These are two very different approaches.
– *Storage*: One of the advantages of a data lake is using open source technologies such as Hadoop that offer extremely low-cost storage and distributed computing capability that support complex data analysis. Hadoop is not only free, but also runs on low-cost commodity hardware. Since Hadoop can scale up as your data grows, you can simply add hardware when you need it. These features provide a lower cost of ownership for storage.
– *Agility*: A data warehouse is inherently designed to be a structured repository. While it's possible to change the data warehouse structure, it will consume a lot of time and resources to do so. In contrast, a data lake relaxes the structural rules of a data warehouse, which gives developers and data scientists the ability to easily load data, configure and reconfigure their models, queries, and data transformations, on the fly.
– *Security*: Data warehouses have matured over the last two decades and are able to offer detailed and robust security and access control features. Data lake technologies are still evolving and hence much of the burden of security and data governance falls on the user community's shoulders. Until the data lake security tools catch up to data warehouse levels, significant effort and attention must be spent toward securing data.
– *Users*: Data warehouses are designed to support hundreds if not thousands of business users and queries as the platform of choice for business intelligence (BI) reports. However, a data lake is typically used by a handful of data scientists and data engineers. While the number of data lake users is steadily on the rise, a data lake is not designed to handle thousands of user accounts and logins without several front-end and middle-ware applications to handle the user interface. The data lake, at this time, is best suited for use by data scientists and data engineers. But as the data lake paradigm shifts from data science to data products, we can anticipate a larger volume of business users utilizing the data lake through new analytics products and data pipelines that are developed by data scientists and data engineers.

Data Governance in the Cloud

The arrival of cloud computing has made a tectonic shift in how IT organizations plan and operate their data assets. The cloud has become a major source of competitive advantage for business. It now impacts and enables every aspect of business in one form or another.

Cloud computing offers many advantages over on-premise infrastructure: lower cost of storage, faster time to implement, high availability, managed services, automated operations, data analytics tools, and BI services. This allows the business to focus on their core business drivers.

Cloud service providers offer a range of services ranging from the most basic tier to the most value-added platforms. There are three major tiers of service:

1. **IaaS:** *Infrastructure as a Service* offers the basic elastic, scalable computing and raw storage. The enterprise is responsible for everything else such as installing the operating system, configurations, applications, monitoring, network implementation, and security.
2. **PaaS:** *Platform as a Service* offers the same capabilities as IaaS, but with added services such as a managed operating system, managed databases, network appliances, data management tools, and security policies.
3. **SaaS:** *Software as a Service* is the higher tier of cloud services. It includes managed applications, tools for data migrations, analytics, and additional security monitoring solutions.

There are many variations to these three tiers that vary by service level and domain-specific configurations. For example, both Amazon and Microsoft are working to differentiate their services by introducing industry-specific cloud solutions. IBM has announced developing the world's first financial services-ready cloud.

Companies are often adopting a hybrid strategy, with part of their applications running on-premise and the rest running in the cloud. Since the public cloud is a shared platform, it's crucial that businesses ensure that cloud service providers adhere to strong data governance practices.

If this wasn't challenging enough, to make data governance more interesting is the prediction that most companies will end up with a hybrid cloud architecture as well as a multi-cloud architecture – meaning a company might be using multiple cloud providers in addition to on-premise (on-prem) infrastructure.

Cloud technology has challenged many of the age-old beliefs and ways of doing things. These days business departments, teams, and even individuals are empowered to set up a computing platform and data environment in a few seconds thanks to cloud technology. With this easy and abundant access to computing resources, any individual can collect, store, and process data with little or no involvement from a central IT organization. The challenge that cloud computing presents is data governance in a distributed computing environment where accountability, regulation, and policies are potentially unclear or fragmented.

Data governance structure and policies must stretch to consider the nuances and challenges presented by cloud environments. The challenges grow as companies plan to migrate and consolidate their data centers into the cloud. Cloud-first has now become the new strategy for IT. This implies that finding applications native to the cloud, such as Software as a Service (SaaS) applications take higher priority over any other off-the-shelf solution.

For a soft landing into cloud computing, a company's security and data governance teams must work much more closely and in concert to ensure the security and privacy of the organization's data. When considering adoption of cloud services, it's imperative that the roles and responsibilities of all parties, the Service Level Agreement (SLA) contract, and policies for security with the cloud vendor are clearly understood and defined before any production use.

It's no longer sufficient for cloud providers to be ISO27001 compliant. They need to be able to support you on a number of security and data management policies.

In a bold move, IBM announced the world's first financial industry-ready cloud initiative.[8] With this initiative, IBM is creating a financial-industry-specific configuration of a cloud environment that implements hundreds of security policies. Such a configuration allows financial industry firms to become compliant right from the start and manage their cloud security and privacy risks more effectively. This initiative has attracted some of the largest financial organizations to join and it's likely that other cloud providers such as Amazon and Microsoft Azure will follow or launch similar initiatives.

The US federal government has adopted a cloud-first strategy and formed the FedRAMP standard for cloud providers who wish to provide cloud platforms for government use. The Federal Risk and Authorization Management Program (FedRAMP) is a government-wide program that provides a standard approach to security assessment, authorization, and continuous monitoring for the cloud and the associated cloud services.

Under FedRAMP regulations, a cloud operator must obtain FedRAMP authorization to operate. FedRAMP authorization requires completing the necessary documentation, implementing the specific security controls, having a plan of action with milestones, as well as having continuous monitoring and remediation plans before they can be assessed by a FedRAMP third-party assessment organization to receive a permit to operate.

[8] *IBM developing world's first financial services-ready public cloud; Bank of America joins as first collaborator*. (n.d.). IBM News Room. https://newsroom.ibm.com/2019-11-06-IBM-Developing-Worlds-First-Financial-Services-Ready-Public-Cloud-Bank-of-America-Joins-as-First-Collaborator.

Implementing Data Governance in the Cloud

According to experts, 95% of cloud security failures are due to customers' operational shortcomings. Most data breaches and security issues arise from a customer's misconfiguration or lack of proper patching of their software. Businesses need to implement data governance because they need a holistic blueprint for business data analytics and data management.

There are serious pressures facing data governance that are further elevated due to the shared resource aspect of cloud environments. We need a new and modern blueprint for data governance that can address these challenges:

- High volumes of data coming from multiple sources resulting in fragmented data sets and data inconsistencies
- Data inconsistencies leading to low quality of data
- Lack of uniform standardized policies that govern data access across hybrid and multi-cloud architectures
- The rise of shadow IT organizations, cloud-based initiatives, SaaS applications driving core business functions, self-service analytics, and data democratization across the enterprise
- The crucial need for common data dictionaries, metadata information, data catalogs across hybrid data storage architecture to enable cross-departmental data analysis

But there are several advantages to cloud computing compared to on-prem infrastructure. For example, consider that consistently updated security patches, regular virus scans, physical security, and much better automated tools for monitoring, business continuity, and disaster recovery are invaluable services in the public cloud solutions that enhance security and governance.

Data governance is the key enabler of cloud strategy because it addresses many of the fears and concerns that keep business leaders from adopting cloud services.

While cloud computing introduces new twists and nuances to a company's data governance, a properly developed governance model should consistently support on-prem as well cloud service implementations.

The Eight Best Practices for Big Data Governance

What are the industry best practices for data governance? There are eight best practices for data governance in cloud and hybrid environments:

1. Identify the primary and secondary data stewards for each data set who are subject matter experts in that data set.
2. Define data dictionary (metadata), data catalog (list of data sets), and data lineage (where the data has come from) across the hybrid environment.

3. Define and apply data lifecycle policies from the early stages of capture and initial transformations all the way to the analysis stage and final retirement (sunset data that expires or data past its retention life).
4. Define policies for data analysis, ethical use, and model risk management.
5. Track and minimize multiple instances of the same data.
6. Define policies for security and privacy to ensure uniform implementation across hybrid cloud architecture.
7. Continuously audit data quality, data usage, and vulnerabilities and report issues to leadership.
8. Define and practice remediation, response, and mitigation plans for data events such as breach of data.

The biggest advantage of cloud computing is that you often have lots of choices about data storage architecture, technology, and configuration. For example, you can choose to implement a Hadoop or Snowflake platform in the cloud. You may select a cloud service provider's data analytics tools or implement your own. You can select the cloud provider's database or implement your favorite DBMS. As you expand your presence in the cloud environment, selecting the right technologies, policies, and architectures becomes more critical.

I've mentioned Hadoop several times in the previous sections and this is a good segue to dig deeper into understanding Hadoop and its architecture.

Chapter 4
NoSQL Storage and Security Considerations

It's a well-known fact that the original developers of Hadoop did not consider security as a key component of their solution at the beginning. The emphasis for Hadoop was to create a distributed computing environment that could manage large amounts of public web data. At the time, confidentiality was not a major requirement. Initially it was argued that since Hadoop uses a trusted cluster of servers, serving trusted users and trusted applications, security would not be a concern. However, at more careful inspection of the architecture, those assumptions were found to be inadequate in a real world where threats are rampant.

The initial Hadoop infrastructure had *no* security model – it did not authenticate users, did not control access, and there was no data privacy. Since Hadoop was designed to efficiently distribute work over a distributed set of servers, any user could submit work to run.

While the earlier Hadoop versions contained auditing and authorization controls, including HDFS file permissions, access control could easily be circumvented since any user could impersonate any other user with a simple command line switch. Since such impersonations were common and could be done by most users, the security controls were quite ineffective.

During that time, organizations concerned with security adopted different strategies to overcome these limitations, most commonly by segregating Hadoop clusters into private networks and restricting user access to the authorized users to those segments.

Overall there were few security controls in Hadoop and as a result many accidents and mishaps occurred. Because all users and programmers had the same level of privilege to all the data in the cluster, any job could access any data in the cluster and any user could potentially read any data set. It was possible for one user to delete massive amounts of data within seconds with a distributed delete command.

One of the key components of Hadoop, called *MapReduce* (we will study MapReduce in more detail later) was not aware of authentication and authorization. It was possible for a mischievous user to lower the priorities of other Hadoop jobs in the cluster to make his job run faster, or even worse, kill other jobs.

Over the years, Hadoop became more popular in commercial organizations and security became a high priority. Security professionals voiced concerns about insider threats by users or applications that could perform malicious functions. For example, a malicious user could write a program to impersonate other users and their Hadoop services, or by impersonating the HDFS or MapReduce jobs, delete everything in Hadoop.

https://doi.org/10.1515/9783110664065-004

Eventually, the Hadoop community started to focus on authentication and a team at Yahoo! chose Kerberos as the authentication mechanism for Hadoop. The Hadoop .20.20x release added several important security features, such as:

- *Mutual authentication with Kerberos:* This incorporates simple authentication and security layer and generic security service APIs (SASL/GSSAPI) to implement Kerberos and mutually authenticate users and their application on remote procedure call (RPC) connections.
- *Pluggable authentication for HTTP Web consoles*: This feature allows implementers of web applications and web consoles to create their own authentication mechanism for HTTP connections.
- *Enforcement of HDFS file permissions*: Hadoop administrators could use this feature to control access to HDFS files by file permissions, namely providing an access control list (ACL) of all users and groups.
- *Delegation tokens*: When a Hadoop job starts after the initial authentication through Kerberos, it needs a mechanism to maintain access on the distributed environment as it gets split over several tasks. The delegation tokens allow these tasks to gain access to data blocks where needed based on the initial HDFS file permissions without the need to check with Kerberos again.

Hadoop Overview

Hadoop is a technology that enables storing data and processing the data in parallel over a distributed set of computers in a cluster. It's a framework for running applications on a distributed environment built of commodity hardware. The Hadoop framework provides both reliability and the data moves necessary to manage the processing in a way that appears to be in a single environment. Hadoop consists of two major components: A file system called Hadoop Distributed File System (HDFS) and MapReduce, a tool to manage the execution of applications over the distributed computing environment.

The following sections offer a more detailed view into Hadoop, MapReduce, and how Hadoop distributed processing works. You can skip these without losing relevant information.

HDFS Overview

HDFS is the utility that stores data on the distributed nodes with inherent redundancy and data reliability. Since HDFS comes with its own RAID-like architecture, you don't need to configure your data in RAID format. The HDFS RAID module allows a file to be divided into stripes consisting of several blocks. This file striping increases protection against data corruption. By using this RAID mechanism, the need for data replication can be lowered while maintaining the same level of data availability. The net result is lower storage space costs.

The NameNode is the main component of the HDFS filesystem. It keeps the directory tree of all files in the filesystem and tracks where the data files are stored across the cluster. It does not itself store the data, but tracks where it is stored.

A user application (client application) talks to the NameNode whenever it wants to locate a file, or when it wants to take any file action (Add/Copy/Move/Delete). The NameNode returns the success of such requests by returning the list of DataNode servers where the data is stored.

The NameNode is a single point of failure in the Hadoop HDFS. Currently, HDFS is not a high availability system. When the NameNode is down, the entire filesystem goes offline. However, a secondary NameNode process can be created on a separate server as a backup. The secondary NameNode only creates a second filesystem copy and does not provide a real redundancy. The Hadoop High Availability name service is currently being developed by a community of active contributors.

It's recommended that the NameNode server be configured with a lot of RAM to allow bigger filesystems. In addition, listing more than one NameNode directory in the configuration helps create multiple copies of the filesystem metadata. As long as the directories are on separate disks, a single disk failure will not corrupt the filesystem metadata.

Other best practices for NameNode include:

- Configure the NameNode to save one set of transaction logs on a separate disk from the image, and a second copy of the transaction logs to network-mounted storage.
- Monitor the disk space available to NameNode. If the available space is getting low, add more storage.
- Do not host JobTracker, DataNode, and TaskTracker services on the same server.

A DataNode stores data in the Hadoop File System. A filesystem has more than one DataNode, with data replicated across them. When a DataNode process starts, it connects to NameNode and spins until that service comes up. Then it acts on requests from the NameNode for filesystem operations.

DataNode instances are capable of talking to each other, which is what occurs when they are replicating data. Because of DataNode, there is typically no need to use RAID storage since data is designated to be replicated across multiple servers.

The JobTracker is the service inside Hadoop that assigns MapReduce tasks to specific nodes in the cluster and keeps track of jobs and the capacity of each server in the cluster. The JobTracker is a point of failure for the Hadoop MapReduce service. If it goes down, all running jobs stop.[1]

[1] https://cwiki.apache.org/confluence/display/HADOOP2/JobTracker

A TaskTracker is a node in the cluster that accepts tasks, such as Map, Reduce, and Shuffle operations from the JobTracker. Every TaskTracker contains several slots, which indicate the number of tasks that TaskTracker can accept. When JobTracker looks for where it can schedule a task within the MapReduce operations, it looks first for an empty slot on the same server that hosts the DataNode containing the data. If no slots are available on the same server, it looks for an empty slot on a server in the same rack.

To perform the actual task, the TaskTracker spawns a separate JVM process to do the work. This ensures that a process failure does not take down the TaskTracker. The TaskTracker monitors these spawned processes, watching the output and exit codes from their execution. When a process completes – whether successful or not – the TaskTracker notifies the JobTracker of the result. TaskTracker also sends heartbeat messages to the JobTracker every few minutes to inform JobTracker that it's still alive. In these messages, TaskTracker also informs the JobTracker of its available slots.

MapReduce

MapReduce is a tool that divides an application into many small fragments, each can be executed on any node on the cluster, and then re-aggregates to produce the final result. Hadoop uses MapReduce to distribute work around a cluster. It consists of a Map process and a Reduce process.

The Map process is a transformation of data containing a row Key and Value to an output Key/Value pair. The input data contains a Key and a Value. The Map process returns the Key-Value pair. It's important to note that 1) the output may be a different key from the input, and 2) The output may have multiple entries with the same key.

The Reduce process is also a transformation that takes all values for a specific Key, and generates a new list of the transformed output data (the reduced output). The MapReduce engine operates in such a manner that all Map and Reduce transformations run independent of each other, so the operations can run in parallel on different keys and lists of data. In a large cluster, you can run the Map operations on servers where the data resides. This saves from having to copy data over the network. Instead you send the program to the servers where the data lives. The output list can be saved by the distributed filesystem, ready for the Reduce process to merge the results.

As we saw earlier, the distributed filesystem spreads multiple copies of data across multiple servers. The result is higher reliability without the need for RAID disk configurations, because it stores data in multiple locations – which comes in handy when running applications in parallel. If a server with the copy of data is busy or offline, another server can run the application.

The JobTracker in Hadoop keeps track of which MapReduce jobs are running and schedules individual Map or Reduce processes and the necessary merging operation on specific servers. JobTracker monitors the success or failure of these tasks to complete the entire client application job.

A typical distributed operation follows these steps:

1. Client applications submit jobs to the JobTracker.
2. The JobTracker communicates to the NameNode to determine the location of the data.
3. The JobTracker then finds the available processing slots among the TaskTracker nodes.
4. The JobTracker submits the task to the selected TaskTracker nodes.
5. The TaskTracker monitors the nodes and the status of the tasks (processes) running in the server slots. If these tasks fail to submit heartbeat signals, they're deemed to have failed and the work is scheduled on a different TaskTracker.
6. Each TaskTracker notifies the JobTracker when a job fails. The JobTracker determines what should be done next, to resubmit the job or mark the data "unusable." This is done by either resubmitting the job on another node, or by marking that specific record as a data item to avoid, or it may block the specific TaskTracker as being unreliable.
7. When the work is finished, the JobTracker updates its status and the client application can poll the JobTracker information about the status of the work.

Therefore, Hadoop is an ideal environment to run applications that can run in parallel. For maximum parallelism, the Map and Reduce operations should be stateless and not depend on any specific data. You can't control the sequence in which the Map or Reduce processes run.

It would be inefficient to use Hadoop to repeat the same searches over and over in a database. A database with an index will be faster than running a MapReduce job over unindexed data. But if that index needs to be regenerated when new data is added, or if data is continuously added (such as incoming streaming data), then MapReduce will have an advantage. The efficiency gained by Hadoop in these situations are measured in both CPU time and power consumption.

Another nuance to keep in mind is that Hadoop does not start any Reduce process until all Map processes have completed (or failed). Your application will not return any results until the entire Map is completed.

Other Hadoop related projects and tools include: HBase, Hive, Pig, ZooKeeper, Kafka, Lucene, Jira, Dumbo, and more importantly for our discussion: Falcon, Ranger, Atlas, and many others. Next, we'll review each of these projects to get a more complete picture of the Hadoop environment.

Security Tools for Hadoop

Solutions like Cloudera Sentry, HortonWorks Ranger, Dataguise, Protegrity, Voltage, Apache Ranger, and Apache Falcon are examples of products that are available to Hadoop system administrators for implementing security measures.

Other open source projects such as Apache Accumulo work to provide additional security for Hadoop. Projects like Project Rhino and Knox Gateway promise to make internal changes in Hadoop itself.

Snowflake Technology Overview

Snowflake is a new cloud-based distributed data warehouse as a service. Unlike Hadoop, Snowflake optimizes performance by separating storage nodes from computational nodes. It uses innovative techniques that allow storage and access to massively parallel clusters, each representing a virtual data warehouse. Each virtual data warehouse is an independent computing cluster that does not share computing resources with other virtual warehouses on the cloud. As a result, the operation of one virtual data warehouse does not impact the performance of other virtual data warehouses.

The Snowflake architecture allows scaling without disrupting the data warehouse operations. It can clone and analyze data quickly. Workloads run concurrently on shared, consistent data that is stored in distributed, virtual data warehouses. You can define the storage size of each data warehouse and how to auto-scale with minimum and maximum cluster sizes. You can start, stop, clone, and scale any virtual data warehouse without impacting queries.

When data is loaded into Snowflake, it reorganizes the data into its internal columnar data format, optimized and compressed for high speed query functions. Snowflake can rapidly store any type of data, of any file size, metadata, statistics, structured or unstructured data.

As a cloud service, Snowflake combines multiple services together, such as: authentication, metadata management, query parsing and optimization, access control, and infrastructure management.

Connecting to Snowflake is easy using a web-based user interface. Data can be stored and accessed using an SQL-like language called SnowSQL, and ODBC and JDBC connections to other tools like Tableau are supported. Python applications can connect to Snowflake using native connectors.

Snowflake delivers a fast and functional cloud-based analytics environment that can scale per users' needs.

Chapter 5
The Key Components of Big Data Governance

There are seven key components to the big data governance framework (see Table 5.1). These include *organization, metadata, compliance, data quality, business process integration, master data management (MDM), and information lifecycle management (ILM)*. Each component is described in more detail here:

1. **Organization:** This component deals with setting up the people, roles, responsibilities, accountability, and ownership for data governance. It includes establishing a data council, data stewards, and data governance steering committee. It might also define the role of chief data officer, the data governance officer, and more strategic roles to lead the organization.
2. **Metadata:** This component focuses on establishing a library of data in the firm, including data dictionary, metadata about data, registration process for data, data lineage, and how to search and identify the data that is needed.
3. **Compliance:** How the firm conforms and adheres to regulatory policies must be covered in this component. Internal policies must support the IT governance and data governance objectives. This component defines the first, second, and third lines of defense on data security and privacy.[1]
4. **Data Quality:** The goal of this component is to establish the standards for data quality. Often a gold standard is defined as data that has been verified to be correct – i.e., it can be trusted. The firm may have different types of standards, such as silver or bronze standards, indicating lower quality data. For most industries, such as the financial industry or where Sarbanes-Oxley rules apply, it's important to certify that data is correct and save an exact copy of the original.
5. **Business Process Integration:** This component defines who is the owner of data and is responsible for acquisition, quality, and approvals for access to that data. Business processes that acquire, transform, and update data are needed to ensure consistency of data treatment across the enterprise.
6. **Master Data Management:** This component is concerned with establishing data classification and categories from multiple perspectives. For example, what is the value and impact of each data set? A firm might define three categories of value as low, medium, and high. They then classify their data sets into these buckets. A firm might define a range of data types from privacy and security perspectives. For example, the firm might establish six types of data, ranging from

[1] As a common best practice, the first line of "defense" against data governance policy violation are the people who work with data directly. The second line of "defense" is typically the data governance organization. The audit and compliance department in most firms represent the third line of defense.

https://doi.org/10.1515/9783110664065-005

type 1 (very sensitive, confidential, and personally identifiable information) to type 6 (non-sensitive, public data).

7. **Information Lifecycle Management:** The objective of this component is to define the entire data life cycle and how data must be treated from inception (acquisition) to removal (purge). It considers policies and guidelines to define how data must be handled along its journey in the organization. It outlines the proper usage patterns of data for operations, commercial purpose, data analytics, and other purposes. The guidelines would include both internal and external data.

Table 5.1: The Seven Components of Big Data Governance.

Organization	Establish Data Council, Data Stewards, Data Governance Steering Committee
Metadata	Data definitions, data lineage, technical metadata, data registration
Compliance	Regulatory audits, policies, compliance with security/privacy policies
Data Quality	Ensure data is complete and correct. Measure, improve, certify data
Business Process Integration	Data procurement, ownership, polices around data frequency, availability, etc.
Master Data Management	Establish data and usage taxonomy: business critical hierarchy; Personas: Admins, Members, Providers, Consumers, etc.
Information Lifecycle Management (ILM)	Data retention, regulatory compliance with data/model snapshots; purge Schedule, storage/archiving

Myths About Big Data and Hadoop Lake

Despite the huge investments made into technology, staffing a data analytics office, and costs of acquiring data, governance is overlooked in many organizations. Part of the issue is that the new technology is a bright shiny object, often distracting the planners and practitioners alike from the challenges of managing the data. Big data has come with some myths which overlook the management issues associated with managing, keeping track, and making sense of all the data.

Here are some examples of myths that are prevalent in the industry: “We’re data scientists . . . we don’t need IT . . . we’ll make the rules as we go.” It’s also common to hear data people say: “We can put ‘anything’ in the lake,” and “We can use any tool to analyze data.”

Other myths are not expressed openly, but seem to frame the wrong conversations around data governance. You might have heard other myths such as: “Data quality is not important since my model can handle anything," and "We'll build the environment first and then apply governance."

But the expectation is that without big data governance from the very start, the Hadoop data lake will be in complete chaos within less than two years – a costly "wild west" experiment! Many organizations that did not plan governance early in their adoption of Hadoop have turned their data lake into a data swamp. After spending millions of dollars on their big data infrastructure, they don't know what data they have, how it's being used, and where to find the data they need.

Data Governance Is More than Risk Management

The risk management perspective of data governance defines data governance as a model to manage risks associated with data management and activity. The risk management approach promotes developing a single governance model for all types of data, processes, usage, and analytics across the enterprise, not just limited to big data or Hadoop infrastructure. It's important to consider that for a big data governance model to be effective in managing overall risk, it must be part of the bigger enterprise-wide data governance and management framework.

First Steps Toward Big Data Governance

Building data governance from the ground up is a daunting task. Most organizations have launched some form of data governance process. But there is a wide range of maturity and capability among different companies and industries. For big data governance to succeed, it needs executive sponsorship and the right skilled resources. The four steps described here provide an outline for starting a big data governance capability:

1. The first step toward a practical data governance model is to recognize and highlight the difference between *traditional data* and *big data* governance policies. Hadoop and open source tools stretch the current envelope on big data governance models such that our current models are not adequate, nor adaptable enough to meet the new challenges introduced by big data. There are serious gaps in data governance when we introduce big data infrastructure.

If you consider a big data repository like Hadoop as being the data lake where all of your enterprise data will eventually reside, you can recognize the overwhelming task of bringing all data activities across the enterprise into a governance structure. This step includes gathering data governance requirements and identifying the gaps for your organization before you get too far into implementing the Hadoop infrastructure.

2. The second step is to establish basic rules of governance and define where they are applied.
3. The third step is to establish processes for governance and then to graduate your data scientists' product into governance. This requires clarity around policies and proper training across the user community in your organization.
4. The final step is to identify the tools for data governance that meet your requirements and can be integrated into your existing infrastructure.

What Your Data Governance Model Should Address

For a big data governance model to be effective, there four elements that it must address. This list provides the fundamental essence of a data governance program:

1. **Accountability:** Your plan should establish clear lines and roles of accountability and responsibility across the enterprise and for each division, affiliate, and country of operation for data governance.
2. **Inventory:** This is one of most important features of a data governance model and one that will have one of the highest returns on your investment. The goal of the inventory is for the data stewards to maintain inventory of all data in the Hadoop data lake using metadata tools. The level of documentation in the metadata tool may vary depending on the importance and criticality of data, but, nevertheless, some documentation is required. The metadata tool is known by many names such as a "data registry" tool or a metadata hub (MDH) among other references. But they all point to the central registry where all data is inventoried. In this book, I'll most often refer to the metadata tool as metadata hub (MDH).
3. **Process:** Establish processes for managing and using data in the Hadoop Data Lake, big data analysis tools, big data resource usage, and quality monitoring. The model must define data lifecycle policies, and procedures for data quality monitoring, data validation, data transformations, and data registry.
4. **Rules:** Establish and train all users on rules and code of conduct. It is important to create clear rules for data usage, such as data usage agreements (DUA), data registry and metadata management, quality management, and data retention practices. I've included a set of rules as examples at the end of this section to illustrate a sample list of rules that your governance program might include.

Data Governance Tools

Hadoop was not designed for enterprise data management and governance. For quite some time there were major gaps in bringing Hadoop into the same level of governance as the rest of the organization. However, there are more tools available today and more are soon to come to market.

All this implies that there are two challenges for the enterprise:

The first challenge is that there are many products, but they do not cover an entire functional span to handle all aspects of governance – security, privacy, data management, quality, and metadata management. That leaves you with choices for which products to select and integrate into a seamless and consistent governance infrastructure. The advice and challenge for you is to select less than two or three tools that completely support the entire span of data governance policies and integrate them. Any more than three products can become hard to integrate, manage, and support.

The second challenge is that the pace of innovation in the big data world is quite rapid. New products emerge almost weekly that supersede prior products. So be prepared to throw away your hard work on the existing tools, revamp them, and adopt new tools every two to three years for some time to come.

Figure 5.1 shows an ecosystem of potential tools that perform certain functional requirements of a big data governance infrastructure. The choice and challenge of selecting the right tool is still on your shoulders.

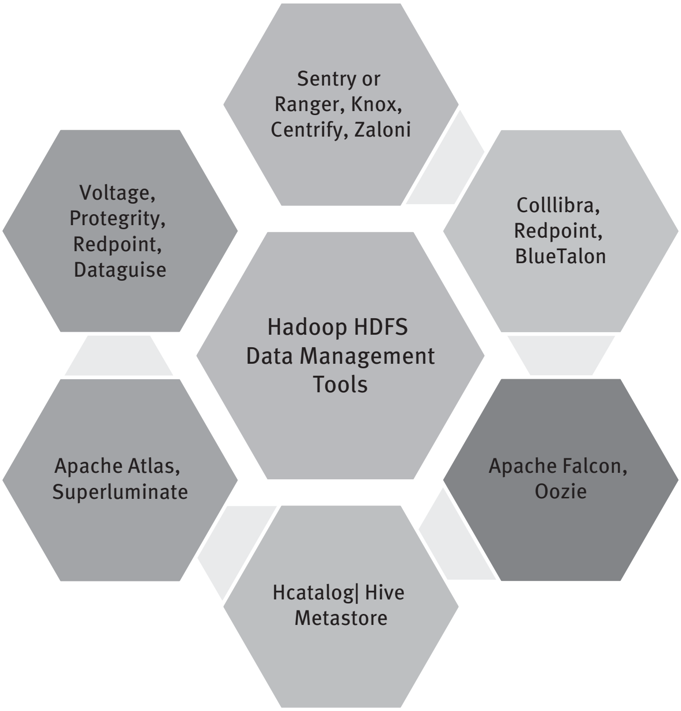

Sentry or Ranger, Knox, Centrify, Zaloni
Colllibra, Redpoint, BlueTalon
Voltage, Protegrity, Redpoint, Dataguise
Hadoop HDFS Data Management Tools
Apache Atlas, Superluminate
Apache Falcon, Oozie
Hcatalog| Hive Metastore

Figure 5.1: The Hadoop and Big Data Governance Tools Ecosystem.

Big Data Governance Framework: A Lean and Effective Model

The major elements of an effective, lean, and agile governance model for big data include four pillars: organization, data quality management, metadata management, and compliance security/privacy policies, shown in Figure 5.2.

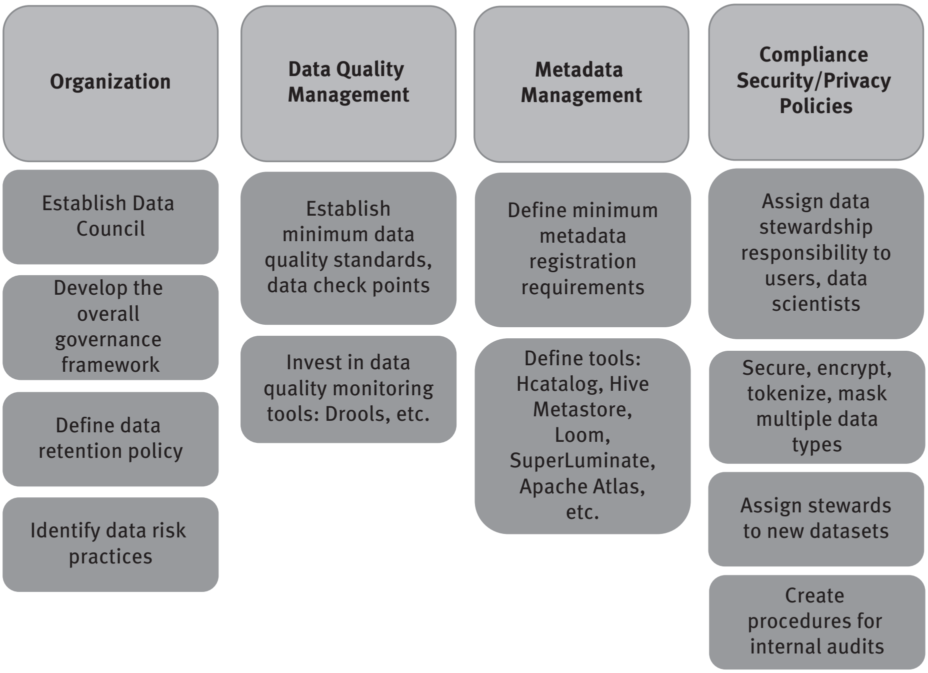

Organization
Establish Data Council
Develop the overall governance framework
Define data retention policy
Identify data risk practices
Data Quality Management
Establish minimum data quality standards, data check points
Invest in data quality monitoring tools: Drools, etc.
Metadata Management
Define minimum metadata registration requirements
Define tools: Hcatalog, Hive Metastore, Loom, SuperLuminate, Apache Atlas, etc.
Compliance Security/Privacy Policies
Assign data stewardship responsibility to users, data scientists
Secure, encrypt, tokenize, mask multiple data types
Assign stewards to new datasets
Create procedures for internal audits

Figure 5.2: The Four Pillars of Big Data Management.

The following are some more points about each pillar.

Organization

The big data governance model must encompass the entire organization. Data governance is not an IT-only concern, nor limited to data scientists in the organization. Everyone in the organization is responsible for good data governance, safeguard of data, and its proper usage. Furthermore, to bring more strategic attention and lines of accountability, we must identify the necessary roles and organizational structures to ensure data governance success.

The first organizational task is to establish a data governance council (*data council*) that is the focal point in the organization to establish and enforce policies. Data councils are typically comprised of IT executives (chief data officer and the chief information officer, to name a few), the *accountable executives* (AE) from various divisions and affiliates, the core data governance leadership, and possibly the chief executive officer (CEO), depending how strategic data management is viewed in the organization. In some organizations, an accountable executive from each line of business may be selected (or appointed) to be an official *data risk officer* (*DRO*) for that division or line of business. The role of DRO is a formal recognition of the importance of data risk management for that line of business.

The other component of organizational structure is to develop the overall governance framework. The framework outlines the expectations, accountability, and elements of data governance. Chapter 10- presents a framework that can be easily adopted as a starting point to form the governance model tailored to your organization.

In addition to the governance framework, we need to establish guidelines around data lifecycle management. Lifecycle management must address the data retention policies, data usage practices, how to handle third party data, and how data should be purged upon its expiration.

Finally, the organization must identify and measure the level of data risks, exposure to data risks, and possible mitigation plans.

Data Quality Management

An activity that's often overlooked in many organizations is oversight of data quality. Across the enterprise, we find either data that's missing, duplicated, or simply filled with errors. The cost of data cleansing and the resources that it consumes are substantial. So why don't we ensure that our data is trusted and of high quality right from the beginning?

We can start anew with the Hadoop Data Lake to enforce the proper data quality and data validation tests to maintain a pristine and trusted data lake.

The data council must define the guidelines for minimum data quality standards and required data check points. Check points in the form of data validation checks are necessary every time data is imported into Hadoop or transformed into a new format.

Process automation can go a long way to effectively manage data quality. We want to implement products such as Drools and comparable solutions to maintain our data engineers' productivity and data analytics momentum. Drools enables data engineers to define rules-based transformations to data. Such rules engines aid in preprocessing data for analysis including converting, normalizing, joining, and cleansing both structured and unstructured data.

Metadata Management

Governance policies must define minimum metadata requirements and rules for registering data in the lake. In the following sections, I'll explain several metadata attributes and a structure that can be adopted as a starting point.

Registering data in a metadata store requires the proper tools. So far, the open source tools like HCatalog, Hive, and Metastore are meeting the very minimum functionality needed to handle metadata management. Other solutions like Loom, SuperLuminate, Apache Atlas, and other tools are available that can offer more utility.

The key to extracting value from data is in the ability to discover the data, finding it, and using it properly. We need to know what data we have and how to find it. We need to maintain information about our data – namely, maintaining metadata.

Metadata is data about data. It's any information that assists in understanding the structure, meaning, provenance, or usage of an information asset. Metadata is sometimes defined as the information needed to make information assets usable, or "information about the physical data, technical and business processes, data rules and constraints, and logical and physical structures of the data, as used by an organization."

It's often said that "you can't manage what you can't measure," which resonates with executive leadership a lot. Similarly, to support the efforts for metadata management, there is plenty of truth to the saying that "you can't govern what you don't understand."

A *business glossary* is a definitive dictionary of business terms, their attributes, and relationships used across an organization. The definitions must be designed to deliver a common and consistent understanding of what is meant by each term for all staff and partners regardless of the business function. For example, a seemingly simple term like "customer" can have many meanings such as "Prospect," "VIP_Customer," "Partner," or "Retailer," where each applies differently with possibly different business rules. Without a consistent and common definition of each customer type, it's difficult to know and segment the market by customers. For example, if your firm intends to send discount coupons to prospects, it should target a different segment and not all customers.

There are three primary categories of metadata:

1. **Business metadata:** Provides the business perspective of data, such as the data dictionary with business definitions for the data items, business processes for data acquisition and transformation, steward and ownership information about the data, regulatory and compliance elements, and data lineage.
2. **Technical metadata:** Provides technical specifications of the data, rules about recovery, backups, transformations, extracting information, audit controls, and version maintenance.

3. **Operational metadata:** Includes operational aspects of data such as data process tracking, data events, results of data operations, dates/times of updates, access times, change control information, and who (or what applications) consume the data.

Metadata management begins with creating a centralized and searchable collection of definitions in a data registry system (the MDH). Providing training to data stewards on how to use the metadata hub so that all data entered into metadata conforms to the necessary quality requirements. In addition, the data stewards are trained to maintain consistent quality of the data – a critical success factor of the program.

Companies such as Kyvos Insights (Los Gatos, CA) offer tools that represent an OLAP cube and dimensional view of your data in Hadoop regardless of the format or the database that it's stored in. Knowing what OLAP cube relations you wish to derive from data is enabled by metadata information that maintains the data relationships with each other.

Compliance, Security, and Privacy Policies

Policies must clearly define accountability in the organization. It's critical that the governance framework identifies the roles and responsibilities of data stewards, data scientists, data engineers, and general users in the organization.

The policies must also define security standards for encryption, tokenization, data masking, and data classification. As part of accountability policies, they must define the process of assigning new data sets to the correct data steward in the organization.

There is no rule of thumb as to how many data stewards are needed for an enterprise. In fact, it's just as difficult to say what the ratio of storage volume per data steward should be. But after grinding into and learning from several big data initiatives around the world, I think you can count on needing 10 data stewards for a petabyte (PB) size data lake, or at least one data steward per 100TB of data.

Finally, be prepared to conduct internal audits. Conducting internal audits is a good measure to manage and identify gaps and risks. Internal audits should be regular (at least once a year). It is best practice that results of internal audits are reported to the data council for review.

Chapter 6
Big Data Governance Framework

Big governance frameworks are crucial to establishing a successful data governance program across the enterprise. The framework must be scalable, extensible, and yet flexible enough to frame the data governance strategy and implementation. Defining the hierarchy of governance imperatives in the form of a priority pyramid can clarify and communicate governance priorities.

One advantage of viewing big data governance as a pyramid is that it exposes governance as a hierarchy and layers of policies. Figure 6.1 shows a four-layer pyramid for governance. At the peak of the pyramid, we enjoy the benefits and value of governance by having fully governed and trusted data. This data is available to the entire enterprise for data scientists and data engineers, subject matter experts and privileged users who can run their data analytics models, business intelligence reports, ad-hoc queries, data discovery, and data insight extraction.

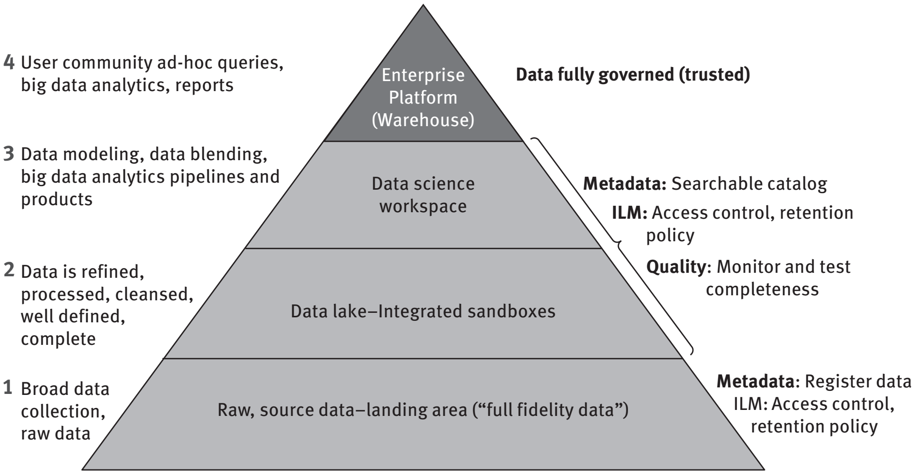

4 User community ad-hoc queries, big data analytics, reports
Enterprise Platform (Warehouse)
Data fully governed (trusted)
3 Data modeling, data blending, big data analytics pipelines and products
Data science workspace
Metadata: Searchable catalog
ILM: Access control, retention policy
Quality: Monitor and test completeness
2 Data is refined, processed, cleansed, well defined, complete
Data lake–Integrated sandboxes
1 Broad data collection, raw data
Raw, source data–landing area ("full fidelity data")
Metadata: Register data
ILM: Access control, retention policy

Figure 6.1: The Big Data Governance Pyramid.

The success of the top layers depends on the effectiveness of the lower layers. The fully governed and trusted enterprise data platform (layer 4) requires a functioning and effective sandbox for data scientists and the user community (layer 3). Layer 4 is the level where users conduct their modeling, big data analytics, set up data analytics pipelines and data products. Layer 3 is the level where users refine their data, process it, cleanse, transform, and curate it – prepare it for analysis.

https://doi.org/10.1515/9783110664065-006

The sandbox functionality requires a robust and well configured data lake to thrive (layer 2). In both layers 2 and 3, it's critical to implement a metadata registry so users can search its catalog and locate their needed data. information lifecycle management (ILM) policies should govern access control, retention policy, and data usage models. In these layers, quality management policies that govern monitoring and data testing should be implemented.

Finally, at the lowest level (layer 1), we have the raw data and landing area from source systems. We want to ensure "full fidelity data," meaning that we want all of the data as-is from the source. In this layer we need policies that define data ingestion pipeline rules, data registration in the metadata hub (MDH), and access controls.

Introduction to Big Data Governance Rules

There are four types of rules (Figure 6.2) that apply to big data in a unique way in addition to the traditional data governance models:

1. Data protection rules
2. Data classification rules
3. Compliance, security, and privacy rules
4. Process (business) rules

Data Protection Rules
Data Classification Rules
Compliance, Security, Privacy Rules
Process (Business) Rules

Figure 6.2: The Four Pillars of Big Data Governance Rules.

I'll discuss each of these rules in more detail in this chapter. Additional policies and configuration recommendations appear in Chapter 9.

Organization

Data Governance Council

In order to manage and enforce governance policies across the enterprise, we need to establish a cadre that supports the operation and implementation of a big data governance framework. The cadre is a core dedicated team of IT engineers who understand Hadoop security, privacy, and compliance standards. Extended from this core team are the "virtual" members of the data council, the data stewards in their respective divisions (or lines of business) who collaborate together collectively under the governance framework. While the data council meets on a regular basis (a monthly schedule is recommended), the data stewards meet more frequently (preferably once a week) to share information and implement data governance policies in a federated model.

Figure 6.3 illustrates a possible organizational chart that defines roles and the interaction model between various data governance roles.

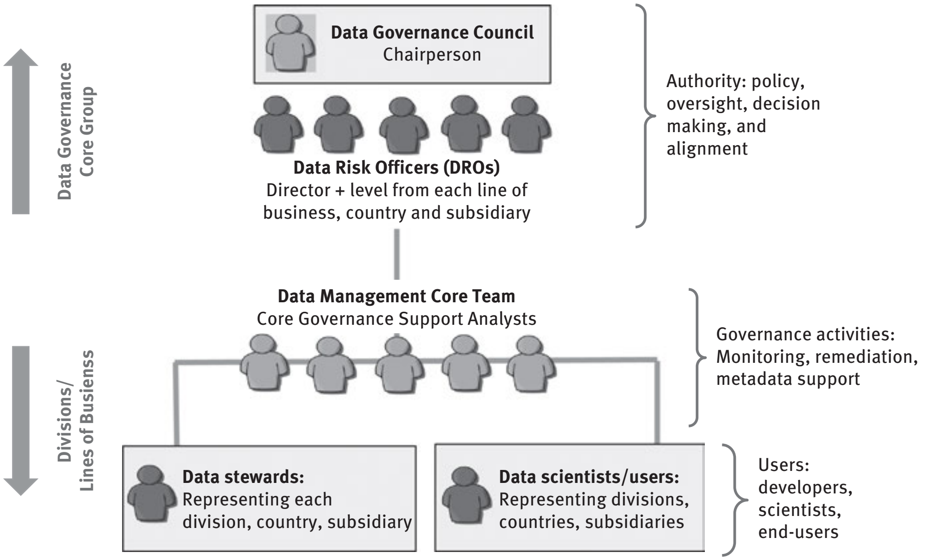

Data Governance Core Group
Divisions/ Lines of Busienss
Data Governance Council
Chairperson
Data Risk Officers (DROs)
Director + level from each line of business, country and subsidiary
Authority: policy, oversight, decision making, and alignment
Data Management Core Team
Core Governance Support Analysts
Governance activities: Monitoring, remediation, metadata support
Data stewards:
Representing each division, country, subsidiary
Data scientists/users:
Representing divisions, countries, subsidiaries
Users: developers, scientists, end-users

Figure 6.3: A Sample Organization Chart for Data Governance.

Here is a brief description of governance job roles:

- *Data council*: A data council is a cross-functional team of business and process managers who oversee the compliance and operational integrity of the data governance framework.

- *Accountable executive*: Also known as AE, an accountable executive is a VP or higher executive responsible for overall compliance with enterprise data governance policies.
- *Data risk officer*: The DRO is a director or higher level manager responsible for identifying and tracking data risk issues related to big data. DROs are appointed by AEs.
- *Data steward*: Data stewards are typically data analysts who represent their respective divisions, country office, subsidiary, or affiliates to apply and monitor data governance policies. Data stewards are the point of accountability for enforcing data governance policies and data quality management. They are often selected by the DROs.
- *Data forum managers*: Some organizations create a data forum management group, but others form the core data management team consisting of IT staff that is responsible for policy updates, technical implementation and operations of systems, tools, and policies of each of the four pillars of the *data governance framework*.

Data Stewardship

Data stewards are the people who define, cleanse, archive, analyze, share, and validate the data that is charted into the metadata of their organization. Data stewards are people who run the data governance operations. Data stewardship is responsible for making certain the information assets of the enterprise are reliable. They update the metadata repository hub, add new databases and applications, introduce new lexicons and glossary definitions, test and validate the accuracy of data as it moves through the organization and transformations.

Data stewardship is the role for practical execution and implementation of policies that the data governance structure mandates. A data steward is accountable for the appropriate, successful, and cost-effective availability and use of a specific part of an organization’s data portfolio.

Data stewards are the agents of trustworthy data within the organization. They’re responsible for tracking, improving, guaranteeing, supporting, producing, purging, and archiving data for their organization. If we visualize the data governance organization as a hub and spokes, the data steward is the hub. They work with the spokes, the other data user roles (data scientist, data engineer, data analysts, etc.) to drive data governance among an organization’s business, IT divisions, and the data council. They help align the data requirements coming from the business community with the IT teams that are supporting their data. Data stewards understand both the technical aspects of the data and business applications of it. So they’re essential to the functional operations of data movement, user access management, and data quality management within their organization (be it an entire company, or a division or country or affiliate, or other sub-segment of the enterprise).

Data stewards oversee the metadata definitions and work daily to create, document, and update data definitions, verifying completeness and accuracy of data assets under their oversight, establishing data lineage, reporting data quality issues, and working with IT staff (or Hadoop technical staff) to fix them, working with data modelers and data scientists to provide context for the data, working to define governance policies, broadcasting and training users on policies for greater data governance awareness.

Data stewards do not have to be in a centralized group under a single manager. In fact, to make the work of data stewards more federated, they can be distributed across the enterprise and report to different managers. It's recommended that each organization within the enterprise (division level or line of business, etc.) fund their staffing resources for data stewards. However, it's conceivable that the IT organization provides such staffing resources as well. In a matrix organization, data stewards may report to the business managers in their organization with a dotted-line to the IT organization.

Authority: Policy, Decision Making, Governance

The data governance framework must establish clear lines of accountability and authority. It will provide the answers to the questions of who will make the decisions related to policy, governance, rules, and process relative to big data governance. So the objective of the discussion around authority must define and regulate policies and answer questions like:

- Which user roles and user groups can access personally identifiable information (PII)?
- What are our data classifications and rules for metadata management related to each class of data?

The framework must also define and create business rules for user access. There is a need for a "controller" who must approve all new employee access to data. In addition, we need to track and report on data lineage to determine (and be able to track and audit) where the data came from and the data source.

Governance polices must also define processes for reporting on data quality rules. They should answer data quality policy questions related to:

- Metadata management and business glossaries
- Data quality profiling
- Master data stewardship

We want to ensure that different departments do not duplicate data across the data lake.

Governance Activities: Monitoring, Support, and More . . .

Let’s address some policies around security and *privacy monitoring*. Implementing a big data governance framework must include measures to:

- Ensure data storage platform access controls and usage patterns are in compliance with data privacy policies.
- Ensure compliance with regulatory standards: HIPAA, HITRUST, and other standards that pertain to data and industry.

When it comes to data management, much of the work of data management is carried out by data stewards. Data stewards will handle lineage tracking. This implies that we need to set up a tagging framework for tracking sources. In addition, there are governance policies related to data jurisdiction, in particular as the organization works on data across country boundary lines. We need to apply data jurisdictional rules by affiliate, country, and data provider agreements.

Compliance audits are important to determine how well the organization is performing against its intended governance policies. The goal of compliance audits is to prove controls are in place as specified by regulations. In some cases, you might want to apply for external certification of your governance framework to assure external organizations that you apply adequate data management and governance policies to your data.

Data stewards perform a lot of metadata scanning. They certify that data has complete metadata information. Furthermore, data stewards perform data quality checks, defect resolution, and escalation of data quality issues to the data governance council.

Finally, data stewards collaborate with the Core Governance Team to define a data usage agreement (DUA) for the organization. The DUA defines and controls the usage model of the data set as documented in the metadata hub (MDH).

Users: Developers, Data Scientists, End Users

The user community of the big data environment can be diverse, consisting of developers, data scientists, data product managers, data engineers, and simply “end users” who deal with data in some form or another.

In order to properly manage access controls, it’s prudent to classify data into several classes, for example, sensitive data (data that includes personally identifiable information), critical data (that is used to make critical business decisions), and normal data (that’s not critical to business operations). For a pharmaceutical company, social media data may be regarded as normal and non-critical data. But, *randomized critical trial data*, collected as part of a new FDA drug trial will be highly likely to be considered critical data.

Users have obligations and responsibilities too. They provide input into the data classification process. They request data access by following a process (submitting requests to the data steward for access to historical data containing PII).

Users are responsible for registering their data upon loading it into the data lake (for example, Hadoop). They're responsible for maintaining the metadata information of their data. They may also request information and verification about their data. An example of requesting data verification is to request data lineage/history if reports appear to be incorrect.

Users may run (or request from data engineers) specific data transformations. However, policies require users to adhere to data lineage rules on where to get data and where to store results. Finally, users are expected to work within their sandbox, where they're allowed to create data analytics pipelines and data analytics products.

Chapter 7
Master Data Management

Master data management (MDM) includes the process of standardizing definitions and the glossary of key business entities about data. There are some very well-understood and easily identified master data items, such as “customer,” “sales,” and “product.” In fact, many define master data by simply using a commonly agreed upon master data item list, such as: customer, product, location, employee, and asset. So the systems and processes required to maintain this data are tools of master data management.

I define *master data management* (MDM) as the technology, tools, and processes required to create and maintain consistent and accurate lists of master data. Many off-the-shelf software systems have lists of data that are shared and used by several of the applications that make up the system. For example, a typical ERP system, at a minimum, will have a customer master, an item master, and an account master.

Master data are the critical nouns of a business and generally cover four categories: *people*, *things*, *places*, and *concepts*. Further sub-categories within those groupings are called *subject areas*, also known as domain areas, or entity types. For example, sub-categories of “people” might include customer, employee, and salesperson. Within “things,” entity types are product, part, store, and equipment. Entity types for “concepts” might include things like contract, warranty, and licenses. Finally, sub-categories for “places” might include store locations and geographic franchises.

Some of these entity types may be further sub-divided. Customer may be further segmented based on incentives and history into “prospect,” “premier_customer,” “executive_customer,” and so on. “Product” may be further segmented by sector and industry. The usage of data, requirements, lifecycle, and the CRUD (create, read, update, destroy) cycle for a product in the pharmaceutical sector is likely to be very different from those of the clothing industry. The granularity of domains is essentially determined by the degree of difference between the attributes of the entities within that domain.

There are many types and perspectives of master data, but here are five types of data in corporations:

– *Unstructured*: This is data found in email, social media, magazine articles, blogs, corporate intranet portals, product specifications, marketing collateral, and PDF files.

https://doi.org/10.1515/9783110664065-007

- *Transactional*: This is data related to business transactions like sales, deliveries, invoices, "trouble tickets,"[1] claims, and other monetary and nonmonetary interactions.
- *Metadata*: This is data about other data and may reside in a formal repository or in various other forms such as XML documents, report definitions, column descriptions in a database, log files, connections, and configuration files.
- *Hierarchical*: Hierarchical data maintains the relationships between other data. It may be stored as part of an accounting system or separately as descriptions of real-world relationships, such as company organizational structures or product lines. Hierarchical data is sometimes considered a super-MDM domain, because it is critical to understanding, and sometimes discovering, the relationships among master data.[2]

Master data can be described by business processes, and the way that they interact with other data. For example, in transactional systems, master data is almost always involved with transactional data. The relationships can be determined via stories that depict the business processes. For example, a *customer* buys a *product*. A *supplier* sells a *part*, and a *vendor* delivers a shipment of materials to a *location*. An *employee* is hierarchically related to their manager, who reports up to a *manager* (another *employee*). A *product* may be a part of multiple hierarchies describing their placement within a *store*. This relationship between *master data* and *transactional* data may be fundamentally viewed as a noun/verb relationship. Transactional data capture the verbs, such as sale, delivery, purchase, email, and revocation; while master data embodies the nouns.

Master data is also described by the way that it is created, read, updated, deleted, and searched through its lifecycle. This lifecycle, called the CRUD cycle for short, defines the meaning and purpose of the data for each stage of the lifecycle, for example, how customer or product data are created depend on a company's business rules, business processes, and industry.

One company may have multiple customer-creation methods, such as through the internet, directly by call-center agent, sales representatives, or through retail stores.

As the number of records about an element decreases, the likelihood of that element being treated as a master data element decreases. But, depending on your business, you may opt to elevate a less frequently used entity to a master data element.

Generally, master data is less volatile than transactional data. As it becomes more volatile, it's considered more transactional. For example, consider an entity

[1] Also referred to as "help-desk" IT tickets. An IT ticket or trouble ticket is issued when IT receives a user problem or issue about data: for example, when certain data is not found or not updated as expected.

[2] Wolter, R., Haselden K., *The what, why, and how of master data management*. (November 2006). Microsoft blog. LinkedIn. https://www.linkedin.com/pulse/what-why-how-master-data-management-manjunath-singh/

like “contract.” Some may consider it a transaction if the lifespan of a “contract” is very short. The more valuable the data element is to the company, the more likely it will be considered a master data element.

One of the primary drivers of master data management is reuse. For example, in a simple world, the CRM system would manage everything about a customer and never need to share any information about the customer with other systems. However, in today’s complex environments, customer information needs to be shared across multiple applications. That’s where the trouble begins because – for a number of reasons – access to a master datum is not always available. People start storing master data in various locations, such as spreadsheets and private application stores, i.e., application-specific OLAP[3] databases.

Since master data is often used by multiple applications, an error in master data can cause errors in all the applications that use it. For example, an incorrect address in the customer master might mean orders, bills, and marketing literature are all sent to the wrong address. One customer intended to send discount coupons to prospects of its product, but mistakenly sent coupons to all customers (including existing customers who had just purchased their product). As a result, the company experienced lower profits when existing customers returned to redeem their coupons. Similarly, an incorrect price on an item master can cause a marketing disaster, and an incorrect account number in an Account Master can lead to huge fines.

Maintaining a high-quality, consistent set of master data for your organization has become a necessity. Master data management must encompass big data and data lake structures like Hadoop. An important step toward enterprise master data management is to create a metadata system in Hadoop that is in sync with the metadata in the rest of the enterprise. Some companies have a single metadata system. Others have created two metadata registries, one that serves the traditional data stores (Oracle, Teradata, MySQL, etc.), and another that is dedicated to Hadoop. However, having two metadata registries out of sync with each other can lead to duplication of efforts and errors. These companies are integrating their two metadata systems into one or enabling them to update each other upon a change in one system or the other.

There are three basic steps to creating master data:

1. Clean and standardize the data.
2. Consolidate duplicates by matching data from all the sources across the enterprise.
3. Before adding new master data from data in the Hadoop data lake, ensure you leverage from existing master data.

[3] Online analytical processing (OLAP) is typically a shadow of an online transaction processing (OLTP) database designed to handle queries and reports so the OLTP database is not impacted. OLTP is transactional, OLAP is informational.

Before cleaning and normalizing your data, you must understand the data model for the master data. As part of the modeling process, the contents of each attribute must be defined, and mapping must be defined from each source system to the master data model. This information is used to define the transformations necessary to clean your source data.

The process of cleaning the data and transforming it into the master data model is very similar to the extract, transform, and load (ETL) processes used to populate the Hadoop lake. If you already have ETL and transformation tools defined, you can just modify these to extract master data instead of learning a new tool. Here are some typical data-cleansing functions:[4]

- *Normalize data formats*. Make all the phone numbers look the same, transform addresses (and so on) to a common format.
- *Replace missing values*. Insert defaults, look up ZIP codes from the address, look up the Dun & Bradstreet number.
- *Standardize values*. Convert all measurements to metric, convert prices to a common currency, change part numbers to an industry standard.
- *Map attributes*. Parse the first name and last name out of a contact-name field, move Part# and Part_No to the Part_Number field.

Most tools will cleanse the data automatically to the extent that they can, and put the rest into an error table for manual processing. Depending on how the matching tool works, the cleansed data will be put into a master table or a series of staging tables. As each source is cleansed, the output should be examined to ensure the cleansing process is working correctly.

Metadata Management

Metadata is not limited to just names, definitions, and labels. In fact, it should include more comprehensive information about the data set, the data table, and the data column (field). The data attributes stored in the metadata should include:

- *Data lineage*: Record the source of the data and the history of its movement and transformations through the data lifecycle.
- *Tracking usage*: Record who creates this data and who (which programs) consume the data.

[4] Wolter, R., Haselden K., *The what, why, and how of master data management*. (November 2006). Microsoft blog. LinkedIn. https://www.linkedin.com/pulse/what-why-how-master-data-management-manjunath-singh/.

- *Relationship to products, people, and business processes*: Define how the data is used in business processes or business decision making. Answer questions like how the data is used in product development and by people.
- *Privacy, access, and confidentiality information*: Record the privacy, regulatory, and compliance requirements of the data.

In order to manage metadata, you must implement a metadata system, also known by other names such as metadata hub (MDH) or metadata registry application. There are several possible choices for implementing a metadata registry system. You may consider open source and proprietary tools such as:

- HCatalog
- Hive Metastore
- SuperLuminate
- Apache Atlas

In addition, you may consider integrating other tools such as Oozie, Redpoint, and Apache Falcon to complement the metadata registry system.

Big Data Classification

The key to successful data management and data governance is data classification. Data may be classified around several topics and contexts. A classification that defines the stages of data in the Hadoop lake is: *raw*, *keyed*, *validated,* and *refined*.

- *Raw data*: When data lands in the Hadoop lake, it's regarded as raw data. Best practices denote that no analysis or reports are to be generated directly from this data. Raw data must have limited access, typically by a few data engineers and system administrators (Admins). Raw data should be encrypted before landing in the Hadoop lake.
- *Keyed data*: This is the data that has been transformed into a data structure with a schema. Keyed data might be put into a *key-value paired structure*, or stored into Hive, Hbase, or Impala structures. Keyed data is derived from raw data, but physically is in a different data file. Access to keyed data is also limited to a few data engineers and administrators. No analysis or reporting is allowed from this data.
- *Validated data*: When data sets go through transformations (such as filtering, extraction, integration and blending with other data sets), the resulting data sets must be reviewed and validated against the original data sets. This is an important step to ensure quality, completeness, and accuracy of data. *Data check points* work to validate data after each transformation. Data engineers and data stewards monitor and pay attention to any error logs from data transformation jobs for failures. There are tests (such as counting the number of records before

and after the transformation) to ensure the resulting data is accurate. Once data passes these tests, it's regarded as validated. Validated data may be used for analysis, reporting, and exporting out of the Hadoop lake.

Governance policies for validated data make it possible for users to access the data (as long as they're privileged to have access to that particular data set) and perform analysis or reporting. Validated data is derived from keyed data but it's in a different file. At this stage the raw and keyed data may be discarded as they've been transformed into the validated stage.

– *Refined data*: Validated data may undergo additional transformations and structure changes. For example, a validated data set may be transformed into a new format and get blended with other data to be ready for analysis. The result is refined data. Refined data may be exported out of Hadoop to be shared with other users in the enterprise.

Figure 7.1 depicts the distinctions and usage rules for each classification of data in Hadoop.

It's important to note that the data files in each stage of this taxonomy are structured in *directory file structures* that are most appropriate for their stage of refinement. As I'll explain in more detail later, in the first two stages, the data directory structure still mimics the directory structure of the source system that it was extracted from. But, in the final two stages (validated and refined), the directory structure is set up according to the user domain or the use case. Per best practices, only data from these final two stages may be published, analyzed, or shared outside of Hadoop.

Security, Privacy, and Compliance

The purpose of *security and privacy controls* is to determine who sees what. This implies certain policies to secure as many data types as possible, and to define data classifications. *Data classifications* have multiple dimensions. For privacy and security controls, consider classifying data into critical, high priority, medium priority, low priority; or a similar model that fits your organization. Critical and high priority data are those data items that are materially significant to your business and business decision making.

Another classification is to determine which data sets are sensitive versus not-sensitive. Sensitive data is a classification that includes personally identifiable information (PII), or patient health information (PHI), and credit card information (subject to PCI security standards).

As part of this process, you must determine which HITRUST controls are required for the Hadoop lake. You must define data encryption, tokenization, and data masking rules. For example, large files that are coarse grained might get encrypted upon

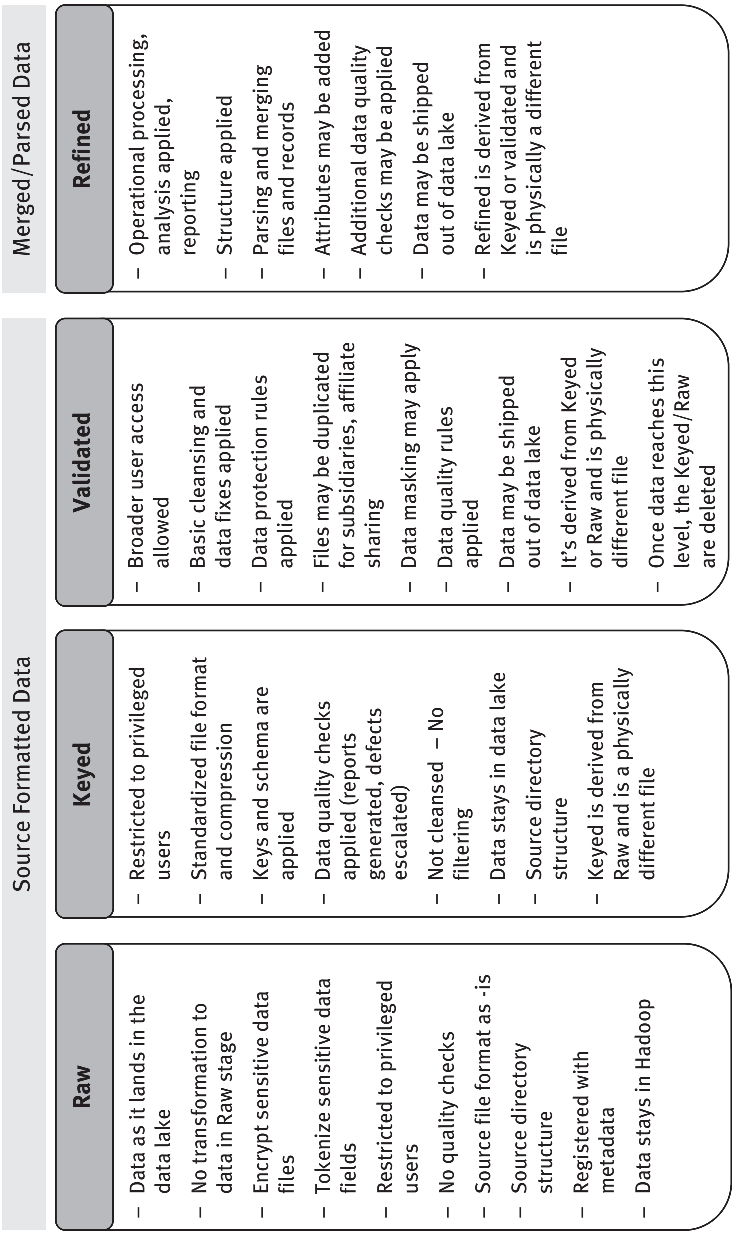

Source Formatted Data
Merged/Parsed Data
Raw
– Data as it lands in the data lake
– No transformation to data in Raw stage
– Encrypt sensitive data files
– Tokenize sensitive data fields
– Restricted to privileged users
– No quality checks
– Source file format as -is
– Source directory structure
– Registered with metadata
– Data stays in Hadoop
Keyed
– Restricted to privileged users
– Standardized file format and compression
– Keys and schema are applied
– Data quality checks applied (reports generated, defects escalated)
– Not cleansed – No filtering
– Data stays in data lake
– Source directory structure
– Keyed is derived from Raw and is a physically different file
Validated
– Broader user access allowed
– Basic cleansing and data fixes applied
– Data protection rules applied
– Files may be duplicated for subsidiaries, affiliate sharing
– Data masking may apply
– Data quality rules applied
– Data may be shipped out of data lake
– It's derived from Keyed or Raw and is physically different file
– Once data reaches this level, the Keyed/Raw are deleted
Refined
– Operational processing, analysis applied, reporting
– Structure applied
– Parsing and merging files and records
– Attributes may be added
– Additional data quality checks may be applied
– Data may be shipped out of data lake
– Refined is derived from Keyed or validated and is physically a different file

Figure 7.1: The Four Stages and Classifications of Data in Hadoop Lake.

landing in the Hadoop lake but more detailed data columns that contain PII, such as Social Security numbers, or credit card numbers would get tokenized. Finally, your model would determine how to mask certain data. For example, if your call center agents need to know the last 4 digits of a credit card number, you can *mask* the first 12 digits of the credit card so only the last 4 digits are visible.

Whether you use an existing tool like Active Directory or make a new tool for user access controls, it's imperative to define user groups and user categories. Using user groups and user roles, you can control access to specific data in the lake. For proper management of security and privacy controls, establish a centralized and integrated security policy tool set and process.

In addition, since Hadoop is a cluster of servers, you must define intra-application security rules on the cluster. To ensure that the security and privacy controls are being properly followed and applied, we need to consider performing periodic internal audits and report gaps to the data council.

Apply Tiered Data Management Model

The data governance policies don't need to apply equally to all data. The low-priority data might include social media data that doesn't rise in importance to other critical data; hence the governance policies might be less stringent. But the critical data such as patient health information will require the most strict of governance policies. Figure 7.2 illustrates the distinctions that you can consider when crafting and administering data governance policies by data classification.

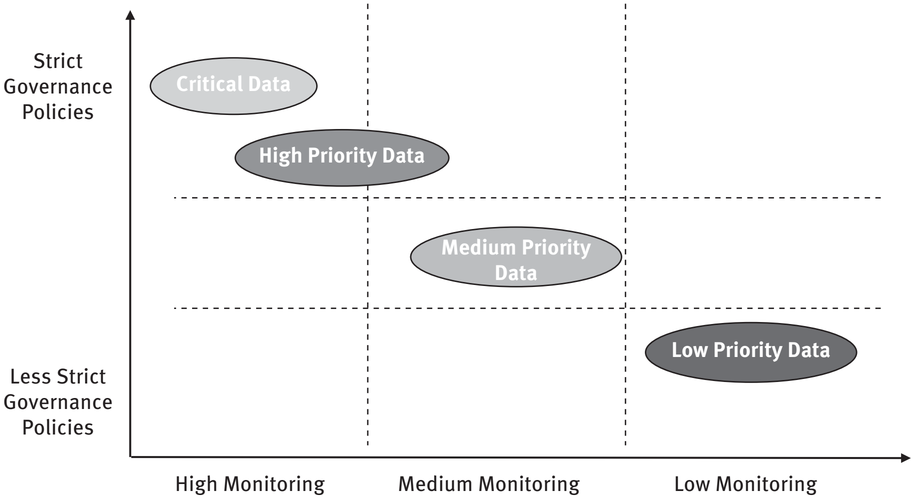


Figure 7.2: More Strict Policies and Monitoring Apply to Critical and High Priority Data.

Big Data Security Policy

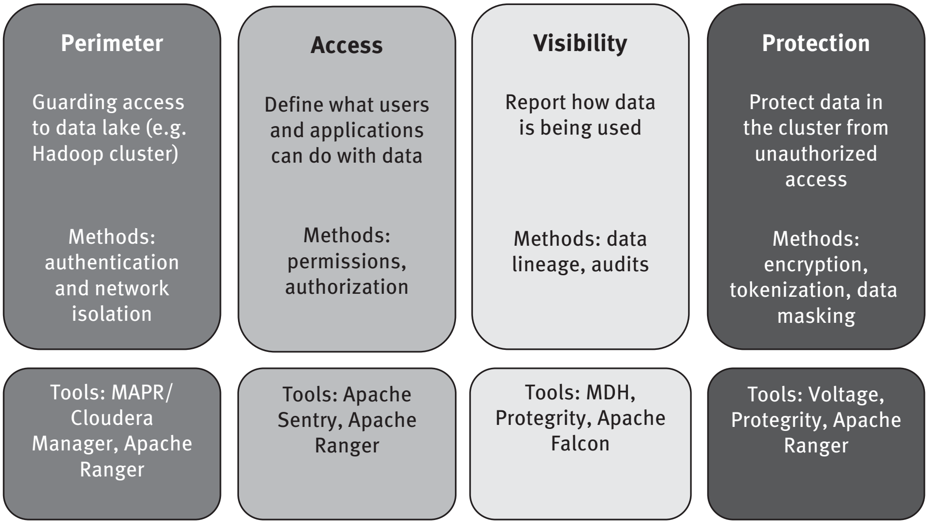
Because Hadoop was not designed with security in mind, two new challenges are added with Hadoop in the cloud: 1) Securing the cloud – whether it's Amazon, S3, Azure, or any other cloud architecture, 2) Securing the Hadoop and open source apps (MapReduce, etc.)

Securing the cloud is made possible by several standards and solutions. For example, you can decide to apply HITRUST, and Cloud Security Alliance STAR certifications.

In order to secure the Hadoop (Cloudera, HortonWorks, MapReduce, etc.) and open source apps, the solution is to apply the four pillars of the Hadoop environment.

1. **Perimeter:** Secure Hadoop at the borders (such as building intrusion detection tools).
2. **Access:** Limit access to data by role and group membership.
3. **Visibility:** Maintain data visibility thru metadata management.
4. **Protection:** Apply encryption, tokenization, data masking.

To illustrate the four pillars of Hadoop security in more detail consider Figure 7.3. I've included the purpose and strategies to implementing each pillar along with tools that make those strategies possible.



Perimeter	Access	Visibility	Protection
Guarding access to data lake (e.g. Hadoop cluster)	Define what users and applications can do with data	Report how data is being used	Protect data in the cluster from unauthorized access
Methods: authentication and network isolation	Methods: permissions, authorization	Methods: data lineage, audits	Methods: encryption, tokenization, data masking
Tools: MAPR/ Cloudera Manager, Apache Ranger	Tools: Apache Sentry, Apache Ranger	Tools: MDH, Protegrity, Apache Falcon	Tools: Voltage, Protegrity, Apache Ranger

Figure 7.3: The Four Pillars of Hadoop Data Lake Security.

Big Data Security: Access Controls

The access control mechanism for the Hadoop lake is best served when it's integrated and compatible with the existing security procedures and policies. A user's access to Hadoop data need not be different from access to conventional data sources.

We can view a user's access as a chain of policies and procedures. To enable proper access controls, we define the user's membership in certain groups. Next, the role of the user is defined. Together group membership and role give a granular control of access (*read, write, delete,* and *execute)* privileges to a particular data element.

Consider a typical user, John Doe, a member of the HEOR group who is a data scientist as shown in Figure 7.4. Based on this information, a Hadoop security tool such as Sentry can be configured to give read-only access to sensitive data. John Doe's access is coordinated with his access privileges in Microsoft Active Directory so it's consistent across the enterprise and centrally managed.

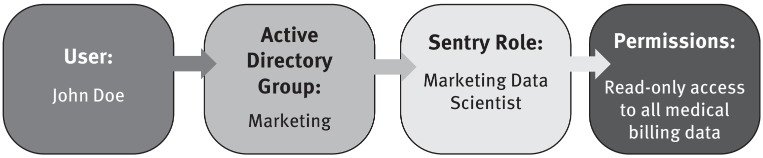


Figure 7.4: Access Control Chain in Hadoop Data Lake.

For example, in Figure 7.4, a user (John Doe) belongs to the marketing access group. The access is read-only and limited to certain data. This type of access could be controlled through active directory membership and Sentry rule sets.

Big Data Security: Key Policies

Let's consider some of the high level and key security policies for big data. Here is a list of sample policies that you can adopt and adapt to your organizational requirements.

1. Security controls on the data lake must be centrally managed and automated.
2. Integrate data lake security with existing IT tools such as Kerberos, Microsoft Active Directory, LDAP, logging, and the governance infrastructure.
3. All intra-Hadoop services must mutually authenticate each other.
4. Apply data access controls to the data lake by using a tool like Apache Sentry to enforce access controls to data before and after any data extract, transform, and load (ETL).

5. Enforce process isolation by associating each task code to a job owner's UID. This ensures that user access privileges are consistently applied to their jobs in the cluster.
6. Maintain full audit history for Hadoop HDFS, for tools and applications used including user access logs.
7. Never allow insecure data transmission via tools such as HTTP, FTP and HSFTP. Enforce IP address authentication on Hadoop bulk data transfers.
8. Apply Hadoop HDFS permissions for coarse data controls: directories, files, and ACLs; for instance, each directory or file can be assigned multiple users or groups with the file permissions Read, Write, and Execute.
9. Apply encryption to coarse-grained data (files) and tokenization to fine-grained data (table columns).

Data Usage Agreement Policy

One of the policies that data governance must establish is a data user agreement. This is an agreement between the data governance council and users that allows users access to data and, in exchange, expects and lists a number of policies that users agree to abide by. Here is a list of policies that can be considered in a data user agreement:

1. All critical and high priority data must be of "trusted source" quality. In other words, users will refrain from conducting analysis or reporting on data that is not trusted. Similarly, users agree to make their data in Hadoop lake "trusted" data, meaning they will complete the metadata for their data and validate their data after each transformation for quality and accuracy.
2. Maintain data asset change control (requiring advanced notice of any data schema changes). This implies that users will notify data stewards if they modify any schema or add data columns to their data table.
3. Data stewards agree to monitor data quality and metadata compliance. The stewards inspect metadata for user data periodically to ensure data with "trusted" status meets metadata specifications for completion.
4. Users agree to register and update the metadata hub (MDH) when loading data into Hadoop. Data stewards may notify users and issue warnings if their data does not have adequate metadata information.
5. Third-party and fourth-party data is to be managed and used in accordance to data agreements with the third-party data provider. If your organization purchases data from external third-party or fourth-party sources, the contract usually comes with certain clauses and usage constraints. For example, one provider of consumer data might allow analysis of data for economic purposes, but not for marketing purposes. This policy ensures that users will abide by these data usage clauses.

6. Data retention policies must be applied according to the DUA[5] policies. Data retention policies vary from country to country and even from state to state in the US. Some states might have retention policies of 10 years and some even longer. Users must be aware and adhere to these retention policies.
7. Users agree to complete the enterprise big data governance annual training. Users and training are often considered as the first line of defense. Data stewards or Hadoop administrators should only provide access to users after users have completed their training in data governance policies.

Security Operations Policies

Operationalizing security policies and coordinating the people, policies, tools, and configurations to deliver a cohesive security policy on open source data lake platforms like Hadoop can be challenging. A slight change to one aspect of the policy, tool, or configuration can introduce huge variations or deviations from the intended course. Hence, it's important to devise operational policies for performing the security measures for big data. I've listed some sample policies that can be adopted for maintaining best in class security operations:

1. Data stewards are tasked with and held responsible for notifying the Hadoop security admin regarding onboarding or off-boarding users for their areas of business.
2. Enable fine grain access controls to data elements. Create user groups and assign user IDs to the appropriate groups.
3. When requesting access to Hadoop data, users must submit an access request form showing they've completed pre-requisites before getting access granted. Access request forms must be approved by the line of business (division, franchise, etc.) data risk officer.
4. Apply random data quality checks for validity, consistency and completeness. Report data quality issues to the data council.
5. All IP addresses must be anonymized to protect the user data.
6. All sensitive data, such as personally identifiable information and data (PII), must be de-identified in raw form or become encrypted. Tokenization is preferred when data columns (fields) are known.

[5] Data Use Agreement (DUA) is an agreement that is required under the privacy rules. This agreement must be entered before using or disclosure of any dataset by either an internal or external entity or both.

7. Develop and monitor performance metrics for services and role instances on the Hadoop clusters. Monitor for sudden deviations. For example, if a user has historically used a small fraction of data and resources in Hadoop, but you discover a sudden increase in data and CPU resource consumption, it should alert you of a possible rogue or insidious attempt to breach your data.

Information Lifecycle Management

Managing data through its lifecycle span is accomplished using a process called CRUD: create, read, update, and delete. There are certain policies that apply to data during its lifecycle that govern allowed usage, handling processes, and retention rules. Here are some high level rules that I recommend as the basis of Information Lifecycle Management (ILM) governance:

1. The rules for minimum data retention may vary from country to country. The longest retention period must be updated in the metadata hub (MDH) tool for all data sets.
2. If your industry requires that you reproduce the data and analytics results at another time, you must take snapshots of data and analytics models and safeguard them. One approach is for you to maintain a separate sandbox for data snapshots that are required for compliance, legal, and medical reasons.
3. The data snapshot sandbox is configured with only *admin* access privileges. In other words, only the Hadoop administrators may have access to this sandbox in order to minimize access to its data.
4. Define Hadoop disaster recovery, downtime, and business continuity policy. Protecting data may require duplication of the data lake across different geographies. For example, storing data in two separate Amazon locations, say Amazon East as well as Amazon Asia, offers redundancy for better business continuity and faster recovery.
5. Data recovery procedures should define how fail-over and fail-back procedures would work, how replication would be restored, the process for restoring data, recovering data, and synchronizing data after the failed site is online.

Quality Management

Big Data Quality and Monitoring

Data quality monitoring involves the inspection and testing of data as it goes through transformation and movement throughout the enterprise. The goal of data quality monitoring is to ensure accuracy and completeness of data.

According to the best practice governance guidelines, all data must be monitored. Consider data extraction, transfer, load (ETL) jobs that run as part of a data processing pipeline. Each step of the data pipeline is carried out by a data process which produces an error log. One data monitoring task is to inspect the error logs to identify if any data processing jobs failed. Other data monitoring techniques including testing, such as conducting basic counts and comparing the number of records or volume of data from source to target (do we have the same number of records before and after the transformation?). Another technique is to conduct boundary value tests on data fields. If a data field is expected to have a range, you can test the new data set against that range. Finally, other techniques look for the suspicious presence of too many NULL values, and other data anomalies to detect quality issues.

The Hadoop ecosystem offers many data pipeline processing tools that report interim error logs for review. These tools and some common practices include:

1. Using Oozie to orchestrate data pipelines and track processing errors at each stage of data processing.
2. Using HBase to record and maintain errors related to metadata errors, data processing job errors, and event errors.
3. Configuring the tools to provide automatic quality error notifications and alerts if errors are reported in the logs.

Data Classification Rules

The key to granular data access management lies with data classification. You can define the data classes for your organization since each organization is different. However, typically, good data classification should consider multiple dimensions and uses of the data, not just format and frequency.

We've seen data already classified into sensitive (PII, PCI, PHI, etc.) and non-sensitive. We also looked at data classes in Hadoop as raw, keyed, validated, and refined. In addition, there is the distinction between "pending" status of data (data that is not primed for usage and analysis) versus "trusted" data (data that has been validated, has complete metadata, has lineage information, and is certified by quality checks). Furthermore, here we can define four additional data classes:

1. **Critical Data:** The type of data that has the highest regulatory, reporting, and compliance requirements defined by FDA and/or GxP[6] data. GxP data apply to key business processes and are materially significant to decision making at an enterprise level.

[6] Good x Practices, where x is one of your key business processes. For example, Good Manufacturing Practices.

2. **High Priority Data:** The type of data that is materially significant to the decision making at an enterprise, and is regarded as classified information and confidential data. This data is required for analytics but not subject to GxP or FDA regulations.
3. **Medium Priority Data:** Data that is typically generated or used as a result of day-to-day operations and is not classified as confidential or classified.
4. **Low Priority Data:** Data that is typically not material and does not have retention requirements.

Data Quality Policies

The partnership between users and data stewards requires policies that define proper quality management of data. Data quality issues may be escalated to the data governance council, but the organization ultimately must commit to track and resolve them. One of the functions that data stewards fulfill is to ensure that their organization's data meet data quality specifications. The following are sample policies that can be used as guidelines for quality-driven standards, strategies, and activities to ensure data can be trusted:

1. *Trusted source* is defined as a critical or high priority data set that has a complete and accurate metadata definition including data lineage and the quality of the data set is known. Sources that are not fully meeting the definition of trusted source will be given "*pending*" status until the data set is brought into compliance. No reporting or analysis of pending status data is allowed.
2. Data stewards can assign the status of pending or trusted to data sets in the metadata. Each accountable executive and data risk officer (DRO) must be aware of and be able to identify which data elements (those that are in their domain of responsibility) are trusted sources or in pending status.
3. Data quality issues shall be defined as gaps between the data element characteristics and deviations from the data governance standard. Data quality gaps typically include: data that has not been validated for accuracy and completeness; data that has incomplete metadata information in the metadata hub (MDH); or data where data lineage is unknown.
4. Policies should require tiered application of data governance rules to data by the data classification. For example, apply data quality monitoring to *critical data* and *high priority* elements within the data assets. Data stewards must document and report to DRO any issues related to data quality.
5. The metadata hub or another database can maintain inventory and records of models used by data scientists. A model may be used by multiple users and data products in the organization. Managing models from a central repository has many advantages. This policy ensures that all models used by the organization are known for tracking purposes, documented for reproducibility at a later time and also to centrally manage and modify them if errors are discovered that require fixing.

6. The data governance council may define data quality metrics for the organization. These metrics might include percent of data in *pending* status versus *trusted* status, the number of open quality issues, and data volume by each division or user group. Data stewards are responsible for providing metrics reports to the DRO, AE, and data council related to the level of compliance of critical data sets. Data stewards must, on a regular basis, provide a data quality issues list with the expected date of remediation and a remediation plan to resolve the issues.
7. Data stewards must maintain evidence of compliance to data management policies and the supporting standards, procedures, and processes.
8. Data stewards along with data officers annually review inventory of critical or high-priority usages and the data usage agreements within their domain of responsibility.
9. For data that is in "pending" status, it can only be moved into "trusted" status (or "certified" status depending on the terminology that you wish to adopt) by the data steward for their area of responsibility if:
   - Metadata is complete in the metadata hub (MDH) tool
   - Lineage is to origination[7] or to the *trusted source*
   - Data quality monitoring and checks are passed
   - The retention period is defined

Metadata Best Practices

The data quality standards at your company under the big data governance policy should require business metadata to be captured and maintained for all important metadata (critical and high priority data).

This section provides guidance on what high-quality metadata looks like. Data stewards are encouraged to use this section to improve their metadata. As your Hadoop data lake grows in size, the value of metadata will grow over time. Descriptive names are required for all important data elements (critical and high-profile data). Generally, data stewards determine the descriptive names.

When completing the metadata, try to answer two key questions:

1. What do I know about this data (about tables/columns or key-value pairs) that someone else would need to know to understand it and properly use it?
2. If you have never seen this data before, what would you want to know to have confidence you were using it properly?

[7] Lineage tracks the journey that a data set has taken from origination or initial acquisition to the current system. The hops and systems from which it has been extracted and copied form the lineage of a data set.

The purpose of metadata is to explain the data so that it can be understood by anyone who is business-aware but does not know the data. The goal is not to give a general brush stroke of information about the data but to give enough specifics so its meaning is clear and it can be interpreted and used. The intent is not to write for someone who already knows the data, but for someone who is new to the company.

Naming a data field (column name) in a way that is self-describing will add clarity to the data naming exercise. For example, a data field that represents customer address is better understood when it has a descriptive name like "CUST_ADDR" than "C_123."

Descriptive names and definitions offer more details and/or more explanation than is present in the physical data element name. While you may not be able to change the database column name, you can change the description of the data element with a descriptive name and be clear in the definition.

Data sets contain codes and flags or indicators. For example, a code may represent gender and a flag may indicate if data was collected or not for a particular time period. You should make proper assignment of descriptive name for *codes* and *indicators*:

- If a physical name contains "_ind" meaning an Indicator, then the possible values should be Yes and No, or True and False, not a list of 5 possible values.
- If a physical name takes more than two values, say 50 values to represent state abbreviations, then the physical data element must contain "_cd" to indicate a code type of data element.

Data stewards must clarify any misleading names. For example, even though the physical data element name is Customer_Identifier, if the field is actually a point of contact, the descriptive name and definitions should be changed to reflect the actual meaning of the data, namely the field name should convey the fact that the data is a point of contact.

Another rule of thumb is to avoid acronyms. Don't assume that everyone will understand what the acronym means. Make an effort to spell out the full name and place the acronym after the name in parentheses.

If the comment with the metadata description is true at that point in time but may change in the future, be specific and indicate the date range for which the description is valid. Avoid using references like "currently" or "now," "before" or "after."

Carefully select the terms that you intend to use. While people use loose terms to mean the same thing conversationally, in metadata descriptions they need to be clarified. Keep in mind that a data element definition must stand on its own. For example, "Customer" and "Account" are two different things. Add qualifiers to make the description more precise. For example, clarify "Customer" by the possible types of customer such as "Account Holder," "Prospect," "Co-Signer," or "Guarantor."

Avoid using pronouns in general. Pronouns are confusing, vague, and can lead to misinterpretation. Best practice guidelines for column and field level metadata include the following policies:

1. Descriptive names should end in a word that classifies the data. For illustration consider date, amount, number, and code which are examples of data classifiers. The classification should immediately follow the name or phrase that it's describing. This implies that descriptive names should be at least two words long. For example, use Customer Balance and Primary Account and not Balance or Account.
2. Descriptive names should be as short as possible while retaining their meaning and uniqueness. But clarity is always more important than brevity.
3. A good descriptive name must reflect the description and definition of the data element. Typically the descriptive name should not contain articles, prepositions, or conjunctions such as "an," "to," "of," "after," or "before."
4. Try to use more generalized definitions and descriptive naming of data elements. For example, if a field is used for both Social Security Number and Employer Identification Number, don't call the field Social Security Number (which is specific to the US). Instead call it Tax Identification Number, or National Tax Identifier, or something more general.
5. Data descriptions, namely data definitions should also contain the business meaning and why the data element is important to the business. Often a data element is computed and if so, it's important to show the math or calculation method in the definition.
6. Each data definition must be independent of any other definition, so that the reader need not jump around to understand the data element. There are however, legitimate situations where it's appropriate to refer readers to other columns/fields. Here are some possible scenarios for cross-referencing descriptions: If there are interrelationships between physical data elements, the relationship would be reflected in the data description; or a hierarchy of data classification exists that forms a tree. For example, when a class of data has sub-classes, the top level column should show how the hierarchy works.

Metadata descriptions must include contextual information. The context should paint a picture of how the data is used and what it means, situated in the context of its usage. The context is often clear to a subject matter expert, but not to the average user. For example, consider a data field name of Interest_Rate. The metadata must inform the reader that the Interest Rate is defined by the 10-year Treasury rate and it changes once a month. It is a real number with 3 digits after the decimal. The metadata must include information about when this rate should be used in calculations and how it is different from other data fields that include other types of interest rates, such as APR (Annual Percentage Rate) or mortgage rate.

If a physical data element is a code, then it must contain a sample set of codes and their meanings in the description. If a data element can have more than one meaning, it implies that another data element determines which definition applies to it. Therefore, the other data element must be mentioned in the description.

If a data element is allowed to contain an anomaly in the data, the meaning of the anomaly must be indicated in the description. For example, if a Social Security field contains a value of -1, as a flag to look up the SSN in a special file, then the description must define that rule.

Furthermore, the organization must maintain a valid value list, namely a code table defining the values for *codes*, *indicators,* and *flags*. If a data element is defined as a code, or indicator or a flag, but does not have a defined value in the code table, the description must provide a valid value list for that element.

If a data element is regarded as critical or high-priority, it should include an indicator to identify it as such. Similarly, if the data element is regarded as sensitive (meaning it contains personally identifiable data), it should also include an indicator to identify it accordingly. The indicator might take values such as “Yes,” “No,” or “Unspecified.”

Finally, data tables and files need to be described in the metadata registry system. The descriptions must identify which applications use the table/file. Table definitions should describe the contents of the data in the table. Table definitions must also include information about uniqueness keys. The uniqueness key is any column or combination of columns which can uniquely identify a row in a table. A table may contain multiple uniqueness keys, so they all need to be defined in the table description.

The primary key (PK) is a uniqueness key that is chosen by designers to enforce uniqueness. A table may have only one primary key either on one column or multiple columns (that together provide a unique identifier) If a primary key involves multiple columns, it’s called a composite key. Multiple columns in a database may be used for a primary key because any one identifier tends to have exceptions. As a result, creating a unique customer id number is a good choice for a primary key.

The contextual information about a table or file is also important. The context should indicate retention period (for how long the table should be maintained) and metadata contact (a person who can explain the meaning of the data) for that table/file. Finally, regular and periodic review of metadata definitions and correcting errors in metadata are important activities for data stewards to improve the overall quality and usage of data in Hadoop.

Chapter 8
Big Data Governance Rules: Best Practices

As a sample of typical big data governance policies and best practices, I've compiled a list of rules. These are, in my opinion, the twelve golden rules to keep in mind when drafting a data governance program.

The goal of these rules is to provide an effective, lean, and tangible set of policies that are actionable and can be put into execution. The data governance council should review these rules at least on annual basis to make changes and modifications. Here is a sample of rules listed below:

- Rule #1: Governed data must have:
  - A known schema with metadata
  - A known and certified lineage (history)
  - A monitored quality test and a managed process for ingestion and transformation
- Rule #2: Governed usage policies are necessary to safeguard data.
- Rule #3: Schema and metadata must be applied before analysis.
  - Even in the case of unstructured data, schema must be extracted before analysis
- Rule #4: Apply the data quality certification program.
  - Apply ongoing data quality monitoring that includes random quality checks and tests
  - Data certification process is applied by data stewards and data scientists
- Rule #5: Do not dump data into the lake without a repeatable process. Establish the process for landing data into the lake (in order to fill the lake). Define guidelines for data transformation and data movement activities within Hadoop.
- Rule #6: Establish data pipeline categories and data classifications that we reviewed. For example, the data stages: raw, keyed, validated, and refined.
- Rule #7: Register data into the metadata registry upon importing it into the lake. In some organizations, it is common practice that Hadoop administrators may issue a warning and purge any data that is not registered in the metadata registry system within 90 days of loading it into Hadoop.
- Rule #8: The higher the data classification and stage, the more complete the metadata must be. For example: A data set in *refined* status is expected to contain all elements of data management (definition, lineage, owner, users, retention, etc.).
- Rule #9: *Refined* data must adhere to a format or structure. Example: Refined data must be either in Hive data format and/or HBase or in a specific schema.
- Rule #10: Classify data with multiple attributes and record the classification in the metadata registry: data owner, type of data, granularity, structure, country jurisdictions, schema, and retention are some examples. The data governance council must define privacy, security, and usage policies for each classification.

https://doi.org/10.1515/9783110664065-008

- Rule #11: One of the cardinal rules of securing data in Hadoop is to encrypt and tokenize. The common rule is to encrypt coarse grained data (these are data files, or unstructured data), and tokenize fine grained data (data sets that include data fields in data tables).
- Rule #12: An alternative policy and extension to Rule #11 for protecting sensitive data is to mandate all sensitive data to be previously encrypted, de-identified, anonymized, and tokenized before getting loaded into Hadoop.

Data protection strategies vary by type and classification of data. Some data my need to be anonymized or de-identified. For example, a company policy might require that names of individual customers, their locations, company names, product names, and so on should never be openly visible. Other types of data such as credit card information should be tokenized while other information may be encrypted.

The implication of these rules in practice is to achieve a policy that never allows any sensitive data into the lake without applying data protection transformations first. Implementing a rule to never load sensitive data as raw into the data lake without first applying encryption and tokenization to the data is a step toward effective data protection.

Sample Data Governance and Management Tools

Fortunately, new data governance tools are emerging almost every month for managing big data in the Hadoop environment. There are possibly three challenges in selecting and implementing tools these days: 1) Identify and select a minimal set of tools that together provide complete coverage of your organization's data governance strategy and requirements; 2) Integrating these tools into a seamless environment; 3) Be prepared to revamp the tool set in two to three years as the pace of technology and innovation in this area are quite dynamic with new tools immediately supplanting the last tool.

Below is a sample of data governance software tools (the tools that appear with a "*" are open source):

- Collibra: Active data governance and data management
- Sentry*: Central access management tool
- Knox*: Central authentication APIs for Hadoop
- Apache Atlas*: Metadata registry and management
- Cloudera Navigator: Central security auditing
- Blue Talon: Central access and security management tool
- Centrify: Identity management
- Zaloni: Data governance and data management tool
- Dataguise: Auto-discovery, masking, encryption

- Protegrity: Data encryption, tokenization, access control
- Voltage: Data encryption, tokenization, access control
- Ranger*: Similar to Sentry but provided by HortonWorks
- Falcon*: Data management and pipeline orchestration
- Oozie*: Data pipeline orchestration and management

Data Governance in the GRC Context

One perspective into big data governance is to view it in the context of governance-risk-compliance (GRC) as shown in Figure 8.1. In this framework, big data governance is viewed as an element of IT governance and must be based on and in alignment with IT governance policies. In turn, IT governance is a sub-segment of, and must be in alignment with, the broader corporate governance policies.

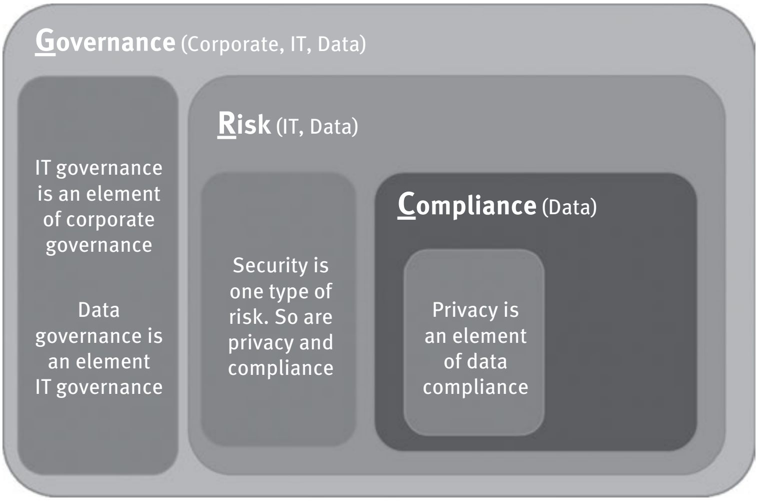

Governance (Corporate, IT, Data)
IT governance is an element of corporate governance
Data governance is an element IT governance
Risk (IT, Data)
Security is one type of risk. So are privacy and compliance
Compliance (Data)
Privacy is an element of data compliance

Figure 8.1: The GRC Perspective of Data Governance.

Whether your organization has substantial experience with corporate and IT governance or not, this book can still provide you with adequate governance strategies and tips that you can use to stand up your own big data governance framework for your organization.

The Costs of Poor Data Governance

The costs of poor or no big data governance can grow substantially and exponentially with the volume of data in your data lake. Aside from the direct and tangible costs that the organization will incur for lack of data governance, there are intangible costs as well.

Poor data governance increases the likelihood of:

- Negative impact on brand quality
- Negative impact on markets
- Negative impact on credit rating
- Negative impact on consumer trust

In summary, poor data governance raises the reputation risk for your organization and can severely tarnish your brand identity.

Big Data Governance Budget Planning

Big data governance relies on staffing, tools, and processes, all of which require an adequate and consistent budget. I raise this issue since many organizations have overlooked the need for data governance investments while planning their big data initiatives.

Some of the line items that a big data governance budget must include are:

- Resource requirements:
  - Big data governance council managers
  - Security/compliance managers
  - IT and technical staff
  - Data stewards
- Budget for tools
  - Consultant support
  - Software tools licensing
  - Software tools integration
  - Software tool maintenance budget
- Training budget
  - Annual training budget for users, data stewards, and data governance core team staff

What Are Other Companies Doing?

Many large enterprises are developing robust data governance policies and frameworks. One example is the Hadoop Security initiative by HortonWorks. This initiative includes Merck, Target, Aetna, SAS, and a few other companies. The goal of this initiative is to bring security policy and tools development to deliver a robust data governance infrastructure for businesses in three phases. This infrastructure is based on Apache Ranger, Hcatalog Metastore, and Apache Falcon.[1]

Other initiatives include IBM's collaboration with financial institutions to develop the world's first financial services-ready public cloud infrastructure. This initiative includes partnership with major US banks and covers over 300 security and privacy policies.[2]

Fortunately, additional initiatives are emerging that include companies like Teradata, Cloudera, and MapR that are forming a viable data governance ecosystem for big data governance. A map of the security components for Hadoop illustrates the need for integration and collaboration among solution vendors to ensure their solutions can easily be integrated.[3]

The National Institute of Standards and Technology (NIST) has initiated several projects and issued a number of crucial publications[4] such as NIST's Cybersecurity Framework which outlines an overarching cyber security framework for the organization. NIST has initiated its Big Data Security and Privacy Working Group to specifically address the standards specific to data management.

Hortonworks (which merged with Cloudera in 2019) and the Apache project are working on additional metadata management tools under project Atlas. I suggest keeping a watch on developments from NIST, the Atlas project, and Rhino Data Protection to stay up to date with the latest releases and functionality from this project. The link is at: http://atlas.incubator.apache.org/

[1] Neumann, S. (2014, November 13). *5 Hadoop security projects*. Xplenty. https://www.xplenty.com/blog/5-hadoop-security-projects/

[2] *IBM developing world's first financial services-ready public cloud; Bank of America joins as first collaborator*. (n.d.). IBM News Room. https://newsroom.ibm.com/2019-11-06-IBM-Developing-Worlds-First-Financial-Services-Ready-Public-Cloud-Bank-of-America-Joins-as-First-Collaborator

[3] Grasso, C. (n.d.). *Hadoop security basics (In under 5 minutes)*. Blog – Dataiku. https://blog.dataiku.com/sound-smart-on-hadoop-security

[4] *SP 1500-4r2, NIST big data interoperability framework security and privacy V3*. (n.d.). NIST Computer Security Resource Center | CSRC. https://csrc.nist.gov/publications/detail/sp/1500-4r2/final

Chapter 9
Big Data Governance Best Practices

Data Governance Best Practices

Hadoop was initially developed without security or privacy considerations. Hence, there has been a huge gap in data management tools, structure, and operations of Hadoop in the past. Without proper data governance tools and measures, the Hadoop data lake can become a data swamp.

Many best practices, tools, and methods are emerging, however. Many companies mentioned in this book are offering solutions to address these challenges and gaps. Hadoop vendors such as HortonWorks, Cloudera, and MapR have released or announced new tools. This is a very dynamic technology area.

This chapter contains some best practices in practical management of data in Hadoop. You'll find many sample policies, configuration rules, and recommendations for best governance of your Hadoop infrastructure. But this is just the tip of the iceberg since technology continuously evolves and security in and of itself is a race – an unending race with hackers.

Data Protection

Data protection is architected as shown in Figure 9.1 to be handled via two access paths: Hadoop HDFS or SQL Access. Both paths must be protected, but each path requires a different design and approach. The HDFS access path includes the physical layer, folders, and files. The SQL access path is concerned with the logical layer and database constructs and tools such as Hive or Impala.

This architecture (Figure 9.1) organizes data classification into several statuses and types. In the Hadoop sandbox, I've shown four data statuses: raw, keyed, validated, and refined. As shown in this architecture diagram, both data paths, HDFS and SQL, utilize the same Hadoop Sandbox. When data is on-boarded (or imported) into the sandbox, it's initially raw data. Then it is transformed into keyed data (in this particular example using a Hive table and format, but it can be done with a variety of other data query and formats). Then the data is validated for quality. Once validated, the data becomes trusted. Finally, data might undergo further transformation and filtering to reach the refined status which is the status that indicates the data is ready for analysis.

At the highest level, data protection involves three categories of considerations: sensitive data protection, data sharing considerations, and adherence to governance policies.

https://doi.org/10.1515/9783110664065-009

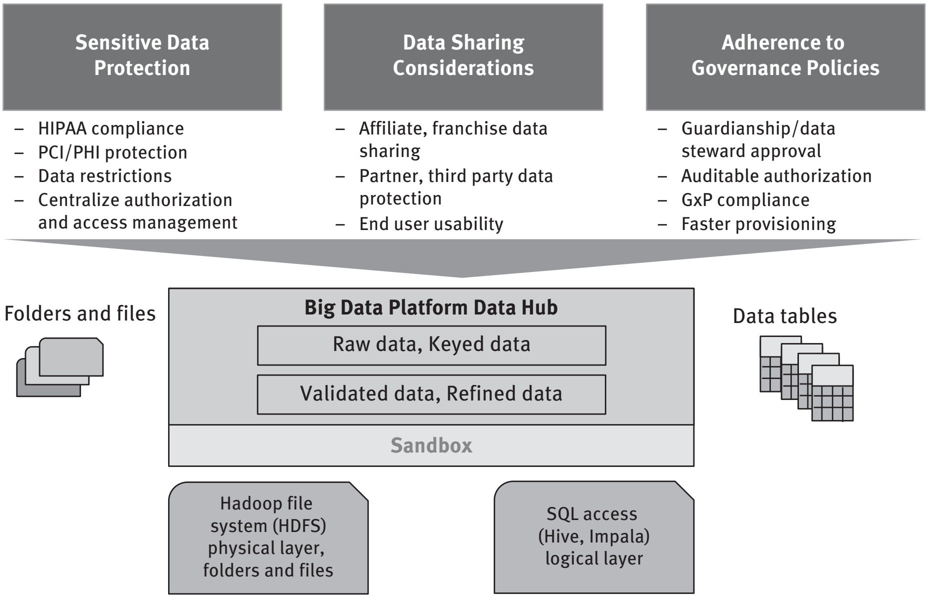
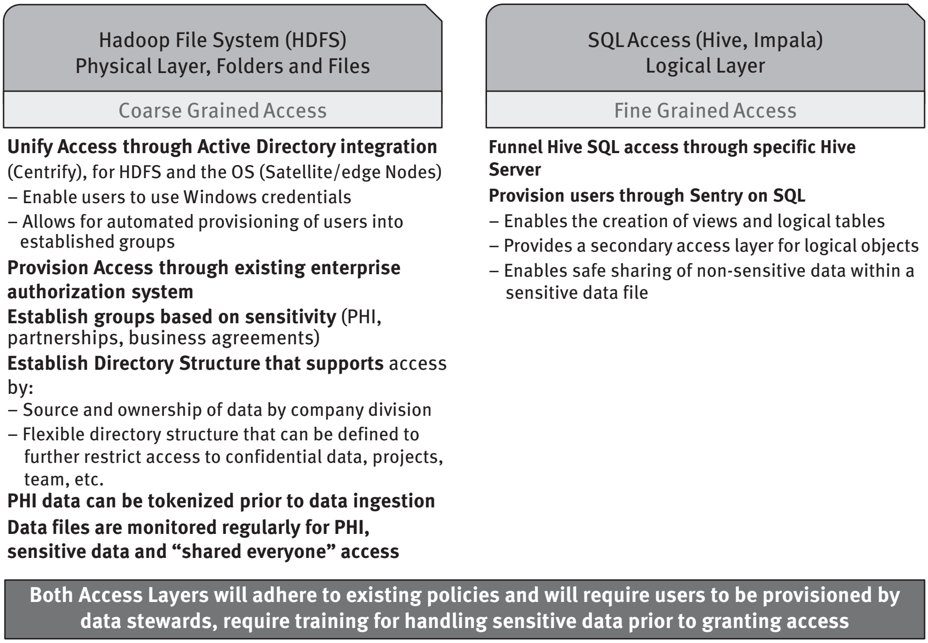

Sensitive Data Protection
– HIPAA compliance
– PCI/PHI protection
– Data restrictions
– Centralize authorization and access management
Data Sharing Considerations
– Affiliate, franchise data sharing
– Partner, third party data protection
– End user usability
Adherence to Governance Policies
– Guardianship/data steward approval
– Auditable authorization
– GxP compliance
– Faster provisioning
Folders and files
Big Data Platform Data Hub
Raw data, Keyed data
Validated data, Refined data
Sandbox
Data tables
Hadoop file system (HDFS) physical layer, folders and files
SQL access (Hive, Impala) logical layer

Figure 9.1: Privacy and Security Architecture Considerations for Hadoop Environment.

Sensitive Data Protection

Sensitive data protection is concerned with policies that ensure your operations are meeting regulatory and compliance requirements for sensitive data. This is predicated on another type of data classification activity: classifying your data into critical, high-priority, medium priority, and low priority. All critical and high priority data are regarded as sensitive. This type of data includes personally identifiable information (PII) and personal health information (PHI) and credit card–payment card industry (PCI) protected information. Some organizations refer to such sensitive information as private information (PI) and non-private information (NPI) that does not contain personally identifiable or sensitive information. In Figures 9.1 and 9.2, I'll refer to various types of sensitive data simply as PHI to represent all sensitive types of data including PII and PCI protected information.

Protecting sensitive data requires implementing centralized authorization and access management, data usage policies, and restrictions on data usage and protection of PHI data.



Hadoop File System (HDFS)
Physical Layer, Folders and Files

Coarse Grained Access

Unify Access through Active Directory integration (Centrify), for HDFS and the OS (Satellite/edge Nodes)
- Enable users to use Windows credentials
- Allows for automated provisioning of users into established groups

Provision Access through existing enterprise authorization system

Establish groups based on sensitivity (PHI, partnerships, business agreements)

Establish Directory Structure that supports access by:
- Source and ownership of data by company division
- Flexible directory structure that can be defined to further restrict access to confidential data, projects, team, etc.

PHI data can be tokenized prior to data ingestion

Data files are monitored regularly for PHI, sensitive data and "shared everyone" access

SQL Access (Hive, Impala)
Logical Layer

Fine Grained Access

Funnel Hive SQL access through specific Hive Server

Provision users through Sentry on SQL
- Enables the creation of views and logical tables
- Provides a secondary access layer for logical objects
- Enables safe sharing of non-sensitive data within a sensitive data file

Both Access Layers will adhere to existing policies and will require users to be provisioned by data stewards, require training for handling sensitive data prior to granting access

Figure 9.2: Managing Security and Privacy on Physical and Logical Layers.

Data Sharing Considerations

Data sharing is a critical governance consideration – in particular in large organizations, multi-division, and global companies whose data generation and consumption spans not only functional lines but countries, divisions, and affiliates such as corporate partners and franchises.

Much of big data comes from external sources and by contracting with third parties (and sometimes fourth-party data providers). These purchases come with usage rules (i.e., do's and don'ts clauses). Data policies related to proper safeguards, access, and usage of third-party data must be implemented to ensure that we're compliant with the terms of the third-party data purchasing agreement.

Data usage agreements (DUA) are policies that define the proper use and access to data. On the other hand, usability of data and making data easily consumable by end users is important.

All access paths (in this example, the SQL and the HDFS access paths) require that users adhere to existing policies and that users are to be provisioned by data stewards, and plan training for handling sensitive data prior to granting access to such data.

Adherence to Governance Policies

The flip side of enforcing governance policies is adherence. Adherence increases with the education and training of all users in the organization. Access to the data lake must be approved by data stewards who are accountable for enforcing data governance policies for their respective group, division, or affiliates.

All authentication tools must support logs for audits. Data stewards are responsible to ensure new users receive proper access and for disabling access for employees who have left the organization.

Certain industries dictate additional standards and compliance requirements regarding data. One of these standards is Good Manufacturing Practices (GMP) and, for that matter, all other practices typically known as GxP (x is for any activity or process) for the organization.

Finally, it is crucial that data stewards and data scientists be empowered to quickly provision storage and data space in their sandbox in the lake.

Data Protection at the High Level

The Hadoop file system, including the data lake, represents data sets at the file level (coarse data) and coarse grained access to data. Best practices recommend that we *unify access* through Active Directory integration. There are tools such as Centrify for HDFS and the operating system security measures (for satellite/edge nodes in the Hadoop cluster). The goal of this approach is to enable users to use Windows credentials for their access to the lake. This approach allows for automated provisioning of users into established groups.

You can provision access through an existing enterprise authorization system such as Microsoft Active Directory. It's recommended that you establish user groups based on data sensitivity (PHI, partnerships, clinical trials, non-PHI, data scientist, etc.) Best practices recommend that you establish *directory structures* that support access by the following classifications of users and usage, role-based access:

- Source and ownership of data by directory and grouped by division, country, affiliate, and similar organizational boundary
- A flexible directory structure that can be defined to further restrict access to confidential data, projects, team, etc.

To protect sensitive data (PHI, PII, PCI) at the coarse grained level, you can and should apply tokenization prior to data ingestion into the lake. Data tokenization is an important strategy to protect data. When you tokenize data, it's advisable to store the token keys on the token servers on your premise in a different subnet, in particular if the data resides in the cloud. In the event that hackers gain access to the data on the cloud, that information is useless to them without the token keys.

Data Tokenization

One best practice is to *double tokenize* the data. Tokenizing sensitive data such as credit card numbers, card expiration dates and security codes, Social Security numbers, and other sensitive information is increasingly a popular technique to protect data from hackers. A financial company might choose to tokenize the first 12 digits of a 16-digit credit card number leaving the last 4 digits available for call center functions such as verification and authentication of customers.

Tokenization replaces the digits with a randomly generated alphanumeric text ID (known as a "token") which cannot be identified without de-tokenization using the initial key that was generated in the token server. A double tokenization essentially tokenizes the token once more using a separate token server. This is an additional safeguard against hackers. In the event that a hacker might gain access to one token server, and even if the hacker could de-tokenize the data once, the ability to de-tokenize a second time is impossible without access to the other token server.

In high level governance, data stewards are vigilant and actively monitor data files regularly for PHI, sensitive data, and data labeled "shared everyone" access.

Low Level Data Access Protection

Representing the fine grained access, this access control involves SQL Access, the logical layer of data management. To ensure proper access and security, you must funnel Hive SQL access through a specific Hive server that is dedicated to control access. Sentry is a tool often used in securing data on SQL. You can provision users through Sentry for SQL query access to data. The advantages of this configuration are that it:

- Enables the creation of views and logical tables
- Provides a secondary access layer for logical objects
- Enables safe sharing of nonsensitive data within a sensitive data file

Security Architecture for the Data Lake

The model diagrams in Figures 9.3 and 9.4 depict a modular design where the functional data security and protection are separated into layers and modules to ensure all aspects of data security are considered. The key aspects of this security design are:

- *Authentication*: Enabled by standards such as Kerberos or similar solutions
- *Authorization*: Enabled by Microsoft Active Directory (AD) and Centrify, or comparable products

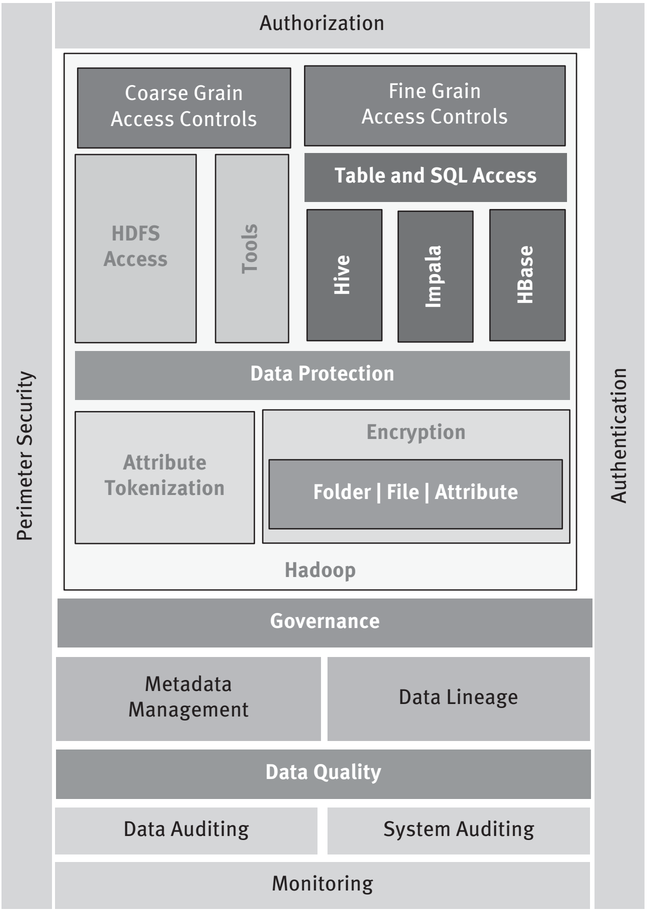


Figure 9.3: A Model Hadoop Security and Privacy Management Stack.

- *Perimeter Security*: Can be provided by Secure Socket Layer (SSL) and Knox Perimeter Security, or a similar solution
- *Monitoring*: Can be enabled by products such as Blue Talon

Within the architecture, the building blocks for governance, data quality management, data protection, and both fine grained and coarse grained access control are shown. The next diagram (Figure 9.4) matches the same functions with specific tools or standards that accomplish that function:

- The quality management requires data auditing and system auditing. These are possible through products such as Drools/Navigator and McAfee or other solutions.
- Governance management can be enabled by solutions such as Zaloni.

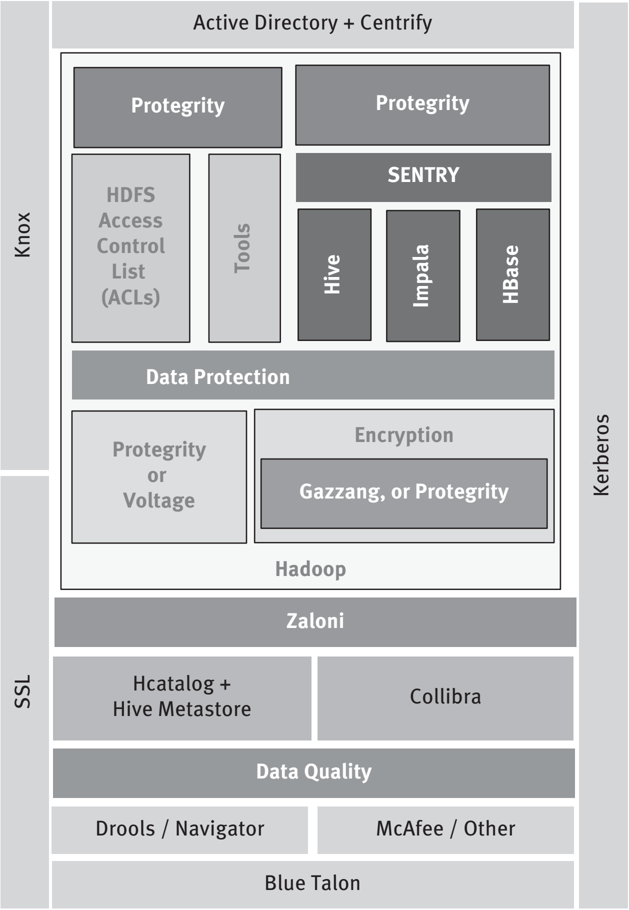


Figure 9.4: A Hadoop Security and Privacy Model Enabled by Tools.

- Data protection includes data encryption and tokenization of the data in the lake. Some of the solutions for tokenization and encryption include Voltage and Protegrity. They apply data protection at the folder, file, and attribute level. Solutions such as Gazzang, Protegrity, and Voltage represent some of the options.
- Metadata management can be implemented using HCatalog and Hive Metastore. While these are somewhat nascent tools, they're becoming more functional and powerful solutions.
- In order to manage data lineage, you may implement solutions such as Collibra.
- Coarse grained and fine grained data protection can be enabled by solutions such as Voltage or Protegrity.

- HDFS Access control can be implemented using Access Control Lists (ACL).
- Access to Hive, Impala, Hbase, and other fine grained database controls can be enabled by a product like Sentry.

Big Data Classification

Data classification promotes better access control, data quality management, and usability features of your big data repository. Here are the four data classifications again, explained in more detail. Within the source formatted data, there are three stages of data formats:

Raw Data

The raw data has no transformation performed on it. It's the data that is on-boarded from the source outside of the lake. All sensitive data is to be encrypted or tokenized before it lands in the lake. Access to raw data is restricted only to privileged users such as data engineers.

There are no quality checks performed on raw data. In order to maintain a catalog of data stored in the data lake, we must *register* data being loaded into the data lake. This registering is done by keeping records of what data is being loaded in the data lake. The raw data represents source files formatted as-is; it still has the source directory structure. Raw data must be registered with metadata even with minimal information about the source, owner, format, and type of data. Any data that is registered in the metadata registry (also referred to as the metadata hub), can stay in Hadoop. There are several software packages available for registering data including Waterline and Collibra.

Keyed Data

Keyed data represents data that has been imported into some data structure such as HBase or Hive or similar data format. Access to keyed data is also restricted to privileged users such as data engineers.

Keyed data is transformed into standardized file formats and compression. It maintains data keys and schema are applied. At this stage, data quality checks are applied. Reports from quality checks are generated, and any defects are escalated to data stewards and even up to the data governance council.

Keyed data is not cleansed and no filtering is applied to it. Keyed data can stay in Hadoop but still maintains its source directory structure. One important fact to note is that keyed data is derived from raw and is a physically different file from the raw file.

Validated Data

Validated data represents the class of data has been through data quality checks and has been fully registered in the metadata registry, basic cleansing and data fixes have been applied, data protection rules are applied, and data quality rules are applied.

In this stage, files may be duplicated across the organization for other divisions, partners and affiliates, and subsidiaries or sharing to other divisions in other countries as long as it is allowed by data jurisdiction policies.

Certain data masking may apply to this data based on data usage policies, but *validated* data may be shipped out of Hadoop. This data is derived from keyed or raw and is a physically different file from its prior stages. Once data reaches this level, the keyed/raw files are deleted. At the validated stage, broader user access is allowed to validated data.

Refined Data

When data can be merged with other data, filtered, and parsed it reaches the *refined* data stage. In this stage, operational processing is performed, analysis is applied, and reports are generated. Users have access to this data for their analytical work. Refined data is typically created from validated data as users apply structure to the data. They may parse and merge files and records, add attributes, and ship the data out of Hadoop.

However, additional data quality checks may be applied or required at this stage. Refined data is derived from keyed or validated and is physically a different file. Figure 9.5 shows the four stages of data in Hadoop again for reference.

Hadoop Lake Data Classification and Governance Policies

Different data governance policies apply to different stages and categories of data. Figure 9.6 shows how various access and usage controls apply to data classes differently. Let’s look at the rules and usage policies for these data classifications in more detail.

Raw Data Usage Rules

Sensitive raw data must be encrypted and/or tokenized before arrival into the lake. Only few, privileged users may access this data; thus the goal is to have very limited access to this data. *Privacy suppression* is needed to access and transform data in this stage, but privacy suppression is granted to a very small group of data engineers or data scientists. All data on-boarded into the Hadoop data lake must be registered in the metadata hub – including data lineage, schema, attribute metadata, and owner

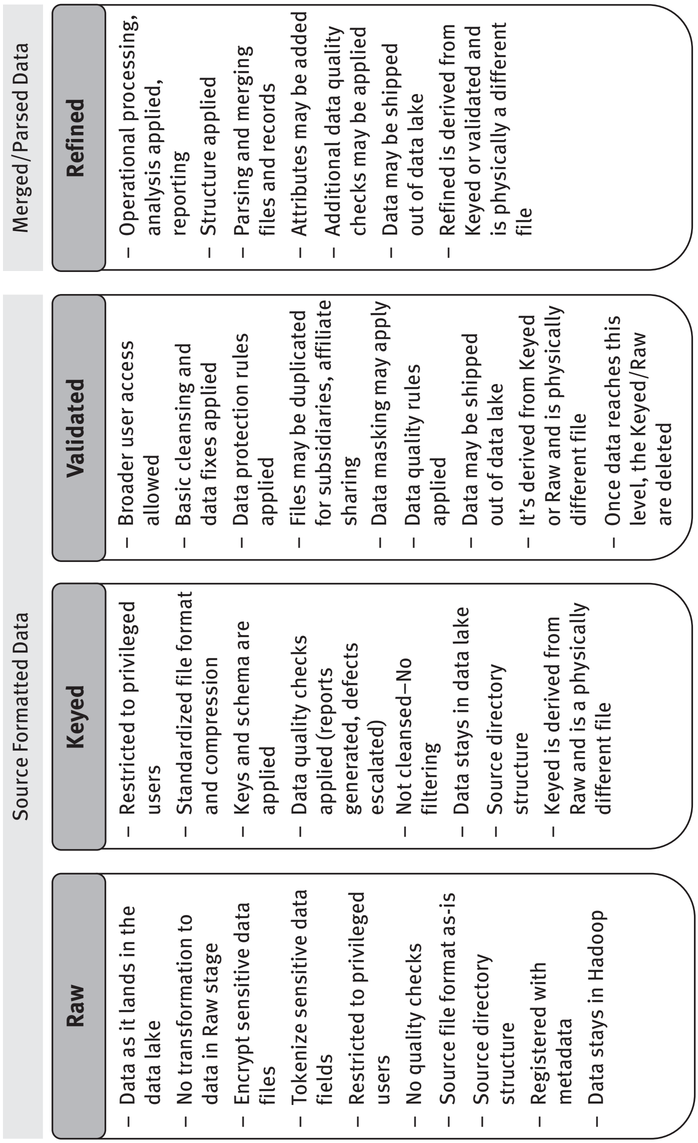
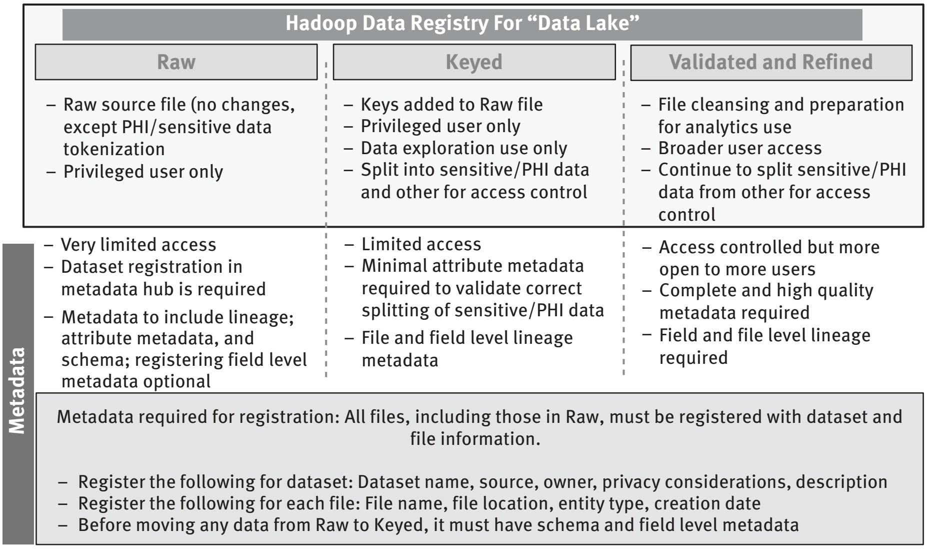

Source Formatted Data
Merged/Parsed Data
Raw
– Data as it lands in the data lake
– No transformation to data in Raw stage
– Encrypt sensitive data files
– Tokenize sensitive data fields
– Restricted to privileged users
– No quality checks
– Source file format as-is
– Source directory structure
– Registered with metadata
– Data stays in Hadoop
Keyed
– Restricted to privileged users
– Standardized file format and compression
– Keys and schema are applied
– Data quality checks applied (reports generated, defects escalated)
– Not cleansed–No filtering
– Data stays in data lake
– Source directory structure
– Keyed is derived from Raw and is a physically different file
Validated
– Broader user access allowed
– Basic cleansing and data fixes applied
– Data protection rules applied
– Files may be duplicated for subsidiaries, affiliate sharing
– Data masking may apply
– Data quality rules applied
– Data may be shipped out of data lake
– It's derived from Keyed or Raw and is physically different file
– Once data reaches this level, the Keyed/Raw are deleted
Refined
– Operational processing, analysis applied, reporting
– Structure applied
– Parsing and merging files and records
– Attributes may be added
– Additional data quality checks may be applied
– Data may be shipped out of data lake
– Refined is derived from Keyed or validated and is physically a different file

Figure 9.5: The Four Stages of Data in Hadoop Data Lake.



Hadoop Data Registry For “Data Lake”

	Raw	Keyed	Validated and Refined
	– Raw source file (no changes, except PHI/sensitive data tokenization – Privileged user only	– Keys added to Raw file – Privileged user only – Data exploration use only – Split into sensitive/PHI data and other for access control	– File cleansing and preparation for analytics use – Broader user access – Continue to split sensitive/PHI data from other for access control
Metadata	– Very limited access – Dataset registration in metadata hub is required – Metadata to include lineage; attribute metadata, and schema; registering field level metadata optional	– Limited access – Minimal attribute metadata required to validate correct splitting of sensitive/PHI data – File and field level lineage metadata	– Access controlled but more open to more users – Complete and high quality metadata required – Field and file level lineage required

Metadata required for registration: All files, including those in Raw, must be registered with dataset and file information.

- Register the following for dataset: Dataset name, source, owner, privacy considerations, description
- Register the following for each file: File name, file location, entity type, creation date
- Before moving any data from Raw to Keyed, it must have schema and field level metadata

Figure 9.6: Metadata Rules Matched to Data Stages in the Lake.

(business division or line of business) of the data. But detailed field level metadata is not always necessary, so you may decide to make it optional.

Keyed Data Usage Rules

Keyed data will also have limited user access available to a few privileged users such as data engineers or IT personnel who manage the data lake. In this stage, keys are added to the data. The application and use cases of this data are for exploration use only. Any sensitive data is split into a separate segment and folder in the lake. *Data splitting* is a key method of keeping sensitive data (PHI, PII, PCI) in a separate container and being able to apply specific access policies to that folder.

Registering this data into a metadata hub is required by including attribute metadata and lineage information to the file and field level in metadata.

Validated and Refined Data Usage Rules

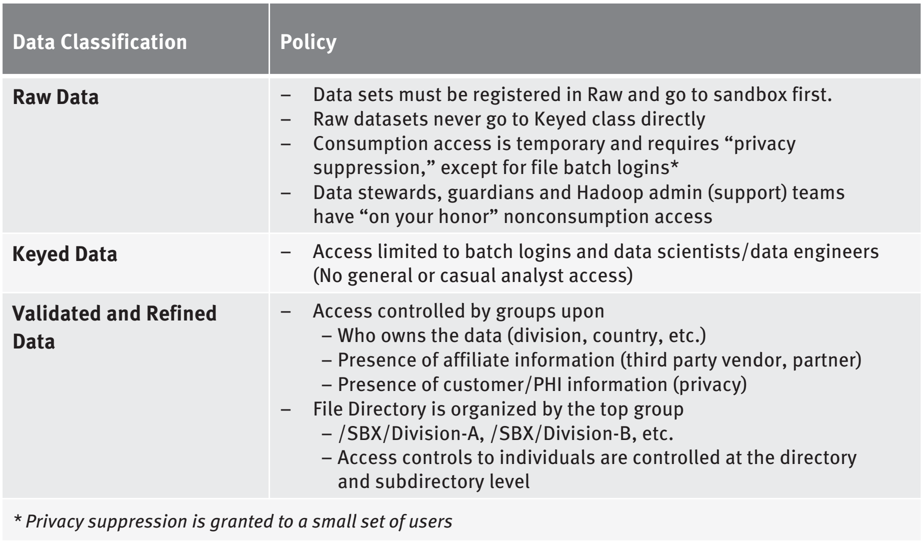
Data functions in both validated and refined stages vary from file cleansing, curation, and preparation of data for analysis to transformations, merging with other data, filtering, and extracting sub-sets of the data sets. This class of data is defined to allow broader user access, but data stewards must be vigilant to continue splitting data into folders for sensitive data to be segregated from the rest. In addition, there can be other reasons for splitting data for specific uses and applications.

Access control to this stage of data is more open to users but still controlled by the designated data stewards for each group in the organization. Keeping metadata for this data at high quality and completeness are important and are expected from users. The metadata requirements for lineage at the coarse level (folder and files) and granular levels (data tables and fields) are required.

Metadata Rules

Some attributes required for metadata registration span *all* files, including those in raw form, which must be registered with data set and file information. Here is a list of some possible attributes and recommendations as illustrated in Figures 9.6 and 9.7.

- Register the following for *dataset*: Data set name, source, owner, privacy considerations, description
- Register the following for *file*: File name, file location, entity type, creation date
- Before moving any data from raw to keyed, it must have schema and field level metadata

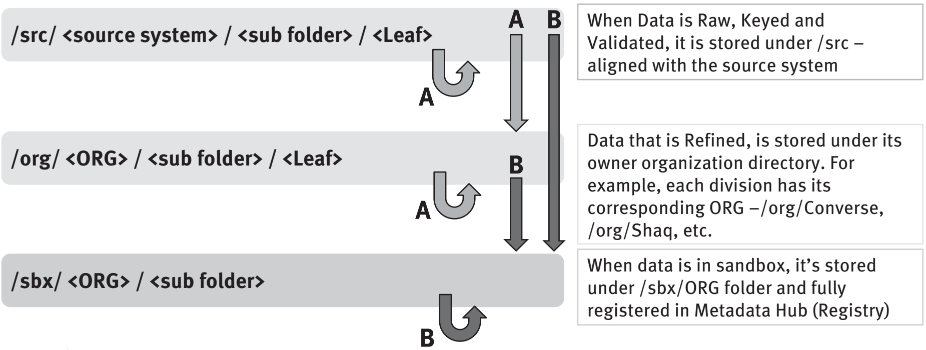


Data Classification	Policy
Raw Data	– Data sets must be registered in Raw and go to sandbox first. – Raw datasets never go to Keyed class directly – Consumption access is temporary and requires "privacy suppression," except for file batch logins* – Data stewards, guardians and Hadoop admin (support) teams have "on your honor" nonconsumption access
Keyed Data	– Access limited to batch logins and data scientists/data engineers (No general or casual analyst access)
Validated and Refined Data	– Access controlled by groups upon – Who owns the data (division, country, etc.) – Presence of affiliate information (third party vendor, partner) – Presence of customer/PHI information (privacy) – File Directory is organized by the top group – /SBX/Division-A, /SBX/Division-B, etc. – Access controls to individuals are controlled at the directory and subdirectory level
** Privacy suppression is granted to a small set of users*	

Figure 9.7: Data Management and Configuration Rules Matched to Data Stages in the Lake.

Data Structure Design

Some of the best practices already discussed include splitting data and storing sensitive data in a separate folder and directory path. Figure 9.8 shows a sample data lake file storage structure.


/src/ <source system> / <sub folder> / <Leaf>
A
B
A
/org/ <ORG> / <sub folder> / <Leaf>
B
A
/sbx/ <ORG> / <sub folder>
B

When Data is Raw, Keyed and Validated, it is stored under /src – aligned with the source system

Data that is Refined, is stored under its owner organization directory. For example, each division has its corresponding ORG –/org/Converse, /org/Shaq, etc.

When data is in sandbox, it's stored under /sbx/ORG folder and fully registered in Metadata Hub (Registry)

Examples:
Source systems: SAS, Oracle, etc.
Classification: Raw, Keyed, Validated, Refined
ORG: Converce, Cole_Haan, Shaq
Sub-folders by group or products: Air_Jordan, Tennis, Soccer, Running (Optional method to create add'l sub directories for more granular access control)
Leaf folder: Directory for a dataset registered in the Metadata Hub (MDH), also Hive Table Folder

Figure 9.8: Data Directory Structure in Hadoop Data Lake.

It's important to note how the folder path at the root is designated and how they branch into leaves that make the path and folders more manageable from an access and privilege perspective. Note that you don't want to build the directory path by data classification. Your directory design for raw data will vary from the refined data.

When data is in raw, keyed, or validated stages, your directory structure reflects the same structure as the source of the data. When data is refined and set up for broader use, your directory structure should be set up by the organization that owns (and uses) that data.

Let's assume you're looking to design a file structure for a large global retail company, say Nike. Nike is the largest athletic footwear and apparel company in the world with sales over $20B and over 30,000 employees worldwide. The company has affiliated brands like Shaq, Hurley, Converse, One Star, Cole Haan, and others.

Let's assume the company has some pre-existing data stored in an SAS server, prior data on an Oracle database, Teradata data, and data at several subsidiaries, divisions, and franchises. The company has three key product categories: let's call them footwear, apparel, and equipment. Footwear products include running, soccer, tennis, golf shoes, etc.

SAS, Oracle, and Teradata are regarded as sources of data in this architecture. However, organizations (ORGs) are the company's subsidiaries and affiliates like Converse, Cole Haan, etc. The product names form the sub-folders in the directory design.

In Figure 9.8, the data movement paths labeled A are production processes, while the data movements labeled B are user data movement processes.

For example, a *refined file folder path* might capture the division as well as the product category as its path. A data file in the refined stage related to tennis shoes affiliated with Converse might have a directory path as: /sbx/Converse/Tennis/file.

The key point from this design to maintain is that data classification is recorded in the metadata registry, not in the directories. This approach allows directory design to be simpler and more consistent across the enterprise with better management functionality for access control.

Sandbox Functionality Overview

The sandbox folder design would follow the <ORG>/<Line of Business>/<Product> path as illustrated in Figure 9.8. Users have read/write access to the sandbox designed for and assigned to their organization. Users may read data from the Hadoop lake and bring it over to the sandbox. This is the common practice in data lake processes. If you are using Active Directory for authentication/authorization, then it's recommended that you implement Active Directory (AD) nesting to simplify access requests for Hadoop.

Detailed Access Policies

Specific access policies can be defined for the sandbox and directory structure. For example, user access to a specific sandbox should be approved by the data steward or data risk officer of each organization (a data risk officer is a VP or higher executive, appointed by a division, country, or similar level line of business to represent that organization on the data governance council).

The data governance best practice policies dictate that there should never by any reporting, analysis or publication directly from raw data. There should be no marketing activities, analysis, or reports generated directly from the raw data.

To gain access to data in a sandbox, privacy rules training is required for users. If the users will access data from another country of jurisdiction, training designed for such countries must be required for users.

Data stewards must on a regular basis (quarterly, semi-annually, or annually) review user training competency and renew access for users. To keep user access consistent, it's common practice and common sense that user access privileges to data lake data sets are identical to their pre-existing data access level in other data stores in the organization.

Top Level Sandbox Protection

The top level sandbox directory is protected by a sandbox specific group such as:

phdp_<resource>_sbx

Example: The /sbx/Converse directory has the phdp_Converse_sbx group assigned. A sandbox group ensures only authorized users have access to the sandbox. To manage sandbox owner-only access, consider the following configuration techniques:

The HDFS Umask value ensures that files are created with “owner only” privileges. Each access control (user permission) to operate HDFS, such as read data, create folder, etc., is defined by a code, i.e., a Umask number. Here are some examples of permission codes:

– Use a default of 600 code for permission on these files as only the file owner has access to the data by default.
– In order to share data with others, the user must change permission to 640 and assign the appropriate group.
– Permission for folder creation has a value of 755.
– Users are expected to be assigned AD Group policy according to the sensitivity of data, similar to the Hadoop lake policy.

A policy that you may want to apply to data in the lake is to require that users must register data in their sandbox within 90 days after the data is arrived or created.

Managing Select Access to Hive Tables

In order to limit SELECT access to Hive tables, users must have READ access on the underlying data in HDFS. Hive databases are to be designed for respective lines of business (division, country, etc.) or by subject areas (products, factories, etc.). Active Directory (AD) group membership is required to get access to the database. For casual users, you can provide READ-only access to tables in the data lake database.

As part of the *data usage agreement*, you may allow users to create or copy Hive tables from the Hadoop lake into their sandbox. Users may run Hive queries using the beeline command line utility or the Hue GUI. However, at the time of this writing, the Hive command line has security vulnerabilities and should not be used.

Split Data Design

Many users ask: “Why data split in Hadoop?” Part of the reason is that there is no concept of “views” in Hadoop to limit access by data type, as we’ve traditionally enjoyed in conventional RDBMS systems.

The open source edition of Hadoop offers data protection at the file level (coarse grained) which may be inadequate for many applications, and inadequate for implementing an enterprise-wide well-governed data lake to support Hadoop and varieties of access mechanisms.

Best practices require files to be split physically in Hadoop to separate PHI/sensitive data from non-PHI/nonsensitive data.

Another reason for physical splitting of sensitive data is that certain legal, regulatory, third-party contractual data and GxP (Good x Practices) requirements or compliance policies apply to sensitive data which therefore must be stored separately.

One of the key data governance designs is to separate data schema and access control to sensitive data. It's recommended that you create a sub-folder in each directory to contain the sensitive split data. Proper data governance policies dictate that strict access controls to this folder are applied by the data lake administrators.

In Chapter 10, I'll introduce a document that can be a starting point and baseline document for your big data governance program.

Chapter 10
Big Data Governance Framework Program

This chapter outlines a framework for implementing data governance for big data. It is intended to be used as a template that you can use as a baseline for your data governance program and modify to suit your business environment. You might find some of the tables or content repeating from prior chapters. But the intent is to keep the document as concise as possible so it can be used as an internal document for your program.

Overview

Data governance refers to the rules, structures, processes, and practices that allocate the roles, responsibilities, and rights of participants in big data analytics. This governance policy is intended to meet all relevant compliance with applicable laws and objectives of operating the big data platform in a safe and sound manner.

This governance framework is a federated model with distributed roles and responsibilities that encompass the big data management framework across franchises, countries, and wholly owned subsidiaries of an organization. The framework is designed to protect the organization's big data assets effectively at the lowest overhead cost. Its policies are intended to conform to the organization's corporate governance policies.

The framework consists of eight capabilities (each discussed in detail later in this chapter):

I. Organization
II. Metadata management standards
III. Data classification
IV. Security, privacy, and compliance standards
V. Data usage agreement (DUA)
VI. Security operations and policies
VII. Information lifecycle management
VIII. Data quality standards

Enterprise data management is responsible for establishing requirements, guidelines, and procedures for proper management and handling of data across the enterprise. The governance framework abides by the enterprise data management guidelines. Since in the Hadoop environment the data and applications are tightly coupled, this framework will include application management policies in addition to data management.

https://doi.org/10.1515/9783110664065-010

The primary governing body of this policy is the data council which consists of executive management from cross-functional organizations (franchises, countries, subsidiaries, affiliates) who are responsible for ensuring your data is reliable, accurate, trustworthy, safe, and protected. The data council makes recommendations to executive management with respect to data management matters.

Integrity of Information, Data Resources, and Assets

Information and data resources are the organization's assets that are used by the organization's personnel to execute critical processes that deliver on its strategies. Because of the risks and costs of using or providing the wrong information, poor-quality data, or a security breach, management must deliberately control the quality and access to data and applications.

Well-designed data and application management controls must enable cost-efficient self-identification and self-correction of data and information risks, such as using inadequate data, or selecting the wrong data for a defined purpose.

This data governance framework and its policies are intended as supporting standards and controls that consistently prevent these breakdowns from happening or provide alerts of breakdowns, plus they identify and remedy control exceptions.

To plan and operate the big data analytics platform and to enable its compliance with laws and regulations of the organization, executive management must identify and control data and information and their usage in a manner that is commensurate with the direction of the enterprise risk management policy. To manage the risks, data and applications must include controls that meet the requirements of this policy and its supporting standards.

The policies and standards in this framework list the requirements that need to be met when designing and operating controls for big data and applications.

An adequately controlled information systems environment includes the following:

1. A method for consistently identifying and defining the information that is important for operating the big data analytics platform.
2. Clear accountability and authority for managing information resources so that their quality is adequate for their intended uses at an appropriate cost to the business.
3. Required and auditable procedures for delivering and publishing information, such as analysis results.
4. Clear lines of responsibility within management regarding what it takes to meet regulatory and legal expectations for data and applications.
5. Policies that ensure data integrity by prioritizing, defining, and documenting requirements.

6. Clear authority to make decisions about data management including a process for applying cost effective controls so that those decisions are acted on, and the ability to demonstrate that the controls are effective.
7. Maintenance of an up-to-date and complete inventory of important data sources for the big data platform and business uses plus data stewardship principles that result in adequate data integrity for the organization.
8. Ongoing monitoring of standards and procedures for managing the big data platform and application assets.

Benefits of Compliance and Risks of Noncompliance

Risks associated with noncompliance with data governance policy must be assessed, recorded, and reported to the data council on a regular basis. Complying with the data governance policy has the following benefits:

- A decreased probability of impact to customers, patients, and business partners and operational loss events.
- An ability to acquire, share, and improve data and information with high efficiency.
- An ability to meet business demands quickly through re-use of accurate, complete, and functioning data resources.
- An ability to reliably improve critical clinical and business processes and leverage from superior information.
- The strategic advantage of better information sources and efficient compliance with regulations.

Failure to comply poses the following risks:

- An increased probability of an unplanned loss or customer impact due to poor quality or wrong use of data or information.
- Waste in critical operations where data is poorly defined, badly documented, or inaccurate. Costly analytical results that lead to inconsistent and misleading information.
- Failed strategic and product objectives due to bad information or use of the wrong information.
- Reduced good will and shareholder value through misrepresented or inaccurate public information.
- Loss of public trust and a tarnished brand in the event of data breach or loss of data.
- Unsatisfactory audit findings, regulatory judgments, or legal actions that pose direct costs.

Definitions: Terms that Your Program Must Define

Accountable Executive (AE): A VP or higher level individual who is designated as the executive responsible for execution and compliance with the governance policies for their franchise.**Date Usage:** A model, report, or analytics process.

Data Council: The primary executive forum for the organization's divisional (franchise/country/subsidiaries) data risk officers. They are responsible for ensuring the company's data is reliable, accurate, trustworthy, properly safeguarded, and used. They communicate, prioritize, assess, and make recommendations to the executive management with respect to data governance and risk management issues.

Data Element: Consists of one or more data sets stored and used in a big data platform for analysis.

Data Levels: Classifications are used to categorize the importance and the regulatory significance of a data element. There are four levels: critical, high priority, medium priority, and low priority.

Data Quality Issue: A data issue is defined as a gap in compliance to the data quality standard, failure in the data usage agreement, or deviation from data governance policy with regard to privacy, security, or audits.

Data Risk Officer (DRO): The executive designated at the division or business unit level who is accountable for ensuring that critical or high priority data is identified, monitored, and kept in compliance with the requirements of the governance policies. The DRO may delegate tasks related to governance standards to data stewards or subject matter experts but the DRO maintains overall accountability for the effective management of such data within their area of responsibility.

Data Steward (DS): Often a manager or higher level employee designated by each AE for each franchise to ensure governance policies are applied, including policies on security, privacy, data quality monitoring, reporting, and certification of compliance. A franchise may have one or multiple data stewards.

Data Usage Agreement (DUA): A contract between big data platform management and the accountable executive (AE) defining the data quality requirements that must be adhered to for the critical or high priority data being consumed.

Subject Matter Experts (SME): Subject matter experts bring their knowledge and expertise in security, privacy, data management, compliance, and technology to the enterprise. SMEs are designated by data stewards in each franchise to operate the data analytics in a manner that is compliant with the data governance framework.

Third-Party Data Provider: A company or individual providing data to the organization to be used for its analytical use. A *fourth-party data provider* is a company or individual providing data to the third party.

Trusted Data: Data elements that meet the data governance and data quality standards defined in this governance framework.

I Organization

Big data governance is federated and organized by the data council. The data council meets regularly, typically monthly, to address changes in policy, manage data issues, audit results and issues related to compliance, security, and privacy.

The data council advises, regulates, and establishes guidelines and policies in the management of big data for the organization and its subsidiaries and affiliates. The data council serves as a forum for discussing and sharing information, escalating key issues, coordinating, and ensuring accountability. Its charter is dedicated to providing consistent well-managed leadership on data quality, governance, and risk for the enterprise. There are four data governance forums that are managed by the data council:

- Metadata management
- Quality management
- Data governance
- Security/privacy and compliance

Each forum will have a designated leader appointed by the data council as listed in Table 10.1.

Table 10.1: Data Governance Forums and Committees.

Role	Metadata Forum	Data Quality Forum	Data Governance Forum	Security/Privacy and Compliance Forum
Leader	Leader Name	Leader Name	Leader Name	Leader Name
Scope	Develop and lead creation of metadata capabilities, tools, and related topics	Develop and lead standards of data quality capabilities, tools, and related topics	Establish data governance capabilities, tools, and related topics	Develop and lead data security, privacy, and compliance capabilities, tools, and related topics
Materials and Tools	Metadata hub (MDH) tool and data registration procedures	Data quality monitoring, testing, and issue remediation procedures	Data governance policies and procedures	Security, privacy and compliance policies, processes, and procedures

The data council members who comprise the decision-making body of the group consist of the IT leadership, big data platform management leaders, and accountable executives from each line of business (franchise, country, subsidiary, affiliate) who participates in using the big data platform.

The federated roles in a data council include the accountable executives, data risk officers, and data stewards. These roles are not full-time positions, they are responsibilities assigned to individuals with vested interest in their use of the big data platform.

The accountable executives (AEs) are VPs or higher level managers who are responsible for the overall data governance for their area of business.

The data risk officers (DROs) are appointed by the accountable executives. They are director+ level managers who ensure data governance policies are followed for their area of responsibility.

The DROs appoint data stewards from their organizations. Data stewards own the responsibility to ensure the users in their areas follow the data governance standards and policies for data quality, data integrity, security, privacy, and regulatory compliance.

Figure 10.1 depicts the data council structure.

II Metadata Management

An inventory of data usage will be maintained in the platform's metadata hub (MDH). The data usage inventory will be reviewed annually for any updates and changes by each DRO for their area of responsibility.

Required metadata information about data elements includes:

- Data asset name (name of the database, data set, spreadsheet, or any other trusted source of data)
- Data Asset/File Type (e.g., flat file, Oracle, SAS, DB2, CSV, XML, JSON, Hive, XLS, etc.)
- Table/File Name (name of physical table/data set, spreadsheet tab where the data field/column is located)
- Column/Field Name
- Data Type/Length (e.g., number, char, varchar, date, and length of the column/field – Numeric (8), Char(25), Varchar (20), Date)
- Name of the analysis (model), report, or process
- Data owner (the business entity, franchise, or country that owns the data)
- Description of data usage (stakeholders, consumers of reports, data scientists using the data)
- Description of AE, DRO, and data steward for the data element
- Criticality of the data (critical, high-, medium-, low priority)
- Frequency of review/refresh (frequency of report creation or analysis)
- Process document location (the physical location of the folder containing the data element processes)
- Primary user (consumer) of the report (or analysis)

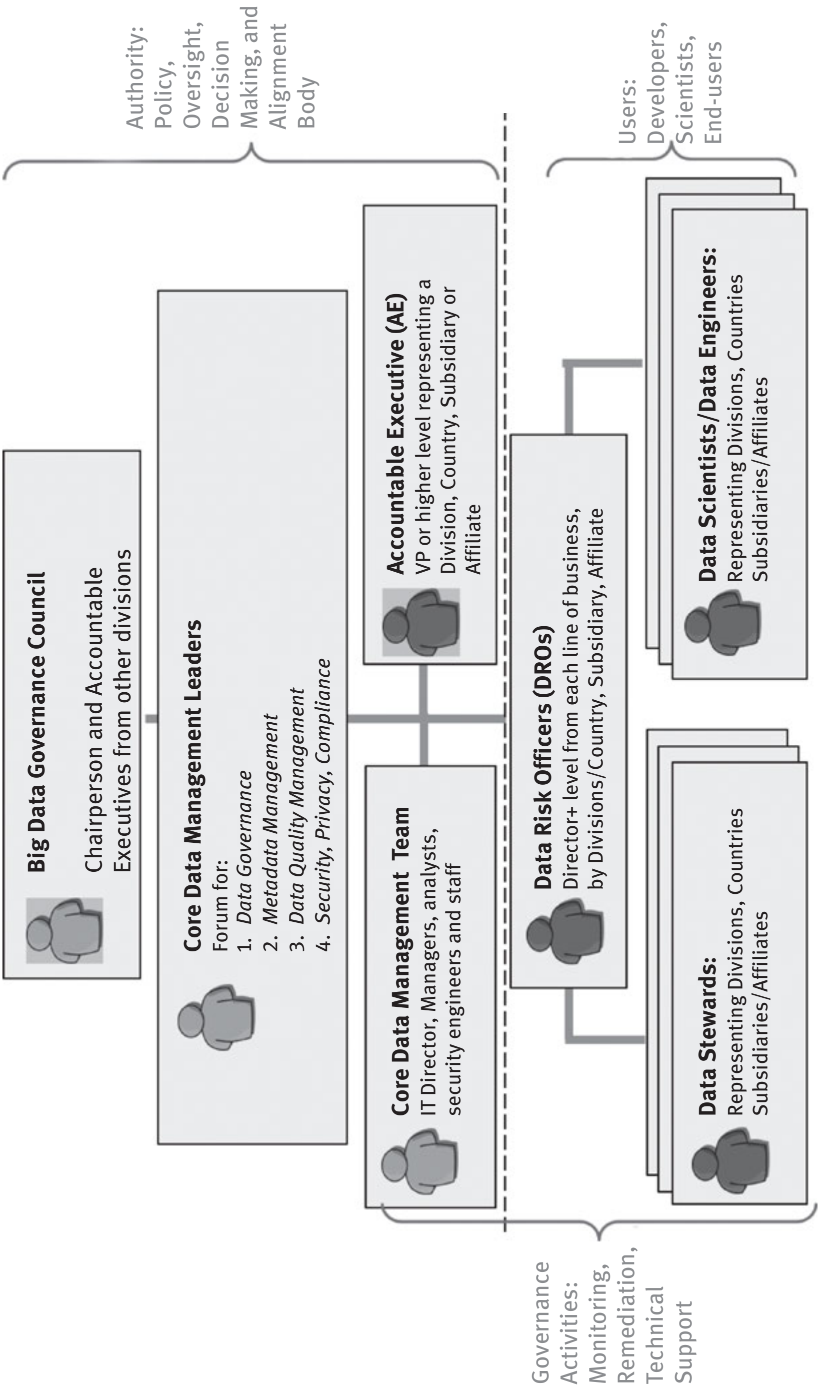

Authority: Policy, Oversight, Decision Making, and Alignment Body
Big Data Governance Council
Chairperson and Accountable Executives from other divisions
Core Data Management Leaders
Forum for:
1. Data Governance
2. Metadata Management
3. Data Quality Management
4. Security, Privacy, Compliance
Accountable Executive (AE)
VP or higher level representing a Division, Country, Subsidiary or Affiliate
Core Data Management Team
IT Director, Managers, analysts, security engineers and staff
Data Risk Officers (DROs)
Director+ level from each line of business, by Divisions/Country, Subsidiary, Affiliate
Users: Developers, Scientists, End-users
Data Scientists/Data Engineers:
Representing Divisions, Countries Subsidiaries/Affiliates
Data Stewards:
Representing Divisions, Countries Subsidiaries/Affiliates
Governance Activities: Monitoring, Remediation, Technical Support

Figure 10.1: Data Governance Council and Stakeholder Organization.

- Date of report publication
- Name of report (or analysis) as registered in the metadata hub (MDH) tool

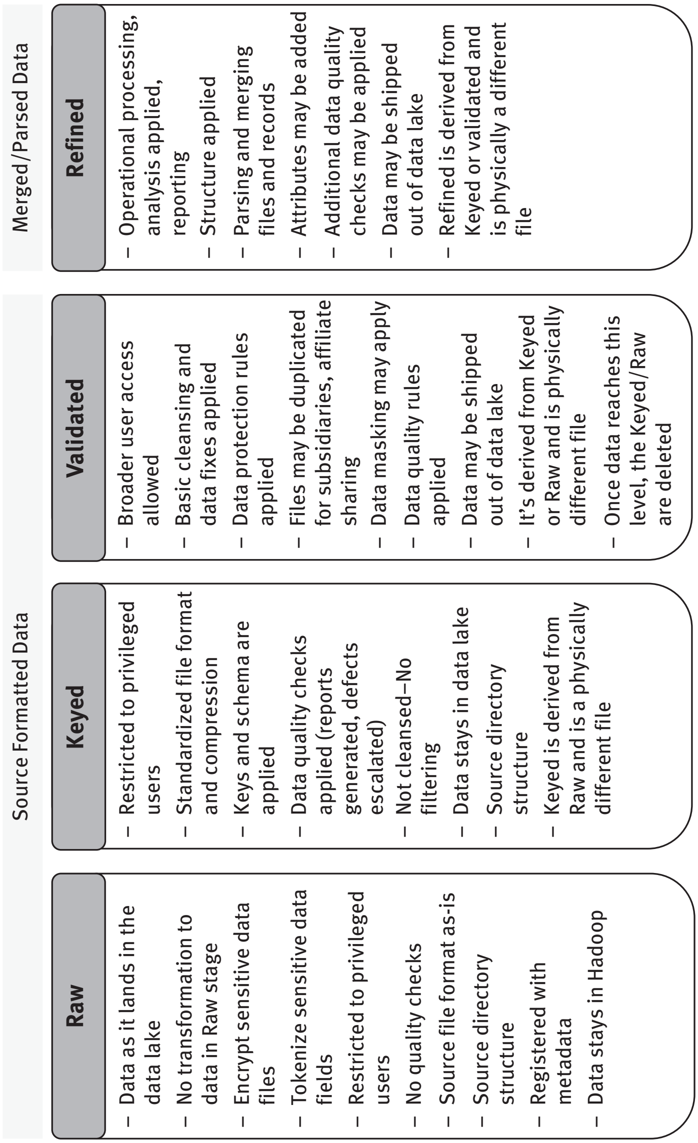
Required metadata information about the data processes (both business and analytics) include:
- Descriptive name (common name for the data element)
- Description (clearly define the data element and intended purpose)
- Valid values, or range of valid values and their meanings for data elements that are codes, flags, or indicators
- Priority level (valid values are critical, high, medium, low, or null priority)
- Source indicator (indicates the trusted source for this data element)
- Associated data usages (identifies the data usages consuming the data element where the data element is identified as high priority)
- Name of AE, DRO, and data steward
- Lineage information (at a minimum, it must indicate the source data asset's name(s), date obtained or most recently updated, and transformation performed on the source data; transformation information includes any processing or calculations performed and expected output on the column/field level if possible).
- Retention period
- Business/regulatory/legal/country-specific laws and regulations that apply to the data element
- Method of archiving/purging the data once retention period is exceeded
- Required level and standard for security (e.g., HITRUST, etc.)
- Required level and standard for privacy (e.g., HIPAA, etc.)

III Data Classification

Data Classification is an important step toward data and application management. The four data classes and their definitions are listed in Figure 10.2.

The data steward for each organization is responsible for monitoring and ensuring that the data is in the proper data classification category in metadata hub (MDH) as well.

Additional security and governance policies related to metadata management apply to the Hadoop environment for all data in the raw, keyed, validated, and refined classifications:
- Location
- Directory owner group
- Resource group
- Classification (raw, keyed, validated, or refined)
- PHI indicator (personally identifiable information)
- Affiliate group (franchise, or country, subsidiary, or affiliate)



Source Formatted Data

Raw

- Data as it lands in the data lake
- No transformation to data in Raw stage
- Encrypt sensitive data files
- Tokenize sensitive data fields
- Restricted to privileged users
- No quality checks
- Source file format as-is
- Source directory structure
- Registered with metadata
- Data stays in Hadoop

Keyed

- Restricted to privileged users
- Standardized file format and compression
- Keys and schema are applied
- Data quality checks applied (reports generated, defects escalated)
- Not cleansed–No filtering
- Data stays in data lake
- Source directory structure
- Keyed is derived from Raw and is a physically different file

Validated

- Broader user access allowed
- Basic cleansing and data fixes applied
- Data protection rules applied
- Files may be duplicated for subsidiaries, affiliate sharing
- Data masking may apply
- Data quality rules applied
- Data may be shipped out of data lake
- It's derived from Keyed or Raw and is physically different file
- Once data reaches this level, the Keyed/Raw are deleted

Merged/Parsed Data

Refined

- Operational processing, analysis applied, reporting
- Structure applied
- Parsing and merging files and records
- Attributes may be added
- Additional data quality checks may be applied
- Data may be shipped out of data lake
- Refined is derived from Keyed or validated and is physically a different file

Figure 10.2: Data Classification Stages and Controls for Hadoop Lake.

- Jurisdictional controls (specify the control and country or geography)
- Additional access restrictions (any specific access restrictions regulated by law or negotiated with the data provider)

IV Security, Privacy and Compliance Standards

>The security policies and processes outlined in this governance framework are aligned with the enterprise data security, privacy, and compliance rules. This governance framework establishes the guidelines to gather and centralize information security from different security capabilities and processes to perform effective risk management.

The additional security domains introduced by big data platform to the enterprise consists of two elements: use of the public cloud (Amazon) and open source big data (Hadoop) environments.

Implement Apache Ranger and monitor Ranger information to maintain a central data security administration capability. Configure Apache Ranger to security tools in order to manage fine grained authorization to specific tasks or jobs in the Hadoop environment. Use Apache Ranger to standardize authentication methods across all Hadoop components. Implement different authorization methods including role-based access control, attribute-based access control, etc., using Ranger.

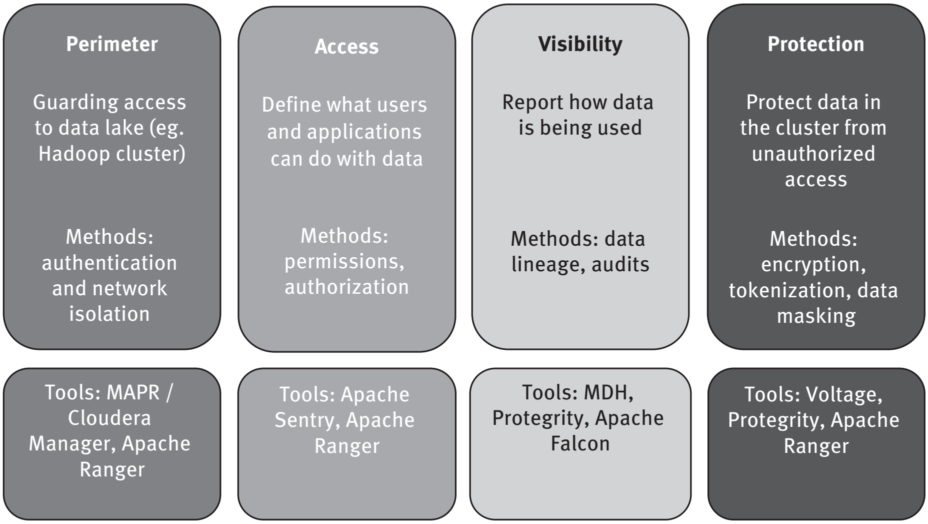
The general security policies applicable to the cloud consist of best practices from ITIL, HITRUST, and Cloud Security Alliance STAR (Security Trust and Assurance Registry) to formal third-party cloud security certification.

The goal of this security policy is to:

- Develop and update approaches to safeguard data at the application, level, sandbox level, file level, and individual field/column level.
- Gather, correlate, and provide threat intelligence to IT security from platform usage and access logs.
- Monitor for advanced threats within the environment using specialized tools and processes conforming with enterprise IT security procedures.
- Develop audit capabilities of all activity at the user and data cell level with the ability to provide data forensics analysis.

The open source Hadoop environment does not come with built-in security and therefore security tools and processes must be bolted-on to this environment. Since data flows in both directions from a traditional data management environment to Hadoop and vice versa, consistent policies are required. The four pillars of big data security (see Figure 10.3) are:

1. **Perimeter:** Guarding access to the Hadoop cluster itself. This is possible via MapR Manager.



Perimeter	Access	Visibility	Protection
Guarding access to data lake (eg. Hadoop cluster)	Define what users and applications can do with data	Report how data is being used	Protect data in the cluster from unauthorized access
Methods: authentication and network isolation	Methods: permissions, authorization	Methods: data lineage, audits	Methods: encryption, tokenization, data masking
Tools: MAPR / Cloudera Manager, Apache Ranger	Tools: Apache Sentry, Apache Ranger	Tools: MDH, Protegrity, Apache Falcon	Tools: Voltage, Protegrity, Apache Ranger

Figure 10.3: Hadoop Data Security and Access Control Considerations.

2. **Access:** Defines what users and applications can do with data. A tool of choice is Apache Sentry, but also Ranger[1] and other tools are possible.
3. **Visibility:** Reporting on how the data is being used. Support for this function comes from a combination of metadata hub (MDH) and Protegrity.
4. **Data protection:** Protecting data in the cluster from unauthorized access by users or applications. The technical methods require encryption, tokenization, and data masking. The tool to perform this function can be a solution like Protegrity or Voltage.

Security controls on Hadoop must be managed centrally and automated across multiple systems and data elements with consistency of policy, configuration, and tools according to enterprise IT policies.

Integrate Hadoop security with existing IT tools such as Kerberos, Microsoft Active Directory, LDAP, logging and governance infrastructure, security information, and event management (SIEM) tools.

[1] Ranger is promoted by HortonWorks, while Sentry was initially promoted by Cloudera. If a different Hadoop infrastructure such as MAPR is selected, then neither tool may be necessary as MAPR includes its internal file structure security tool, but you can opt to use Sentry as the central authorization tool as well.

Authentication

You can manage access to your Hadoop environment via a single system such as LDAP and Kerberos. Kerberos features include the capability to intercept authentication packets unusable by the attacker, eliminating the threat of impersonation, and never sending a user's credentials in the clear over the network. The security policy requires using Kerberos for authentication and LDAP for user access privilege control.

All intra-Hadoop services must mutually authenticate each other using Kerberos RPC. This internal check prevents rogue services from introducing themselves into the cluster activities via impersonation. Ensure that the Kerberos Ticket Granting Token feature is property configured.

Authorization

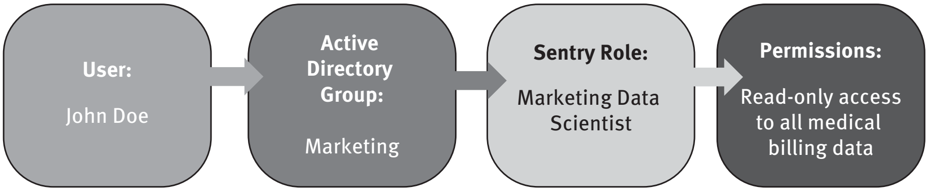
Apply data access controls in Hadoop consistent with the traditional data management policies using Apache Sentry (or Ranger) to enforce user access controls to data in Hadoop before and after any data extract, transform, and load (ETL).

Specific security rules listed below apply to big data management:

- To apply user access controls to Hadoop data, apply POSIX-style permissions on files and directories, access control lists (ACL) for management of services and resources, and role-based access control (RBAC) for certain services that access the data.
- Configure Apache Sentry for consistent access authorization to data across shared jobs and tools including Hive, MapReduce, Spark, Pig, and any other direct access to HDFS (specifically to MapR RDFS) to ensure a single set of permissions controls for that data are in place.
- Security controls must be consistent across the entire cluster to ensure that intra-process accesses to data on various VMs are enforced. This is of particular importance within MapReduce since the task of a given job can execute UNIX processes (i.e., MR Streaming), individual Java VMs, and arbitrary code on the host server.
- Ensure Hadoop configuration is complete for internal tokens including Delegation Token, Job Token, and Block Access Token; and all are enabled.
- Ensure Simple Authentication and Security Layer (SASL) is enabled with an RPC[2] Digest mechanism configuration.
- Never allow insecure data transmission via tools such as HTTP, FTP, and HSFTP. Find alternate tools to HSFTP since Hadoop proxies use the HSFTP protocol for bulk data transfers.

[2] Remote Procedure Call.

- Enforce IP address authentication on Hadoop bulk data transfers.
- Enforce process isolation by associating each task code to the job owner's UID. MapReduce offers such features to isolate a task code to a specific host server using the job owner's UID, thus providing reliable process isolation and resource segmentation at the OS level of the cluster.
- Apply Sentry tool configurations to create central management of access policies to the diverse set of users, applications, and access paths in Hadoop. Apply Sentry's fine-grained role-based access controls (RBAC) on all Hadoop tools consistently such as Impala,[3] Hive, Spark, MapReduce, Pig, etc. This creates a traceable chain of controls that governs access at the row and column level of data, as shown in Figure 10.4 which is repeated from Figure 7.4 so that this template is complete.
- Apply Sentry's features to use Active Directory (AD) to determine user's group assignments so that any changes to group assignment in AD are automatically updated in Sentry – resulting in updated and consistent role assignments.
- Apply RDFS permissions for coarse data controls, including directories, files, and ACLs such that each directory or file can be assigned multiple users – group with file permissions are read, write, and execute. Additionally, configure RDFS directory access controls to determine the access permissions to child directories.
- Apply additional access control lists (ACLs) to control data-level authorization of various operations (read, write, create, admin) by column, column family, and column family qualifier, to specific cell-level (A cell is the intersection of a column and row). Configure ACL permissions at both group and user levels.



Example: user John Doe belonging to Marketing Group, having Read-only access in Data Scientist Role to medical billing data, controlled access thru Active Directory membership and Sentry rule-sets

Figure 10.4: Access Control Model Example for a Typical User.

[3] Impala was initially developed by Cloudera and requires Sentry for application token security. Hence, if you select Impala, then you'll be required to implement Sentry. However, if you choose MAPR and Drill (instead of Impala) then Sentry is optional and not mandatory.

Visibility and Audit Reports

The goal of visibility is transparency and the goal of audits is to capture a complete and immutable record of all data access activity in the big data platform. The rules for audit and transparency include:

- Implement access log audit trails on log-ins and data access in order to track changes to data by users.
- Define data categories that contain personally identifiable information (PII) and are subject to HIPAA or PCI Data Security standards. Ensure the privacy category is complete in the metadata hub (MDH) tool for all data sets.
- Implement the capability to furnish audit log reports of all users including admin activity within the context of historical usage and data forensics.
- Maintain full audit history for RDFS, Impala, Hive, Hbase, and Sentry in a central repository, including user ID, IP address, and full query text.
- Periodically (on a monthly basis) review access logs and identify anomalies, issues, and gaps associated with security policies to the data council.
- Adopt work flow management tools such as Oozie as standard to maintain error and access logging at each stage of data processing, transformation, and analysis. Combine Oozie with Apache Falcon to track and manage job streams within specified data governance policies.
- Store Hadoop job tracker and task tracker logs for all users and review the logs on a regular (monthly) basis.

V Data Usage Agreement (DUA)

The data usage agreement is a set of rules and policies that apply to uses of data elements, whether they are internal or externally obtained. All critical and high priority data are expected to be of “trusted source” data quality. The best practice policies and rules associated with DUA specifically include:

- Usage of all critical and high priority data will be subject to data quality monitoring, testing, and data certification.
- The data usage agreement defines appropriate and acceptable use of critical and high priority data for each data element.
- The DUA may indicate the data retention period and acceptable usage time period.
- Data asset change control management requires advanced notice of any data changes, review, and approval and a defined escalation process.
- SMEs, data engineers, and data scientists (“users”) agree to maintain and update the metadata hub (MDH) by registering their data with the latest information about the data elements that they load into the Hadoop lake, or consume or produce as a result of their analysis for their area of responsibility.

- SMEs, data engineers, and data scientists ("users") agree to follow privacy, confidentiality, and security standards for the data in their areas of responsibility.
- SMEs, data engineers, and data scientists ("users") agree to complete big data governance annual training.
- Third-party data is monitored and to be managed in accordance with the third party data agreement specified with the third party data provider. Lineage of data elements must indicate the third party and fourth party data providers. (Fourth party data providers are those entities that provide data to the third party data provider).

VI Security Operations and Policies

Anonymization has been applied and studied extensively by leading IT management experts. A number of general open source tools for anonymization are available, such the Cornell Anonymization Toolkit and ARX. Toolkits that work with big data tools, like the Hadoop Anonymization Toolkit have emerged such as Protegrity and Voltage. Policies should determine and offer guidance on anonymizing IP addresses and data in the cloud, and monitor and measure the quality of anonymization regularly.

Both encryption and tokenization are techniques widely used in securing big data in the cloud. Data governance security best practices should indicate that while tokenized data may be stored in the cloud, the tokens must be maintained on premise and not in the same logical and physical domain as the data itself.

Medical data has traditionally been anonymized to enable research while protecting privacy. Policies must indicate whether you can bring anonymized data sources together and join the information while the data is anonymized. Similarly, the policies should indicate when we can use simple key encryption and what type of data must be tokenized.

Additional policies and operational rules are as follows:

- Ensure all users have completed pre-required training, such as the Hadoop security and governance training material, before granting access to them.
- Data stewards are responsible for notifying the Hadoop security admin regarding onboarding or off-boarding users for their area of business.
- Create user groups and assign user IDs to the appropriate groups.
- Users must submit an access request form showing they've completed prerequisites before getting access granted.
- Access request forms must be approved by the LOB (line of business) data risk officer.
- Create training materials for users to include, at a minimum using the following outline: A data lake environment overview, data and application governance policies, policies about PII and sharing data and reports with affiliates, information

security of PII, compliance and regulatory compliance rules specific to the big data platform.
- Streaming data into Hadoop can produce large and fast data loads that make data quality checks of every data item difficult. Instead, apply random data quality checks for validity, consistency, and completeness.
- Ensure data quality metrics are reported for streaming data including metadata information and lineage information.
- All IP addresses must be anonymized to protect the user data.
- All personally identifiable data must be de-identified in raw or become encrypted. Tokenization is preferred.
- Develop and monitor performance metrics for services and role instances on the Hadoop clusters. These metrics include HDFS I/O latency, number of concurrent jobs, average CPU load across all VMs, CPU/memory usage of the cluster, disk usage, etc. Other metrics include activity metrics, agent (application) specific metrics, fail-over metrics, and attempt metrics.
- Apply the highest security control to the Kerberos Keytab file that contains the pairs of Kerberos principals and tickets to access Hadoop services.
- Configure Hadoop clusters to monitor NameNode resources and generate alerts when any specified resources deviate from the desired parameters.
- Use Hadoop metrics to set alerts to capture sudden changes in system resources.

VII Information Lifecycle Management

Information lifecycle management governs the standards and policies of data from creation and acquisition to deletion. Due to regulatory and well-managed practices, several rules for retention and management of data are applicable to big data governance.

Some of the best practice policies that govern information lifecycle management are included here:

- The retention rules for data may vary from country to country. The longest retention period must be updated in the metadata hub (MDH) tool for all data sets.
- Certain analytics data snapshots are necessary for compliance, legal, and medical reasons. Such snapshots may include a mirror copy of the data, the algorithms, models, the version, and results. The snapshot data will be saved in a separate protected Hadoop sandbox with Admin access privileges.
- Users may request copies of data snapshot by requesting such information from the Hadoop Admin.
- Data that reaches its end of life (upon maturity of the retention period) should be purged according to the IT data purging procedures.

- Define the Hadoop disaster recovery, downtime, and business continuity policy. A downtime means that the data lake (for example, Hadoop) infrastructure is not accessible. The policy would indicate how replicated sites will be managed and the process for either automatic or manual fail-over in the event of a downtime. The business recovery plan should indicate how replication will resume, the process for restoring data, recovering data, and synchronizing data after the failed site is online.

VIII Data Quality Standards

The data quality standard defines requirements to classify, document, and manage data to ensure that the big data platform's critical and high priority data meets established quality requirements. The objective of this standard is to ensure that the platform's data is managed and is of sufficient quality.

The data quality standard requires the data steward to manage data issues. This responsibility includes identifying data issues, assigning severity, and notifying impacted stakeholders; developing data remediation plans; providing regular status updates on data remediation efforts; communicating data issues involving the remediation plans and their status with the DRO, the AE, and the data council.

Data issues can be detected from various sources, including but not limited to compliance monitoring activities, control testing, data quality and Metadata monitoring, project work, quality assurance testing and audits.

All data whether internally or externally obtained is to be classified into four categories:

1. **Critical Data**: The type of data that has specific regulatory and compliance requirements such as HIPAA, FDA, FISMA, and financial is materially significant to decision making in the organization.
2. **High Priority Data:** The type of data that is materially significant to decision making in the organization regarding classified information and confidential data. This data is required for analytics but does not carry the high compliance requirements such as FDA, CDI, or HIPAA.
3. **Medium Priority Data:** Data that is typically generated or used as a result of day-to-day operations and is not classified as confidential or classified.
4. **Low Priority Data:** Data that is typically not material to decision making and analytics and does not have specific retention or security/privacy requirements.

The data quality standard applies to critical data and high priority data that:

- Is applicable to data usage of critical and high priority data elements as designated by each franchise, country, or division account executive.
- Sets criteria for the identification of data deemed most significant and impactful to the business.

- Defines the accountable executive and performer roles responsible for data management across the organization.
- Sets the minimum requirements for managing risk due to data-related issues.

Enforcing the consistent creation, definition, and usage of critical and high priority data reduces reputational, financial, operational, and regulatory risks and prevents subsequent costs of presentation and clinical errors as well as the potential for making decisions based on incorrect, inconsistent, or poor-quality data. Applying these requirements enables the organization to effectively demonstrate the quality of the data, raising confidence with regard to regulatory compliance in the data used.

The data quality standard for each division, franchise, or country would consist of the following:

- Critical and high priority data are to be maintained and managed as data assets, which are collections of one or more data sets. A data set includes a data table, file, spreadsheet, or any collection of data elements.
- Trusted source is defined as a critical or high priority data set that has a complete and accurate metadata definition including data lineage, and the quality of the data set is known.
- Sources that are not fully meeting the definition of trusted source will be deemed in “pending” status until the data set is brought into compliance.
- A plan, with commitment dates for compliance with the governance standards are in place and are being executed.
- Each AE and DRO must identify which data elements contained in data assets in their area of responsibility are trusted sources or in “pending” status.
- Validate lineage from trusted sources to the point of origination for critical or high priority data elements.
- Maintain accurate and complete metadata for critical and high priority data elements.
- Apply data quality monitoring to critical data elements within data assets. Document and report to the DRO any issues related to data quality.
- Data quality issues are defined as gaps between the data element characteristics and any deviations from the data governance standard document. For example, if the data governance standard document states that there should be no data duplicates, no NULL values for certain data fields, and data must be within certain range defined in its metadata definition; then a data set that contains duplicate data records, missing data, or incorrect data indicates data quality gaps between the data element and the data governance standard.
- Understand how the critical and high priority data within each area of responsibility is being consumed and inform the DRO of material changes that impact their specific usages as defined in the data usage agreement.
- Create and maintain data retention plans according to data lifecycle management policies.

- Maintain inventory and certification of models used by data scientists.
- Provide metrics reports to the DRO, AE, and data council related to level of compliance of critical data sets. Report issues list with the expected date of remediation and a remediation plan.
- Provide evidence of compliance with the data management policies described in this document and supporting standards, procedures, and processes.
- Manage and prioritize activities to mitigate data quality risks and issues, including the oversight by the AE and DRO of an execution plan to apply these requirements across critical or high priority data usages within their domain.
- Annually review inventory of critical or high priority usages and data usage agreements within their domain.
- Document processes for identification and classification of data elements by criticality level and adherence to quality standards.
- Critical and high priority data are to be inventoried, maintained, and linked to appropriate model(s), process(es), and report(s) in the metadata hub (MDH) tool.

For data that is in “pending” status, it can only be moved into “certified” status by the data steward for its area of responsibility if it meets the following criteria (compliance requirements):

- Metadata is complete in the metadata hub (MDH) tool
- Lineage is tracked to origination or to the trusted source
- Data quality monitoring and checks passed
- The retention period is defined

Data Quality Reporting

It is a requirement that the AE, DRO, and data stewards regularly monitor and report on the quality of critical and high priority data elements within their assigned areas of responsibility on a quarterly basis to the data council.

Each data quality report must contain the following components:

- Results of monitoring
- List of issues identified (i.e., data quality failures)
- Pass/fail based on thresholds
- Assessment of the reliability of the data for the usages as specified in the data usage agreement

The following metrics are used when defining data quality rules for critical and high priority data elements:

1. **Accuracy:** Measures the degree to which data correctly reflects its true value. It’s a determination of how correctly the data reflects the real world object or

event being described. Data accuracy is most often assessed by comparing data with a reference source.

2. **Validity:** Validity measures the degree to which data is within an appropriate range or boundaries, conforms to a set of valid values, or reflects the structure and content of a value which could be valid.
3. **Completeness:** Measures the degree to which data is not missing and is of sufficient breadth and depth for the task at hand. In other words, it's fulfilling formal requirements and expectations with no gaps.
4. **Timeliness:** Measures the degree to which data is available when it is required. Additionally, this measure ensures data is combined together with consistent time periods.
5. **Consistency:** A measure of information quality expressed as (a) the degree to which redundant instances of data, present in more than one location, reconcile, and (b) the degree to which the data between a trusted source and usage(s) can be reconciled during the movement and/or transformation of the data.

When data issues are detected from various sources including compliance reporting, control testing, audit findings, data quality and metadata monitoring activities, or

Table 10.2: Data Issue Severity Rubric.

Severity Level	**Extreme (Level 4)**	**High (Level 3)**	**Medium (Level 2)**	**Low (Level 1)**
Criteria for triage of data quality issues	Issue has or will prevent timely preparation of business or regulatory reports or will negatively impact business by > $1M	Issue impacts business processing or reporting. Remediation activity is required.	Issue has minimal impact to business decision making or reporting. Data remediation activity is recommended.	Issue has minor impact to business decision making or reporting. Remediation is recommended but not time sensitive.
Issue remediation plan document within (time):	1 week	2 weeks	3 weeks	4 weeks
Recommended timeline for remediation	< 1 month	1–3 months	3–6 months	6–12 months
Recommended communication of status frequency	Weekly	Bi-weekly	Monthly	Quarterly

from third party data management audits and activities, they must be documented and reported by the data steward for their area of responsibility. Such reports for critical and high priority data must include the remediation plan and target date for remediation.

Data issues come in different severity levels. The rubric in Table 10.2 contains the criteria for triage of data quality issues.

When an issue is resolved, the data steward is responsible for closure of data issues to the impacted executives, DRO, and AE in their area of responsibility. However, in the event that the issue is not resolved in a timely fashion, the DRO and AEs are responsible for escalating the issue to the data council and including it in the enterprise risk management quarterly risk report.

Part 3: **Big Data and Model Risk Management**

Chapter 11
Why Data and Model Risk Management?

Big data governance refers to the rules, structures, processes, and practices that allocate the roles, responsibilities, and rights of participants in big data analytics. This governance policy is intended to meet all relevant compliance with applicable laws and objectives of operating the big data platform in a safe and sound manner.

As we reviewed earlier, data risk is the exposure to loss, financial or reputational, caused by issues related to an organization's handling of data assets. Lack of consistent policies, planning, and operations with respect to data can lead to operational and strategic issues, such as duplication of data, conflicting data interpretations, inability to access data, and inability to monetize data.

Data is growing in volume, velocity, and value. When properly managed, data can drive tremendous business value and competitive advantage. Lack of proper governance, disciplined and strategic data management, can pose a huge obstacle to business survival. The two extremes of thrive-on-data versus get crushed by it, have never been so diametrically opposite to each other. The middle ground of giving lip service and minimal investment in data governance can no longer sustain an organization. So where is the future of data governance headed given the current socio-technical trends?

The Future of Big Data Governance

Data risk and its implications are now a topic of boardroom discussions. Board members are now much more engaged in data governance and their responsibilities toward data management. They want to have confidence in the integrity of data management processes and policies. Topics range from understanding the health of the organization's data infrastructure to how data is used, as well as compliance with regulations and ethics issues.

A strategic data review should include data risk assessment that may help uncover systemic gaps and issues. The issues might include data risk identification, access and usage policies, monitoring and decision-making procedures, risk training, and governance.

Leaders are concerned with five questions that a robust data governance structure should be prepared to respond to today and in the future:

1. **How is data ownership assigned and managed?** As corporations manage data as a strategic asset, their investment in agile infrastructure, cloud computing, and nimbler data storage technologies must increase. Establishing the right organizational and cultural structure for data governance is critical. This requires proper investment and raising visibility of data governance in the

https://doi.org/10.1515/9783110664065-011

leadership ranks. Investing in the proper organization includes forming a data governance team, establishing a chief data officer, chief risk officer, chief model risk officer, data steward, and data compliance and audit roles; these are necessary to define clear accountability and ownership across the enterprise.

2. **How is data aligned across the organization?** Data classification must be enterprise-wide, consistent, and aligned with corporate strategy. Increasingly, as data is collected and accumulated at a rapid rate, organizations must maintain and monitor regulatory compliance and processing activities across the enterprise. Interoperability and re-use of data becomes a critical component of data lifecycle management. Consistent and common metadata and data descriptions across different domains and business units must be maintained and updated in a continuous manner.
3. **How confident are we in our data and its protection?** The technical and procedural aspects of how the company manages and interacts with data are foundational to good data governance. Confidence in data starts with safeguarding data – protecting it from internal redundancies, lack of controls, and external attacks. Automated data security audits will be necessary to continuously ensure data is protected and there are no data leakages from the cloud. Steps to encrypt, mask, anonymize, and pseudonymize data based on proper data classification are key to implementation of a sound data practice. Consistent data policies across the enterprise must be in place that drive accountability, consistency, compliance, and ethical conduct in handling data.
4. **What are the risks associated with use of data?** Risk assessment of big data must implant ethics, culture, usage, and privacy into the design of data management. As organizations modernize their data infrastructure and work to accommodate big data, they must also invest in data governance enhancements and maturity. Data classification and risk assessments must include standard data models, reference architecture, integration patterns, ownership, interoperability, a risk rating framework, and risk mitigation plans. Using Kanban (visual displays of data issues} and real time notifications of data risks will become best practice. Automated dashboards that display data issues and analysis summaries in real time will provide quick awareness of risks to leadership.
5. **How is the data lifecycle managed?** Big data has amplified the complexity and diversity of the data lifecycle. Data is now being collected in real time. Hence we need a more agile process to define data governance attributes (metadata, master data, quality, controls, privacy, etc.). This leads to the need for automation of data governance policies using bots or automated data lifecycle management tools. Policies to govern data usage, business intelligence and analytics tools to generate insight must be clearly communicated to all data professionals including those who develop machine learning, cognitive science, and AI data products.

Examples of Data and Model Bias

Accuracy and fairness of models are crucial to drive correct decisions and business operations. Consider predictive or machine learning models. If the data has inherent biases, then the model is likely to mimic those biases or make it worse by amplifying them. With biased data, we're likely to make the wrong conclusions and cause disastrous outcomes for business and even for society at large. One of the key mantras in data management and modeling is "do no harm." Biases in data and models can target certain people adversely and result in discrimination that cause more harm.

In most instances, collected data for analysis must represent the population of subjects adequately and yet ensure that the data is not unbalanced. For example, one of my colleagues developed a program that predicted job candidate performance for a company. The machine learning model was trained on the existing collection of employee resumes and subsequent performance reviews. The model predicted attrition and high performers with excellent accuracy. But there was a problem. After further analysis, it became clear that the model was working against women applicants, even though gender was not a data variable. This was due to the fact that the company's prior hiring practices were biased against women, and there were far more male employees than females. Hence the data scientist had to take steps to balance the data to make the model equitable.

Amazon developed a tool to automatically screen resumes and hiring. They scraped the project since they found bias against women. In other words, if the data is biased, the model will find a way to encode that information.

On the surface, the most obvious approach to removing bias from a model seemingly is to explicitly remove variables that are associated with bias. But that's not enough and you need to consider two other remedies:

First, test your model against bias. You should insert those variables back into the analysis of results to determine that the model is not biased against any of those removed variables.

Second, ensure that some of the remaining variables are not highly correlated with the removed variables that cause bias. For example, if the model predicts who would be a good hire from a list of candidates, you may include relevant inputs like skill and experience, while excluding irrelevant variables such as gender, race, and age. But, it's possible that even when you exclude data such as "race," the other information used in the model can be encoded to represent race. This happens often when a candidate's location or father's education level are included as input. You can just exclude race, but fail by including other variables that may encode someone's race.

Recently a company conducted research into software used to predict recidivism rate. They found that the model had much higher false positive rates for the

African American population than the White population.[1] It turned out that the company that developed the survey of 137 questions included race as input. But other questions could have also encoded race. For example, if a question asks whether the person's parents were separated while growing up, it's likely to indirectly encode their race because there are different divorce rates among different races.

Increasingly such mistakes could result in significant fines and compliance investigations. All data managers, data scientists, and data workers should consider bias and the ethical impact of their models.

From a data and model governance perspective, it's important to understand the levels of risk associated with the data and its use cases. Data can be misused and abused for unethical and illegal purposes. Recently two data scientists at a credit card company were indicted for insider training. With access to the company's credit card data, they were able to predict how clients will perform based on their product sales reflected in credit card transactions. They extrapolated what the total revenue of the company could be based on a smaller credit card sample. Armed with this analysis they engaged in a lucrative insider trading business. Eventually the law caught up with them. But the credit card company was unaware of the misuse of its data by its employees.

We need comprehensive data and model governance and risk management to prevent such damaging incidents. There are three concepts to managing bias when building data analytics products:

1. **Examine the data to ensure it's not biased.** Your training data set must be representative of the population that you are analyzing. There are at least 12 types of bias in data analysis that you should steer away from. Data scientist have tools to test the data against these biases. The most common biases are:
   a) *Confirmation bias*. This occurs when the person collecting data wants to prove a predetermined decision.
   b) *Selection bias*. The data collected ignores a certain population(s) on purpose or by negligence.
   c) *Overfitting and underfitting*. Overfitting happens when the data scientist includes too many data variables which could introduce noise into the model, diminishing the importance of the key predictive variables. Underfitting occurs when the data scientist ignores some key data variables that have predictive power and significance.
   d) *Outlier data*. Collecting data that includes extreme data values that are outside of the normal range.

[1] *Machine bias* (2016, May 23). ProPublica. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

   e) *Simpson's paradox*. The data includes separate trends when individual data variables are considered, but overall, the data patterns point to the opposite trend.
   f) *Availability bias*. Data scientists analyze readily available data without considering if it is a good representation of the entire population.
2. **Choosing the right model.** Require your data scientists to select the best model that fits the problem and the data. No single model can fit all. For example, using unsupervised learning models can reduce data dimensions and cluster data subjects into different segments. This approach shows different groups of data subjects and reveals bias. Supervised learning can offer control over bias in data selection. But it can also introduce bias during model training. The choice of modeling algorithm is an important decision with impacts that determine bias and ethical application of the data and the model. Model risk management (MRM) reviews are critical in two ways: They identify the risk associated with the model and usage patterns of data; and they identify mitigation solutions to minimize risks of bias and ethical use of data.
3. **Test and monitor your model's performance.** A model might work well under a controlled data environment. But how will it perform over time as data and population demographics change? It's important to monitor a model's accuracy and ethical review on a regular basis. Examine and audit the results of models periodically. Look for simple, often-consistent patterns that represent bias. For example, if you're analyzing housing, are you finding consistent patterns to certain groups (single parents, employed or unemployed, recent migrants, etc.)?

Model Risk Management Overview

Just as data volumes are increasing, the number of models are increasing dramatically. McKinsey estimates that the number of models are increasing at 10% to 25% annually and they're becoming more complex using advanced analytics techniques from machine learning to predictive AI and time series analysis.

Regulatory and ethical requirements, in particular for the financial industry, are mandating specific review processes and approvals before release into production. Big data ushers in more advanced analytics models such as credit scoring, loan default prediction, anti-money laundering, fraud detection, pricing, customer personalization, and asset-liquidity prediction.

As the risks associated with models increases, more scrutiny from the public and regulators should be expected. Companies are increasingly adding model risk management (MRM) to the list of governance issues. Model risk lies in defective models, wrong interpretation of model results, model misuse, and inappropriate model selection for a given data set. Model risk can expose a company to tremendous liability that can be avoided through proper model risk management.

In the financial industry, the US Board of Governors of the Federal Reserve System published the Supervisory Guidance on Model Risk Management in 2011. The guidance states that the use of models presents risks of potential adverse consequences from decisions based on incorrect or misused model outputs and reports. The guidance requires models to be identified, mapped, tested, and reviewed.

Model risk management can be implemented in three evolutionary stages:

Stage 1: Establish the Foundation of MRM

This stage develops the MRM program and the foundations for a model risk management program to:

- Define model management policy and procedures, model governance, and standards.
- Define the model lifecycle risk management. The lifecycle should consider the model from inception to retirement. It should define the process for MRM model intake, documentation, testing, model validation, review, and approval. It must address policies and processes for post release monitoring and maintenance of the model.
- Define a model catalog to inventory all models.
- Define the MRM organization including the governance team, validation team, and model risk officer.

Stage 2: Operationalize and Execute

This stage implements and operationalizes MRM practices, including staffing, establishing regular governance meetings, reviews, and infrastructure to support the operations.

- Implement the controls and processes defined in Stage 1.
- Provide stakeholder training and internal audits to ensure the policies are implemented and are followed properly.

Stage 3: Enhance and Optimize

This stage offers opportunities to gather feedback from stakeholders and improve on policies, issue resolution, and operations. It can include:

- Developing a center of excellence for model risk management.
- Enhancing model validation, monitoring, and quality.
- Defining ways to extract value from models.

Every MRM consists of two levels of governance: one level governs the model lifecycle and another covers leadership communication and compliance reporting.

A typical model risk management process might follow the steps shown in Figure 11.1.

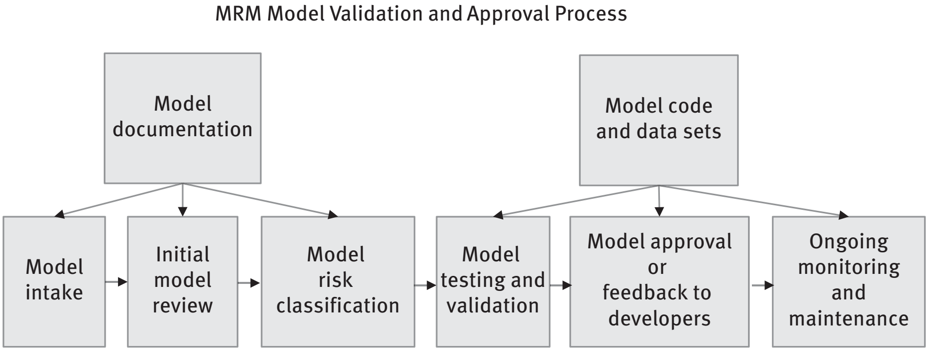

MRM Model Validation and Approval Process
Model documentation
Model code and data sets
Model intake
Initial model review
Model risk classification
Model testing and validation
Model approval or feedback to developers
Ongoing monitoring and maintenance

Figure 11.1: Model Risk Management Process.

A robust model risk management and standard governance practice will enable more responsible and practical modeling products with fewer issues and resources. Using MRM gives organizations the confidence that their models meet a higher standard of quality, reliability, accuracy, and ethical use.

Summary

This book has covered a wide range of topics and areas, including modern data governance principles for more effective management of big data. While the amount of effort and scope is substantial, the art and science of this field is continuously changing and new tools and methods emerge almost daily. It's important to continue staying on top of the latest techniques in data governance to deliver effective and best practice methods to your organization.

Similarly, best practices in big data governance are evolving. But this book has laid a foundational and practical groundwork to design and implement a data governance structure for any industry and company of any size.

Big data will only grow bigger and with it the importance and visibility of data governance will increase. The leading C-suite executives have come to realize investing in data governance will protect their investments in data in the long run. Fortunately, building a big data governance structure doesn't need to be expensive or complex. It starts with implementing the basic, common sense practices explained in this book.

https://doi.org/10.1515/9783110664065-012

Index

access controls 53, 58, 59, 63, 74, 78, 87, 89, 91, 101, 108–111, 114, 117–119, 129–132
Amazon.com 4, 5, 13, 19, 40, 41, 54, 55, 88, 92, 129, 145
analytics layer 8, 10, 13–16
Analytics Platform Framework 8–18
Apache Atlas 68, 69, 71, 84, 100
artificial intelligence 7, 14, 15, 19, 23

best practices 18, 29, 32, 43, 44, 47, 56, 57, 60, 64, 72, 84, 85, 93, 95–119, 129, 133–135, 144, 151
big data 3–22, 24, 28, 29, 31, 39, 41, 43–46, 56, 57, 64–78, 82, 84–140, 143–151
big data governance rules 74, 99–103
big data security 88, 89, 103, 129
business glossary 35, 44, 47, 71
business intelligence (BI) 4, 7, 8, 13, 21, 23, 31, 53, 54, 73, 139, 144

California Consumer Privacy Act (CCPA) 10, 26–29
Capability Maturity Model Integration (CMMI) 32
Cassandra 11–13
chief data officer 19, 47, 64, 70, 144
chief information officer 70
classification engines 14, 15
cloud 3, 4, 10, 12, 17, 19, 21, 39, 41, 54–57, 63, 88, 103, 107, 129, 134, 143, 144
cloud initiative 55
CMMI Institute 32
coarse grain data 85, 90, 100, 107, 109, 110, 119
Collibra 100, 110, 111
Common Security Framework (CSF) 10
compliance audits 78
compliance, security, privacy 44, 72, 74
corporate governance 22, 43, 45, 101, 120
critical data 26, 78, 87, 93–95, 136–138
CRUD cycle 81

data access protection 108
data breaches 23, 26, 39–46, 56, 122
data classification 15, 64, 72, 74, 77, 79, 84–94, 97, 99, 104, 105, 111–117, 120, 127–129, 144
data classification rules 74, 93
data cleansing 10, 35, 70, 83
data connection layer 8, 9
data de-duplication 9
data dictionary 23, 56, 64, 71
data forum managers 76
data governance 4–6, 17–19, 22–35, 39, 41–57, 64–79, 84, 87, 90, 91, 94, 95, 99–140, 143, 144, 151
data governance council (data council) 47, 64, 70, 72, 75, 76, 78, 87, 90, 91, 94, 95, 99, 102, 111, 117, 121–126, 133, 136, 138, 140
data governance software tools 100
data lake 5, 9, 12, 17, 18, 20, 21, 39, 52, 53, 66, 67, 70, 72, 74, 77, 79, 82, 89, 91, 92, 95, 100, 102, 104, 107, 108–119, 134, 136
Data Management Body of Knowledge (DMBOK) 18, 29–32
Data Maturity Model (DMM) 18, 32–35
data mining 3, 9, 14
data normalization 4, 7
data objects 12, 21, 22
data quality 20–24, 31, 32, 34, 35, 44, 57, 64, 65, 67, 69, 70, 76, 77, 78, 90–92, 94, 95, 99, 109, 111, 112, 120, 123–125, 133, 135–140
data quality management 24, 31, 32, 69, 70, 76, 109, 111
data risk officer (DRO) 47, 70, 76, 91, 94, 117, 123, 125, 134
data sharing 104–106
data stewardship 43, 76, 77, 122
Data Usage Agreement (DUA) 67, 78, 90, 95, 106, 118, 120, 123, 133, 134, 137, 138
Data Usage Agreement Policy 90
data volume, velocity, and variety 3
data warehouse 4, 5, 8, 11–13, 17, 18, 20, 21, 28, 52, 53, 63
de-normalized data 11
dimension tables 11
DMBOK2 29, 30

https://doi.org/10.1515/9783110664065-013

enterprise data governance 24–35, 76
extract, load, transform (ELT) 9, 17
extract, transfer, load (ETL) 9, 17, 20, 31, 45, 83, 89, 93, 131

Facebook 3, 4, 12, 13, 41
fact tables 11
FedRamp 10, 55
fine grain data 90, 91, 100, 108–111, 129, 132
forecasting 15

General Data Protection Regulation (GDPR) 10, 19, 23–28
governance framework 24, 43, 64, 69–78, 101, 120–140
governance-risk-compliance (GRC) 101

Hadoop 3–5, 8–13, 17, 18, 20, 31, 41, 46, 53, 57–68, 70, 72, 75, 77, 79, 82–93, 95, 98–100, 102–107, 109–120, 127, 129–136
Hadoop Data File System (HDFS) 12, 58–60, 90, 104–107, 111, 118, 131, 135
HBase 62, 84, 93, 99, 111, 133
HCatalog 71, 84, 103, 110
hierarchical data 81
high priority data 85, 90, 94, 95, 105, 123, 133, 136–138, 140
HIPAA 10, 23, 28, 78, 127, 133, 136
history of Hadoop 59
HITRUST 10, 78, 85, 88, 127, 129
Hive 13, 17, 62, 71, 76, 84, 99, 104, 108, 110, 111, 118, 125, , 131–133
HiveQL 13

IaaS: Infrastructure as a Service 54
Impala 84, 104, 111, 132, 133
inference 15
information integrity 30, 44, 121
information lifecycle management (ILM) 64, 65, 74, 92, 120, 135

Kerberos 59, 89, 108, 130, 131, 135
keyed data 84, 85, 104, 111, 114, 115
Kimball 11
Kyvos Insights 72

Loom 71
Lucene 11, 13, 62

machine learning 7, 13, 14, 16, 19, 23, 144, 145, 147
MapReduce 12, 58–62, 88, 131, 132
master data management (MDM) 31, 64, 80–98
metadata 9, 22, 23, 24, 31, 32, 34, 35, 44, 47, 56, 60, 63–65, 67–69, 71–79, 81–84, 88, 90, 92–100, 103, 110–117, 120, 124, 125, 127, 130, 133, 135–139, 144
metadata best practices 95–98
metadata hub (MDH) 67, 72, 74, 78, 84, 90, 92, 94, 95, 111, 112, 114, 125, 127, 130, 133, 135, 138
metadata management 23, 24, 31, 32, 34, 35, 63, 67–69, 71, 72, 77, 83, 88, 103, 110, 120, 124–127
Metastore 71, 84, 103, 110
model bias 145
MongoDB 5, 12
monitoring 3, 28, 31, 32, 41, 42, 45, 54–56, 67, 74, 78, 87, 92–95, 99, 109, 122, 123, 127, 133, 136–139, 143, 148, 149
MySQL 82

natural language processing (NLP) 5, 7, 14, 17
Netflix 13
New York Shield Act 27, 28
non-SQL databases 11

OLAP 72, 82
OLTP 82
open source tools 5, 9, 15, 17, 18, 66, 71, 134
optimization engine 14
Oracle 5, 82, 116, 125

PaaS: Platform as a Service 54
pattern recognition 14
personally identifiable information (PII) 25, 27, 28, 41, 65, 77–79, 85, 87, 91, 93, 105, 107, 114, 127, 133–135
Power BI 8
predictive analytics 3, 4, 15
predictive modeling 7, 14, 23
presentation layer 8, 16
private health information (PHI) 85, 93, 105, 107, 108, 114, 119, 127

QlikView 8, 16

RAID storage 60
raw data 3, 7, 17, 74, 84, 104, 111, 112, 115–117

real-time data 6, 18, 21
refined data 52, 85, 99, 112, 114, 116
relational databases 4, 11, 12
return on data (ROD) 5, 44
risk management 35, 39–57, 66, 70, 121, 123, 129, 140, 143–149
risks of non-compliance 122

SaaS: Software as a Service 10, 54–56
sandbox 17, 73, 74, 79, 92, 104, 107, 117, 118, 129, 135
schema-on-read 53
schema-on-write 53
semantics 4, 5, 14, 15, 22, 47
sensitive data 78, 85, 89, 91, 100, 104–108, 111, 114, 115, 119
Snowflake 4, 5, 11, 12, 57, 63
Snowflake schema 11
SOLR 11, 13, 17
Spark 4, 5, 11, 13, 131, 132
split data design 118, 119
SQL databases 5, 11, 12
star schema 11
statistical analysis 15, 16
structured data 7, 11, 14, 52
subject matter experts 56, 73, 97, 123

Tableau 8, 16, 63
Tamara Dull 52
Teradata 82, 103, 116
third-party data provider 90, 123
TOGAF 47, 48–51
tokenization 72, 85, 88, 90, 91, 100, 101, 107, 108, 110, 130, 134, 135
trusted data 70, 73, 90, 93, 123
trusted source 90, 94, 95, 125, 127, 133, 137–139
Twitter 4

unstructured data 4, 5, 11–15, 17, 63, 70, 99, 100

validated data 84, 85, 112