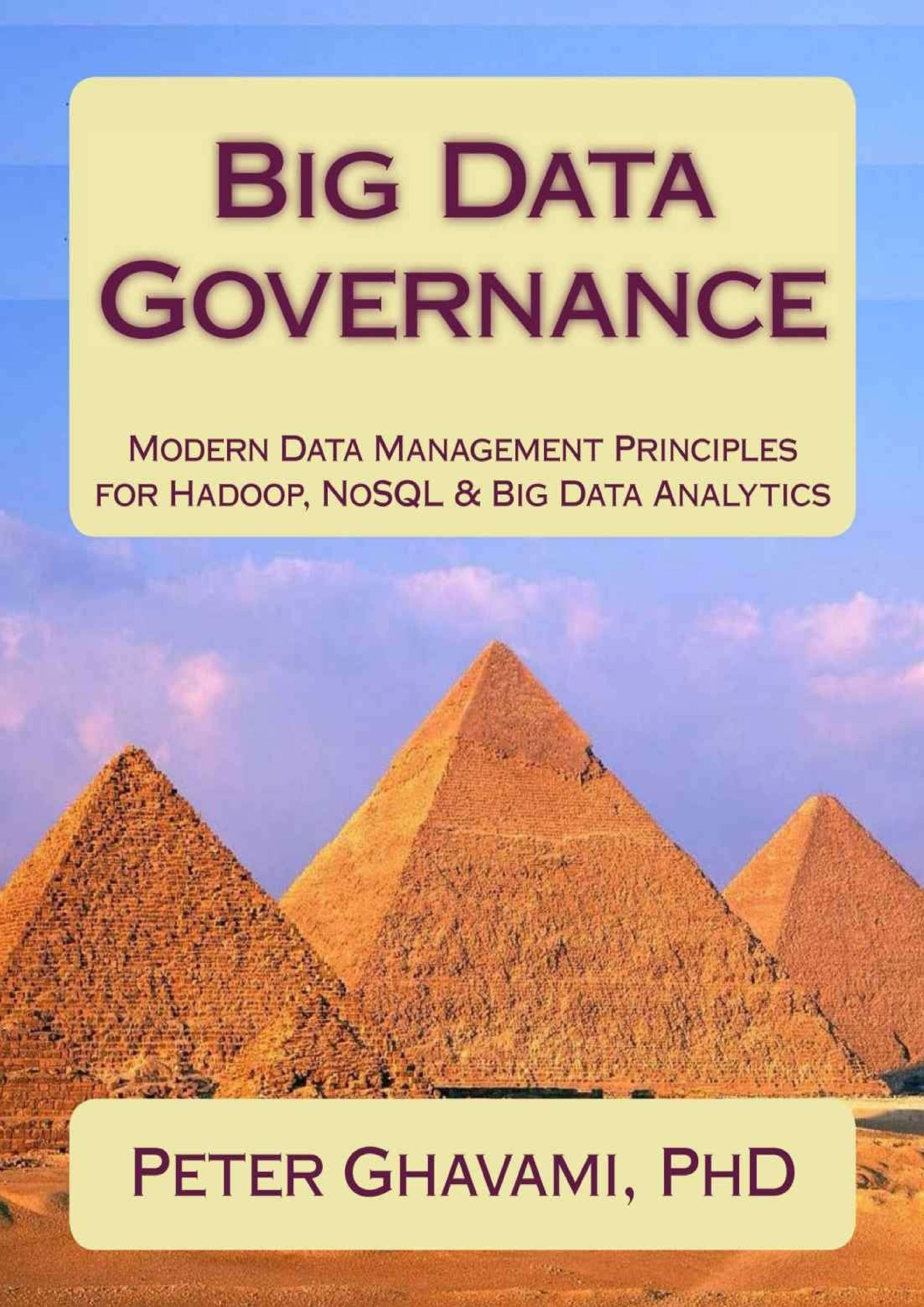
BIG DATA GOVERNANCE

MODERN DATA MANAGEMENT PRINCIPLES
FOR HADOOP, NOSQL & BIG DATA ANALYTICS



PETER GHAVAMI, PHD

BIG DATA GOVERNANCE:

Modern Data Management Principles for Hadoop, NoSQL & Big Data Analytics

Peter K. Ghavami, PhD

ISBN: 1519559720

ISBN-13: 978-1519559722

BIG DATA GOVERNANCE:

Modern Data Management Principles for Hadoop, NoSQL & Big Data Analytics

Peter K. Ghavami, PhD

Peter K. Ghavami, Ph.D.

Peter.Ghavami@Northwestu.edu

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.

Keywords: 1. Big Data Analytics, 2. Data governance, 3. Hadoop, 4. Data Security, 5. Data Management, 6. Data Lifecycle Management

ISBN: 1519559720

ISBN-13: 978-1519559722

BIG DATA GOVERNANCE

Modern Data Management Principles for Hadoop, NoSQL & Big Data Analytics

Peter K. Ghavami, PhD

First Edition

2016

Acknowledgements

This book was only 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 deal 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

January 2016

To my beautiful wife Massi,

whose unwavering love and support make these accomplishments possible and worth pursuing.

CONTENTS

INTRODUCTION

Purpose

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 it harnesses value from it.

While Hadoop promises 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 its 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 to 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 life cycle management, the chance of harnessing this value and leveraging from data will be very limited.

According to expert reports, in 2010 there were around 1.2 zetta bytes (1.2 billion gigabytes) of data worldwide. In 2011 that number jumped to 1.8 zetta bytes. It's expected that the global volume of storage will double every 12-18 months. In 2015, it's expected that the world will accumulate near 8 zetta bytes of data. The majority of this data is unstructured in form of PDFs, spreadsheets, images, multimedia (audio, video), 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 cost to implement. 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 other companion books by the same author: 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: https://www.createspace.com/4772104 . The second companion book is titled "Big Data Analytics Methods: Modern Analytics Techniques for the 21st Century". It can be found on Amazon.

This book consists of four major sections: Section I offers an overview of big data and Hadoop. Section II is an overview of big data governance concepts, structure, architecture, policies, principles and best practices. Section III presents the best practices in big data governance policies. Finally, Section IV includes a ready-to-use template for governance structure written in a flexible format such 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. There is an electronic slide deck available which can be purchased online at Amazon.com. You may adapt these slides and modify them to fit your particular enterprise and situation making it your own.

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

SECTION I:

INTRODUCTION TO BIG DATA

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 a big data analytics, the study of large data sets was called *data mining*. But, big data analytics has come a long way since a decade ago and is now gaining popularity thanks to 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 amount of data is being generated.

Big Data is known by its three key attributes, known as the three V's: Volume, Velocity, and Variety. The world 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 volume of data collected from people, internet and machines we're challenged with making increasingly analytical decisions quickly from a large set 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.

It's reported that on average most large companies in the U.S. have accumulated an average of 1.5 PetaBytes (1.5PB) of data by 2015. A few companies now offer Hadoop data vaults and warehousing products including Cloudera, MapR and HortonWorks. The NoSQL (Not only SQL) and non-relational database movement such as Hadoop, Lucene and Spark offer new data management tools that allow storage and management of large structured and non-structured data sets. But the biggest challenges, big data governance remains mostly uncharted territory at this time.

The variety of data is also increasing. For example, the 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) to include 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 burn and wounds, digital images of dental x-rays, waveforms of brain scans, electro cardiogram (EKG) and the list goes on.

Big Data is characterized by 3V's: Velocity, Volume and Variety. This underscores the large volume of data that is being collected and stored by our digital society; the rapid pace at which data is generated by consumers, smart devices, sensors and computer; the variety of data formats which in almost every industry spans a wide range of data from text to images, web and multimedia formats.

IDC[1] predicts that the worldwide volume of data will increase by 50X from 2010 to 2020. The world volume of data will reach 40ZB (Zetta bytes)[2]. The firm predicts that by 2020, 85% of all this data will be new data types and formats. There will be a 15X growth in machine generated data by 2020. The notion of all devices and appliances generating data has led to the idea of the internet of things, where all devices communicate freely with each other and to other applications through the internet. McKinsey & Company predicts that by 2020, Big Data will be one of the five game changers in US economy and 1/3rd of the world data will be generated in the US.

How will we manage and govern this vast and complex sea of data? What will be the costs of poor or no 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 will 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 steaming data and customer sentiment data about you. It includes social media data including Twitter, Facebook and local RSS feeds about your organization, your people and products. All these varieties of data types can be harnessed to provide a more complete picture of what is happening in delivery of 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 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 form of narrative text, emails or audio recording. 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 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 are going to be stored using open source and Hadoop technology. However, understanding sentence structure, context and relationships between business terms are critical to detect the semantics and meaning of data.

Given the rising volume of data and the demand or high-speed data access that can handle analytics some IT leaders are contemplating to invest in dedicated analytics platform like Hadoop, Non-SQL databases and open source tools for data transformation. Transactional systems and reporting platforms are 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.

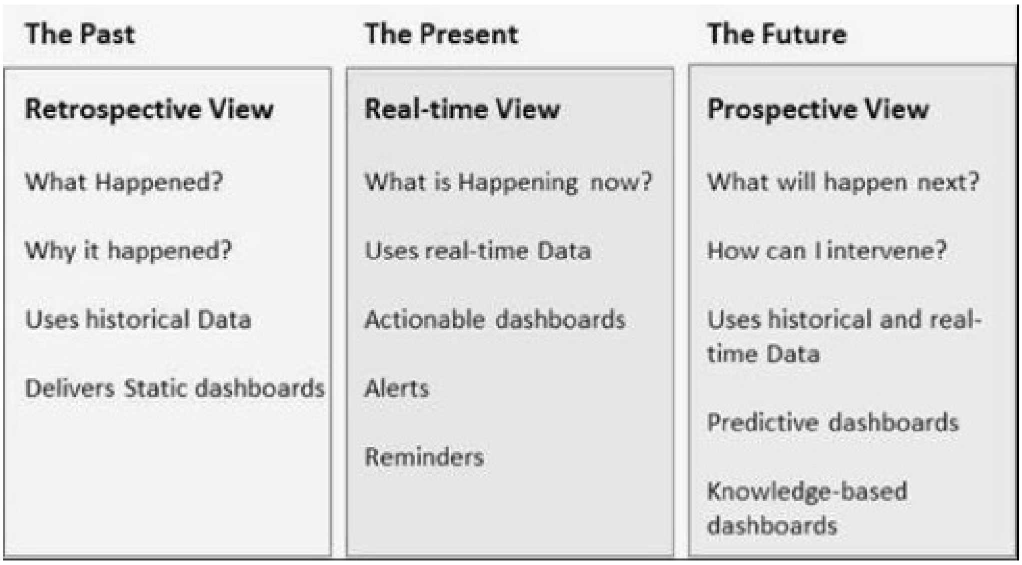
While Big Data Analytics promises phenomenal improvements in every industry, as with any technology acquisition, we want to take prudent steps towards adoption. We want to define criteria for success, gauge Return on Investment (ROI), and its data-centric metric that I define as Return on Data (ROD). A successful project must demonstrate palpable benefits and value derived from new insights. Implementing Big Data Analytics will be necessary and standard procedure for most organizations as they strive to identify any remaining opportunities in improving efficiency, strategic advantage and cutting costs. Managing data and data governance are critical to success of big data analytics initiative 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 titled: “Lean, Agile and Six Sigma IT Management”, successful implementation of big data analytics also requires team effort combined with lean and agile practices. 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 in needed for successful implementation. 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 which are 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 termporal dimensions of analysis: The retrospective analysis, the Real-time (current time) analysis and Predictive analysis. The retrospective analytics can explain and provide knowledge about the events of the past, show trends and help find root-causes for those events. The real-time analysis shows what is happening right now. It works to present situational awareness, alarms when data reaches certain threshold or send reminders when a certain rule is satisfied. The prospective analysis presents a view in to the future. It attempts to predict what will happen, what are the future values of certain variables. The figure below shows the taxonomy of the 3 analytics dimensions.



The Past	The Present	The Future
Retrospective View	**Real-time View**	**Prospective View**
What Happened?	What is Happening now?	What will happen next?
Why it happened?	Uses real-time Data	How can I intervene?
Uses historical Data	Actionable dashboards	Uses historical and real-time Data
Delivers Static dashboards	Alerts	Predictive dashboards
	Reminders	Knowledge-based dashboards

The 3 temporal dimensions in data analysis

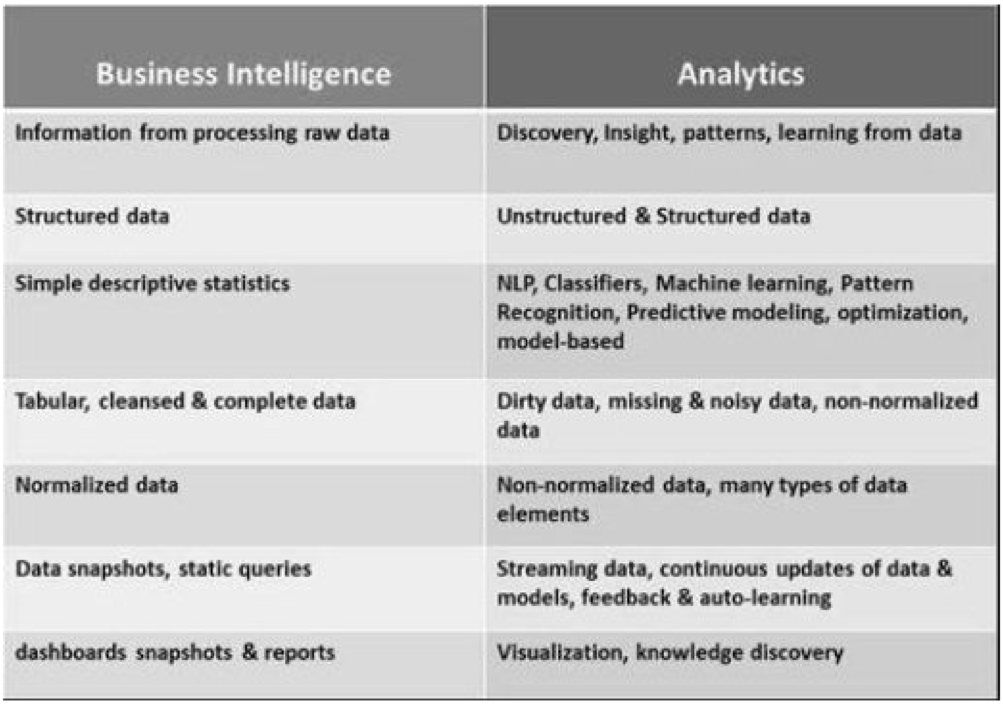
The Distinction between BI and Analytics

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. Often, data that is out of range or outlier are removed before processing. Data processing uses simple, descriptive statistics such as mean, mode and possibly trend lines and simple data projections to extrapolation 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 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 view 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 discovery and knowledge. The patterns have a geometric shape and these shapes as some data scientists believe, have mathematical representations that explain the relationships and associations between 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, Analytics is highly model-based. A data scientist builds models from data to show patterns and actionable insight. Feedback and machine learning are concepts found in Analytics. The table below illustrates the distinctions between BI and Analytics.

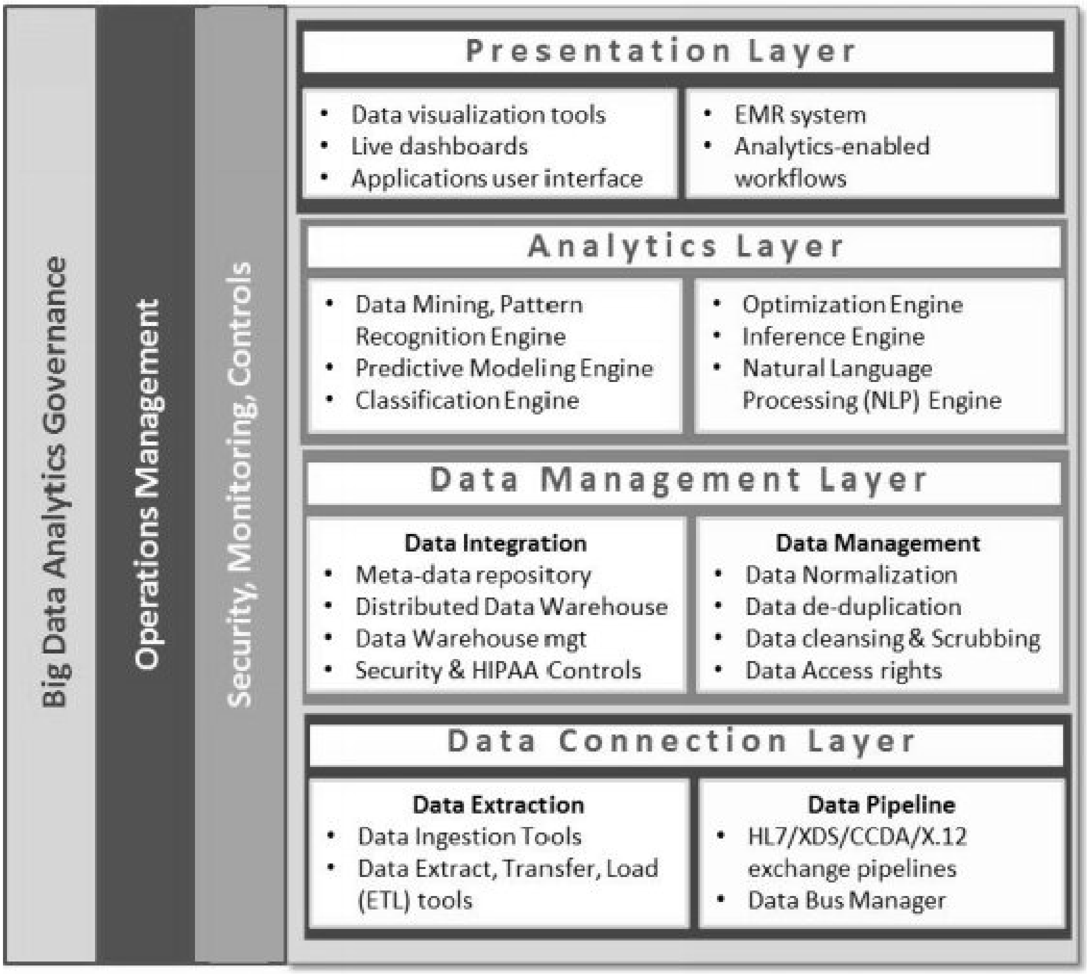


Business Intelligence	Analytics
Information from processing raw data	Discovery, Insight, patterns, learning from data
Structured data	Unstructured & Structured data
Simple descriptive statistics	NLP, Classifiers, Machine learning, Pattern Recognition, Predictive modeling, optimization, model-based
Tabular, 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

The differences between Business Intelligence and Data Analytics

Analytics Platform Framework

When considering building analytics tools and applications, a data analytics strategy and governance is recommended. One of the strategies 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 the entire enterprise analytics applications. The 4-layer framework consists of a data management layer, an analytics engine layer and a presentation layer as shown in the figure below.


Big Data Analytics Governance
Operations Management
Security, Monitoring, Controls
Presentation Layer
• Data visualization tools
• Live dashboards
• Applications user interface
• EMR system
• 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
• Meta-data repository
• Distributed Data Warehouse
• Data Warehouse mgt
• Security & HIPAA Controls
Data Management
• Data Normalization
• Data de-duplication
• Data cleansing & Scrubbing
• Data Access rights
Data Connection Layer
Data Extraction
• Data Ingestion Tools
• Data Extract, Transfer, Load (ETL) tools
Data Pipeline
• HL7/XDS/CCDA/X.12 exchange pipelines
• Data Bus Manager

The 4-layer Data Analytics Framework

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

The analytics layer may be implemented using SAS and 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, SpotFire, or other data visualization 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) are critical to successful implementation.

The key management framework that spans the entire 4 layers is data management and governance. This framework which is the subject of this book 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 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 meta-data in all source data repositories. This layer starts with making an inventory of where the data is created and stored. The data analysts might implement Extract, Transfer, and 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.

Data Management layer

Once the data has been extracted, data scientist must perform a number of functions that are grouped under 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 architecture in the upcoming sections.

In this layer, we must pay attention to HIPAA standards for security and privacy. 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.

Data Scientists may apply other data cleansing programs in this layer. They might write tools to de-duplicate (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 source foundation have developed alternate database architectures that allow parallel execution of queries, read, write and data management.

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

The traditional data management approaches have adopted centralized data warehouse architectures. The traditional data warehouses, 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 data base and tabular structures and data to be normalized. 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.

Star Schema: The more advanced data warehouses have adopted Kimball's Star Schema or Snowflake Schema to overcome the 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 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). Snowflake schema is similar to Star schema but its tables are normalized.

Star schemas are de-normalized data 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 (discussed below), the rules of data integrity are not strongly enforced.

Non-SQL Database schema: In order to liberate data from relational constraints, several alternate architectures have been devised in the industry as explained in the next sections. 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 like NoSQL, Hadoop, Cassandra, Lucene, SOLR, Spark and other commercial adaptations of these solutions.

NoSQL database means Not Only SQL. A NoSQL database provides storage mechanism for data that is modeled 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 either in a tree, columnar, graph or key-value pair. However, a NoSQL database can support SQL-like queries.

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.

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 .pdfs, HTML, Microsoft Word, and OpenDocument can be indexed by Lucene as long as their 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 include Netflix and Amazon. It supports a 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. Its 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 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: Many 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 carry 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.

Optimization engine is used to optimize and find the best possible solution to a given problem. 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 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 data as data 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 structured text. When combined with other engines, the extracted text can be analyzed for a variety of applications.

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 prediction about patient re-admissions 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. These tools enable data scientists to identify shape and patterns in data, perform correlation analysis and clustering of 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 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 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.

Machine Learning: Machine learning is a discipline outgrowth of Artificial Intelligence. Machine learning methods learn from data patterns and can imitate intelligent decisions, perform complex reasoning and make predictions. Predictive analytics use machine learning techniques to make predictions. These algorithms process data at the same rate as they receive data in real time, but they also have a feed-back loop mechanism. The feedback loop takes the output of the analytics system and uses it as input. By processing and comparing their output to the real world outcome, they can fine-tune their learning algorithms and adapt to new changes in data.

Statistical Analysis: Statistical Analysis tools would 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 would include Regression (Trending) analysis, correlation, Chi-square, maxima and minima calculations, t-Test and F-test, and other methods. For more advanced tools, one can use an open source tool developed at Stanford called MADLIB, (www.madlib.net). Also, Apache Mahout includes a library of open source data analytics functions.

Forecasting & 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 regression line and the parameters of the equation 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 Multi-variable Analysis of variance (MANOVA), Mean Square Error (MSE) calculations, and residual calculations.

Predictive Analytics is intended to provide insight into future events. It is model-driven and includes methods that produce predictions using supervised and unsupervised learning. Some of the methods include neural networks, PCA 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 provides a rich set of analytics functionality. These algorithms include Logistic Regression, Naive Bayes, Decision trees and Random forest, regression trees, linear and non-linear regression, Time Series ARIMA[3], ARTXp, 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 and Ridge regression, Conditional Random Fields (CRF) methods, and Cox Proportional Hazard models. For more details refer to the Data analytics book written by the same author.

Pattern Analysis: Using machine learning algorithms researchers can detect patterns in data, perform classification of patient population, and cluster data by various attributes. The algorithms used in this analysis include various neural networks methods, Principal Component Analysis (PCA), Support Vector Machines, 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 for dashboards.

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, variety of pie charts and parallel charts and many more visualization forms.

Advanced presentation layers include data visualization tools that allows data scientist to easily visualize 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 which are revolutionizing how we view data.

Now we’re ready to turn our attention to data governance and best practices in big data management.

The Diverse Applications of Big Data

While Hadoop was designed to handle big data and distributed processing of massive amounts of data, it has other diverse use cases. In fact there are 6 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 (like system logs, network security logs and access logs and so on). Since storage is inexpensive and Hadoop does not mandate any relational structure upfront, you can be free to bring any data into the Hadoop.

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 data?

The ability and confidence to answer these questions lies in data governance and is the premise of this book.

2. **Stage Data.** Use Hadoop as 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 transform raw data into relational data formats that fit the data warehouse structure.

3. **Process Unstructured & 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 a common practice now for many organizations. In particular when you want to process social media data which is typically semi-structured, Hadoop offers the ideal tools to parse and process textual information that deliver customer sentiment analysis.

4. **Process any Data.** Use Hadoop to combine both internal and external data for processing. 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 & 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 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 petabyte scale and analyze it on the fly as it gets stored in Hadoop in real time.

6. **Virtualize Data.** The goals 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 big data lake or even spread among many excel spreadsheets. This vision is intended to address an endemic IT problem of not knowing where does the information in the organization reside and how can we access it 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 overtime provide robust solutions for data virtualization. Keeping a 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 segments, I'll review these two references briefly to set the stage for the next sections.

Data Management Body of Knowledge (DMBOK)

Data Management Association (now referred to as DAMA International) is an organization that devotes its energy to advance 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 international scale. It has published a set of guidelines that are publically available under Data Management Body of Knowledge Version 2.0 (DMBOK2)[4].

DMBOK2: What is it?

DMBOK2 is a collection of processes, principles & 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 division level. It was developed by Data Management Association (DAMA) which has providing data management guidance since 1991. The PMBOK1 guide was released in 2009 and PMBOK2 in 2011. The guidelines were developed for IT professionals, consultants, data stewards, analysts, and data managers.

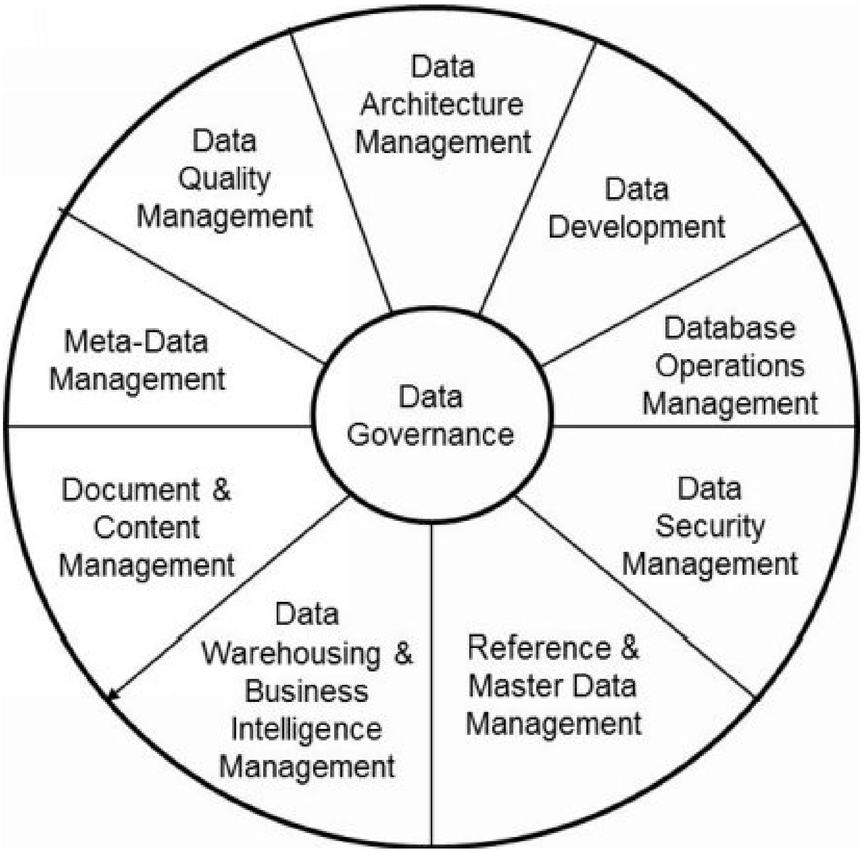
There are eight key reasons to consider DMBOK2 in your organization. The key words are in Caps and italicized:

Eight Reasons for DMBOK2

1. It brings *Business* and *IT* together at the table with a common understanding of data management principles and vocabulary.
2. It connects *Process* and *Data* together so the organization can develop the appropriate processes for managing and governing its data.
3. Defines *Responsibility* for data and accountability for safeguard, protection and integrity of data.
4. Establishes guidelines and removes ambiguity around change and *Change* management.
5. Provides *Prioritization* framework to focus the business on what matters the most.
6. Considers regulatory and compliance requirements to provide intelligent *Regulatory* response for the organization.
7. Helps identify *Risks* and risk mitigation plans.
8. It considers and promotes the notion of *Data as an Asset* for the organization.
9. It links data management guidelines to process *Maturity and Capability*.

The DMBOK2 covers 9 domains of data management as shown in the next diagram. These 9 domains are:

1. Data Architecture Management
2. Data Development
3. Database Operations Management
4. Data Security Management
5. Reference & Master Data Management
6. Data Warehousing & Business Intelligence Management
7. Document & Content Management
8. Meta-Data Management
9. Data Quality Management


Data Architecture Management
Data Development
Database Operations Management
Data Security Management
Reference & Master Data Management
Data Warehousing & Business Intelligence Management
Document & Content Management
Meta-Data Management
Data Quality Management
Data Governance

DAMA International DMBOK2

The contents of this book all related to these 9 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.

Document & Content Management deals with data storage principles such as acquisition & 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.

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

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.

Meta-data management is about integrating, controlling and delivering meta-data. It also includes guidelines on meta-data 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.

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.

Reference & Master Data Management focuses on managing gold versions of data, replicas of data and data lineage. It provides guidelines for standardized catalogues for external codes, internal codes, customer data, product data and dimension management.

Data Maturity Model (DMM)

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

In 2014, CMMI Institute released a model with a set of principles to enable organizations improve data management practices across the full spectrum of their business. 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 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 CMMI maturity model to measure process maturity of an organization, improve efficiency and productivity, 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 is your organization at? 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 5 levels of maturity as shown in the following table. The source of the table is CMMI Institute's standard for Data Maturity Model. For more information, I recommend that you visit their website[5].

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 standardized set of processes, roles and responsibilities and 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 business needs change.

Level	Maturity Levels	Key Characteristics of Processes and Functions
Level 1	Initial	• Ad hoc, inconsistent, unstable, disorganized, not repeatable • Any Success achieved through individual level
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 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 on overall organizational performance • Defined feedback loop between measurement and process change

The DMM model offers six domain (principle areas) for maturity assessment and process improvement. The six domains are: Data Management Strategy, Data Governance, Data Quality, Data Operations, Platform & Architecture, and Supporting Services.

The following table represents the highlights of each domain and a brief description of each domain is provided below:

Data Management Strategy
• Data management strategy • Data management function • Business case for data management, • Funding, • Communications
Data Governance
• Meta-data 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 & 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 Management Strategy is concerned with how the organization views data management function, business case for data management, funding, communications among data management roles and adherence to data management strategy.

Data Governance considers principles such as Meta-data management, implementation of 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 definition and provider management.

Platform & Architecture is concerned with activities and procedures associated with data integration, overall data architecture creation, architecture standards, 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.

SECTION II:

BIG DATA GOVERNANCE FUNDAMENTALS

Introduction

Big Data analytics is now a common strategy for many corporations to identify hidden patterns and insight from data. The 3 V's of Big Data Analytics are stretching the limits if not breaking the traditional frameworks around data governance and data management and data lifecycle management.

In the current era of Big Data, organizations are collecting, analyzing and differentiating themselves based on analysis of massive amounts of information, data that comes in many formats, from various sources and at very high 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.

Top 10 Data Breaches

The list of security breaches is a changing and dynamic list as everyday some new breach is reported in the news. 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 healthcare sector. Here are some of the most egregious breaches that led to substantial data loss:

BlueCross 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 Jan. 2014.

In early 2009, Heartland Payment Systems, a payment processing company in New Jersey reported the largest ever data breach that affected an American company. Around 130 million credit and debit cards were lost to hackers. A malware planted in the company's network recorded card data that arrived from more than 250,000 retailers across the country so the impact was substantial.

Sony Online Entertainment Services and Sony experienced repeated breaches. In 2011, more than 102 million records were compromised when hackers attached the PlayStation Network that links Sony's home gaming consoles and the servers that host the large multi-player online PC games. After the breach was discovered, the Sony PlayStation Network went dark worldwide for more than 3 weeks. Some 23 million accounts including credit card, phone numbers and addresses were compromised in Europe alone. Soon after, about 65 class action lawsuits were brought against the company at the cost of $171 million.

This is not to mention 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.

In 2008, the National Archive and Records Administration experienced a data breach that affected 76 million records of government employees, of U.S. military veterans including their Social Security Numbers. The organization shipped a failed disk drive containing this information out to be scrapped but the drive was never destroyed. Since then NARA has changed its policies about protecting PII (Personally Identifiable Information).

May 2014 – Thieves stole 8 computers from an office which gave them access to personal information of roughly 350,000 individuals. The data was not encrypted.

Feb. 2015- Anthem, formerly known as WellPoint, the 2nd largest health insurer in the U.S. reported 80 million records stolen by hackers. Stolen data included names, addresses, date of birth, employment histories and Social Security numbers.

In 2011, Epsilon reported a breach that compromised 60 million to 250 million records. The Texas-based marketing firm which handles email communications for more than 2,500 clients worldwide reported that its databases related to 50 of its clients had been stolen.

Email addresses of over 60 million of clients was stolen which included customers from a dozen key banks, retailers and hotels including Best Buy, JPMorgan Chase and Verizon.

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.

Similar to HomeDepot, Target reported a breach of its databases which caused tremendous customer complaints and damage to the company's image and consumer trust.

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 massive 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.

One of the worst security breach cases occurred in 2007 at TJX Companies, the parent company of the retailer, TJ Maxx and HomeGoods. At least 45 million credit and debit card numbers were stolen over an 18 month period, though some put the estimate 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.

With the advent of Hadoop and Big Data, the accumulation of vast amount 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.

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 that serves the business needs but a clear and hard defense against hackers. For example, policies might state that data at rest and 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 organization. Studies of prior breaches have shown that if the affected organizations had applied simple and basic principles of security and data governance, they could have prevented 80% of their data loss events.

By the time you read this book, you can think of many additional examples – unfortunately. I raise this point early in this chapter to highlight the much needed focus and investment in data governance across organizations globally.

What is Governance?

There are many definitions for data governance. One popular definition is this: 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 with in the IT Governance framework.

Along with data governance, you might have seen references to data stewardship and 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 which 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 they treat data. Another definition describes a job role of data steward is 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) 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 heavy, iron-fist 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 right behaviors towards data governance in an organization. A data governance framework must support existing and new processes to ensure the proper management of data, proper production and usage of data through its life cycle.

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 A) Transparent, B) Supportive and C) Collaborative.

Why Big Data Governance?

Another perspective of big data governance is to think of it as 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 & consistently managed through enterprise. It ensures data can be trusted and people are held accountable for low data quality.

Big data governance brings people & technology together to prevent costly data issues and ensure enterprise data is more efficiently managed and utilized. Finally, it enables using 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:

- Compliance and less regulatory problems. Without metadata compliance and regulatory reports are unreliable without quality data.
- Increased assurance and dependability of knowledge assets. With governance, you can trust your data and place decisions on solid data.
- Improved information security & 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 raise the value and return on data (ROD).

Data Steward Responsibilities

Data Stewards are formally accountable for defining rules about new data vs. existing data. Their responsibilities include defining rules around logical and physical data modeling, meta data definition, 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.

Another key responsibility includes defining and administering Data Usage rules and enforcing them, in particular defining business rules, classification of data, protection of data, retention rules around data, compliance and regulatory requirements.

Corporate Governance

Corporate governance refers to the mechanisms, processes, policies and controls 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 that big data governance must be in alignment.

Big Data Governance Certifications

Many organizations opt to apply and receive certification of their data governance structure. Certification can be received 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 trail 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 the company who can't govern and manage its big data.

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 organization $5.4M dollars. Another study claims that the cost of cybercrime to the US economy is $140B 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 $24B 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 to enterprise governance and IT governance, both. 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 data, 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 sent 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.

Strategic Data Governance, or Tactical Data Governance?

With Big Data and 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 & value but with it comes responsibility, and bigger need for governance. There are 3 primary levels or perspectives to data governance: Tactical, Strategic and Operational.

Operational data governance focuses on daily operations and 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 stewards or data guardians in the organization.

Tactical Data Governance is concerned with the current and immediate issues associated with management of data and implementation of governance, 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 steward 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 it will encounter as the data grows throughout the organization and the maintenance of it; the long term and future perspective of data governance. Strategic data management treats data as a corporate asset and works to add as much 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 including 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 latest standard, TOGAF Version 9, released in 2011. The goal of TOGAF is to provide guidance and open forum to enable interoperability and boundary-less flow of information among vendor applications and systems.

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

- The Preliminary phase
- Architecture vision
- Business architecture
- Information systems architecture
- Technology architecture
- Opportunities & 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 has provided a set of sample principles which include a wide range of information architecture topics, but for 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 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 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.

Data Lake vs. Data Warehouse

We often hear references to the data lake and data warehouse together as being synonymous and data lake is just another incarnation of a data warehouse . The truth is that there are major differences between data lake and data warehouse. The data lake can be thought of 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, semi-structured data. The data structure and requirements are not defined until the data is needed.

A data lake is not a data warehouse. They're optimized for different purposes and to provide the best tool for that purpose.

Tamara Dull in one of her blogs provides a table to highlight differences as shown below:

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 our choice of data lake vs. data warehouse? Here is a discussion of these points:

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

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 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 to load the data as is, raw-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 advantage of 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 only 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, 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, 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 efforts and attention must be spent towards 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, data lake is typically used by a handful of data scientists and data engineers. While the number of data lake users are steadily on the rise, a data lake is not designed to handle thousands of user accounts and log-ins 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 utilize the data lake through new analytics products and data pipelines that are developed by data scientists and data engineers.

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.

History of Hadoop

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 time. 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 trusted cluster of servers, serving trusted users and trusted applications, security would not be a concern. However, the more careful inspection of the architecture has shown that all those assumptions are 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 really 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 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 read 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 professional voiced concerns about insider threats by users or applications that could perform malicious use. For example, a malicious user could write a program to impersonate other users and their Hadoop services, or impersonating the HDFS or MapReduce jobs, deleting everything in Hadoop.

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 – Incorporates Simple Authentication & 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 – The Hadoop administrators could use this feature to control access to HDFS files by file permissions, namely providing Access Control List (ACLs) 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 were 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 data moves necessary to manage the processing to appear as 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 segments offer a more detailed view into Hadoop, MapReduce and how Hadoop distributed processing works. You can skip the following segments without loss of 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. 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 HDFS filesystem. It keeps the directory tree of all files in the filesystem and tracks where the data files is stored across the cluster. It does not itself store the data but tracks where they're stored.

A user's 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 returns the list of DataNode servers where the data is stored.

The NameNode is a single point of failure in 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 a network mounted storage.

- Monitor the disk space available to NameNode. If the available space is getting low, add more storage.
- Do not host Job Tracker, DataNode and Task Tracker services on the same server.

A DataNode stores data in the Hadoop File System. A file system has more than one DataNode, with data replicated across them. When 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 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.

A TaskTracker is a node in the cluster that accepts tasks, such as Map, Reduce and Shuffle operations from the Job Tracker. Every TaskTracker contains several slots, which indicate the number of tasks that TaskTracker can accept. When JobTracker looks to find where it can schedule a task within the MapReduce operations, it first for an empty slot on the same server that hosts the DataNode containing the data. If the 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. 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 the application into many small fragments, each can be executed on any node on the cluster, and then get re-aggregated 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 A) the output may be a different key from the input; and, B) 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 the new list of the transformed output data (the reduced output). The MapReduce engine such 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, as it stores data in multiple locations which comes in handy to run 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 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 data
3. The JobTracker then finds the available processing slots among TaskTracker nodes
4. The JobTracker submits the task to the selected TaskTracker nodes
5. The Task Tracker 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: it may resubmit the job on another node or may mark that specific record as a data 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 an 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.

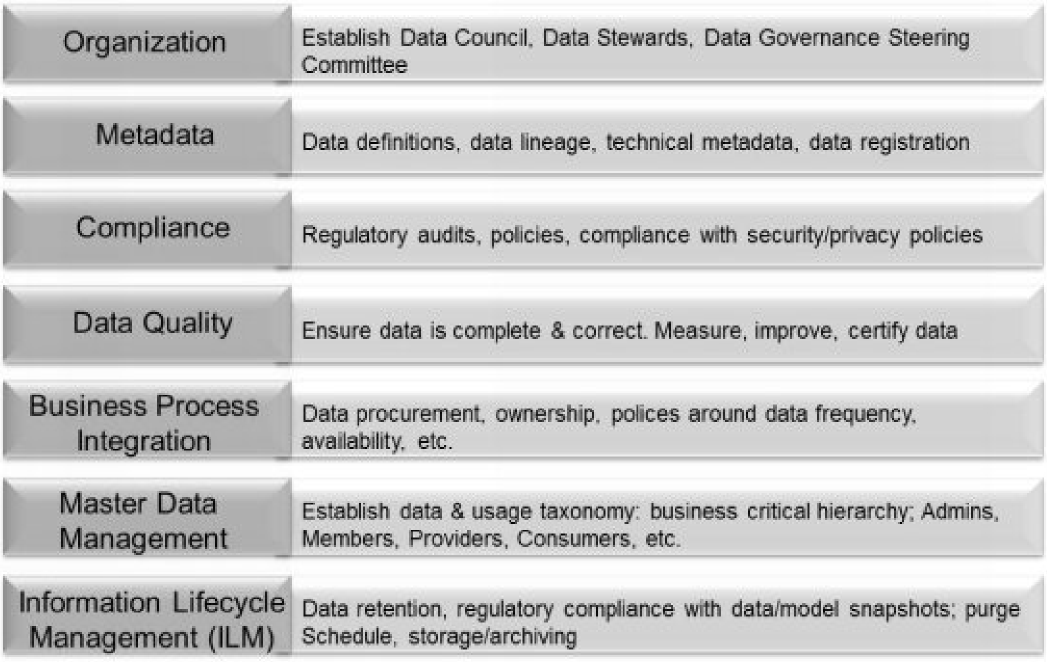
Security Tools for Hadoop

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. Other projects like Project Rhino and Knox Gateway promise to make internal changes in Hadoop itself.

The Components of Big Data Governance

There seven key components of big data governance framework. These include Organization, Metadata, Compliance, Data Quality, Business Process Integration, Master Data Management (MDM), and Information Lifecycle Management (ILM).



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 & correct. Measure, improve, certify data
Business Process Integration	Data procurement, ownership, polices around data frequency, availability, etc.
Master Data Management	Establish data & usage taxonomy: business critical hierarchy; Admins, Members, Providers, Consumers, etc.
Information Lifecycle Management (ILM)	Data retention, regulatory compliance with data/model snapshots; purge Schedule, storage/archiving

The 7 Components of Big Data Governance

Myths about Big Data & Hadoop Lake

Despite the huge investments made into technology, staffing 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”. Or it’s common to hear data people say: “We can put “anything” in the lake” and “we can use any tool to analyze data”.

Perhaps not expressed openly, but other myths 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”; “We’ll build the environment first and then apply governance”.

But the truth is an important position and expectation that without Big Data Governance from the very start, within less than 2 years the Hadoop data lake will be in a complete chaos … 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.

Enterprise Data Governance Directive:

The goal of this book is to show the components and tips on establishing both effective and lean data governance. 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?"

The answer is in this book. I've looked to bring an agile and lean data governance together in this book. To reiterate our goals a lean and minimally viable governance framework must include:

1. Organization
2. Metadata management
3. Security/Privacy/Compliance Policies
4. Data Quality Management

I'll cover these components plus more in this book starting from a tactical data governance framework and build toward a strategic data governance structure.

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 & analytics across the enterprise, not just limited to big data or Hadoop infrastructure. It's important that we 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

The first step towards 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 envelop on big data governance models such that our current models are not adequate, nor adaptable 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, 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 the entire data activities across the enterprise into governance. This step includes gathering data governance requirements and identifying the gaps for your organization before you get too far into implementing the Hadoop infrastructure.

The second step is to establish basic rules of governance and define where they are applied. The next 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.

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:

1. **Accountability:** Your plan should establish clear lines and role 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 data governance model and one that will have one of the highest returns on your investment. The goal of inventory is for 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 "data registry" tool or Meta Data 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 meta data tool as Meta Data Hub (MDH).

3. **Process:** Establish processes for managing and usage of data in Hadoop Data Lake, Big Data analysis tools, Big Data resource usage & quality monitoring. The model must define data lifecycle policies, procedures for data quality monitoring, data validation, data transformations and data registry.

4. **Rules:** Establish & train all users on rules & code of conduct. Creating clear rules for data usage, such as Data Usage Agreements (DUA), data registry and meta data management, quality management and data retention practices are important. I've included a set of rules as examples to the end of this section to illustrate a sample list of rules that your governance model might include.

Data Governance Tools

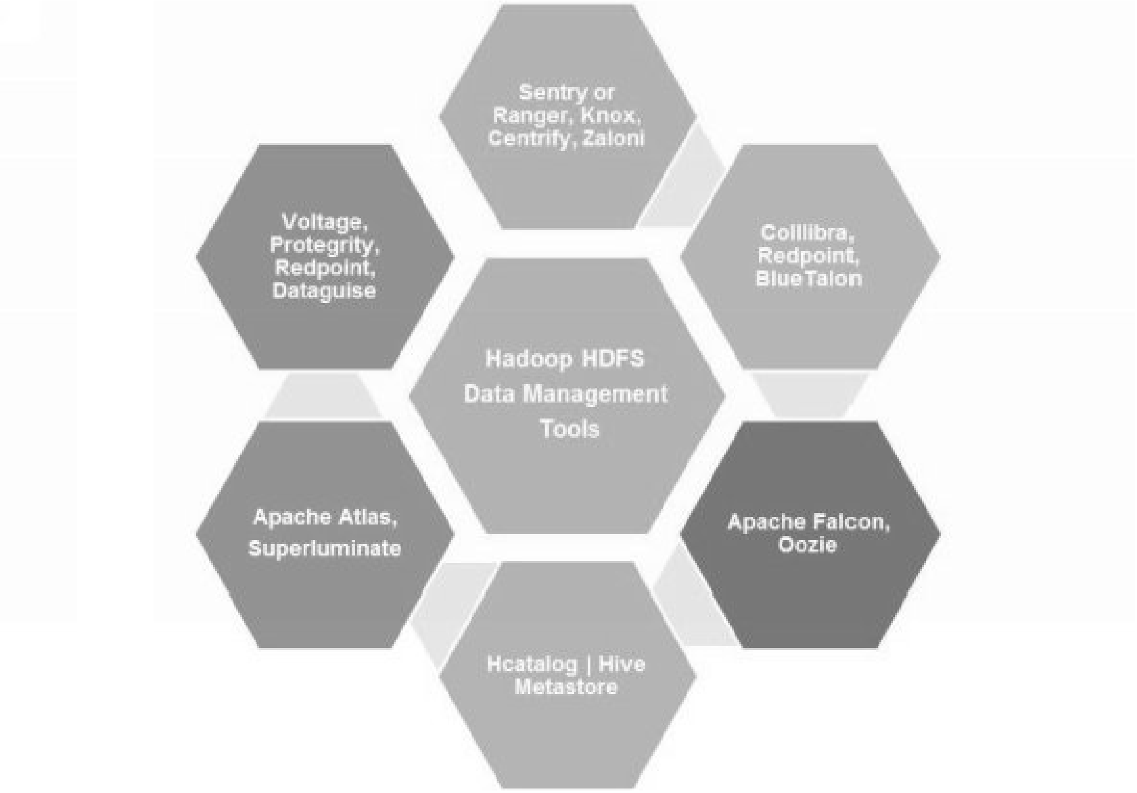
Hadoop was not designed for enterprise data management and governance. For quite some time there were major gaps in bringing Hadoop into the same level of governance as the rest of the organization. However, there are more tools available today and more tools are about to come to market soon.

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 that cover security, privacy, data management, quality and meta data management. That leaves you with choices of which products to select and integrate them into a seamless and consistent governance infrastructure. The advice and challenge for you is to select less than 2-3 tools that completely support the entire span of data governance policies and integrate them. Any more than 3 products can become hard to integrate, manage and support.

The second challenge is that the pace of innovation in big data world is quite rapid. New products emerge almost weekly that supersede prior products. So, be prepared to throw away your hard earned work on the existing tools, revamp them and adopt these new tools every 2-3 years for some time to come.

The diagram below 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

The Hadoop & Big Data Governance Tools Ecosystem

Big Data Governance Framework: A Lean & Effective Model

As alluded before the major elements of an effective, lean and agile governance model for big data includes four pillars: Organization, Data Quality Management, Meta data management and Policies. Here 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 organization, 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 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 components of the governance. Section IV, presents a framework that can be easily adopted as a starting point to form your customized 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 3rd party data and how the data must be purged upon it 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 Data Quality. Across the enterprise, we find either data that's missing, duplicated or simply filled with errors. The cost of data cleansing and resourced that it consumes are substantial. So, why don't 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 form of data validation checks are necessary every time data is imported into Hadoop, or transformed into a new format.

To effectively manage data quality, process automation can go a long way. So, we want to implement products such as Drools and comparable solutions to maintain data engineers' productivity and data analytics momentum.

Metadata Management

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

Registering data in a metadata store requires proper tools. So far, the open source tools like HCatalog, Hive and Metastore are meeting the very minimum functionality needed to handle meta data management. But, 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 useable", 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 the 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".

Business Glossary is a definitive dictionary of business terms, their attributes and relationships used across the 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", "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 3 primary categories of metadata:

1. Business metadata: Consists of business perspective of data, like business definitions, business processes, 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, extract information, audit controls, and

version maintenance.

3. Operational metadata: Includes operational aspects of data such as data process tracking, data event, 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 which I call Meta Data Hub (MDH). Providing training to Data Stewards on how to use the metadata hub so all data entered into metadata conforms to the necessary quality requirements. In addition the Data Stewards were 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 drive from data is enabled by metadata information that maintains the data relationships with each other.

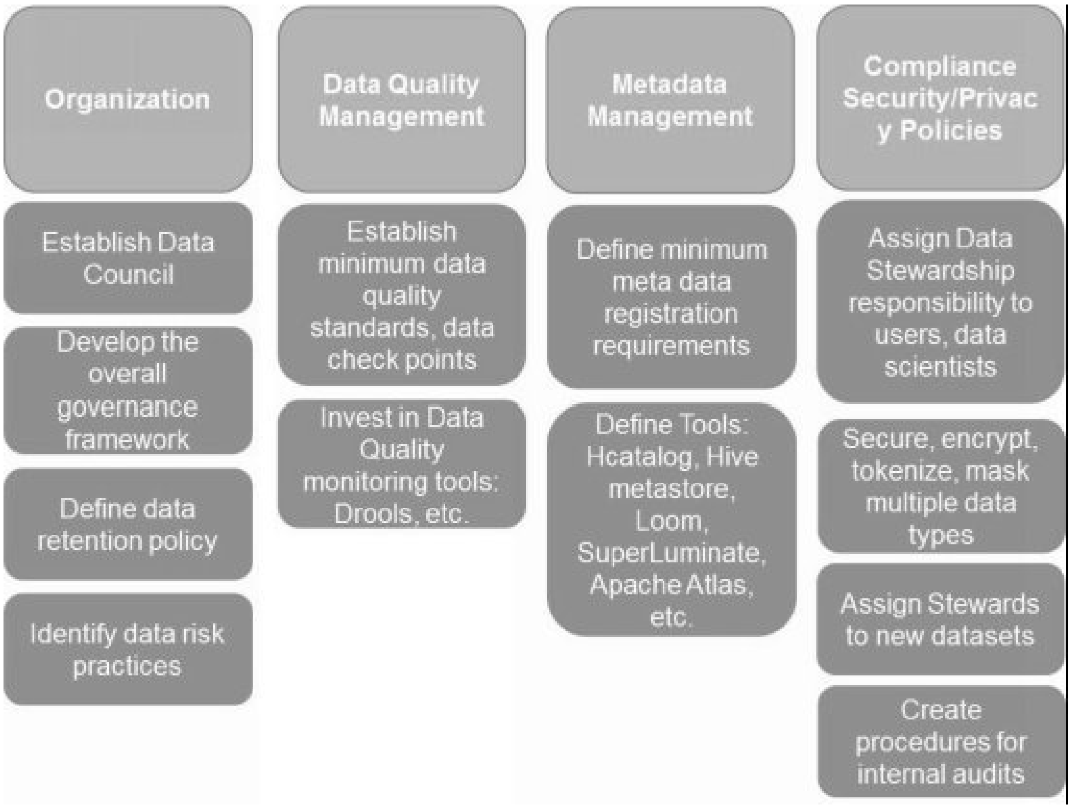
Compliance, Security, Privacy Policies

Policies must clearly define the accountability in the organization. It’s critical that the governance framework identifies the role and responsibilities of Data Stewards, data scientists, data engineers and general users for the organization.

The policies must also define security standards for encryption, tokenization, data masking and data classification. As part of accountability policies, it must define the process of assigning new datasets to the right 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 should be the ratio of storage volume per Data Steward. 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, namely at least 1 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.



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

The Four Pillars of Big Data Management

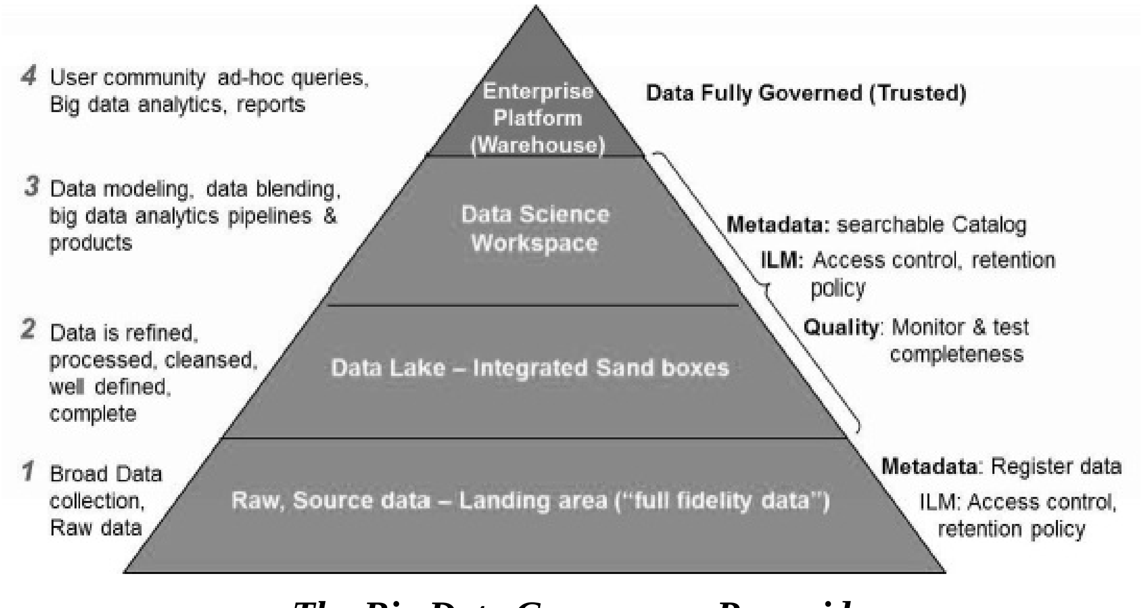
The Enterprise Big Data Governance Pyramid

One advantage of viewing big data governance as a Pyramid is that it exposes governance as a hierarchy and layers of policies. Consider the following diagram that 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.

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 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.

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 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 Meta Data Hub (MDH) and access controls.


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 &
products
Data Science Workspace
Metadata: searchable Catalog
ILM: Access control, retention policy
2 Data is refined,
processed, cleansed,
well defined,
complete
Data Lake – Integrated Sand boxes
Quality: Monitor & test completeness
1 Broad Data
collection,
Raw data
Raw, Source data – Landing area ("full fidelity data")
Metadata: Register data
ILM: Access control, retention policy

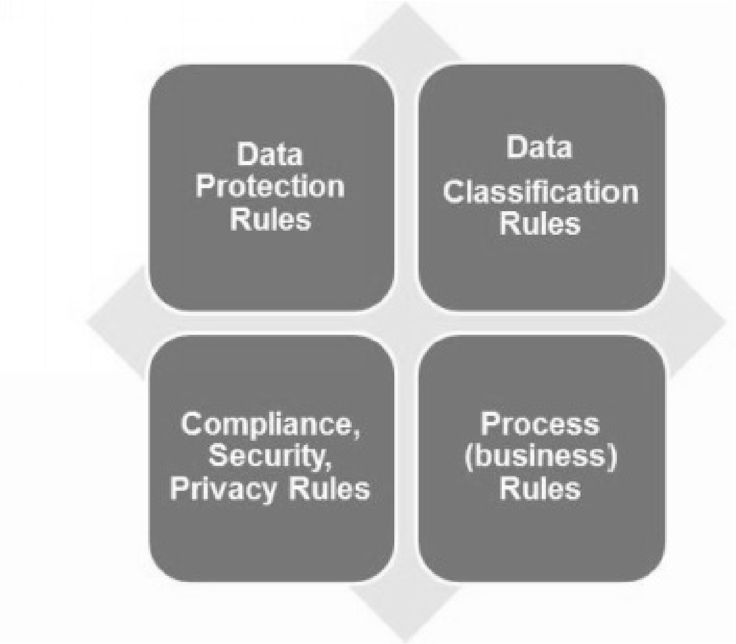
The Big Data Governance Pyramid

Introduction to Big Data Governance Rules

There are 4 types of rules 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 Rules
4. Process (Business) Rules

I'll discuss each of these rules in more detail in this section. Additional policies and configuration recommendations appear in Section IV.


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

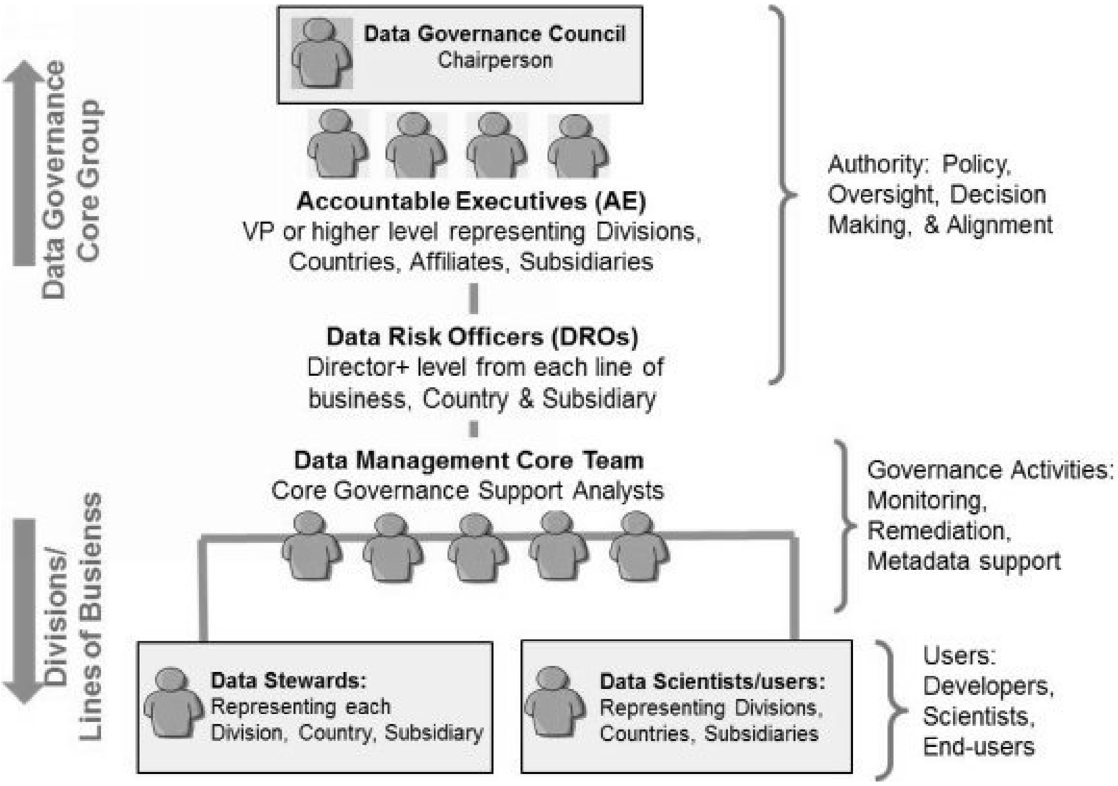
The Four Pillars of Big Data Governance Rules

Organization

Data Governance Council

In order to manage and enforce governance policies across the enterprise, we need to establish a cadre that supports the operational and implementation of big data governance framework. The cadre is a core team 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) but collaborate together collectively under the governance framework. While the Data Council meets on a regular basis (monthly cadence is recommended), the Data Stewards also meet but more frequently (preferably once a week) to share information and implement data governance policies in a federated model.

The diagram below illustrates a possible organizational chart that defines roles and interaction model between various data governance roles.


Data Governance Council
Chairperson
Accountable Executives (AE)
VP or higher level representing Divisions, Countries, Affiliates, Subsidiaries
Data Risk Officers (DROs)
Director+ level from each line of business, Country & Subsidiary
Data Management Core Team
Core Governance Support Analysts
Data Stewards:
Representing each Division, Country, Subsidiary
Data Scientists/users:
Representing Divisions, Countries, Subsidiaries
Data Governance Core Group
Divisions/ Lines of Busienss
Authority: Policy, Oversight, Decision Making, & Alignment
Governance Activities: Monitoring, Remediation, Metadata support
Users: Developers, Scientists, End-users

A Sample Organization Chart for Data Governance

Here is a brief description of the job roles:

- Data Council: 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, 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 4 pillars of 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 meta data 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 practical execution and activities of what the Data Governance structure is enabling. A Data Steward is accountable for the appropriate, successful and cost-effective availability and use of some part of an organization's data portfolio.

Data Stewards are the agents of trustable data within the organization. They're responsible for tracking, improving, guaranteeing, supporting, producing, purging and archiving data for their organization. They work with other data user roles (data scientist, data engineer, data analysts, etc.) in the user community as the spokes for data communication between 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, 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 resource as well. In a matrix organization, Data Stewards may report to their business managers in their organization with a dotted-line to the IT organization.

Authority: Policy, decision-making, governance

Data governance framework must establish clear lines of accountability and authority. Who will make the decisions related to policy, governance, rules and process for decision making relative to Big Data governance. So, the objective of 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 classification & rules for meta data management related to each class of data?

It 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 what was the data source.

Governance polices must also define processes for reporting on data quality rules. Data quality policies 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 measure to:

- Ensure Hadoop access controls and usage patterns are in compliance
- Ensure compliance with regulatory standards: HIPAA, HITRUST, and other standards that pertain to your data and industry

Data Management: Much of work of data management is carried out by Data Stewards. Data Stewards will handle lineage tracking. This implies that we need to set up 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, 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 polices 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 and 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 usage model of the data sets as documented in the Meta Data Hub (MDH)

Users: Developers, Data Scientists, End-users

The user community of 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 data), critical (data 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 a Critical data.

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

Users are responsible to register their data upon loading into Hadoop. They're responsible to maintain the metadata information for their data. Also, they may request information and verification about their data. An example of requesting data verification is to request data lineage/history if reports appear 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, 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.

Master Data Management

Master Data Management (MDM) is the process of standardizing definitions and glossary of 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. The systems and processes required to maintain this data are known as *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 as 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, Warrantee, 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, life cycle, and CRUD (Create, Read, Update, Destroy) cycle for a product in pharmaceutical sector is likely very different from those of the clothing industry. The granularity of domains is essentially determined by the degree of differences between the attributes of the entities within that domain.

There is a lot of types and perspectives of master data but here are five types of data in corporations:

- **Unstructured**—This is data found in e-mail, social media, magazine articles, blogs, corporate intranet portals, product specifications, marketing collateral, and PDF files.

- **Transactional**—This is data related to business transactions like sales, deliveries, invoices, trouble tickets, claims, and other monetary and non-monetary 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 between master data[7].

Master data can be described by business processes, and they way that it interacts 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 life cycle. This life cycle, called the CRUD cycle for short, defines meaning and purpose of that data for each stage of the lifecycle. For example, how a 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 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 application private stores.

Since maser 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 Hadoop. An important step towards an enterprise master data management is to create a metadata system in Hadoop that is in synch 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 synch 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, and 3) before adding new master data from data in the Hadoop Lake, ensure you leverage from existing master data.

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 a 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[8]:

- **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.

Meta Data Management

Meta data is not limited to just names, definitions & labels. In fact it should include more comprehensive information about the dataset, 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 history of its movement and transformations through the data lifecycle.
- Tracking usage – Record who creates this data and who (which programs) consume the data.
- Relationship to products, people & business processes – Define how is the data used in business processes or business decision making. Answer questions like how the data is used in the product development and by people?
- Privacy, access & 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 Meta Data 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.

Lake Data Classification

The key to a successful data management and data governance is data classification. Data may be classified around several topics and contexts as you've seen so far in this book. A classification that defines the stage of data in Hadoop lake is the 4 stages of: Raw, Keyed, Validated and Refined.

Raw Data: When data lands in the Hadoop lake, it's regarded 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.

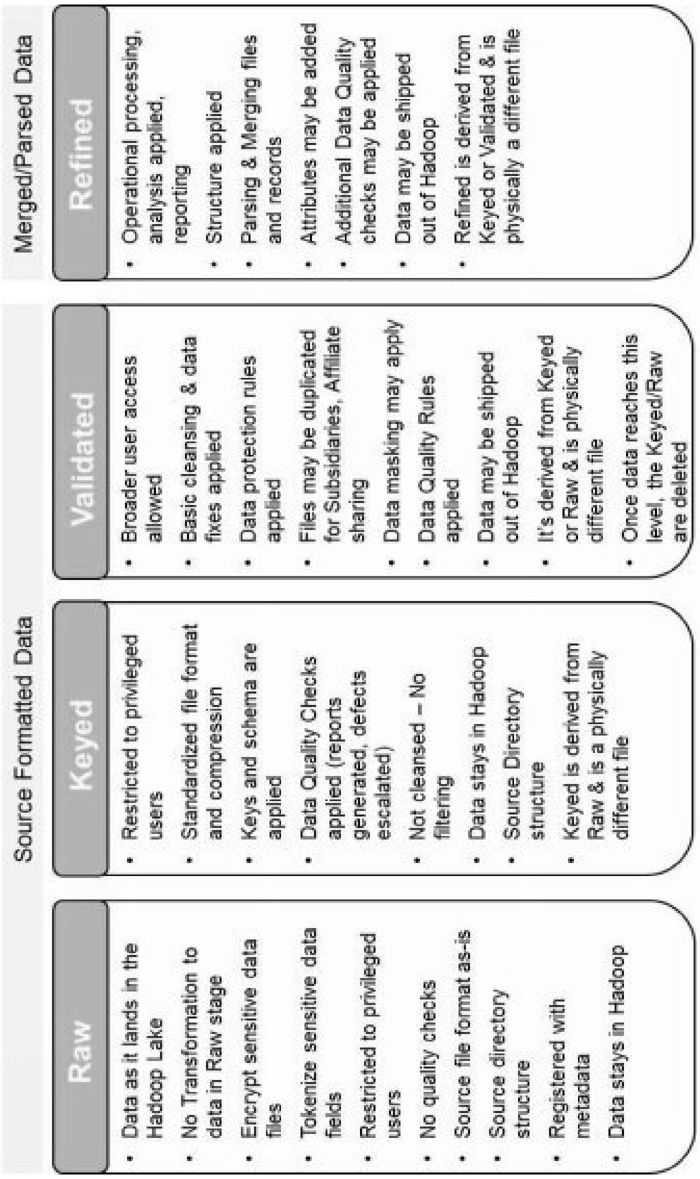
Keyed Data: This is the data that has been transformed into a data structure with schema. A Keyed data might be put into a Key-Value paired structure, or stored into Hive, HBase or Impala structures. Keyed data is derived from Raw data, but physically a different data file. Access to Keyed data is also 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 datasets), they 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 validated. Validated data may be used for analysis, reporting and exporting out of 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 a different file. At this stage the Raw and Keyed data may be discarded as they've transformed into the validated stage.

Refined Data: A 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 and be shared with other users in the enterprise.

The diagram below depicts the distinctions and usage rules for each classification of data in Hadoop.


Source Formatted Data
Merged/Parsed Data
Raw
• Data as it lands in the Hadoop 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 Hadoop
• Source Directory structure
• Keyed is derived from Raw & is a physically different file
Validated
• Broader user access allowed
• Basic cleansing & 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 Hadoop
• It's derived from Keyed or Raw & is physically different file
• Once data reaches this level, the Keyed/Raw are deleted
Refined
• Operational processing, analysis applied, reporting
• Structure applied
• Parsing & Merging files and records
• Attributes may be added
• Additional Data Quality checks may be applied
• Data may be shipped out of Hadoop
• Refined is derived from Keyed or Validated & is physically a different file

The Four Stages & Classifications of Data in Hadoop Lake

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 & 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 vs. 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).

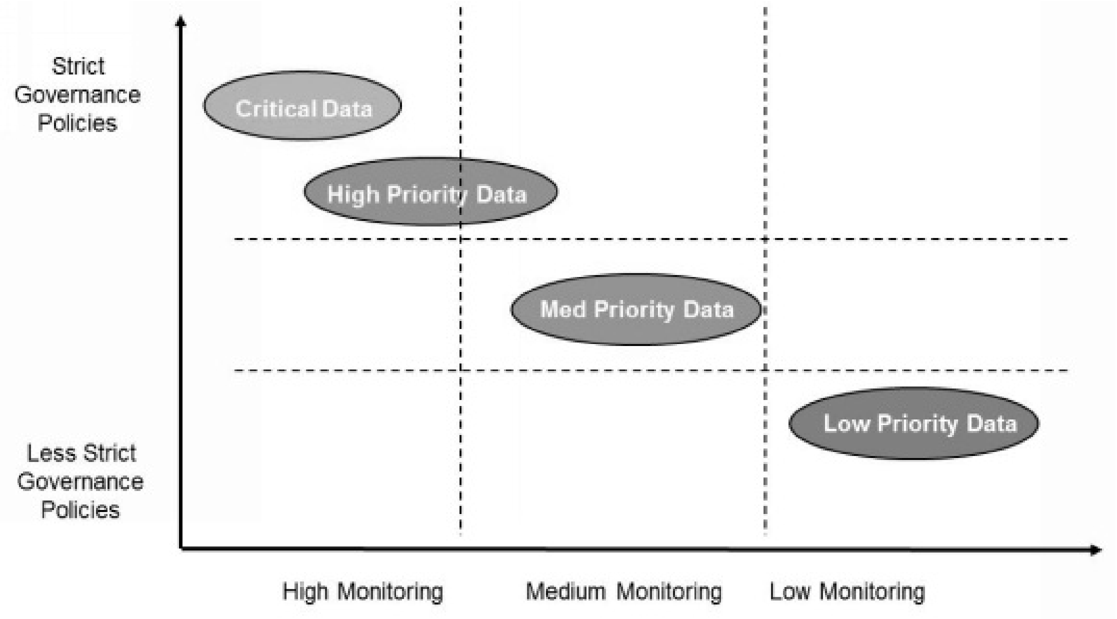
As part of this process, you must determine which HITRUST controls are required for the Hadoop Lake. You must define data encryption, tokenization & data masking rules. For example, large files that are coarse grained might get encrypted upon landing into the Hadoop Lake. But, more detailed data columns that contain PII, such as Social Security number, or Credit Card number 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 don't rise in importance to other critical data; hence their governance policies might be less stringent. But, the critical data such as patient health information will require most strict of governance policies. The diagram below illustrates the distinctions that you can consider when crafting and administering data governance policies by data classification.



More Strict Policies & Monitoring Apply to Critical & High Priority Data

Big Data Security Policy

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

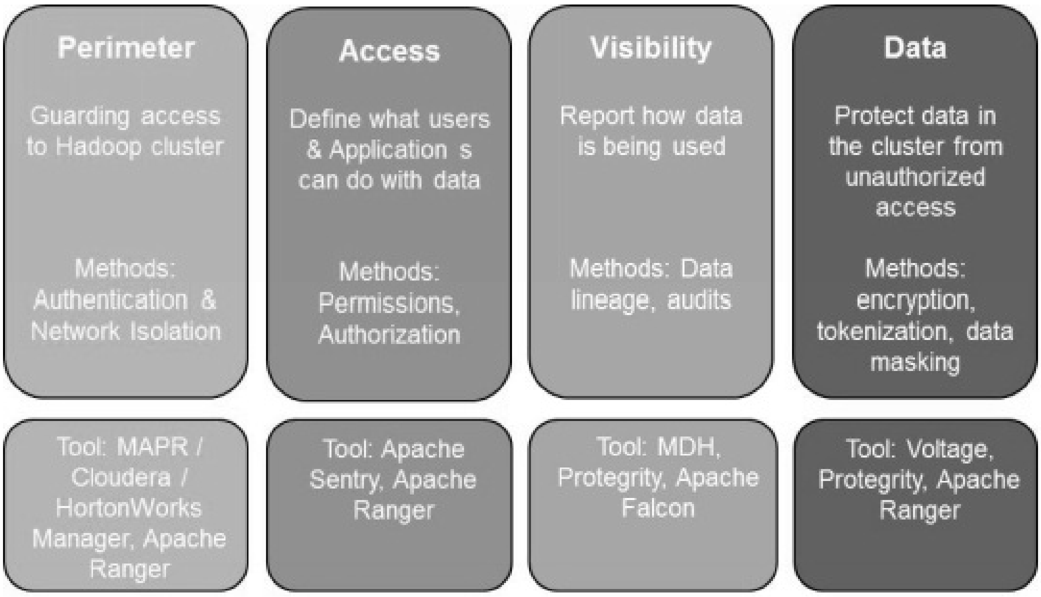
Securing the cloud is made possible by several standards and solutions. For example, one can decide to apply HITRUST, and Cloud Security Alliance STAR certifications.

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

Big Data Security Policy

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


Perimeter
Guarding access to Hadoop cluster
Methods: Authentication & Network Isolation
Access
Define what users & Application s can do with data
Methods: Permissions, Authorization
Visibility
Report how data is being used
Methods: Data lineage, audits
Data
Protect data in the cluster from unauthorized access
Methods: encryption, tokenization, data masking
Tool: MAPR / Cloudera / HortonWorks Manager, Apache Ranger
Tool: Apache Sentry, Apache Ranger
Tool: MDH, Protegrity, Apache Falcon
Tool: Voltage, Protegrity, Apache Ranger

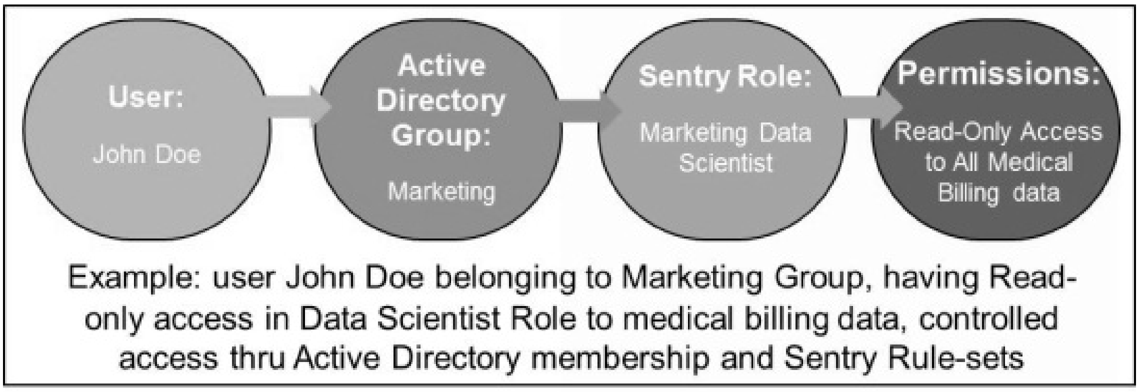
The Four Pillars of Hadoop Data Lake Security

Big Data Security – Access Controls

The access control mechanism for Hadoop Lake is best served when it’s integrated and compatible with the existing security procedures and policies. A users’ access to Hadoop data need not be different that 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. Based on this information, a Hadoop security tool such as Sentry can be configured to give a 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.



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

Access Control Chain in Hadoop Data Lake

Big Data Security: Key Policies

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

1. Security controls on Hadoop must be managed centrally and automated
2. Integrate Hadoop security with existing IT tools such as Kerberos, Microsoft Active Directory, LDAP, logging and governance infrastructure
3. All intra-Hadoop services must mutually authenticate each other
4. Apply data access controls to Hadoop 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 the Job owner's UID. This ensures that the 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 such as each directory or file can be assigned multiple users, group with file permissions are 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 Data Governance Council and users that allows users access to data in exchange expects and lists a number of policies that users agree to abide. Below is a list of policies that can be considered in a Data User Agreement:

1. All critical & 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 meta data compliance. Data 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 meta data Hub (MDH) when loading data

into Hadoop. Data Stewards may notify users and issue warning if their data does not have adequate metadata information.

5. Third-party and 4th party data is to be managed and used in accordance to data agreements with the 3rd party data provider. If your organization purchases data from external 3rd party or 4th party sources, the contract usually comes with certain clauses and usage constraints. For example, one data 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 applied according to the DUA policies. Data retention policies vary from country to country and even from state to state in the U.S. 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 Steward or Hadoop Administrators should only provide access to users after users have completed their training of data governance policies.

Security Operations Policies

Operationalizing security policies and coordinating the people, policies, tools and configurations to deliver a cohesive security on Hadoop is 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 operating 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 and held responsible for notifying the Hadoop security Admin regarding onboarding or off-boarding users for their area 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 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, a la Personally Identifiable Information and data (PII) must be de-identified in Raw or become encrypted. Tokenization is preferred when

data columns (Fields) are known.

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 breech your data.

Information Lifecycle Management

As I mentioned earlier, managing data through its life cycle spans a process of CRUD: Create, Read, Update and Delete. There are certain policies that apply to data during its lifecycle that govern allowed usage, handling process and retention rules. Here are some high level rules that I recommend as the basis of an Information Lifecycle Management (ILM) governance:

- The rules for minimum data retention may vary from country to country. The longest retention period must be updated in the Meta Data Hub (MDH) Tool for all datasets.
- If your industry requires that you reproduce the data and analytics results at another time, you must take snapshot 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.
- 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.
- 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.
- 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 & Monitoring

Data quality monitoring involves 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 & 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 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 suspicious presence of too many NULL values, and other data anomalies to detect quality issues.

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

1. Use Oozie to orchestrate data pipelines and track processing errors at each stage of data processing

2. Use Hbase to record and maintain errors related to metadata errors, data processing job errors and event errors

3. Configure 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 as each organization is different. But, 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) vs. “Trusted” data (data that has been validated, has a complete metadata, lineage information and is certified by quality checks). Furthermore, here we can define four additional data classes:

- **Critical Data –** The type of data that has the highest regulatory, reporting and compliance requirements defined by FDA and/or GxP (Good Manufacturing Practices, Good x Practices, just place your key business process for x) and is materially significant to the decision making at Enterprise.

- **High Priority Data –** the type of data that is materially significant to the decision making at Enterprise, regarded classified information and confidential data. This data is required for analytics but not subject to GxP or FDA regulations.

- **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.

- **Low Priority Data –** Data that is typically not material and does not have retention requirements.

Data Quality Policies

Policies that define proper quality management of data require partnership between users and Data Stewards. 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 as guidelines for quality-driven standards, strategies activities to ensure data can be trusted:

- *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 *"Pending"* Status until the data set is brought into compliance. No reporting or analysis of Pending Status data is allowed.

- 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 and be able to identify which data elements (those that are in their domain of responsibility) are Trusted Sources or in Pending Status.

- Data quality issues are defined as gaps between the data element characteristics and deviations from this data governance standard. Data quality gaps typically include: data that has not been validated for accuracy and completeness; data that is incomplete metadata information in the Meta Data Hub (MDH); and data lineage is unknown.

- 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 of any issues related to data quality.

- The Meta Data Hub or another database can maintain inventory and record 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 fix.

- Data Governance Council may define data quality metrics for the organization. These metrics might include percent of data in *Pending* status vs. *Trusted* status, Number of open quality issues, and data volume by each division or user group. Data Stewards are responsible to provide metrics reports to the DRO, AE and Data Council related to 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.
- Data Stewards must maintain evidence of compliance to the data management policies and supporting standards, procedures and processes.
- Data Stewards along with Data Risk Officers annually review inventory of critical or high-priority usages and Data Usage Agreements with in their domain of responsibility.
- 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 Meta Data Hub (MDH) Tool
    - Lineage to origination or to the Trusted Source
    - Data Quality Monitoring & checks passed
    - Retention period is defined

Metadata Best Practices

The data quality standards at your company under the big data governance policy should require the 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 Steward 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 pair) that someone else would need to know in order to understand it and property 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?

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.

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.

You should make proper assignment of descriptive 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 this.

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

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

Carefully select the terms that you intend to use. While conversationally people use loose terms to mean the same thing, 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. As an 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:

Descriptive names should end in a word that classifies the data. For illustration consider, date, amount, number, code which are examples of data classifiers. The word 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.

Descriptive names should be as short as possible too while retaining their meaning and uniqueness. But, clarity is always more important than brevity.

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 “an, to, of, after, before”.

As needed, modifiers may be used to create independent attributes. For example, Application Status Code uses the modifier that allows the data element to stand independent of the table/file which it’s stored.

Try to use more generalized definition and descriptive 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. Instead call it Tax Identification Number or National Tax Identifier or something more general in description.

Data Descriptions 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.

Each definition must be independent of any other definition, such that the reader should not need to 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 which forms a tree. For example, when a class of data has sub-classes the top level column should 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.

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 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 that 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 so. The Indicator might take values such as “Yes”, “No” or “Unspecified”.

Finally, data tables and files need to be described in 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. It 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. If a uniqueness key involved multiple columns, each column must refer to the combination which produces the uniqueness.

The contextual information about table/file is also important. The context should indicate retention and metadata contact for that table/file. Furthermore, the table/file descriptor must include an Indicator to determine if the table/file is a reference file. When the Indicator is “Y”, it indicates that the table/file contains data that explains the meaning of another data. Typically, this is used as a code table. If the Indicator is “N”, or “Unspecified”, then the data table/file signifies that the element is not a reference table or 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.

Big Data Governance Rules: Best Practices

As a sample of typical big data governance policies and best practices, I’ve compiled a list in form 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. Data Governance Council may 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
  - A monitored, quality test, managed process for ingestion & 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 Data Quality Certification program
  - Apply ongoing Data Quality monitoring that includes random quality checks, tests
  - Data certification process by Data Stewards & Data Scientists
- Rule#5: Do not dump data into the lake without 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 with in Hadoop.
- Rule#6: Establish data pipeline categories and data classifications that we reviewed already. For example, the data stages: Raw, Keyed, Validated and Refined.
- Rule#7: Register data into 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 meta data registry system within 90 days of loading 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 by multiple attributes and record the classification in metadata registry: data owner, type of data, granularity, structure, country jurisdictions, schema and retention are some examples. Data Governance

Council must define privacy, security and usage policies for each classification.

- 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 or their location, company names, product names and such 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 achieve a policy that never allows any sensitive data into the lake without applying the 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 towards effective data protection.

Sample Data Governance & Management Tools

Fortunately, new data governance tools are emerging almost every month for managing big data in Hadoop environment. There are possibly three challenges in selecting and implementing tools these days: 1) Identify and select a minimal set of tools which together provide a complete coverage of your organization's data governance strategy and requirements; 2) Integrating these tools into a seamless environment; and 3) Be prepared to revamp the tool set in 2-3 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):

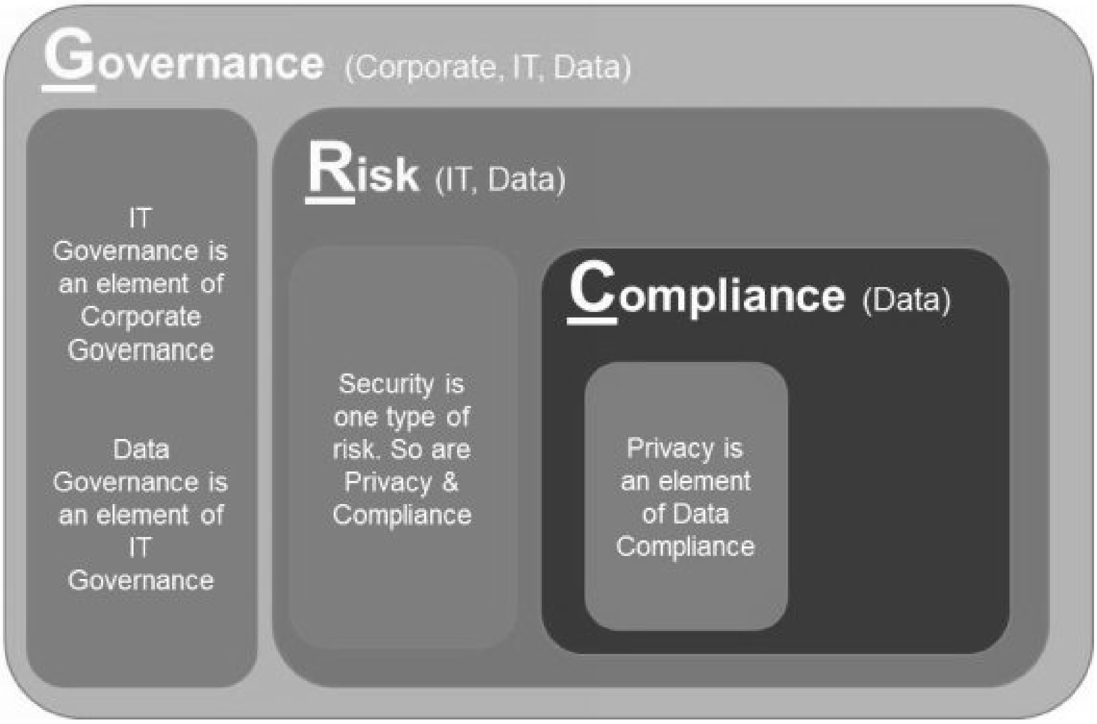
- Collibra – Active Data Governance and data management
- Sentry* – Central access management tool
- Knox* – Central authentication APIs for Hadoop
- Apache Atlas* - Metadata registry and management
- Cloudera Navigator – Central security auditing
- BlueTalon – Central access & security management tool
- Centrify – Identity management

- Zaloni – Data governance & 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 & pipeline orchestration
- Oozie* – Data pipelines 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). 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.

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 stand up your own big data governance framework for your organization.


Governance (Corporate, IT, Data)
IT Governance is an element of Corporate Governance
Data Governance is an element of IT Governance
Risk (IT, Data)
Security is one type of risk. So are Privacy & Compliance
Compliance (Data)
Privacy is an element of Data Compliance

The GRC Perspective of Data Governance

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 branch 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 adequate and consistent budget. I raise this issue since many organizations have overlooked the need for data governance investment 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 Other Companies are doing?

Many large enterprises are developing robust data governance policies and framework. 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 & tools development to deliver a robust data governance infrastructure for businesses in 3 phases. This infrastructure is based on Apache Ranger, Hcatalog Metastore & Apache Falcon.

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.

Hortonworks and the Apache project are working on additional metadata management tools under the project Atlas. I suggest keeping a watch on the Atlas project to stay up to date about the latest releases and functionality from this project. The link is at: http://atlas.incubator.apache.org/

SECTION III:

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.

However, many best practices, tools & methods are emerging. 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 section 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 only the tip of the iceberg as technology continuously evolves and security in and by itself is a race, an unending race with hackers.

Data Protection

Data Protection is architected as shown in the diagram below 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 concerned with Logical layer and database constructs and tools such as Hive or Impala.

This design promotes data classification into several status and types. In the Hadoop sandbox, I've shown four data statuses: Raw, Keyed, Validated, and Refined. Both data paths come to the Hadoop Sandbox. When data is on-boarded (or imported) into the sandbox, it's initially Raw Data. Then it is transformed into Keyed such as in a Hive table and format. Then the data validated for quality. At this stage the data is validated and Trusted. (I'll explain the Trusted data classification later). Finally data might get 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 3 categories of considerations: Sensitive Data Protection, Data Sharing Considerations, and Adherence to Governance Policies.

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 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 the following diagrams, I'll refer to various types of sensitive data simply by 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.

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 3rd party (and sometimes 4th party data providers). These purchases come with usage rules, do's and don't clauses. Data policies related to proper safeguard, access and usage of 3rd party data are needed to ensure that we’re compliant with the initial 3rd party data purchasing agreements and contracts.

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.

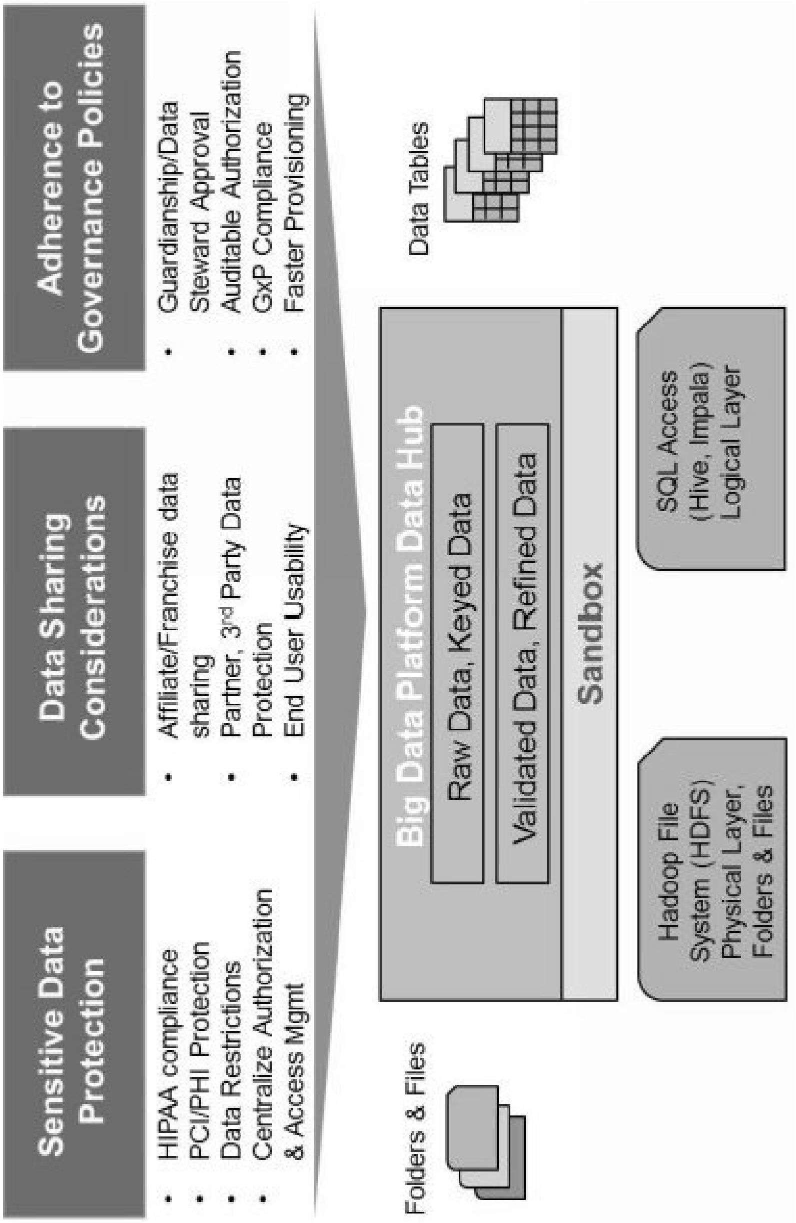
Both access paths require that users adhere to existing policies and that users are to be provisioned by Data Stewards, plan training for handling sensitive data prior to granting access to such data.

Adherence to Governance Policies

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

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

Certain industries dictate additional standards and compliance requirements regarding data. One of these standards is the 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.


Sensitive Data Protection
HIPAA compliance
PCI/PHI Protection
Data Restrictions
Centralize Authorization & Access Mgmt
Data Sharing Considerations
Affiliate/Franchise data sharing
Partner, 3rd Party Data Protection
End User Usability
Adherence to Governance Policies
Guardianship/Data Steward Approval
Auditable Authorization
GxP Compliance
Faster Provisioning
Folders & Files
Big Data Platform Data Hub
Raw Data, Keyed Data
Validated Data, Refined Data
Sandbox
Data Tables
Hadoop File System (HDFS) Physical Layer, Folders & Files
SQL Access (Hive, Impala) Logical Layer

Privacy & Security Architectural Considerations for Hadoop Environment

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 datasets at file level (coarse data) and coarse grained access to data. The 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 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.)

The best practices recommend that you establish Directory Structure that supports access by the following classifications of users and usage, role-based access:

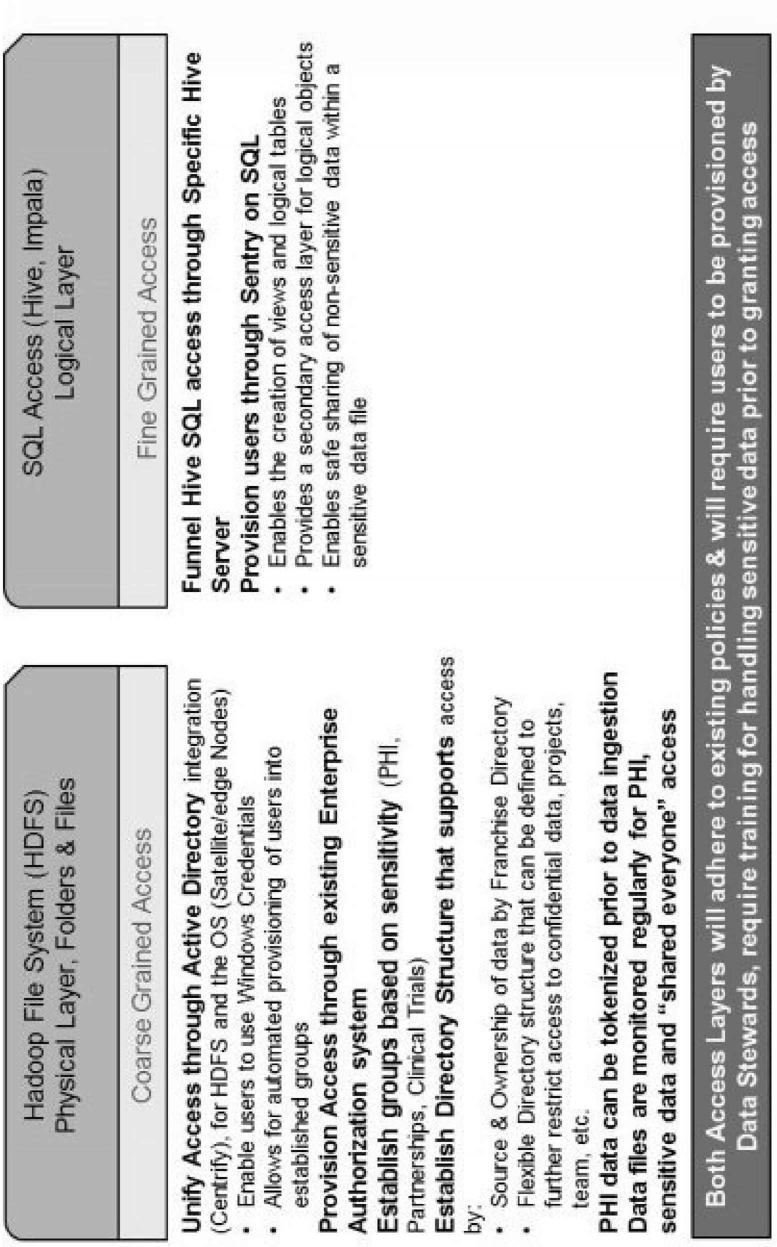
- Source & Ownership of data by Directory grouped by Division, Country, Affiliate and similar organizational boundary
- Flexible Directory structure that can be defined to further restrict access to confidential data, projects, team, etc.

To protect sensitive data, at the coarse grained level of data , one can and should apply tokenization to sensitive (PHI, PII, PCI) data 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 even that hackers gain access to the data on the cloud, that information is useless to them without the token keys.

One of the best practices 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's 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 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, 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 the high level governance, data stewards are vigilant and actively monitor data files regularly for PHI, sensitive data and data labeled "shared everyone" access.



Hadoop File System (HDFS)
Physical Layer, Folders & 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, Clinical Trials)

Establish Directory Structure that supports access by:

- Source & Ownership of data by Franchise Directory
- 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 & will require users to be provisioned by Data Stewards, require training for handling sensitive data prior to granting access

Managing Security & Privacy on Physical & Logical Layers

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 Specific Hive Server that controls access. Sentry is a tool often used in securing data on SQL. You can provision users through Sentry on SQL. The advantages of this configuration include:

- 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

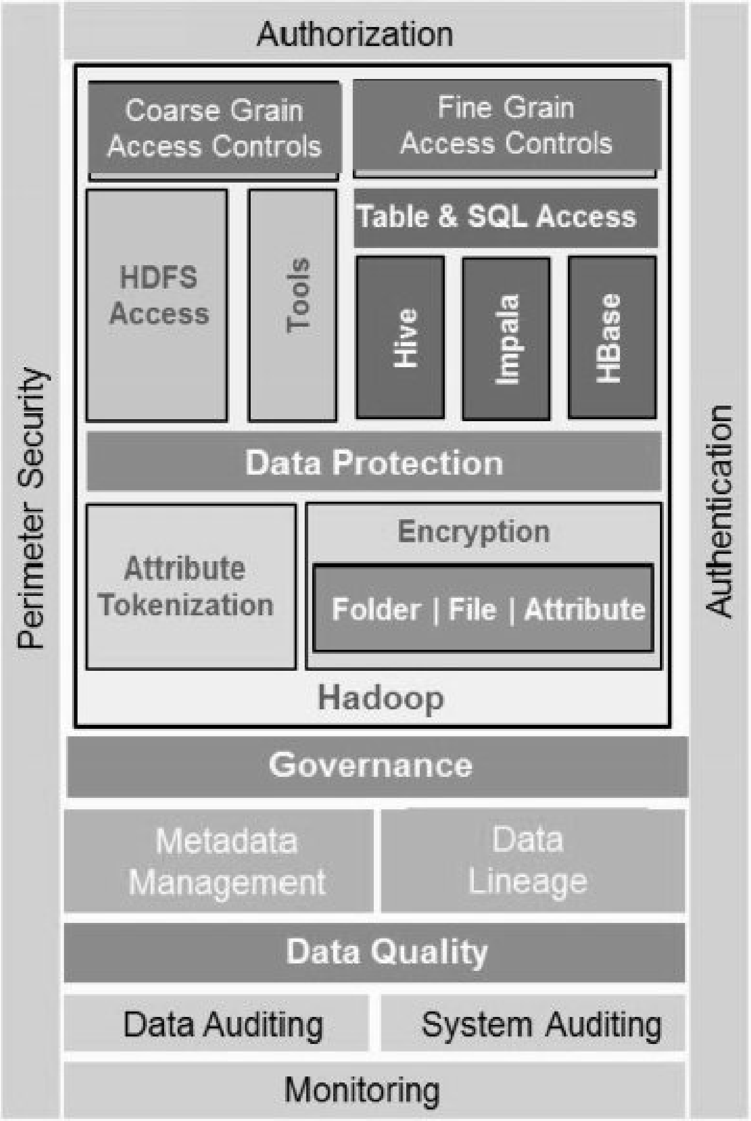
Security Architecture for Data Lake

The following architectural diagrams depict a modular design where the functional data security and protection are separated by design into layers and modules to ensure all aspects of data security are considered. The key aspects of this security design are:

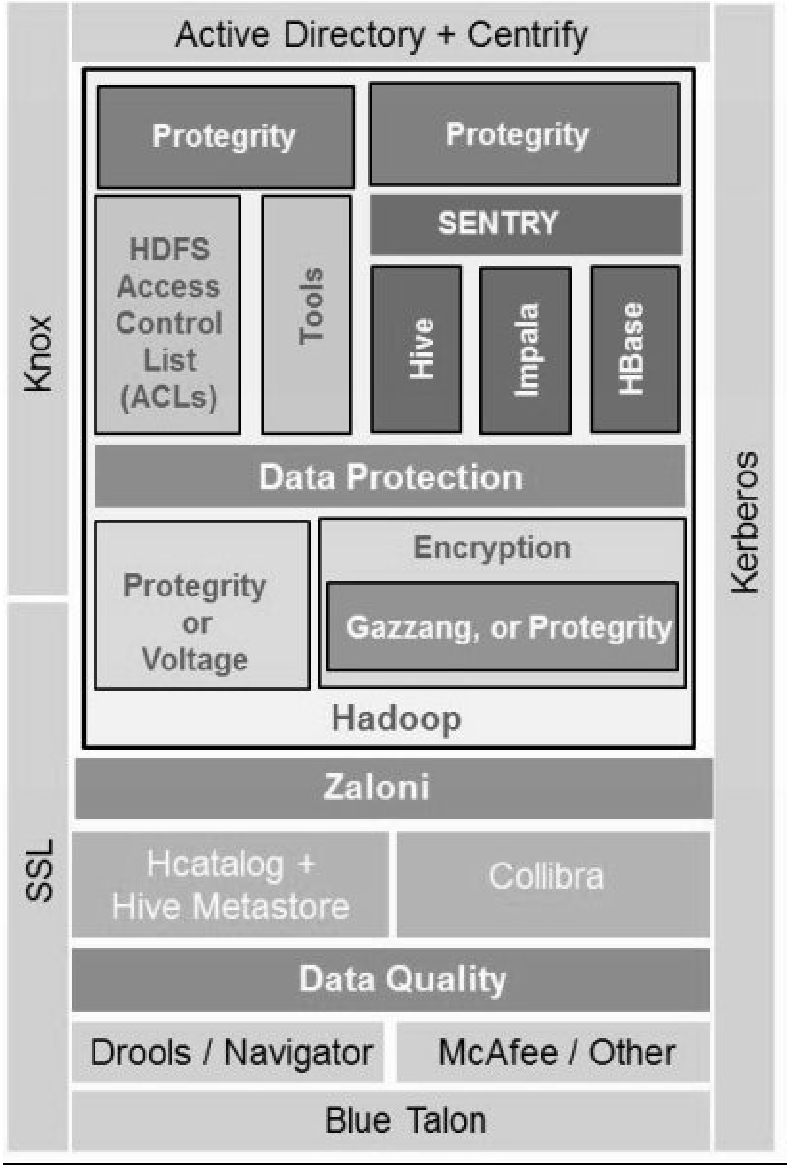
- Authentication, enabled by standards such as Kerberos or similar solutions
- Authorization, enabled by Microsoft Active Directory (AD) and Centrify or comparable product
- Perimeter Security, can provided by Secure Socket Layer (SSL) and Knox product or similar solution
- Monitoring, can be enabled by products such as Blue Talon

Within the architecture, the building blocks for governance, data quality management, data protection, both fine grained and coarse grained access control are shown. The next diagram 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 or similar solutions
- Data Protection included data encryption and tokenization of the data in the lake. Some of the solutions for tokenization & encryption include Voltage and Protegrity. The apply data protection at the folder, file and attribute level, solutions such as Gazzang, Protegrity and Voltage represent some of the options.
- Meta data 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 grain and fine grain data protection can be enabled by a solution 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.



A Model Hadoop Security & Privacy Management Stack



A Hadoop Security & Privacy Model Enabled by Tools

Data Lake Classification

Lake 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 3 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 this data is restricted to privileged users such as data engineers only.

There are no quality checks performed on raw data. This data represents source file format as-is; it had 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 Meta Data

Hub – MDH), can stay in Hadoop.

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 represents standardized file format 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 & 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 & data fixes have been applied; Data protection rules are applied and Data Quality Rules are applied.

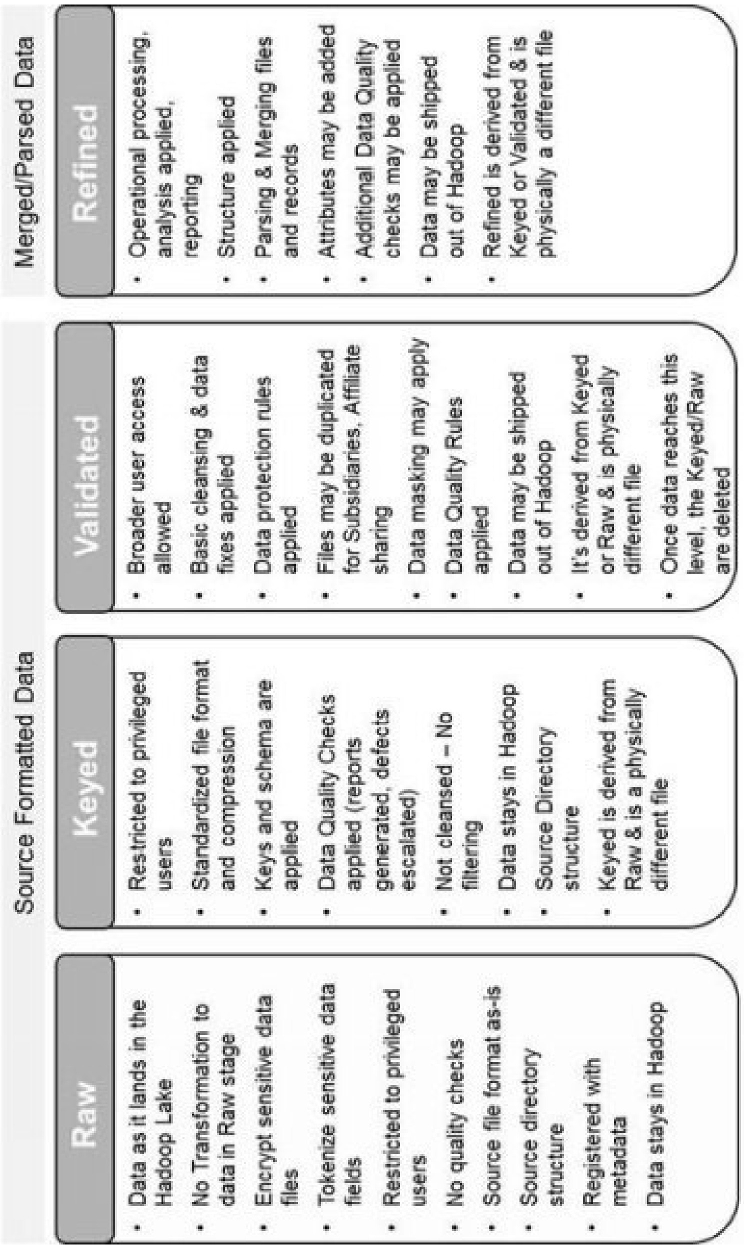
In this stage, files may be duplicated for across the organization for other divisions, partners and affiliates, Subsidiaries or sharing to other divisions in other countries as long allowed by data jurisdiction policies.

Certain data masking may apply to this data based on data usage policies. But, Validated may be shipped out of Hadoop. This data is derived from Keyed or Raw & is physically different file from its prior stages. Once data reaches this level, the Keyed/Raw files are deleted. At this 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 reporting is generated. Users have access to this data and typically create it from Validated data as they apply structure to the data. They may parse & 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 & is physically a different file. The diagram that shows the four-stages of data in Hadoop is shown again below for reference.



Raw

- Data as it lands in the Hadoop 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 Hadoop
- Source Directory structure
- Keyed is derived from Raw & is a physically different file

Validated

- Broader user access allowed
- Basic cleansing & 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 Hadoop
- It's derived from Keyed or Raw & is physically different file
- Once data reaches this level, the Keyed/Raw are deleted

Refined

- Operational processing, analysis applied, reporting
- Structure applied
- Parsing & Merging files and records
- Attributes may be added
- Additional Data Quality checks may be applied
- Data may be shipped out of Hadoop
- Refined is derived from Keyed or Validated & is physically a different file

The Four-stages of Data in Hadoop Data Lake

Hadoop Lake Data Classification and Governance Policies

Different data governance policies apply to different stages and categories of data. As the next diagram shows various access and usage controls apply to data classes differently. Let's look at that 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, prevailed users may access this data thus the goal is to have a 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 Hadoop data lake must be registered in the meta data hub including data lineage, schema, attribute metadata and owner (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 is 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 information) in a separate container and be able to apply specific access policies to that folder.

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

Validated/Refined data Usage Rules

Data functions in this stage vary from file cleansing, curation and preparation of data for analysis to transformations, merging with other data, filtering and extracting sub-sets of the datasets. This class of data is defined to have broader user access. But, data stewards must be vigilant to continue splitting data into folders for sensitive data from the rest.

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 & files) and granular levels (data tables and fields) are required.

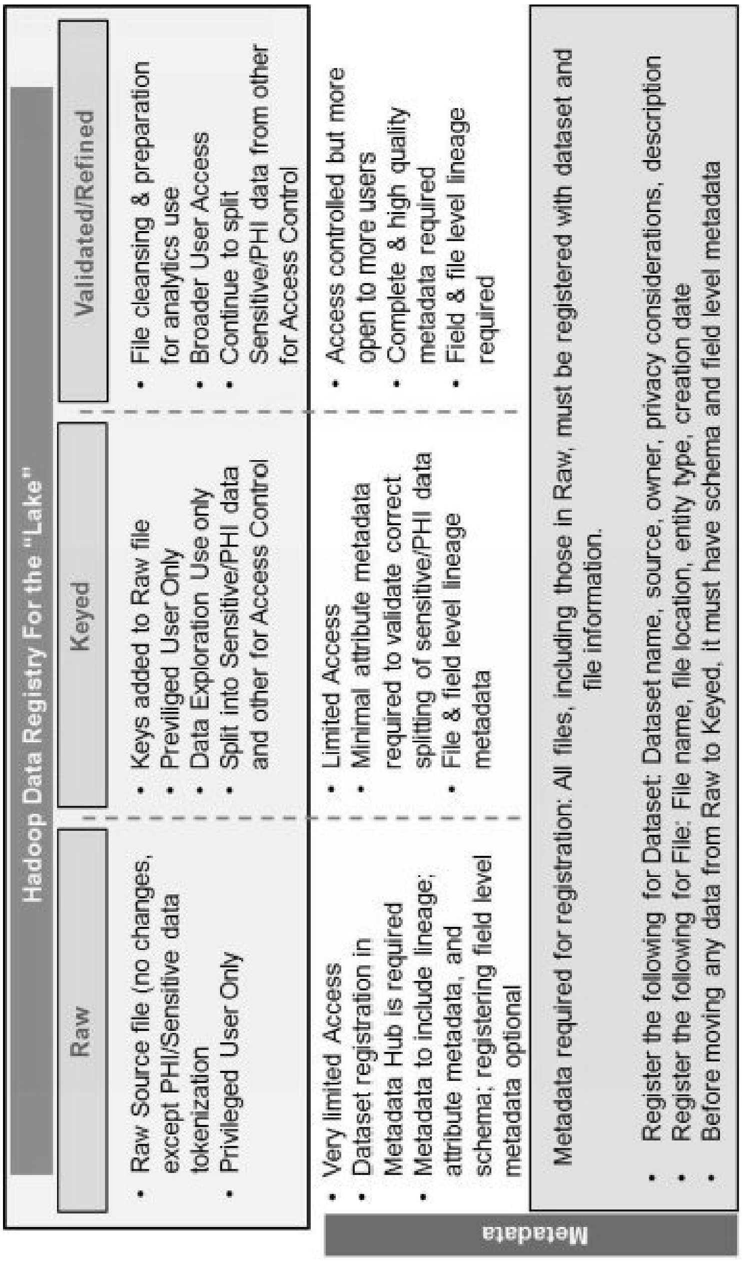
Metadata Rules

Some attributes required for Metadata registration span All files, including those in Raw, which must be registered with dataset and file information. Here is a list of some possible attributes and recommendations as illustrated in the next two diagrams:

- Register the following for Dataset: Dataset 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



Hadoop Data Registry For the "Lake"

Raw
- Raw Source file (no changes, except PHI/Sensitive data tokenization
- Privileged User Only

Keyed
- Keys added to Raw file
- Previliged User Only
- Data Exploration Use only
- Split into Sensitive/PHI data and other for Access Control

Validated/Refined
- File cleansing & 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

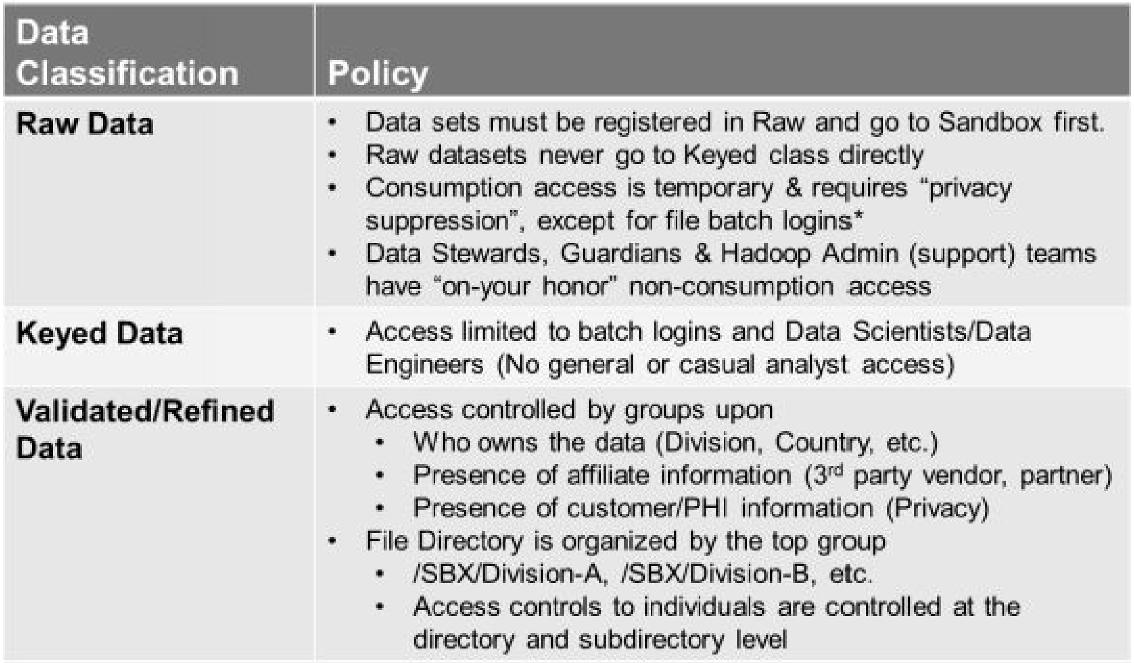
- Limited Access
- Minimal attribute metadata required to validate correct splitting of sensitive/PHI data
- File & field level lineage metadata

- Access controlled but more open to more users
- Complete & high quality metadata required
- Field & 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 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

Metadata Rules Matched to Data Stages in the Lake



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 & requires "privacy suppression", except for file batch logins* • Data Stewards, Guardians & Hadoop Admin (support) teams have "on-your honor" non-consumption access
Keyed Data	• Access limited to batch logins and Data Scientists/Data Engineers (No general or casual analyst access)
Validated/Refined Data	• Access controlled by groups upon • Who owns the data (Division, Country, etc.) • Presence of affiliate information (3rd 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

Data Management & Configuration Rules Matched to Data Stages in the Lake

Data Structure Design

Some of the best practices -which we've already discussed- include splitting data and storing sensitive data in a separate folder and directory path. The following diagram shows a sample data lake file storage structure.

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 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 should start first by source of data

When data is Refined and set up for broader use, your directory structure should be set up by organization that owns (and uses) that data.

Let's assume you're looking to design file structure for a large global retail company, say Nike. Nike is the largest of athletic footware and apparel companies 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 other brands.

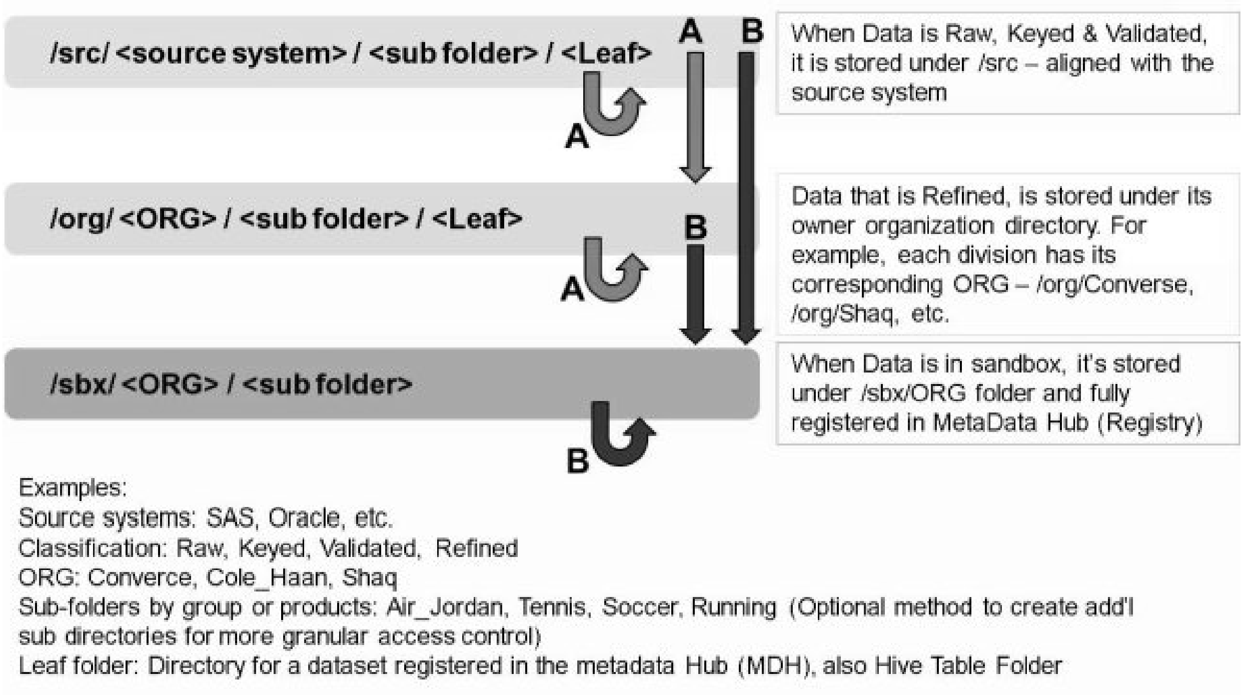
Let's assume the company has some pre-existing stored in a SAS server, prior data on an Oracle database, Teradata data and several subsidiaries, divisions and franchises. The company has 3 key product categories like Footwear, Apparel, and Equipment. Footwear products include Running, Soccer, Tennis, Golf shoes, etc.

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

In the diagram below, 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 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 metadata registry not in the directories. This approach allows directory design to be simpler, more consistent across the enterprise with better management functionality for access control.



/src/ <source system> / <sub folder> / <Leaf>

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

/org/ <ORG> / <sub folder> / <Leaf>

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

/sbx/ <ORG> / <sub folder>

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

Data Directory Structure in Hadoop Data Lake

Sandbox Functionality Overview

The sandbox folder design would follow the <ORG>/<Line of Business>/<Product> path. As illustrated in the diagram above. Users have read/write access to the Sandbox designed and assigned to their organization. Users may read data from Hadoop lake and bring over to 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 Director (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 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 sandbox Privacy rules training are required for users. Depending if the users will access data from US, Europe, Canada or another country of jurisdiction, training designed for such countries must be required for users.

Data Stewards must on a regular basis (quarterly, half-yearly 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 Hadoop data is identical to their pre-existing data access level to other data stores in the organization.

Top level Sandbox Protection

Top level Sandbox directory is protected by a Sandbox specific group such as:

phdp_<resource>_sbx

For example: /sbx/Converse directory has phdp_Converse_sbx group assigned. Sandbox group ensures only authorized users access to Sandbox. To manage Sandbox Owner-only access consider the following configuration techniques:

- HDFS Umask value ensures that files are created with “owner only” privileges
- Example: Use default 600 permissions on files as only file owner has access to data by default
- Example: In order to share data with others, user must change permission to 640 and assign appropriate group
- Users expected to assign AD Group policy according to the sensitivity of data,

similar to the Hadoop Lake policy

A policy that you can apply to data in the lake is to require 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 Line 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 Lake database.

As part of Data Usage Agreement, you may allow users to create or copy Hive Tables from the Hadoop lake into their Sandbox. In order to run Hive queries, users may run Hive queries using beeline command line utility or Hue GUI. However, at the time of writing this manuscript, 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.

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

It’s encouraging to note that a 2015 Hadoop enhancement is expected to deliver “views” capability to separate data files logically. Until this “view” capability is implemented, physical separation of sensitive is necessary.

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

Another reason for physical splitting of sensitive data is that certain legal, regulatory, 3rd party contractual data and GxP (Good x Practices) requirements or compliance policies apply to sensitive data which 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 sub-folder in each directory to contain the sensitive split data. Proper data governance policies dictate that strict access controls to this folder is applied by the data lake administrators.

In Section IV, I’ll introduce a document that can be a starting point and baseline document for your big data governance program.

SECTION IV:

BIG DATA GOVERNANCE FRAMEWORK PROGRAM

Big Data Governance Framework 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 the 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 four pillars:

1. Organization
2. Metadata management Standards
3. Security/Privacy/Compliance Standards
4. Data Quality management

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

The primary governing body of this policy is the Data Council ("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 and Data Resource and Assets

Information and data resources are the organization's assets, used by the organization's personnel to execute critical processes that deliver on its strategies. Because of risks and costs of using or providing the wrong information, poor-quality data or a 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 to 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 to operate 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, analysis results.
4. A clear line of responsibility between management and what it takes to meet regulatory and legal expectations for data and applications.
5. Policies that ensure data integrity by prioritizing, defining, 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. Maintain 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 procedure for managing big data platform and application assets.

Benefits of Compliance and Risks of non-compliance

Risks associated with non-compliance 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
- Reduce good will and shareholder value through misrepresented or inaccurate public information
- Loss of public trust and 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

Date Usage – A model, report or analytics process.

Data Levels – Describes the importance and the regulatory significance of a data element. There are four levels: Critical, High-priority, Medium priority and low priority.

Data element – Consists of one or more data sets stored and used in big data platform for analysis.

Data Usage Agreement – Defined as a contract between big data platform management and Accountable Executive defining the data quality requirements that must be adhered to for the critical or high-priority data being consumed.

AE – Accountable Executive, a VP or higher level individual who is designated as the executive responsible for execution and compliance with the governance policies for their franchise.

DRO – Data Risk Officer is the executive designated at the division or business unit level who is accountable for ensuring 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.

DS – Data Steward is designated by each AE for each franchise to ensure governance policies applied including policies on security, privacy, data quality monitoring, reporting, and certification of compliance. A franchise may have one or multiple Data Stewards. A Data Steward is a manager or above.

SME – Subject Matter experts that bring their knowledge and expertise in security, privacy, data management, compliance, and technology. SMEs are designated by Data Stewards in each franchise to operate the data analytics in a manner that is compliant with this data governance framework.

Data Quality Issue – A data issue is defined as a gap in compliance to the Data Quality standard, or failure in Data Usage Agreement, deviation from this data governance policy, privacy, security or audits.

Third-party Data Provider – A company or individual providing data to the organization to be used for its analytical use. 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.

Data Council – The primary executive forum for the organization’s divisional (franchise/country/subsidiaries) data risk officers – who are responsible for ensuring the company’s data, is reliable, accurate, trustworthy, properly safeguarded and used -to communicate, prioritize, assess and make recommendations to the executive management with respect to data governance and risk management issues.

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 & 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 & Compliance

Each forum will have a designated leader appointed by the Data Council as listed on the next page.

Role	**Metadata Forum**	**Data Quality Forum**	**Data Governance Forum**	**Security, Compliance Privacy Forum**
Leader	Leader Name	Leader Name	Leader Name	Leader Name
Scope	Develop & Lead creation of metadata capabilities, tools & related topics	Develop & Lead standards of data quality capabilities, tools & related topics	Establish data governance capabilities, tools & related topics	Develop & Lead data security, compliance & privacy capabilities, tools & related topics
Materials & Tools	Meta Data Hub (MDH) tool & data registration procedures	Data quality monitoring, testing and issue remediation procedures	Data Governance policies and procedures	Security, privacy and compliance policies, processes & 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.

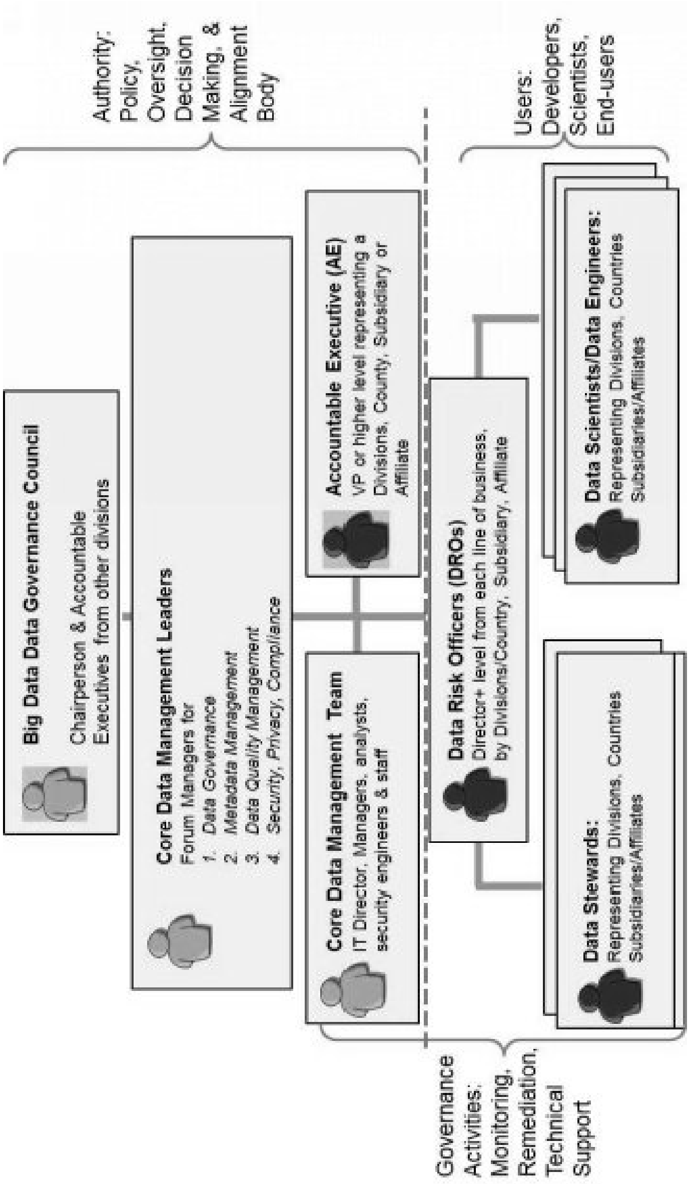
The federated roles in Data Council include the Accountable Executives, Data Risk Officers and Data Stewards. These roles are not full-time positions, rather responsibilities assigned to individuals with vested interest in their use of big data platform.

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 organization. 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.

The following diagram depicts the Data Council structure.


Authority: Policy, Oversight, Decision Making, & Alignment Body
Big Data Data Governance Council
Chairperson & Accountable Executives from other divisions
Core Data Management Leaders
Forum Managers for
1. Data Governance
2. Metadata Management
3. Data Quality Management
4. Security, Privacy, Compliance
Accountable Executive (AE)
VP or higher level representing a Divisions, County, Subsidiary or Affiliate
Core Data Management Team
IT Director, Managers, analysts, security engineers & 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

Data Governance Council and Stakeholder Organization

II. Metadata Management

An inventory of data usage will be maintained in the platform's metadata management Hub (MDH: Meta Data Hub). 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 the data element 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, i.e. 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-, Med-, 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)
- Date of report publication
- Name of report (or analysis) as registered in the Meta Data Hub (MDH) tool

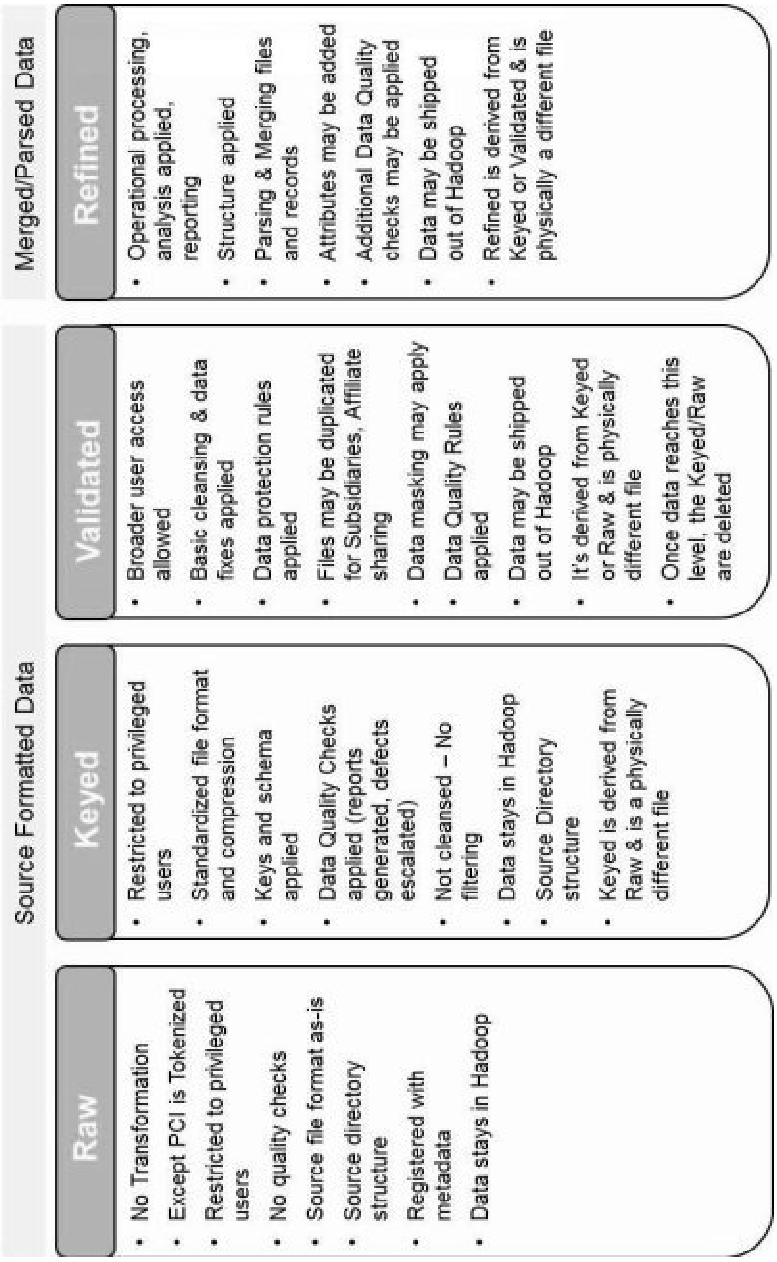
Required metadata information about the data processes (both business and analytics) includes:

- 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 – As 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.)
- Requires 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 the next diagram:



Source Formatted Data

Raw

- No Transformation
- Except PCI is Tokenized
- 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 applied
- Data Quality Checks applied (reports generated, defects escalated)
- Not cleansed – No filtering
- Data stays in Hadoop
- Source Directory structure
- Keyed is derived from Raw & is a physically different file

Validated

- Broader user access allowed
- Basic cleansing & 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 Hadoop
- It's derived from Keyed or Raw & 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 & Merging files and records
- Attributes may be added
- Additional Data Quality checks may be applied
- Data may be shipped out of Hadoop
- Refined is derived from Keyed or Validated & is physically a different file

Data Classification Stages & Controls for Hadoop Lake

The Data Steward for each organization is responsible to monitor and ensure the data is in proper data classification category in Meta Data Hub (MDH) too.

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 or Subsidiary or affiliate)
- 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. Big Data Security, Privacy & Compliance

The security policies & processes outlined in this governance framework are aligned with the enterprise data security, privacy and compliance rules. This governance framework establishes a framework 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 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 Hadoop environment. Use Apache Ranger to standardize authentication method 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.

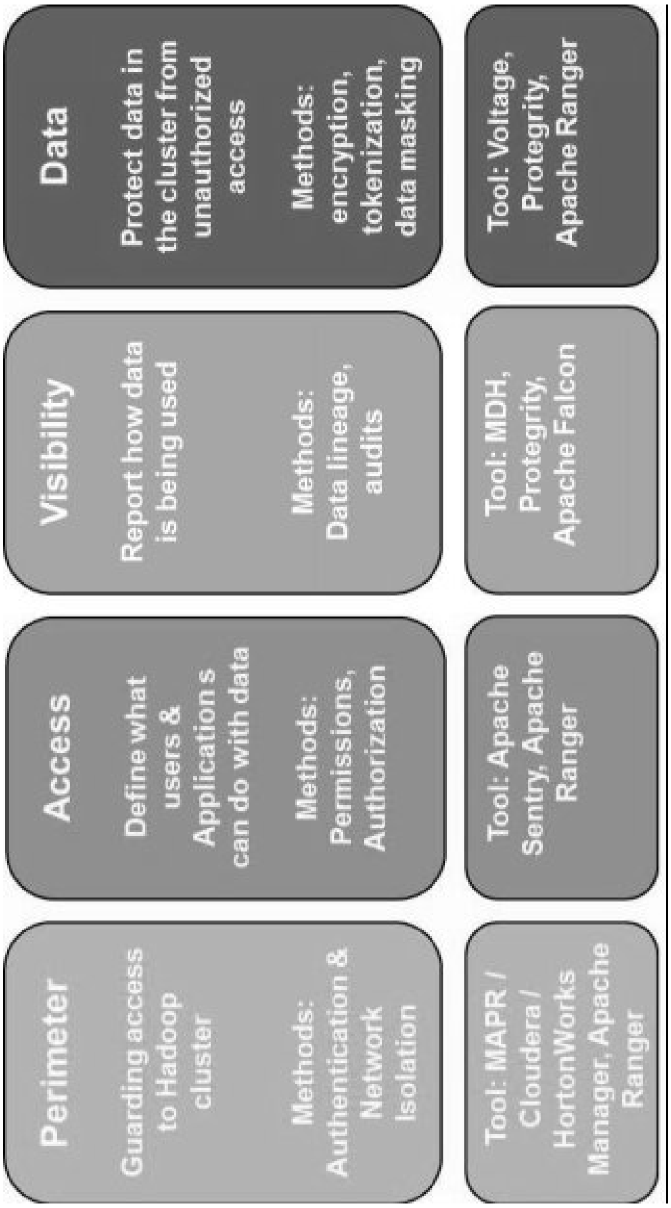
The goal of this security policy is to:

- Develop and update approaches to safeguard data at the application, level, sandbox level, File level and individual field/column level.
- Threat intelligence – gather, correlate, and provide intelligence to IT security from platform usage and access logs.
- Advanced threat monitoring – Monitor for advanced threats within the environment using specialized tools and processes conformant with enterprise IT security procedures.
- Develop audit capability of all activity at the user and data cell level with ability to provide data forensics analysis.

The Hadoop environment does not come with built-in security and therefore security tools and processes must be bolt-on to this environment. Since data flows in both directions from traditional data management environment to Hadoop and vice versa, consistent policies are required. The four pillars of Big Data security are:

1. Perimeter – Guarding access to the Hadoop cluster itself. This is possible via MapR Manager.
2. Access – Defines what users and applications can do with data. A tool of choice is Apache Sentry, but also Ranger[9] and other tools are possible.
3. Visibility – Reporting on how the data is being used. Support for this function comes from a combination of Meta Data 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.



Perimeter	Access	Visibility	Data
Guarding access to Hadoop cluster	Define what users & Application s can do with data	Report how data is being used	Protect data in the cluster from unauthorized access
Methods: Authentication & Network Isolation	Methods: Permissions, Authorization	Methods: Data lineage, audits	Methods: encryption, tokenization, data masking
Tool: MAPR / Cloudera / HortonWorks Manager, Apache Ranger	Tool: Apache Sentry, Apache Ranger	Tool: MDH, Protegrity, Apache Falcon	Tool: Voltage, Protegrity, Apache Ranger

Hadoop Data Security & Access Control Considerations

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.

Authentication – Manage access to 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.

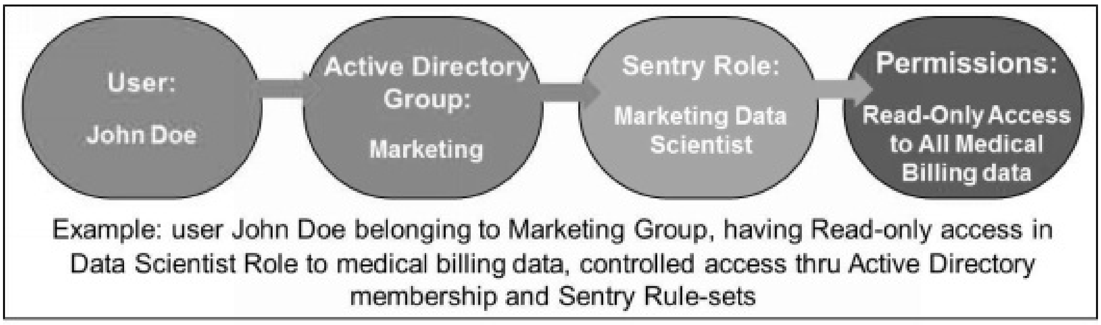
All intra-Hadoop services must mutually authenticate each other using Kerberos RPC. This internal check prevents rogue services to introduce themselves into the cluster activities via impersonation. Ensure Kerberos Ticket Granting Token feature is property configured.

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 directors, 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 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 intra-process accesses to data on various VMs are enforced. This is of particular importance within MapReduce as 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 are enabled.
- Ensure Simple Authentication and Security Layer (SASL) is enabled with RPC 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 HSFTP protocol for bulk data transfers.
- 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 a 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[10], 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 below:


User:
John Doe
Active Directory Group:
Marketing
Sentry Role:
Marketing Data Scientist
Permissions:
Read-Only Access to All Medical Billing data
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

Access Control Model Example for a Typical User

- Apply Sentry's features to use Active Directory (AD) to determine user's group assignments so 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: directories, files and ACLs such as 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 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, column family qualifier, to specific cell-level. Configure ACL permissions at both Group and User levels.

Visibility & 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 data (PII) and are subject to HIPAA, or PCI Data Security standards. Ensure the privacy category is complete in Meta Data Hub (MDH) tool for all data sets.
- Implement 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 manager 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.
- Data Usage Agreement defines appropriate and acceptable use of Critical and High-priority data for each data element.
- Data retention requirements – 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 defined escalation process.
- SMEs, Data Engineers and Data Scientists ("Users") agree to maintain and update the Meta Data Hub (MDH) (by registering their data) with the latest information about the data elements that they load into 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 the 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 3rd party data provider. Lineage of data elements must indicate the Third party and Fourth party data providers. (4th party data providers are those entities who provide data to the third party data provider).

VI. Security Operations Considerations & 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 guide 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 cloud. Data governance security best practices should indicate that while tokenized data may be stored in the cloud, the tokens must be maintained on premises 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 if 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 Steward 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 Access Request Form showing they've completed pre-requisites before getting access granted.
- Access Request forms must be approved by the LOB Data Risk Officer
- Create training material for users to include as minimum the following: Hadoop environment overview, data and application governance policies, polices about PII and sharing data with affiliates and publication, 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 meta data information, 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, No. 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 highest security control to the Kerberos Keytab file that contains the pairs of Kerberos principals and tickets to access Hadoop services.
- Configure Hadoop cluster to monitor NameNode Resources and generate alerts when any specified resources deviate from 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 below:
- The retention rules for data may vary from country to country. The longest retention period must be updated in the Meta Data Hub (MDH) Tool for all datasets.
- 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, their 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 retention period) should be purged according to the IT data purging procedures.
- Define Hadoop disaster recovery, downtime and business continuity policy. A downtime means that the 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 with resume, the process for restoring data, recovering data and synchronizing data after the failed site is online.

VIII. Quality Standards

Data Quality standard defines requirements to classify, document and manage data to ensure big data platform’s critical and high priority data meets established quality requirements. The objective of this standard is to ensure that platform’s data is managed and if 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 remediation plans and status with DRO, 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 Meta data monitoring, project work, quality assurance testing and audits.

All data whether internal or externally obtained is to be classified into 3 categories:

Critical Data – The type of data that has specific regulatory and compliance requirements such as HIPAA, FDA, FISMA, and Financial is materially significant to the decision making in the organization.

High Priority Data – the type of data that is materially significant to the decision making at the organization, regarded 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.

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.

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:

- Is applicable to data usage of critical and high priority data elements as designated by each franchise or country Account Executive.
- Sets criteria for the identification of data deemed most significant and impactful to 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 & 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 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 DRO of any issues related to data quality.
- Data quality issues are defined as gaps between the data element characteristics and deviations from this data governance standard.
- Understand how the critical and high-priority data with in each area of responsibility is being consumed and informing 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.
- Evidence of compliance to the data management policies described in this document and supporting standards, procedures and processes.
- Manage and prioritize mitigation activities to mitigate data quality risks, issues, including oversight by 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 with 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 Meta Data 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 Meta Data Hub (MDH) Tool
- Lineage to origination or to the Trusted Source
- Data Quality Monitoring & checks passed
- Retention period is defined

Data Quality Reporting

It's a requirement that AE/DRO/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:

- **Accuracy –** Measure the degree to which data correctly reflects its true value. It's a determination on 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.
- **Validity –** Validity measure the degree to which the 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
- **Completeness –** Measures the degree to which data is not missing and is of sufficient breadth and depth for the task in hand. In other words, it's fulfilling formal requirements and expectations with no gaps.
- **Timeliness –** Measures the degree to which data is available when it is required. Additionally, this measure ensures data is combined together from consistent time periods.
- **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 from Third-party data management audits and activities, the 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 below contains the criteria for triage of data quality issues:

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 by not time sensitive.
Issue Remediation Plan document within:	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

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 include it in the enterprise Risk Management Quarterly Risk Report.

Summary

This book has covered a wide range of topic and areas including modern data governance principles for a 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 ground work to design and implement a data governance structure for any industry and company of any size.

Big data will grow only 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.

ABOUT THE AUTHOR

Peter K. Ghavami received his PhD in Systems and Industrial Engineering from University of Washington in Seattle, specializing in big data analytics. He then served as Head of Data Analytics at Capital One Financial, Bank line of business. He received his BA from Oregon University in Mathematics with emphasis in Computer Science. He received his M.S. in Engineering Management from Portland State University. His career started as a software engineer, with progressive responsibilities at IBM Corp., as Systems Engineer. Later he became responsible as Director of engineering, Chief Scientist, VP of Engineering and Product Management at various high technology firms.

Before coming to Capital One Financial, he was Director of Informatics at UW Medicine leading numerous clinical system implementations and new product development projects. He has been 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: *Clinical Intelligence-The Big Data Analytics Revolution in Healthcare*, which has become an industry standard for big data analytics in biosciences.

Peter is on the advisory board of several clinical analytics companies and often invited as 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 HIMSS Data Analytics Task Force and advises Fortune 500 executives on data strategy.

[1] International Data Corporation (IDC) is a premier provider of research, analysis, advisory and market intelligence services

[2] Each Zetta byte is roughly 1000 Exabytes and each Exabyte is roughly 1000 Petabytes. A Petabyte is about 1000 TeraBytes.

[3] Auto Regressive Integrated Moving Average.

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

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

[6] http://pubs.opengroup.org/architecture/togaf9-doc/arch/

[7] “The What, Why and How of Master Data Management”, by roger Wolter, Kirk Haselden, Microsoft Corporation, Nov. 2006

[8] “The What, Why and How of Master Data Management”, by roger Wolter, Kirk Haselden, Microsoft Corporation, Nov. 2006

[9] 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.

[10] Impala initially developed by Cloudera requires Sentry for application token security. Hence, if you select Impala then you’ll require implementing Sentry. However, if you choose MAPR and Drill (instead of Impala) then Sentry is optional and not mandatory.