CHAPTER 1

An Introduction to Big Data Governance

We are drowning in data today. This data comes from social media, telephone GPS signals, utility smart meters, RFID tags, digital pictures, and online videos, among other sources. IDC estimates that the amount of information in the digital universe exceeded 1.8 zettabytes (1.8 *trillion* gigabytes) in 2011 and is doubling every two years.[1] Much of this data can be characterized as big data. Big data is generally referred to in the context of the "three Vs"—volume, velocity, and variety. We add another "V" for value. Let's consider each of these terms:

- *Volume (data at rest)*—Big data is generally large. Enterprises are awash with data, easily amassing terabytes and petabytes of information, and even zettabytes in the future.
- *Velocity (data in motion)*—Often time-sensitive, streaming data must be analyzed with millisecond response times to bolster real-time decisions.
- *Variety (data in many formats)*—Big data includes structured, semi-structured, and unstructured data such as email, audio, video, clickstreams, log files, and biometrics.
- *Value (cost effectiveness)*—Organizations are looking to gain insights from big data in a cost-effective manner. This is where open-source technologies such as Apache Hadoop have become extremely popular. Hadoop is software that processes large data sets across clusters of hundreds or thousands of computers in a cost-effective way.

Organizations must govern all this big data, which brings us to the subject of this book. We define big data governance as follows:

> **Big data governance** is part of a broader information governance program that formulates policy relating to the optimization, privacy, and monetization of big data by aligning the objectives of multiple functions.

Let's decompose this definition into its main parts:

- *Big data is part of a broader information governance program.*
  Information governance organizations should incorporate big data into their existing information governance frameworks by doing the following:
  - Extend the scope of the information governance charter to include big data governance.
  - Broaden the membership of the information governance council to include power users of big data such as data scientists.
  - Appoint stewards for specific categories of big data such as social media.
  - Align big data with information governance disciplines such as metadata, privacy, data quality, and master data.
- *Big data governance is about policy formulation.*
  Policy includes the written or unwritten declarations of how people should behave in a given situation. For example, a big data governance policy might state that an organization will not integrate a customer's Facebook profile into his or her master data record without that customer's informed consent.
- *Big data must be optimized.*
  Consider how organizations might apply the principles of the physical world to their big data. Companies have well-defined enterprise asset management programs to care for their machinery, aircraft, vehicles, and other assets. Similar to cataloging physical assets, organizations need to optimize their big data as follows:
  - *Metadata*—Build information about inventories of big data.
  - *Data quality management*—Cleanse big data just as companies conduct preventive maintenance on physical assets.
  - *Information lifecycle management*—Archive and retire big data when it no longer makes sense to retain these massive volumes.

- *Privacy of big data is important.*
  Organizations also need to establish the appropriate policies to prevent the misuse of big data. Organizations need to consider the reputational, regulatory, and legal risks involved when handling social media, geolocation, biometric, and other forms of personally identifiable information.
- *Big data must be monetized.*
  Monetization is the process of converting an asset such as data into money by selling it to third parties or by using it to develop new services. Traditional accounting rules do not allow companies to treat information as a financial asset on their balance sheets unless purchased from external sources. Despite this conservative accounting treatment, organizations now recognize that they should treat big data as an enterprise asset with financial value. For example, operations departments can use sensor data to increase equipment uptime based on preventive maintenance programs. Call centers can analyze agents' notes to reduce call volumes by understanding why customers call. In addition, retailers can use master data to power Facebook apps that drive customer loyalty.
- *Big data exposes natural tensions across functions.*
  Big data governance needs to harmonize competing objectives across multiple functions. For example, the wireless marketing department at a telecommunications carrier might be interested in using geolocation data to drive new revenue streams, such as when a subscriber receives coupons from retailers that are in close proximity. However, the wireline business might be concerned about the reputational hazard associated with reusing subscribers' geolocation data without their consent. Meanwhile, the network management team might want to use this information to address any issues with network performance, such as a large number of dropped calls at a specific wireless tower. Finally, the chief privacy officer might have concerns about the potential for regulatory backlash. In this situation, big data governance needs to bring all the parties together to determine whether the potential revenue upside from the new services outweighs the associated reputational and regulatory risks. The usage of geolocation data for internal network analytics is probably okay, but the other business uses might not be.

Case Study 1.1 reviews the unfortunate events surrounding the Mars Climate Orbiter. We would not consider this volume of data to be "big" by today's standards. However, NASA likely produced the navigation commands by crunching some very big numbers with complex mathematics. If commercial organizations do similar crunching of big data to score a risk, fraud, or propensity to buy, they might incorrectly reject credit card applications or miss customer churn events because scores are misunderstood or applied incorrectly.

Case Study 1.1: Big data governance and the Mars Climate Orbiter[2,3,4]

Any effort to launch objects into space requires immense amounts of data. The ill-fated mission by NASA to launch the Mars Climate Orbiter is a good example of the lack of governance over big data.

In 1999, just before orbital insertion, a navigation error sent the satellite into an orbit 170 kilometers lower than the intended altitude above Mars. One of the most expensive measurement incompatibilities in space exploration history caused this error. NASA's engineers used English units (pounds) instead of NASA-specified metric units (newtons). This incompatibility in the design units resulted in small errors being introduced in the trajectory estimate over the course of the nine-month journey and culminated in a huge miscalculation in orbital altitude. Ultimately, the orbiter could not sustain the atmospheric friction at low altitude. It plummeted through the Martian atmosphere and burned up.

This relatively minor mistake resulted in the loss of $328 million for the orbiter and lander, in addition to setting space exploration back by several years in the United States.

In a typical information governance project, the team identifies a business problem, develops a business case, obtains an executive sponsor, defines a technical architecture, and proceeds with the rest of the initiative. However, big data projects are different because of the following characteristics:

- Projects are driven by early adopters.
- The business problem needs to be discovered.
- IT is often at the forefront with technologies such as Hadoop.
- The business case has not been developed.
- The characteristics of the data are unclear.

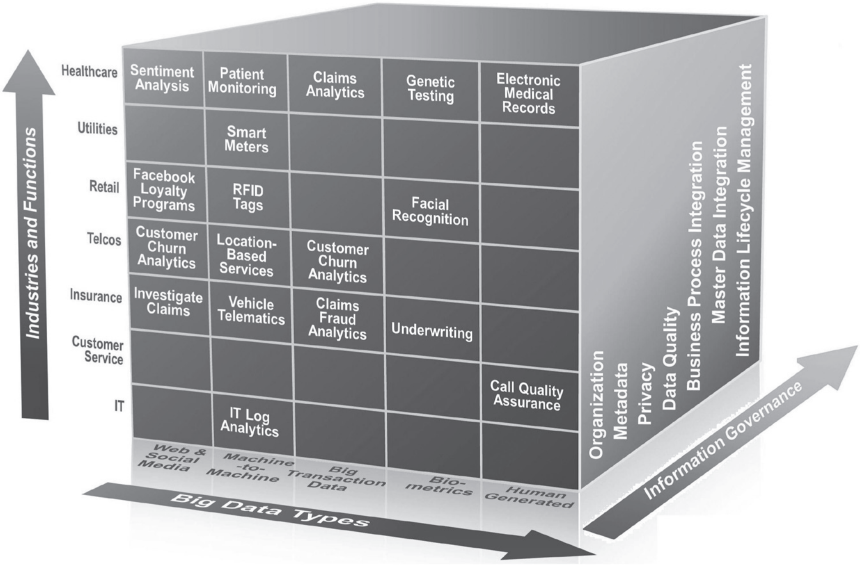
As of the publication of this book, governance has taken a backseat to the analytics and technologies associated with big data. However, as big data projects become mainstream, we anticipate that privacy, stewardship, data quality, metadata, and information lifecycle management will coalesce into an emerging imperative for big data governance.

CHAPTER 2

THE BIG DATA GOVERNANCE FRAMEWORK

This chapter provides a framework for big data governance. As shown in Figure 2.1, this framework consists of three dimensions:

- *Big data types*—Big data governance needs a heightened focus on the data itself. We have classified big data into five distinct types: web and social media, machine-to-machine, big transaction data, biometrics, and human generated.
- *Information governance disciplines*—The traditional disciplines of information governance also apply to big data. These disciplines are organization, metadata, privacy, data quality, business process integration, master data integration, and information lifecycle management.
- *Industries and functions*—Big data analytics are driven by use cases that are specific to a given industry or function. Given space limitations, we have only included a handful in Figure 2.1. Big data analytics can be leveraged by many other industries and functions, including marketing, risk management, customer service, information security, information technology, and human resources.



	Web & Social Media	Machine-to-Machine	Big Transaction Data	Bio-metrics	Human Generated
Healthcare	Sentiment Analysis	Patient Monitoring	Claims Analytics	Genetic Testing	Electronic Medical Records
Utilities		Smart Meters			
Retail	Facebook Loyalty Programs	RFID Tags		Facial Recognition	
Telcos	Customer Churn Analytics	Location-Based Services	Customer Churn Analytics		
Insurance	Investigate Claims	Vehicle Telematics	Claims Fraud Analytics	Underwriting	
Customer Service					Call Quality Assurance
IT		IT Log Analytics			

Figure 2.1: A framework for big data governance.

We will discuss each dimension in this chapter.

2.1 Big Data Types

As shown in Figure 2.2, big data can be broadly classified into five types.

Let's consider each type in more detail:

1. *Web and social media*—This includes clickstream and social media data such as Facebook, Twitter, LinkedIn, and blogs. Big data governance programs will increasingly be required to integrate this data with master data and with core business processes such as customer loyalty programs. The big data governance program needs to establish policies regarding the acceptable use of social media data, especially since regulations and precedents are continually evolving. The program also needs to establish guidelines regarding the acceptable use of cookies, especially third-party cookies, to track users and to personalize their web interactions. Metadata is also critical to web and social media. For example, two sites might measure the term "unique visitors" differently for clickstream analytics. One site might measure unique visitors within a month, while the other might measure unique visitors within a week.