PART ONE

Getting Started

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, discussed in detail

1 *The 2011 Digital Universe Study: Extracting Value from Chaos* (IDC, 2011).

later in this book, is software that processes large data sets across clusters of hundreds or thousands of computers in a cost-effective manner.

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.

2 http://en.wikipedia.org/wiki/Mars_Climate_Orbiter.

3 "Mars Climate Orbiter Fact Sheet." http://mars.jpl.nasa.gov/msp98/orbiter/fact.html.

4 "Mars Climate Orbiter Mishap Investigation Board Phase I Report." November 1999.

- 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 uses 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.

We will discuss each dimension in this chapter.

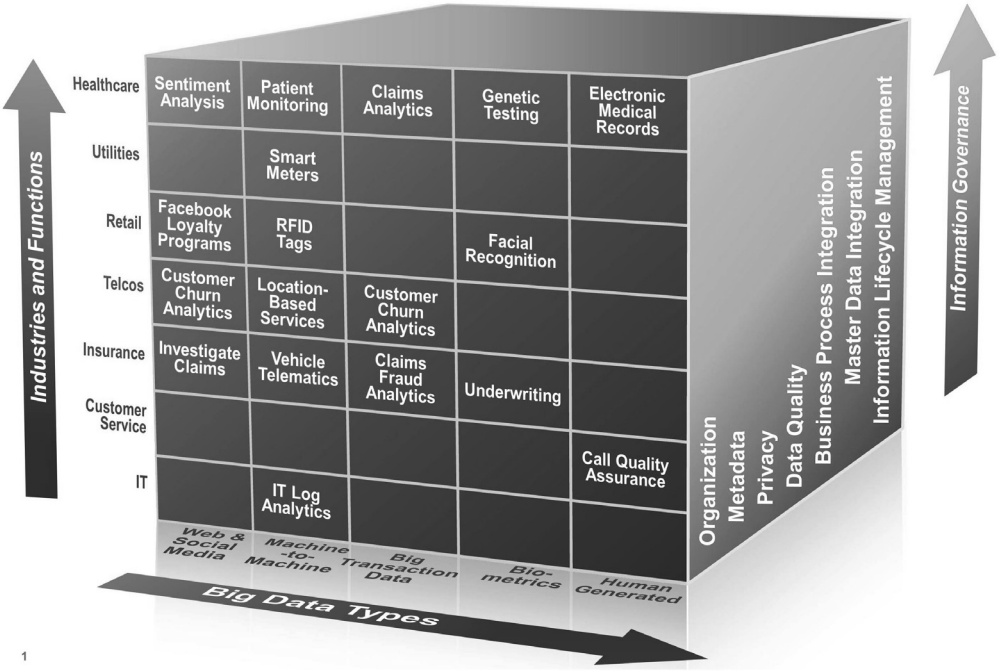

Figure 2.1: A framework for big data governance.

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

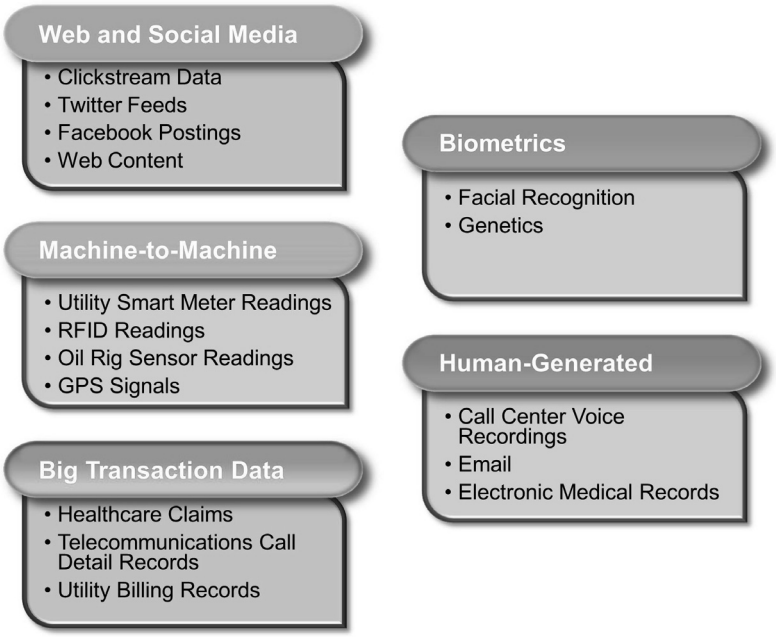

Figure 2.2: Big data types.

analytics. One site might measure unique visitors within a month, while the other might measure unique visitors within a week.

2. *Machine-to-machine data*—Machine-to-machine, or *M2M*, refers to technologies that allow both wireless and wired systems to communicate with other devices. M2M uses a device such as a sensor or meter to capture an event such as speed, temperature, pressure, flow, or salinity. This event is relayed through a wireless, wired, or hybrid network to an application that translates the captured event into meaningful information. M2M communications create the so-called "internet of things." The big data governance program needs to establish a number of policies around M2M data. For example, the program needs to draw up guidelines around the acceptable use of geolocation and RFID data that can be used to build a profile of individuals and potentially violate their privacy. The program needs to establish retention policies around the massive volumes of M2M data, which can easily overwhelm IT budgets if not properly controlled. The big data governance program needs to address any data quality concerns such as RFID read rates in environments with high moisture content

and lots of congestion. Finally, the big data governance program needs to secure the Supervisory Control and Data Acquisition (SCADA) infrastructure from vulnerability to cyber attacks.

3. *Big transaction data*—This includes healthcare claims, telecommunications CDRs, and utility billing records. Big transaction data is increasingly available in semi-structured and unstructured formats. Information governance challenges such as metadata, data quality, privacy, and information lifecycle management also apply to this data.
4. *Biometrics*—Biometric recognition, or biometrics, refers to the automatic identification of a person based on his or her anatomical or behavioral characteristics or traits.[1] Anatomical data is created from the physical characteristics of a person including a fingerprint, an iris, a retina, a face, an outline of a hand, an ear shape, a voice pattern, DNA—even body odor. Behavioral data includes handwriting and keystroke analysis.[2] Advances in technology have vastly increased the available biometric data. Law enforcement, the legal system, and intelligence agencies have been using this information for a long time. However, biometric data is increasingly available in the commercial arena, where it can be combined with other types of data such as social media. This opens up new business opportunities as well as several governance issues relating to privacy and data retention.
5. *Human-generated data*—Human beings generate vast quantities of data such as call center agents' notes, voice recordings, email, paper documents, surveys, and electronic medical records. This data might contain sensitive information that needs to be masked. It might also contain insights that can improve the quality of structured data sets and integrate with MDM. In addition to dealing with these issues, organizations need to establish policies regarding the retention period for this data to adhere to regulations and manage storage costs.

2.2 Information Governance Disciplines

The seven core disciplines of information governance also apply to big data:

1 "An Overview of Biometric Recognition." http://biometrics.cse.msu.edu/info.html/.

2 "Biometrics in the workplace." Data Protection Commissioner of Ireland. http://dataprotection.ie/viewdoc.asp?DocID=244.

1. *Organization*—The information governance organization needs to consider adding big data to its overall framework, including the charter, organization structure, and roles and responsibilities. The information governance council might seek new members who can provide a unique perspective on big data, such as data scientists. It might also decide to appoint stewards for social media, RFID, and other types of big data. Finally, the information governance program might add additional responsibilities to the job descriptions of existing stewards. For example, the customer data steward might be accountable for the Twitter handles and Facebook accounts within the master data repository.
2. *Metadata*—The big data governance program needs to integrate big data with the enterprise metadata repository. This involves the following activities:
   - Include big data terms within the business glossary. For example, add the term "unique visitor" to support clickstream analytics.
   - Import technical metadata from Hadoop into the metadata repository.
   - Ensure that the data lineage administrator is able to import flows from Hadoop into the technical metadata repository.
   - Manage data lineage and impact analysis within the big data environment.
3. *Privacy*—As far back as 1890, Louis Brandeis (later a justice of the United States Supreme Court) and Samuel Warren published an article called "The Right to Privacy" in the *Harvard Law Review*. This article defined privacy as the "right to be left alone."[3] Subsequent regulations and legislation around the world have formalized and expanded this theory of privacy. Big data governance needs to identify sensitive data and establish policies regarding its acceptable use. These policies need to consider regulations that vary by big data type, industry, and country. Given the many headlines on the subject, the big data governance program needs to establish guidelines regarding the acceptable use of social media and geolocation data, if applicable.
4. *Data quality*—Data quality management is a discipline that includes the methods to measure, improve, and certify the quality and integrity of an organization's data. Because of its extreme volume, velocity, and variety, big data quality needs to be handled differently than traditional data types. For example, big data quality might need to be handled in real-time and address

3 Warren, Samuel and Brandeis, Louis. "The Right to Privacy." *Harvard Law Review*, Vol. IV No. 5, December 15, 1890.

issues relating to semi-structured and unstructured data. Big data needs to be "good enough" because poor data quality does not necessarily impede the analytics that are required to derive business insights.

5. *Business process integration*—The program needs to identify key business processes that require big data. The program then needs to define key policies to support the governance of big data. For example, drilling and production are key processes within oil and gas. The big data governance program must establish policies around the retention period for sensor data such as temperature, flow, pressure, and salinity on an oil rig. Not only is this data costly to store, but it might also be required by regulators to justify an operator's actions in case of an oil spill. In another example, a retailer might establish a policy that it will access a customer's Facebook profile, including his or her list of friends, only if it has obtained informed consent via a Facebook app. The retailer will obtain the informed consent from the customer in exchange for discounts on certain products as part of an overall loyalty program.
6. *Master data integration*—The big data governance program needs to establish policies regarding the integration of big data into the master data management environment. As discussed above, a retailer needs to first define policies for the acceptable use of social media. The retailer then needs to deploy the appropriate data stewardship policies and tools to determine if the "Susie Smith" on Facebook is the same as the "Susan Smith" in the customer master.
7. *Information lifecycle management*—Because of the massive increase in big data volumes, organizations will be challenged to understand the regulatory and business requirements that determine what data to retain in operational and analytical systems, what data to archive, and what data to delete. Without a high level of specificity regarding the legal and regulatory obligations of information, IT must manage all data as if it had high value and ongoing obligations, or the company faces a very high risk of improper disposal. With IT budgets continuing to be under pressure, over-managing information is a gross waste of capital resources. The program needs to expand the retention schedule to include big data based on regulations and business needs. The big data governance team needs to create pointers to the physical repositories of big data to facilitate records retention and e-discovery activities. The big data governance program needs to leverage compression and archiving policies, tools, and best practices to reduce storage costs and to improve application performance. Finally, the

organization needs to defensibly dispose of big data that is no longer required based on regulations and business needs.

2.3 Industry and Functional Scenarios for Big Data Governance

Before going any further, it's important to discuss experimentation in big data projects where the killer use case might not be known upfront. Tom Deutsch argues that organizations should adopt experimentation as a strategic process, along with a tolerance (and even encouragement) of failure as a way of advancing the business. This tolerance of experimentation is rooted in the understanding that failure is often the single best way to learn something. Deutsch further argues that organizations that are not trying and failing at validating their analytic hypotheses are actually not being aggressive enough in working with their data.[4] Notwithstanding all of this, big data needs to be governed with specific reference to the analytic and operational requirements for applications that vary by industry and function. We discuss these scenarios by industry and function in the following pages.

Healthcare Industry, Scenario 1

Solution: *Sentiment analysis*
Big data type: *Web and social media (health plans)*
Disciplines: *Privacy*

Because of privacy regulations such as the United States Health Insurance Portability and Accountability Act (HIPAA), health plans are somewhat limited in what they can do online.

If someone posts a complaint on Twitter, the health plan might want to post a limited response and then move the conversation offline.

Healthcare Industry, Scenario 2

Solution: *Patient monitoring*
Big data type: *M2M data (healthcare providers)*
Disciplines: *Data quality, information lifecycle management, privacy*

A hospital leveraged streaming technologies to monitor the health of newborn babies in the neonatal intensive care unit. Using streaming technologies, the hospital was

4 Deutsch, Tom. "Failure Is Not a Four-Letter Word." *IBM Data Management*, April 17, 2012. http://ibmdatamag.com/2012/04/failure-is-not-a-four-letter-word/.

able to predict the onset of disease a full 24 hours before the onset of symptoms. The application depended on large volumes of time series data. However, the time series data was sometimes missing when a patient moved, which caused a lead to disengage and stop providing readings. In these situations, the streaming platform used linear and polynomial regressions to fill in gaps in the historical time series data. The hospital also stored both the original and modified readings in the event of a lawsuit or medical inquiry. Finally, the hospital established policies around safeguarding protected health information.

Healthcare Industry, Scenario 3

Solution: *Claims analytics*
Big data type: *Big transaction data (health plans)*
Disciplines: *Data quality*

A large health plan processed over 500 million claims per year, with each claims record consisting of 600 to 1,000 attributes. The plan was using predictive analytics to determine if certain proactive care was required for a small subset of members. However, the business intelligence team found that physicians were using inconsistent procedure codes to submit claims, which limited the effectiveness of the predictive analytics. The business intelligence team also analyzed the text within claims documents. The team used terms such as "chronic congestion" and "blood sugar monitoring" to determine that members might be candidates for disease management programs for asthma and diabetes, respectively.

Healthcare Industry, Scenario 4

Solution: *Employee credentialing*
Big data type: *Biometrics (healthcare providers)*
Disciplines: *Privacy*

The use of biometric information continues to expand in the marketplace and the workplace. For example, hospital staff members might use biometrics so that they do not have to retype their user names and passwords. Instead, they can use a biometric scan to quickly log into the system when they arrive at a bed or station. Hospitals need to work with legal counsel, however, to conduct privacy impact assessments regarding the use and retention of biometric data about their employees.