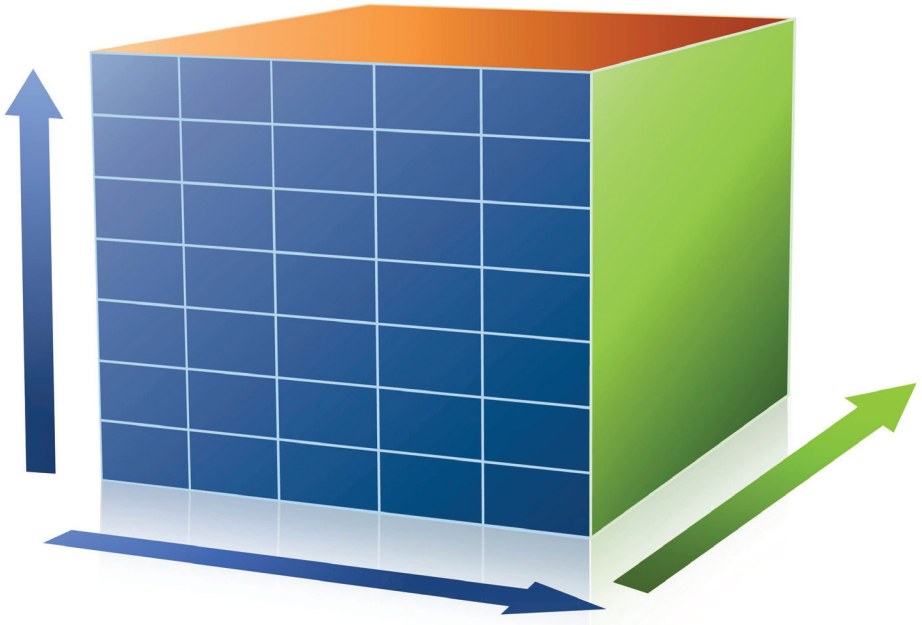
IBM InfoSphere:

A Platform for Big Data Governance and Process Data Governance



Sunil Soares

IBM InfoSphere:

A Platform for Big Data Governance and Process Data Governance

Sunil Soares

MC Press Online, LLC
Boise, ID 83703

IBM InfoSphere: A Platform for Big Data Governance and Process Data Governance
Sunil Soares

First Edition
First Printing — February 2013

MC Press Online, LLC, 3695 W. Quail Heights Court, Boise, ID 83703-3861 USA
Customer Service: (208) 629-7275 ext. 500; service@mcpressonline.com
Permissions and Special/Bulk Orders: mcbooks@mcpressonline.com
On the Web: www.mc-store.com

ISBN: 978-1-58347-382-5

To my beautiful daughters
Lizzie and Maya

About the Author

Sunil Soares is the founder and managing partner of Information Asset, LLC, a consulting firm that specializes in helping organizations build out their information governance programs. Prior to this role, Sunil was the director of information governance at IBM, where he worked with clients across six continents and multiple industries.

Sunil's first book, *The IBM Data Governance Unified Process* (MC Press, 2010), details the 14 steps and almost 100 sub-steps to implement an information governance program. The book has been used by several organizations as the blueprint for their information governance programs and has been translated into Chinese. Sunil's second book, *Selling Information Governance to the Business: Best Practices by Industry and Job Function* (MC Press, 2011), reviews the best practices to approach information governance by industry and function. Sunil's third book, *Big Data Governance: An Emerging Imperative* (MC Press, 2012), discusses the governance of different types of big data.

Prior to joining IBM, Sunil consulted with major financial institutions at the Financial Services Strategy Consulting Practice of Booz Allen & Hamilton in New York. Sunil lives in New Jersey and holds an MBA in finance and marketing from the University of Chicago Booth School of Business.

Contents

Foreword

by David Corrigan

Big data is certainly a hot topic right now. Most organizations are only beginning their adventure with big data. Many are experimenting with new technology, but few are thinking ahead for long-term success. And while most organizations are pursing big data in order to improve their competitiveness, the fact is fewer still will truly succeed in gaining competitive advantage. The reason is simple: you cannot compete on analytics alone. After all, what does analytics analyze? Information. Information needs to be trusted in order to be acted upon. That is the role of Information Governance: to create trusted information to make big data analytics more successful.

First, there is the technology aspect of Information Governance. IBM® InfoSphere® is the leading platform for creating trusted information. It has deep capabilities in data quality, privacy and security, master data management, information integration, information lifecycle management, metadata management, and policy management. Those capabilities have been well integrated to address multiple requirements. InfoSphere is a key component of IBM's big data platform, and it provides a foundation of trusted information for big data analytics. It is integrated with other big data components for Hadoop-based analytics, streaming analytics, data warehousing, and federated discovery and navigation. The importance of this integration cannot be emphasized enough, as it helps to integrate big data technologies within the enterprise and also to supply them with a steady flow of trusted information. IBM has many clients in multiple industries and has deep industry experience in implementing big data and Information Governance.

Second, there is the process, or strategy, of Information Governance. The importance and the role of a governance strategy are still not well-understood. Information Governance is a business strategy that has a series of IT deliverables. Sunil has been one of the pioneers in this area, defining the Unified Information Governance Process several years ago. He defined several key steps, such as identifying a business problem and executive sponsor, setting up cross-functional governance boards, and measuring and communicating success. He has applied this process at hundreds of clients and has helped them achieve successful implementations. His approach can also be applied to governing big data. It has helped many organizations get the business involved

in governance and establish trusted information for a key enterprise application. In short, this process helps you move beyond an IT project toward a true business strategy. It helps by getting business executives and owners involved in the process of governing data. It helps ensure successful outcomes.

Sunil, thank you for continuing to contribute to the discipline of Information Governance and move it into the new era of computing—the era of big data. And to the readers of this book, remember that the competitive advantage you seek from insights garnered from big data has *two* components: big data analytics and trusted information. Information Governance creates trusted information from very uncertain sources, enabling you to trust and act upon the insights from analytics. I wish you well in your big data strategy.

David Corrigan
Director, Product Marketing, InfoSphere
IBM

Foreword

by Inderpal Bhandari

Editor's Note: *These remarks were originally published in the book* Big Data Governance: An Emerging Imperative *(Sunil Soares, MC Press, October 2012), from which some of the material in this book has been excerpted.*

We now live in an age of seemingly unlimited data. It has slowly, but surely, pervaded our lives. We rely on it to accomplish all manner of tasks, ranging from governing economies and advancing science, to maintaining an electronic record of our health. We have come to realize that its value must be truly understood and unlocked by deriving insights that are revealed through analysis and then translating those insights into information, knowledge, and ultimately action. Recently, with the advent of social media, sensor networks, streaming technology, and the like, the sheer scale of work required to unlock this value has increased beyond the scope of traditional database and data warehousing technologies. In other words, we have entered the age of Big Data.

What is Big Data? What is Big Data Governance? And how does one create business value through Big Data? These and many more questions about Big Data are answered within *Big Data Governance: An Emerging Imperative* (MC Press, October 2012). If you have not yet asked these questions, you *need* to read this book or miss out on the opportunity to differentiate and grow your business by leveraging big data.

Through detailed case studies, Sunil does an excellent job of educating the reader on a complicated and still-evolving subject. For example, he shares a story where a retail outlet uses social media data to analyze consumer awareness and sentiment for a wide range of products. That understanding was used by the retailer to optimize discounts for the products, greatly increasing the firm's profitability.

Sunil shares over twenty such case studies in his book. While reading these case studies, I encourage you to go through a two-step mental exercise, as I did. First, consider how one might attempt to attain the same results as described in the case study, but by using traditional data technologies and processes. Returning to the retail example, by commissioning market research surveys, one can understand the awareness and market sentiment for a range of products, but that process will take months. The research results, when available, are no

longer valid for setting discounts effectively, as awareness and sentiment may be completely different by then. We are now in an era where social media data must be harnessed, using Big Data approaches, to provide business intelligence that is far more timely than is possible through traditional approaches. Second, consider the gaps that exist in current data management processes and systems. Again in the retail example, it is clear that we must be able to match social media data to the products, but the matching rules in traditional master data management systems might not be equal to that task. For instance, these matching rules require natural language processing capability. That is a gap. We then need to provide governance around such rules. Yet another gap. And so on.

Sunil makes this comparative exercise easy by identifying and discussing several such gaps for each case study. The end result is that one comes away with a sound understanding of Big Data Governance, along with answers to the questions I raised earlier in this foreword.

Enough said. Read this book. It will help you enter the age of Big Data.

Inderpal Bhandari
Vice President and Chief Data Officer
Express Scripts

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.

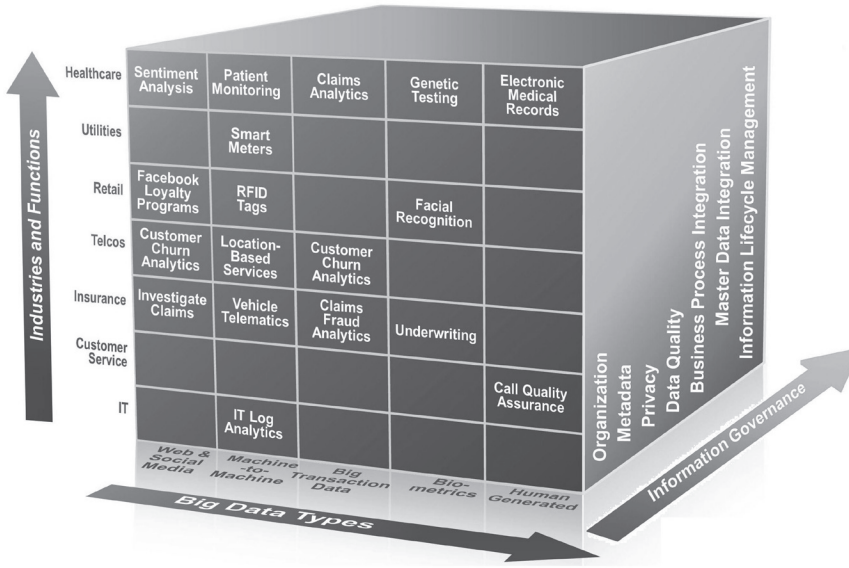


Figure 2.1: A framework for big data governance.

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.

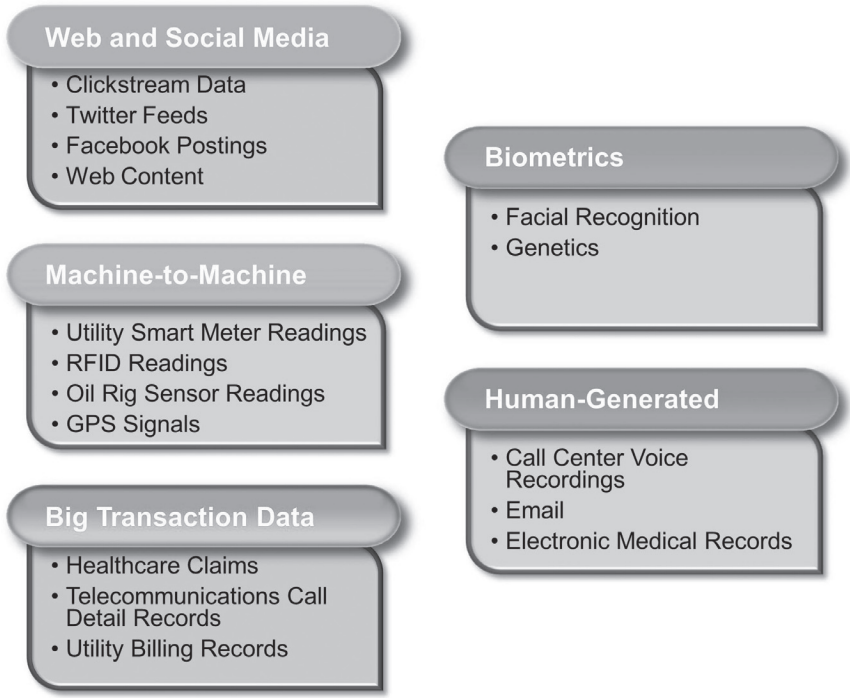


Figure 2.2: Big data types.

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 call detail records (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.[5] 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.[6] 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. *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."[7] 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 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 eDiscovery 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

We discuss these scenarios by industry and function in the following pages.

Healthcare Industry

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.

Utilities Industry

Solution: *Smart meters*
Big data type: *M2M data*
Disciplines: *Privacy, information lifecycle management*

Several utilities are rolling out smart meters to measure the consumption of water, gas, and electricity at regular intervals of an hour or less. These smart meters generate copious amounts of "interval" data that need to be governed appropriately. Utilities need to safeguard the privacy of this interval data because it can potentially point to a subscriber's household activities, as well as the comings and goings from his or her home. In addition, utilities need to establish policies for the archival and deletion of interval data to reduce storage costs.

Retail Industry

Solution: *Facebook loyalty app*
Big data type: *Web and social media*
Disciplines: *Privacy, master data integration*

The marketing department at a retailer might want to use master data on customers, products, employees, and store locations to enrich its Facebook app. The success of the Facebook app depends on a strong foundation of master data management and policies pertaining to social media governance. For example, the retailer needs to adhere to the Facebook Platform Policies by not using data

on a customer's friends outside of the context of the app. In addition, the retailer needs to leverage a consistent set of identifiers to link a customer's Facebook profile with his or her MDM record. Finally, the retailer needs to establish a robust product hierarchy to enable product comparisons. As a simple example, the retailer would need to know that a customer who purchased a "Whirlpool GX5FHDXVY" already had a product in the "refrigerator" hierarchy.

Telecommunications Industry

Solution: *Location-based services*
Big data type: *M2M data*
Disciplines: *Privacy*

Let's consider an example of the marketing department at a telecommunications operator that wants to unlock new sources of revenue. The marketing team wants to sell geolocation data to third parties, who can offer coupons to subscribers based on their proximity to certain retailers. However, the privacy department is concerned about the reputational and regulatory risks associated with sharing subscriber geolocation data. The big data governance program needs to balance the revenue potential from a new revenue source with the privacy risks involved.

Insurance Industry

Solution: *Claims investigation, underwriting*
Big data type: *Web and social media*
Disciplines: *Privacy, business process integration*

Many insurance carriers now use social media to investigate claims. However, most regulators still do not permit insurers to use social media to set rates on policies during the underwriting process. For example, can a life insurer use the fact that an applicant's Facebook profile indicates that she is a skydiver to increase her premiums because of her higher risk profile?

Oil and Gas Industry

Solution: *Rig and environmental monitoring*
Big data type: *M2M data*
Disciplines: *Information lifecycle management*

In the past, an oil rig might have had only about 1,000 sensors, of which only about 10 fed databases that would be purged every two weeks due to capacity limitations. Today, a typical facility might have more than 30,000 sensors.

Oil and gas companies also need to retain sensor data for a longer period. This information needs to be maintained well after the lifetime of the rig itself, to demonstrate adherence to environmental regulations. As a result, this information might need to be stored for 50 to 70 years, and up to 100 years in some cases. While storage is cheap, it is not free. The big data governance program needs to establish retention schedules for sensor data, and archiving policies to move information on to cheaper storage, if possible.

Banking Industry

Solution: *Risk management*
Big data type: *Web and social media (web content)*
Disciplines: *Master data integration*

Risk management departments need to update their customer hierarchies based on up-to-date financial information. As an example, when Tata Motors acquired Jaguar, the risk management department needed to update the risk hierarchy for Tata Motors to also include any exposure to Jaguar. In another example, a bank has developed an economic hierarchy to aggregate its overall exposure to a car manufacturer, its tier-one and tier-two suppliers, and the employees of the manufacturer and its suppliers. The risk management department might want to update its economic hierarchy in the event of consolidation between suppliers. All these hierarchies depend on up-to-date financial information. The risk management department might use text analytics to crawl through unstructured financial information such as U.S. Securities and Exchange Commission (SEC) 10-K and 10-Q filings, and U.S. Federal Deposit Insurance Corporation (FDIC) call reports to dynamically update changes in company ownership, directors, and officers within its MDM hierarchies.

Information Technology Function

Solution: *Log analytics*
Big data type: *M2M data*
Disciplines: *Metadata*

IT departments are turning to big data to analyze application logs for slivers of insight that can improve the performance of systems. Because application vendors' log files are in different formats, they need to be standardized first, before anything useful can be done with them.

Information Security Function

Solution: *Network analytics*
Big data type: *M2M data*
Disciplines: *Metadata*

Security Information and Event Management (SIEM) tools aggregate log data from systems, applications, network elements, and security devices across the enterprise. SIEMs correlate the aggregated information to determine if a security incident might be taking place. SIEMs then determine the origin of the security incident and apply recourse techniques to interrupt the flow of information to prevent further data leakage. From a big data governance perspective, security professionals need to grapple with inconsistent nomenclature across network elements such as firewalls, virtual private networks, bridges, and routers from different vendors. It is highly likely that the log files from two network elements will refer to the same event using different codes. Security professionals need to normalize these event codes prior to SIEM analytics.

Summary

There are five different types of big data: web and social media, M2M, big transaction data, biometrics, and human generated. In addition, there are seven information governance disciplines: organization, metadata, privacy, data quality, business process integration, master data integration, and information lifecycle management. This chapter provides a framework so that organizations can tailor their governance programs by big data type, information governance discipline, industry, and function.

CHAPTER 3

THE IBM BIG DATA PLATFORM[8]

This chapter includes contributions from Jim Hare (IBM) and Lawrence Weber (IBM).

IBM has developed an enterprise-class big data platform that allows organizations to fully address the spectrum of big data business challenges. The platform blends traditional technologies that are well-suited for structured tasks with complementary new technologies that address speed and flexibility and are ideal for ad hoc data exploration, discovery, and unstructured analysis.

IBM's big data strategy includes placing the analytics as close as possible to the big data sources. This structure enables organizations to cost-effectively manage and analyze data-at-rest and data-in-motion in its native format, whether structured, semi-structured, or unstructured.

Figure 3.1 gives an overview of the IBM® big data platform.

Figure 3.1: The IBM big data platform.

3.1 IBM Big Data Products

The IBM big data platform encompasses the following products and capabilities.

IBM InfoSphere Data Explorer

Vivisimo, acquired by IBM in May 2012, was a leading provider of federated discovery and navigation software to help organizations access and analyze big data. Based on IBM Vivisimo® Velocity™, IBM InfoSphere® Data Explorer automates the discovery of big data, regardless of its format or where it resides, providing a federated view of key business information necessary to drive

new initiatives. The technology is characterized by its unique index and search capabilities that uncover data from multiple repositories, making it valuable to clients in any industry.

IBM InfoSphere BigInsights

IBM InfoSphere BigInsights™ is IBM's mature Hadoop-based solution for big data analytics. IBM InfoSphere BigInsights is available in two editions: Basic and Enterprise.

Basic Edition

IBM InfoSphere BigInsights Basic Edition, the entry-level offering, is an integrated, tested, and pre-configured, no-charge download for anyone who wants to experiment with Hadoop.

Enterprise Edition

IBM InfoSphere BigInsights Enterprise Edition combines the power of Apache Hadoop with IBM-unique innovations and enhancements to make Hadoop enterprise-ready. These enhancements include:

- *Advanced analytics*—Sophisticated text analytics developed by IBM research with a vast library of extractors enabling actionable insights from large amounts of native textual data.
- *Multi-distribution support*—Ability to use the built-in Hadoop distribution provided with IBM InfoSphere BigInsights or any other distribution, including Cloudera's Distribution Including Apache Hadoop (CDH). IBM InfoSphere BigInsights sits on top of and enhances any Hadoop open-source distribution to make it enterprise-ready.
- *Performance optimizations:*
  - *Adaptive MapReduce*—Adapts to user needs and system workloads automatically to improve performance and simplify job tuning without the need for users to understand and manipulate the many tuning knobs in Hadoop.
  - *Workload Scheduler*—Provides optimization and control of job scheduling based on user-selected metrics and large-scale indexing for low-latency access to data and job results by external applications.
- *Visualization*—Includes IBM BigSheets, a spreadsheet-like tool, allowing data scientists and business users to explore IBM InfoSphere BigInsights collections and discover new insights without writing any code.

- *Enterprise integration*—High-speed data connectors for IBM DB2®, IBM Netezza®, IBM Smart Analytics System (ISAS), IBM InfoSphere Warehouse, IBM InfoSphere Streams, and IBM InfoSphere Information Server to enhance the value of big data by combining it with other data within the enterprise to gain even more insight.
- *Application connectors*—Provides access to any JDBC-compatible data store, making IBM InfoSphere BigInsights data accessible to a variety of applications, including business intelligence offerings such as IBM Cognos® Business Intelligence.
- *Multi-data type query*—Supports Jaql, an open-source declarative language that provides the capability to process both structured and non-traditional data. Jaql is extensible via modules, and IBM is leveraging this capability to include pre-built Jaql modules within IBM InfoSphere BigInsights, enabling integration with text analytics, HBase, and IBM Netezza.
- *Security*—Built-in Lightweight Directory Access Protocol (LDAP) support, enabling administrators to restrict access to only those users who possess the appropriate authorization.

IBM InfoSphere Streams

IBM InfoSphere Streams is a high-performance computing platform that allows user-developed applications to rapidly ingest, analyze, and correlate information as it arrives from thousands of real-time sources. It can handle incredibly high data throughput rates that can range to millions of events or messages per second.

IBM InfoSphere Streams enables organizations to leverage massively parallel processing (MPP) capabilities to analyze big data in motion. As opposed to storing large volumes of structured, unstructured, and semi-structured data on disk for analytics, IBM InfoSphere Streams applies analytics to the data while it is in motion.

IBM InfoSphere Streams grew out of early work with the United States government starting in 2003. The tool is currently deployed in a number of different industries. For example, in the financial services sector, an IBM InfoSphere Streams–based application analyzes and correlates over five million market messages per second to execute algorithmic option trades, with an average latency of 30 microseconds. In another example, a hospital is using the tool to analyze information from multiple sensors in a neonatal intensive care unit, with the goal being to detect changes in patients' medical conditions up

to 24 hours earlier than before.[9] IBM InfoSphere Streams includes a connector to allow users to read and write data to the file system in IBM InfoSphere BigInsights.

IBM InfoSphere Streams includes the following key components:

- *Streams Studio*—An Eclipse-based interactive development environment (IDE) that supports application development with editors, wizards, application structure graphs, and runtime application monitoring.
- *Streams Runtime*—A single server or a cluster of servers that have no limit on cluster size. High-availability features include the ability to detect failing process elements, relocate, restart, and optionally restore state.
- *Integrated toolkits and sample applications*—Facilitate the development of solutions for particular industries or functions, including the Mining Toolkit and the Financial Services Toolkit, which encompass an array of the most commonly utilized operators.

Data Warehouse Solutions

The data warehouse category includes the following products:

- *IBM Netezza*—A family of high-performance data warehouse appliances purpose-built to make advanced analytics on large data volumes simpler, faster, and more accessible. IBM Netezza leverages parallel in-database analytics for fast queries for very large datasets.
- *IBM Smart Analytics System*—A comprehensive portfolio of packaged data management, hardware, software, and services capabilities that predates the Netezza acquisition and modularly delivers a wide assortment of analytics.
- *IBM InfoSphere Warehouse*—A comprehensive data warehouse software platform that delivers access to structured and unstructured information in real-time.

Information Integration and Governance

This category includes the following products:

- *IBM InfoSphere Information Server*—Offers comprehensive data integration and data quality capabilities to ensure delivery of trusted information to a wide variety of IT systems.

- *IBM InfoSphere Master Data Management (MDM)*—Enables organizations to maintain a single version of the truth across multiple domains, including customers, materials, vendors, assets, and locations.
- *IBM InfoSphere Optim*™—Controls data growth, streamlines test data management, and supports data masking.
- *IBM InfoSphere Guardium*®—Provides a robust solution for assuring the privacy and integrity of trusted information and reducing costs by automating the entire compliance auditing process in heterogeneous environments.

Supporting Platform Services

The IBM big data platform includes the following supporting services:

- *Visualization and discovery*—Discover, understand, search, and navigate federated sources of big data while leaving that data in place.
- *Application development*—Streamline the process of developing big data applications.
- *Systems management*—Monitor and manage big data systems for secure and optimized performance.
- *Accelerators*—Prepackaged analytical and industry-specific content for faster deployment.

3.2 IBM Big Data Platform Differentiators

The IBM big data platform has five distinct differentiators:

- *Comprehensive platform*—A complete platform for managing and analyzing the volume, variety, and velocity of big data while ensuring veracity.
- *Enterprise-class capabilities*—Delivers the management, security, reliability, and usability features necessary for large-scale deployments.
- *Advanced analytics*—Analytic engines pre-integrated and optimized for big data, and pre-built accelerators for industry-specific and cross-industry applications.
- *Visualization tools*—Professional-grade tools to explore all available data for ad hoc analysis.
- *Integration and governance*—Simplifies integration of big data technologies with existing IT architectures to leverage big data as another source to enhance strategic initiatives. Information integration

and governance capabilities ensure the veracity of big data for trusted decision-making.

Summary

IBM has developed an enterprise-class big data platform that allows organizations to fully address the spectrum of big data business challenges. The platform blends traditional technologies that are well-suited for structured, semi-structured, and unstructured data-at-rest and data-in-motion. The remainder of Part I of this book will focus on the information integration and governance capabilities of the IBM big data platform.

CHAPTER 4

BIG DATA INTEGRATION

This chapter includes contributions from Aarti Borkar (IBM), Tony Curcio (IBM), and Wolfgang Nimfuehr (IBM).

There are three key styles of big data integration:

- Bulk data movement
- Data replication
- Data virtualization

Each style is discussed in detail in the rest of this chapter.

4.1 Bulk Data Movement

Bulk data movement includes technologies such as Extract, Transform, and Load (ETL) to extract data from one or more data sources, transform the data so that it is prepared for an alternative use, and then load the data into a target database. There are several variations of this paradigm, including Extract, Load, and Transform (ELT), which extracts data from one or more data sources, loads the data into a target database, and then transforms the data using functions native to the target database. ELT approaches are often used in Hadoop to leverage its massive parallel processing power. From an architectural perspective, organizations may move large warehouse workloads into Hadoop using ETL, conduct data-intensive analytical processing in Hadoop, and then use ETL to transfer the results into the data warehouse.

IBM InfoSphere DataStage® is IBM's flagship platform for bulk data integration. IBM has made the following enhancements to IBM InfoSphere DataStage to support big data integration:

- *Big Data File Stage moves data in and out of Hadoop.*
  IBM InfoSphere DataStage Version 8.7 included the new Big Data File Stage, which supports reading and writing multiple files in parallel from and to the Hadoop Distributed File System (HDFS). The Big Data File Stage leverages the parallel engine within IBM InfoSphere DataStage to provide massive scalability. The Big Data File Stage mirrors the Sequential File stage experience so that users find it intuitive to get started. The Big Data File Stage provides reading from multiple files in parallel (either listed specifically or through file patterns) to simplify how data is merged into a common transformation process.

  In addition, the IBM InfoSphere DataStage massively parallel processing engine can be used to mimic the same degree of partitioning as the Big Data File Stage, or the engine can dynamically repartition that data "on the fly" as may be required to meet business requirements. These features offer both flexibility and optimization as organizations address their big data challenges. Figure 4.1 shows a sample configuration of the Big Data File Stage within IBM InfoSphere DataStage Version 8.7.

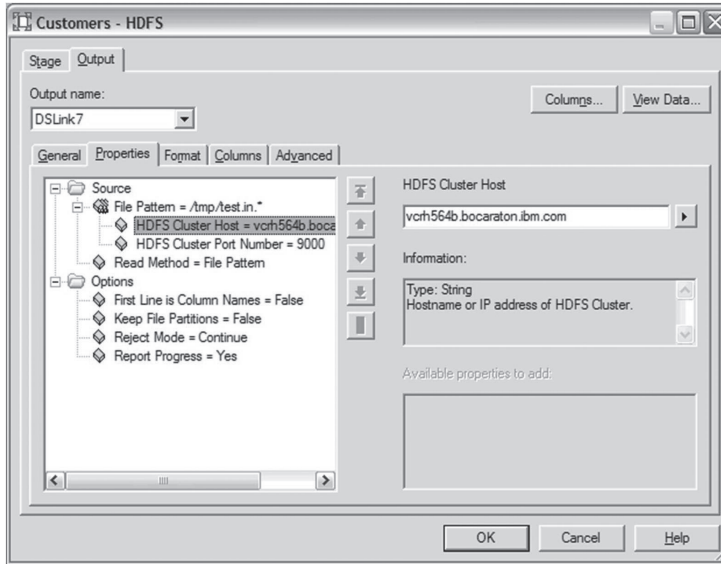


Figure 4.1: Configuration of the Big Data File Stage within IBM InfoSphere DataStage Version 8.7.

- *Job Sequencer combines DataStage and Hadoop for end-to-end workflow.*
  IBM InfoSphere DataStage provides a graphical Job Sequencer, which allows developers to run a sequence of parallel or server jobs. In Version 9.1, Oozie workflows may be included within the Job Sequencer so that MapReduce jobs from IBM InfoSphere BigInsights or Cloudera Enterprise can be represented alongside InfoSphere DataStage jobs for an end-to-end big data workflow.
- *Use ELT capabilities to leverage the processing power of Hadoop.*
  As shown in Figure 4.2, IBM InfoSphere DataStage offers Balanced Optimization capabilities so that developers can design a job through the same, simple data flow paradigm and deploy it in full or in part to IBM InfoSphere BigInsights, Cloudera Enterprise, IBM DB2, Oracle®, Teradata, or the ETL engine itself. This means that, within a single job flow, developers combine the advantages of both ETL and ELT styles of processing. This gives organizations the ability to avoid moving data across the network for big data sources that may be more efficiently processed inside Hadoop.

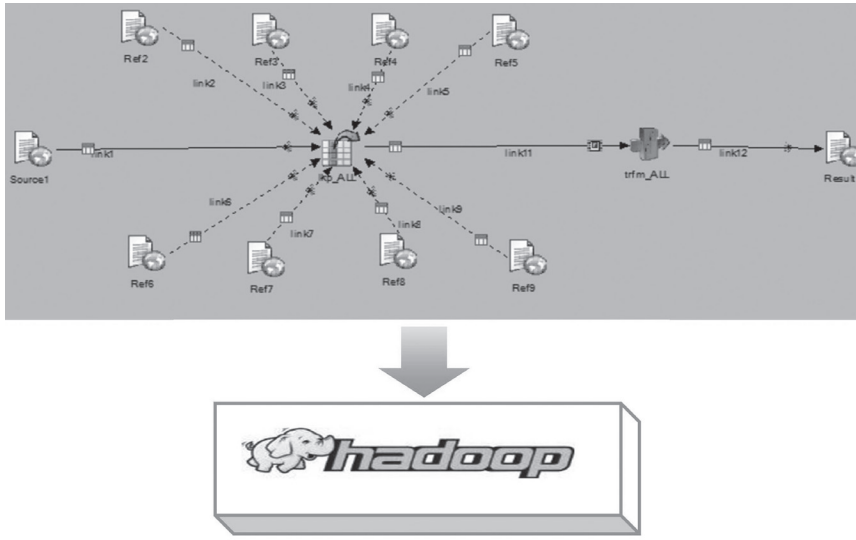


Figure 4.2: A developer can design a job in the IBM InfoSphere DataStage and QualityStage® Designer client and deploy all or part on IBM InfoSphere BigInsights or Cloudera Enterprise.

Figure 4.3 shows a portion of the MapReduce code that is automatically generated for developers using this Balanced Optimization technology. So rather than learning Java® and MapReduce programming and standards, IBM InfoSphere DataStage developers can leverage their existing skills to become instantly productive in transforming data and optimizing processing for big data projects.

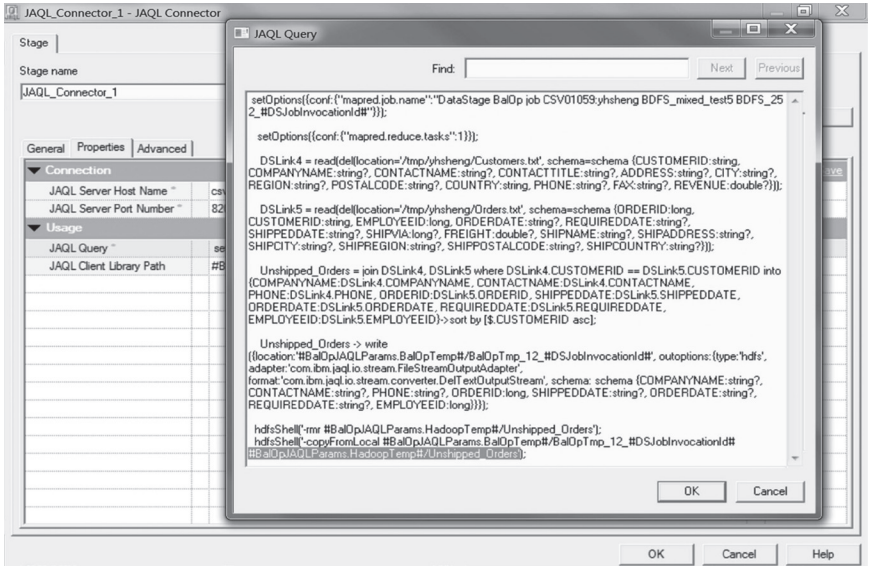


Figure 4.3: IBM InfoSphere DataStage auto-generates Jaql for MapReduce through Balanced Optimization technology.

- *Combine real-time analytical processing with data integration.* IBM InfoSphere DataStage can now move data in and out of IBM InfoSphere Streams to combine the power of both platforms. As organizations analyze millions of records per second using IBM InfoSphere Streams real-time analytical processing, they can now send the data directly to IBM InfoSphere DataStage in order to transform and load discovered insights to a broad range of enterprise applications and databases.

4.2 Data Replication

According to *Information Management* magazine (*www.information-management.com*), data replication is the process of copying a portion of a database from one environment to another and keeping the subsequent copies of the data in sync with the original source. Changes made to the original source are propagated to the copies of the data in other environments. Replication technologies such as Change Data Capture (CDC) allow the capture of only change data and transfer it from publisher to subscriber systems. Rather than performing queries directly against the database, CDC tools improve system performance by capturing inserts, updates, and deletes directly from the database transaction (redo) logs. As a result, CDC technology can capture big data, such as utility smart meter readings, in near real-time, with minimal impact to system performance.

IBM InfoSphere Data Replication is the company's flagship data replication platform that includes CDC technology. As shown in Figure 4.4, IBM InfoSphere Data Replication supports low-latency capture of real-time information by reading data directly from the database transaction logs. It can then distribute that data to relational databases, data warehouses, IBM InfoSphere Streams, and IBM InfoSphere DataStage. Through the integration with IBM InfoSphere DataStage, it can also distribute the data to IBM InfoSphere BigInsights or IBM InfoSphere Streams.

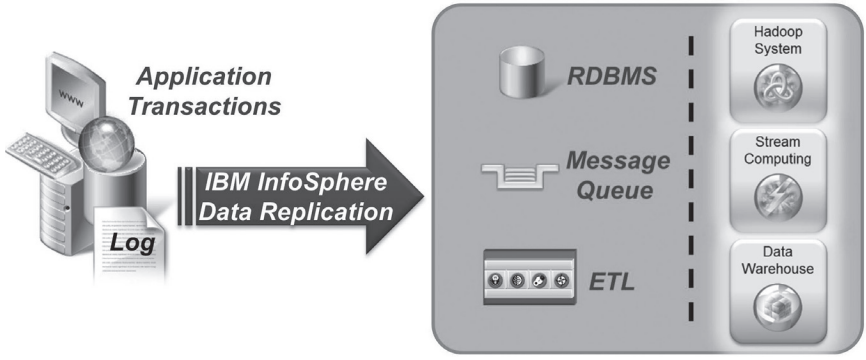


Figure 4.4: Big data integration with IBM InfoSphere Data Replication.

Figure 4.5 illustrates the basic workflow of an IBM InfoSphere DataStage job that incorporates change data capture using the Flat File connection method:

1. On the computer where the source database is installed, the IBM InfoSphere CDC (now IBM InfoSphere Data Replication) service for the database reads the transaction log to capture changes.
2. IBM InfoSphere Data Replication transfers the change data according to the replication definition.
3. The IBM InfoSphere Data Replication server hardens the files and deposits them in the flat file location.
4. The IBM InfoSphere DataStage sequential file reader retrieves the flat files as part of a data integration job and transforms them.
5. IBM InfoSphere DataStage then transforms the data and deposits it in a new location. This target can be another sequential file reader, as shown in the workflow or, alternatively, Hadoop, enterprise applications, data warehouses, or a variety of other destinations.

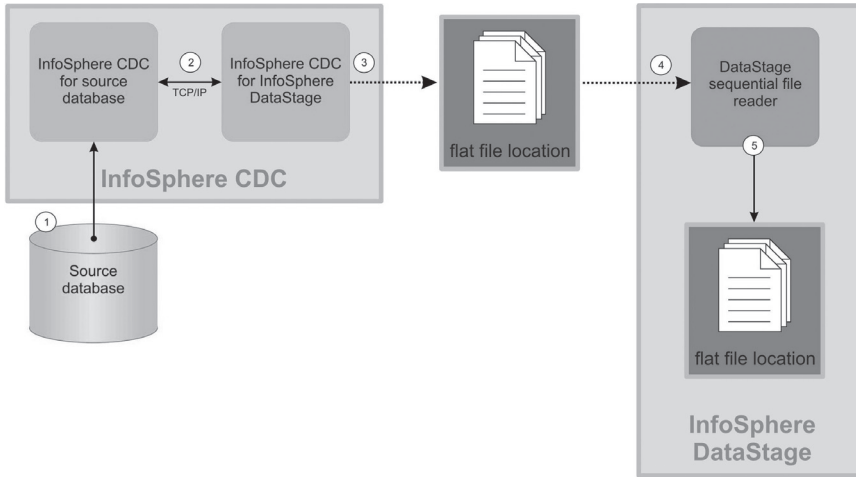


Figure 4.5: Basic workflow of an IBM InfoSphere DataStage job with change data capture using the Flat File connection method.

4.3 Data Virtualization

Data virtualization is also known as *data federation*. According to *Information Management*, data federation is the method of linking data from two or more physically different locations and making the access/linkage appear transparent, as if the data were co-located. This approach is in contrast to the data warehouse

method of housing data in one place and accessing data from that single location.

As shown in Figure 4.6, IBM InfoSphere Federation Server is a data virtualization solution that allows an application to issue SQL queries against a virtual view of data in heterogeneous sources such as in relational databases, XML documents, and on the mainframe.

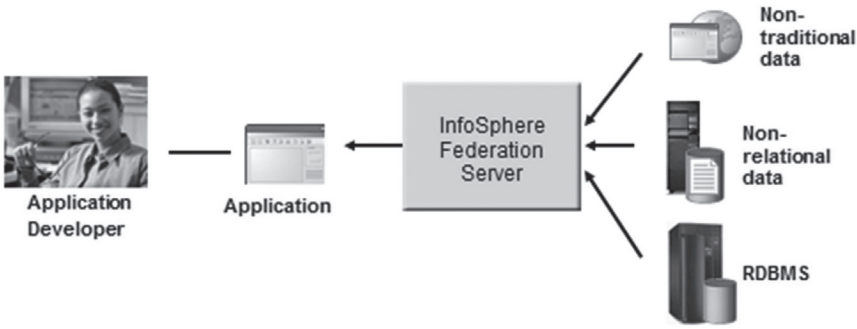


Figure 4.6: IBM InfoSphere Federation Server lets an application access heterogeneous data sources.

As shown in Figure 4.7, IBM InfoSphere Federation Server can be another source of enterprise data for IBM InfoSphere Data Explorer (Vivisimo Velocity). This capability allows organizations to combine both unstructured data sources with the enterprise's structured data stores in a seamless user experience.

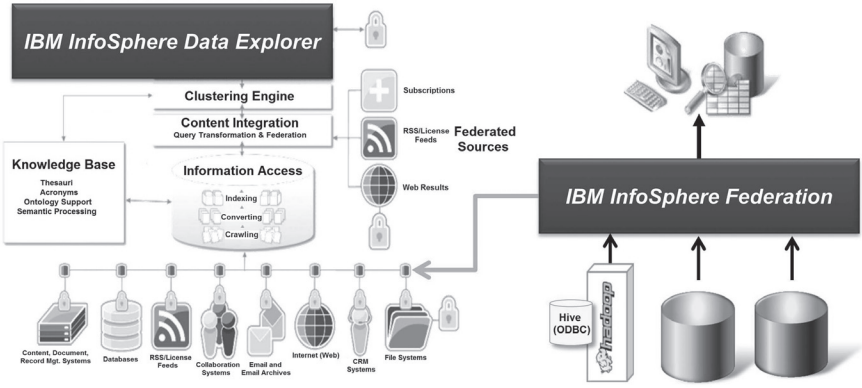


Figure 4.7: IBM InfoSphere Federation Server can be another source of enterprise data for IBM InfoSphere Data Explorer.

Summary

There are three styles of big data integration: bulk data movement, data replication, and data virtualization. Organizations need to consider all three styles depending on the needs of the application and user requirements. IBM InfoSphere supports all three big data integration styles with a robust portfolio that includes IBM InfoSphere DataStage, IBM InfoSphere Data Replication, and IBM InfoSphere Federation Server.

CHAPTER 5

METADATA

This chapter includes contributions from Aarti Borkar (IBM), Roger Hecker (IBM), and Wolfgang Nimfuehr (IBM).

Metadata is information that describes the characteristics of any data artifact—such as its name, location, perceived importance, quality, or value to the enterprise—and its relationships to other data artifacts that the enterprise deems worth managing. Because metadata determines how the information architecture satisfies the needs of the business, the management of enterprise metadata is a key element of any successful information governance program. As analytics and warehousing requirements have grown, so has the importance of a strong metadata infrastructure to deliver consistent, high-quality data that meets those requirements. However, as any experienced information professional knows, building and maintaining a strong metadata infrastructure is a significant challenge.

Big data expands the volume, velocity, and variety of information while adding new challenges in building and maintaining a coherent metadata infrastructure. Metadata programs in traditional data warehousing help to ensure that the proper data is used, and reused, to satisfy defined requirements. For big data, any failure of the metadata program to deliver these information management capabilities can result in data duplication, poor data quality, and lack of access to key information. Notwithstanding all of this, metadata is still low on the priority list for customers who are beginning their big data journey. However, it is important to plan now because metadata will gain in importance once the business starts to sponsor big data projects.

To ensure enterprise accessibility to trusted information, the big data governance program needs to adopt the following best practices to address metadata issues:

- ✓ Establish a glossary that represents the business definitions for key big data terms.
- ✓ Tag sensitive big data within the business glossary.
- ✓ Maintain technical metadata to support data lineage and impact analysis.
- ✓ Gather metadata from unstructured documents to support enterprise search.

Each best practice is discussed in detail in the rest of this chapter.

5.1 Establish a Glossary That Represents the Business Definitions for Key Big Data Terms

A business glossary includes language that the enterprise uses to communicate its understanding of information. Establishing and maintaining this layer of business metadata is critical to express the meaning of requirements and to describe the information available in IT systems. By deploying these terms in a manner that is easily accessible by all relevant enterprise users, including database administrators and executives, the business glossary ensures accurate and highly accelerated information development.

IBM InfoSphere Business Glossary helps organizations create, manage, and share an enterprise-wide common business terminology. This authoritative source of information promotes better communication among business and technical teams and aligns cross-team efforts. The line of business uses this centralized information source as a gateway to all information assets to support data governance initiatives. Critical business concepts, their interrelationships and association with key business rules, technical assets, and responsible data stewards are visible within the tool. Accessible through a web browser or directly from any application, IBM InfoSphere Business Glossary brings understanding, consistency, and trust in information to any application or context. IBM InfoSphere Business Glossary includes business definitions for big data–related terms such as the ones below.

Web Clickstream Analytics

Business term

Unique visitor

Sample definition

A unique visitor is a unit used to count individual, different users of a website.

Implications

This term is a fundamental building block of clickstream analytics. Based on this definition, the web analytics team might create additional terms that represent a granular understanding of user behavior. However, two sites might measure unique visitors differently. For example, one might measure unique visitors within a month, while another one might measure unique visitors within a week.[10]

Figure 5.1 shows the editing interface for a business term in IBM InfoSphere Business Glossary.

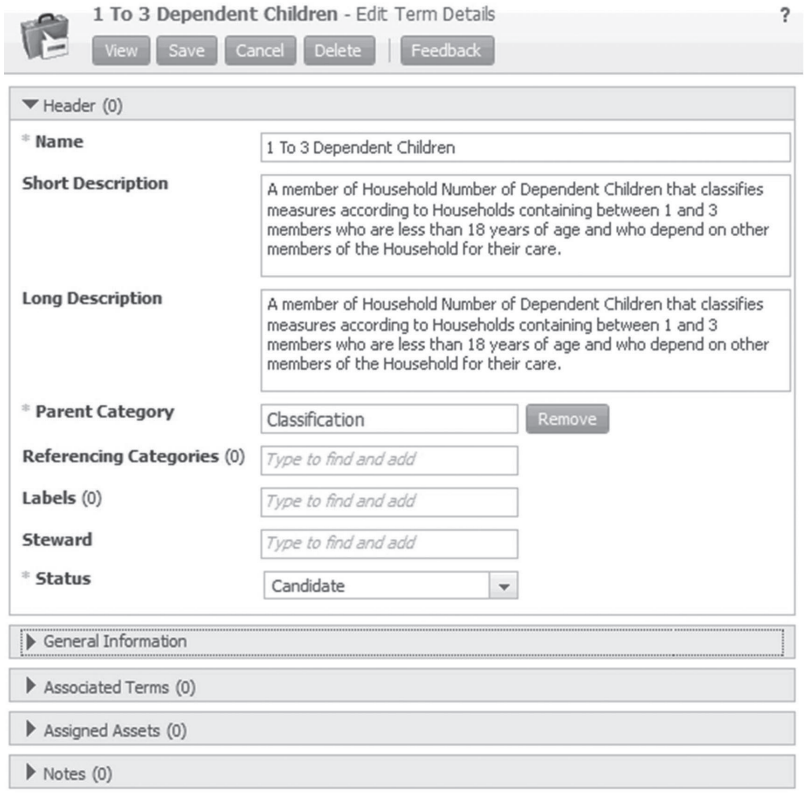


Figure 5.1: Editing interface for a business term in IBM InfoSphere Business Glossary.

5.2 Tag Sensitive Big Data Within the Business Glossary

Big data can be highly sensitive and may constitute personally identifiable information when combined with other data sources. For example, a phone company can combine Global Positioning System (GPS) data with geospatial information to produce a highly detailed account of a person's activities and lifestyle. The big data governance program needs to adopt the following sequence of activities relating to privacy and metadata:

- *Classify sensitive big data.*
  The big data governance program needs to classify sensitive data such as Social Security numbers. The classification should contain the various levels of sensitivity, where possible.
- *Discover sensitive data.*
  Sensitive big data might be hidden within unstructured text. For example, call center agents might enter Social Security numbers within unstructured fields. IBM InfoSphere Discovery can help big data governance programs automate the discovery of sensitive data within unstructured fields.
- *Tag sensitive data within the business glossary.*
  Chief information security officers set policies around sensitive data. Tagging sensitive data in the business glossary is critical because an organization can only enforce policies when it identifies the location of sensitive data.
- *Enforce big data privacy policies.*
  Taking the earlier example further, the organization might insist that call center agents delete any sensitive information from unstructured fields after completing each call. The big data governance team can monitor adherence to this policy by using IBM InfoSphere Discovery to discover any instances when call center agents' notes contain sensitive data.

5.3 Maintain Technical Metadata to Support Data Lineage and Impact Analysis

Data lineage provides an audit trail for data movement through integration processes. The data lineage answers basic questions such as "Where did this data come from?", "Where does this data go?", and "What happened to it along the way?" The ability to understand how a change to one data artifact affects other data artifacts is called *impact analysis* and is crucial functionality within IBM

InfoSphere Metadata Workbench. Figure 5.2 shows the browser-based interface of IBM InfoSphere Metadata Workbench.

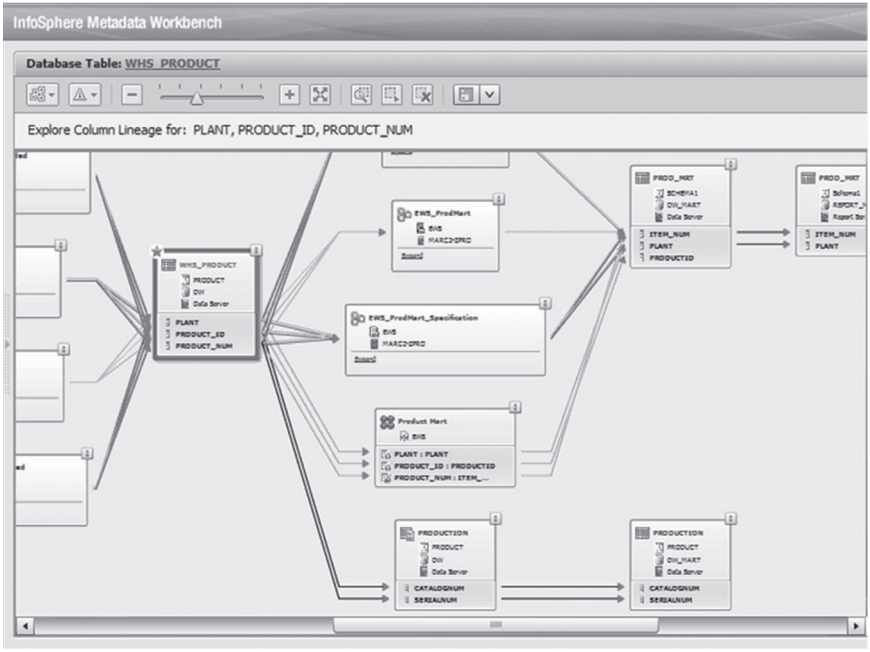


Figure 5.2: The IBM InfoSphere Metadata Workbench canvas showing a data lineage report.

As discussed in chapter 4, IBM InfoSphere DataStage Version 8.7 adds the new Big Data File Stage that supports reading and writing multiple files in parallel from and to the Hadoop Distributed File System (HDFS). If the Big Data File Stage is used to move data in and out of HDFS, IBM InfoSphere Metadata Workbench will be able to provide data lineage in and out of Hadoop. Added lineage information about the Hadoop processing can be captured using the metadata extensibility feature of IBM Information Server.

As organizations use Hadoop for mission-critical applications such as fraud analytics and risk calculations, their information management systems need to support data lineage and impact analysis that includes big data. Case Study 5.1 describes a large transportation services provider that was looking to integrate Hadoop into its existing data warehousing environment.

Case Study 5.1: A blended Hadoop and data warehousing environment at a large transportation services provider

As shown in Figure 5.3, a large transportation services provider generated significant volumes of clickstream data from its web presence. The clickstream data had the following characteristics:

- The data was in XML format.
- Each user session generated large volumes of data.
- The data was sparse, and there was only a small amount of insight to be gained from vast quantities of information.
- Licensing fees made it cost-prohibitive to handle the raw clickstream data within the data warehouse.
- The business intelligence team found it difficult to parse the XML data with their current ETL tool.

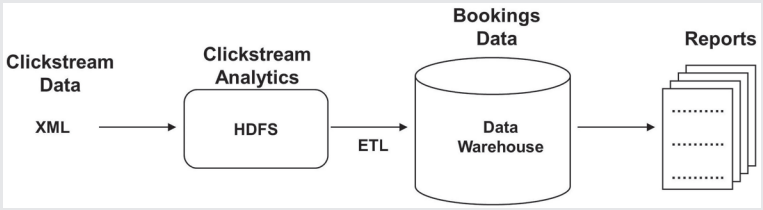


Figure 5.3: A proposed technical architecture for a blended Hadoop and data warehousing environment at a large transportation services provider.

The business intelligence team began exploring Hadoop to analyze user browsing patterns within the clickstream data. However, the team needed to marry the browsing data with the bookings data (sales information) in the limited number of cases where the user actually made a purchase. Because the bookings information was in the data warehouse, the business intelligence team decided to use ETL to move the clickstream data for actual buyers from Hadoop into the data warehouse.

As of the publication of this book, the organization was in the evaluation process for Hadoop. Once Hadoop is implemented, the business intelligence team will need a strong metadata foundation to answer questions such as the following:

- Marketing: "Can we verify the primary landing page for shoppers who spent more than $500?"
- Finance: "Are we sure that our most profitable shoppers came from the sites that are listed in this report?"
- IT: "What will be the impact to downstream reports if we drop these files from HDFS?"

5.4 Gather Metadata from Unstructured Documents to Support Enterprise Search

The challenge of "information silos" is not going away any time soon. Each of the systems in an enterprise was designed to serve a critical business function, whether managing customer data, overseeing the supply chain, securing sensitive content, or any of a myriad of different functions. While these applications can be coordinated and improved with effective master data management, they remain silos that must be accessed separately. IBM InfoSphere Data Explorer enables everyone—from management through knowledge workers to front-line employees—to access all the information they need in a single view, regardless of format or where it is managed. Creating search indexes for unstructured documents is also a form of metadata. IBM InfoSphere Data Explorer supports use cases such as the following:

- *Insurance*—Reduction of three seconds of average handling time, and millions of dollars annually, by providing call center agents with searchable access to multiple document repositories for customer care, alerts, policies, and customer information files[11]
- *Pharmaceuticals*—Accelerating research by providing quick access to customer, patient, and research data within content management, file systems, Microsoft® SharePoint®, intranet pages, and external databases[12]
- *Healthcare*—Providing clinicians with access to the latest research from medical journals and other document repositories[13]

IBM InfoSphere Data Explorer's foundation is enterprise search that is applied to business problems. The tool brings the right information to the right person regardless of where the information resides and regardless of format (while supporting security protocols). The IBM InfoSphere Data Explorer solution "sits on top" of all data sources and brings them together in one unified view via a dashboard/console user interface.

Summary

Metadata best practices are well-established for traditional data projects. However, big data presents unique metadata challenges because of the high velocity and variety of data. IBM InfoSphere provides strong capabilities to manage business and technical metadata for big data.

CHAPTER 6

Big Data Security and Privacy

This chapter includes contributions from Aarti Borkar (IBM), Michael Eggloff (IBM), Eberhard Hechler (IBM), Wolfgang Nimfuehr (IBM), and Brian Roosevelt (IBM).

Over 100 years ago, 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*. As discussed in chapter 2, this article defined privacy as the "right to be left alone." Regulations and legislation around the world have since formalized and expanded this theory of privacy.

According to IDC, only about 50 percent of the information in the digital universe that should be protected actually is protected.[14] Although big data has many exciting applications, privacy concerns continue to pose a major challenge. Regulations and best practices are evolving very rapidly, as are consumers' expectations about how their data is collected, kept, used, protected, and destroyed. Scarcely a day goes by without newspaper headlines reporting new and scary ways that organizations are using big data to track members of the public. Telephone companies can use signals from a GPS to track the movement of individuals and use their calling patterns to understand their social relationships. Case Study 6.1 discusses the privacy implications of geolocation data in telecommunications.

Case Study 6.1: A German politician demonstrates how mobile phones can become tracking devices[15]

Malte Spitz from the German Green party wanted to make a clear and convincing case for greater regulations around the privacy of big data. Mr. Spitz decided to publish the data collected from his mobile phone over a period of six months. Mr. Spitz sued the phone company and received a giant file with 35,000 data points showing his location over a period of six months. These data points represented GPS signals that were transmitted by his mobile phone even when it was not in use. Mr. Spitz then made this information publicly available.

Taken individually, these data points were meaningless. However, researchers were able to construct a detailed profile of Mr. Spitz's life by aggregating this data, integrating it with other publicly available information, and placing it on a visual map. For example, researchers combined his GPS latitude/longitude coordinates and timing data to pinpoint when he had visited political rallies, when he was on a train, which cities he visited, where he slept, where he worked, and which beer gardens he visited.

We cover best practices relating to big data security and privacy in this chapter:

- ✓ Identify sensitive big data.
- ✓ Flag sensitive big data within the metadata repository.
- ✓ Mask sensitive big data in production and non-production environments.
- ✓ Monitor access to sensitive big data by privileged users.

Each best practice is discussed in the sections that follow.

6.1 Identify Sensitive Big Data

As a first step, the big data governance program needs to identify sensitive data. In general, big data that can be linked to an individual should be treated as *personally identifiable information* (*PII*). Each jurisdiction may have its own regulations that govern the usage of this data. For example, the European Union requires a *privacy impact assessment* (*PIA*) regarding radio-frequency identification (RFID) data.

Big data is now blurring the lines between PII and non-PII. For example, browser fingerprinting does not rely on cookies and can uniquely identify computers and other personal devices. The term "browser fingerprint" refers to the specific combination of characteristics such as system fonts, software, and installed plug-ins that are typically made available by a consumer's browser to any website visited.[16] In another example, the country of birth, country of residence, and account balance might not constitute PII when viewed separately. However, the combination of these three attributes might create sensitive data to uniquely identify a person born in Venezuela, living in Kazakhstan, and holding more than two million Euros in his bank account.

6.2 Flag Sensitive Big Data Within the Metadata Repository

Sensitive data can be stored in different parts of the organization. Big data stewards need to ensure that this data is appropriately classified in the metadata repository such as IBM InfoSphere Metadata Workbench. Once the appropriate metadata is in place, the application can enforce the appropriate privacy policies, such as "If any application wants to access sensitive data, it needs be approved by the appropriate party."

Certain data fields with personally identifiable information might not be subject to the required privacy safeguards. For example, a customer's Social Security number (SSN) might reside in a field called "EMP_NUM," and the last four digits of the SSN might be part of another field called "PIN." As a result, just looking at the column headers is not sufficient. IBM InfoSphere Discovery can discern that the `EMP_NUM` field in one table actually relates to the SSN in another table and to the `PIN` column. Figure 6.1 shows the column analysis feature of IBM InfoSphere Discovery.

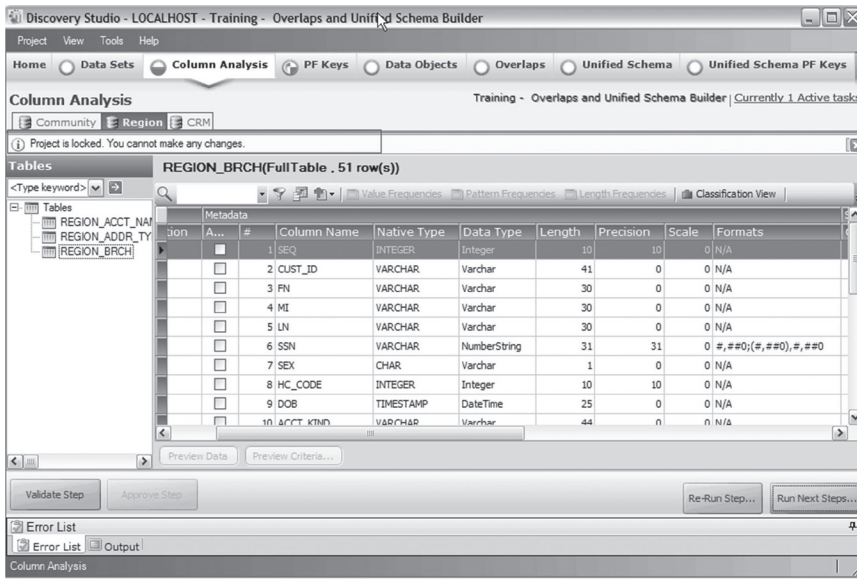


Figure 6.1: Column analysis with IBM InfoSphere Discovery.

6.3 Mask Sensitive Big Data in Production and Non-Production Environments

Sensitive data embedded within production, test, training, and business intelligence environments represents a potential exposure for organizations. Commonly, live production systems, which can include confidential data, are cloned to a test, or training, environment. Developers and quality assurance testers find it easy to work with live data that produces test results that everyone can understand. *Data masking* is the process of systematically transforming confidential data elements, such as trade secrets and personally identifiable information, into realistic, but fictionalized, values.

Data masking represents a simple concept, but it is technically challenging to execute. Finding and masking data is part of the solution. However, there is an added complication. You need the capability to propagate masked data elements to all related tables in the database, and across databases, to maintain referential integrity. For example, suppose a masked data element, such as a telephone number, is a primary or foreign key in a database table relationship. This masked data value must be propagated to all related tables in the database,

or across data sources. If the data is a portion of another row's data, it must be updated with the same data as well.

Figure 6.2 offers an example of data-masking capabilities with IBM InfoSphere Optim Data Masking solution. In this example, the customer named Elliot Flynn has a customer identifier of 27645 and a street address of 96 Avenue in the Customers table. Elliot Flynn has also placed two orders in the Orders table. In the masked data on the right, Elliot Flynn's name, customer identifier, and street address have been changed to Pablo Picasso, 10002, and Saturn25, respectively, in the Customers table. In addition, the customer identifier in the Orders table has also been changed to 10002 to maintain referential integrity between the Customers and Orders tables.

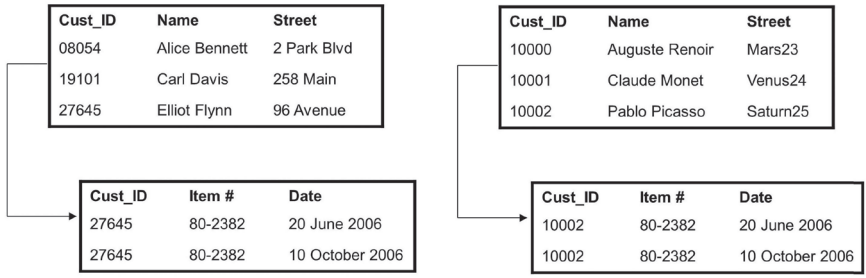


Figure 6.2: Data masking with IBM InfoSphere Optim Data Masking solution.

IBM InfoSphere Optim Data Masking on Demand is a newly released feature that allows organizations to invoke masking algorithms in real-time irrespective of the data type and for data-at-rest and data-in-motion. Developers can invoke this data masking functionality directly from MapReduce routines and Jaql scripts. In addition, this feature has been packaged up as database-specific user-defined functions (UDFs) so that data moving in and out of HDFS, HBase, IBM InfoSphere BigInsights, IBM InfoSphere Streams, IBM Netezza, IBM DB2, IBM DB2 for z/OS®, and Oracle can be masked on demand. As shown in Figure 6.3, the chief financial officer is abled to view unmasked data, but IBM InfoSphere Optim Data Masking on Demand masks the data for marketing and other business users.

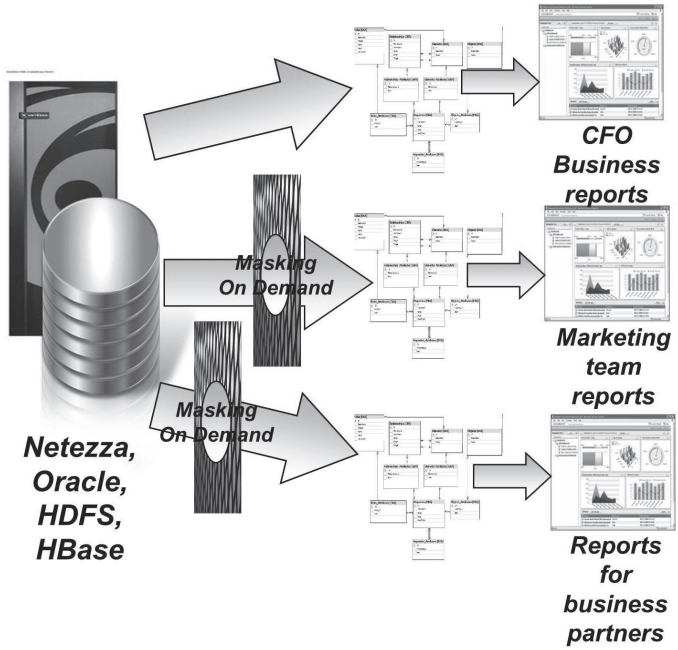


Figure 6.3: IBM InfoSphere Optim Data Masking on Demand.

IBM InfoSphere Optim Data Masking on Demand supports several options to mask data: lookup, rule-based, and JavaScript. The lookup option uses a lookup table to provide masked data; the rule-based option uses functions to generate masked data; and the JavaScript option uses JavaScript expressions to define data transformations.

6.4 Monitor Access to Sensitive Big Data by Privileged Users

Most organizations have policies regarding access to sensitive data by privileged users such as database administrators, call center agents, and help desk personnel. However, many organizations do not have effective mechanisms to enforce these policies. To make matters worse, some of these privileged users might reside outside the country, in the case of applications that are outsourced. Here are some examples of how privileged users can potentially abuse sensitive big data:

- *Financial services*—Privileged users viewed and shared financial information with unauthorized individuals and organizations during blackout periods.
- *Telecommunications*—Database administrators looked up the calling patterns of a vice president of marketing—a married man—who was making and receiving repeated calls and texts from a female staffer outside business hours over several months.
- *Health insurance*—Call center agents reviewed claims records and discovered that the CEO of the company had recently undergone surgery.
- *Cruise ship operator*—A vacation planner used information from the reservation system to alert organized crime to the periods when people would not be in their homes.
- *Utilities*—Database administrators could potentially leak smart meter readings to burglars. Burglars could use hourly meter data to determine when people would not be in their homes.

Given all of this, the big data governance program needs to define sensitive data and establish policies to monitor access to this data by privileged users. From a Hadoop perspective, a database monitoring solution needs to address the following questions, which are very similar to traditional database environments:

- Who is requesting access?
- Are the requests for sensitive information authorized and valid?
- Are we seeing abnormal error patterns that point to a possible attack?
- Are privileged users or application IDs accessing data from known clients?
- Are there any new requests that were not previously scheduled or known?

At the very minimum, Hadoop Activity Monitoring (HAM) should support functionality that includes the following:

- Real-time activity monitoring of HDFS, MapReduce, Hive, and HBase data sources
- Automated compliance controls
- Complete integration with any existing enterprise database activity monitors to leverage the same policies, processes, and procedures
- Integrated view of Hadoop systems with other data sources

IBM InfoSphere Guardium creates a continuous, fine-grained trail of database activities that is contextually analyzed and filtered in real-time to

implement controls and produce the specific information required by auditors. The resulting reports demonstrate compliance by making it possible to view database activities in detail, such as login failures, escalation of privileges, schema changes, access during off-hours or from unauthorized applications, and access to sensitive tables.

As shown in Figure 6.4, IBM InfoSphere Guardium offers a scalable, multi-tier architecture that supports large and small environments, with centralized aggregation and normalization of audit data and centralized management of security policies enterprise wide. S-TAPs are lightweight, host-based probes that are able to monitor database traffic, including local access by privileged users, and relay it to IBM InfoSphere Guardium collector appliances for analysis and reporting.[17]

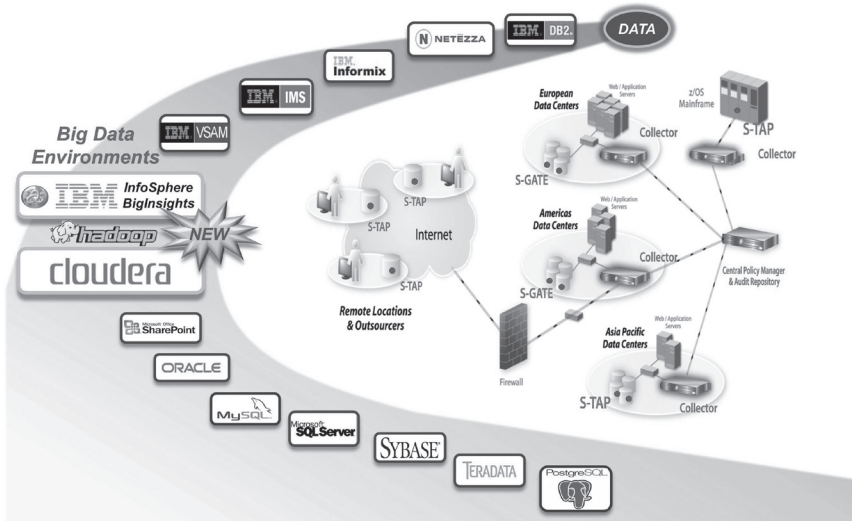


Figure 6.4: Database activity monitoring with IBM InfoSphere Guardium.

IBM InfoSphere Guardium provides real-time monitoring and auditing for Hadoop. In addition to support for existing structured data sources, IBM InfoSphere Guardium monitors Hadoop traffic, including HDFS, MapReduce, Hive, and HBase. The HAM feature of IBM InfoSphere Guardium provides S-TAP support for Hadoop. HAM securely collects audit data for Hadoop as follows:

- *HDFS*—Captures user, IP address, and actions such as `open`, `create`, `delete`, `rename`, `setOwner`, `setPermission`, and `listStatus`, as well as the source and target of these actions and related permissions.
- *MapReduce*—Captures activities such as operations, target, permissions, and description.
- *Oozie*—Captures `jobId`, `appName`, operations, and parameters.
- *Hive/HBase*—Monitors actions such as `alter`, `count`, `create`, `drop`, `get`, `put`, and `list`.

Summary

Organizations need to identify sensitive big data and flag it within their metadata repositories. They also need to mask sensitive big data in production and non-production environments and monitor access by privileged users. IBM InfoSphere Metadata Workbench, IBM InfoSphere Discovery, IBM InfoSphere Optim Data Masking on Demand, and IBM InfoSphere Guardium Hadoop Activity Monitoring provide robust solutions to help organizations improve the security and privacy of their big data.

CHAPTER 7

BIG DATA QUALITY

This chapter includes contributions from Aarti Borkar (IBM), Roger Rea (IBM), Randy Schnier (IBM), and Brian Williams (IBM).

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 volumes, velocity, and variety, big data quality needs to be handled differently from quality for traditional information governance programs. Table 7.1 compares and contrasts the differences between traditional and big data quality programs.

Table 7.1: Traditional vs. Big Data Quality Programs

Dimension	Traditional Data Quality	Big Data Quality
Frequency of processing	Processing is batch-oriented.	Processing is both real-time and batch-oriented.
Variety of data	Data format is largely structured.	Data format may be structured, semi-structured, or unstructured.
Confidence levels	Data needs to be in pristine condition for analytics in the data warehouse.	"Noise" needs to be filtered out, but data needs to be "good enough." Poor data quality might or might not impede analytics to glean business insights.
Timing of data cleansing	Data is cleansed prior to loading into the data warehouse.	Data may be loaded as-is because the critical data elements and relationships might not be fully understood. The volume and velocity of data might require streaming, in-memory analytics to cleanse data, thus reducing storage requirements.

Table 7.1: Traditional vs. Big Data Quality Programs (continued)

Dimension	Traditional Data Quality	Big Data Quality
Critical data elements	Data quality is assessed for critical data elements such as customer address.	Data may be quasi- or ill-defined and subject to further exploration, hence critical data elements may change iteratively.
Location of analysis	Data moves to the data quality and analytics engines.	Data quality and analytics engines may move to the data, to ensure an acceptable processing speed.
Stewardship	Stewards can manage a high percentage of the data.	Stewards can manage a smaller percentage of data, due to high volumes and/or velocity.

Given the differences highlighted in the table, the big data quality program has to adopt a somewhat different approach than traditional projects while still adhering to key best practices. As a result, the big data governance program needs to adopt the following best practices to address data quality issues:

- ✓ Leverage semi-structured and unstructured data to improve the quality of sparsely populated structured data.
- ✓ Use streaming analytics to address data quality issues in-memory without landing interim results to disk.
- ✓ Cleanse big data before or after processing in Hadoop.

Each of these sub-steps is discussed in detail in the rest of this chapter.

7.1 Leverage Semi-Structured and Unstructured Data to Improve the Quality of Sparsely Populated Structured Data

The big data governance program might be faced with situations where the structured data is sparsely populated. In that case, the big data governance team might leverage semi-structured and unstructured sources to improve data quality. Case Study 7.1 describes the use of text analytics at a hospital system to predict the likelihood that a patient with congestive heart failure would be readmitted within 30 days.

Case Study 7.1: The use of text analytics to predict the likelihood that a patient with congestive heart failure would be readmitted within 30 days

A hospital system consisted of 15 facilities that offered a broad range of services, including emergency care. A significant portion of its patient population was indigent. The hospital implemented a pilot program to leverage big data analytics, aimed at reducing the readmission rate of patients with congestive heart failure. The analytics department built a predictive model in IBM SPSS® based on 150 variables and 20,000 patient encounters over five years. This data was sourced from a variety of applications, including the electronic medical records package, the admissions system, and the cost accounting database.

The hospital system's analytics team determined that a number of variables were significant predictors of a patient's readmission rate, including smoking status and drug or alcohol abuse.

Smoking Status

Smoking status is a significant factor associated with heart disease. Surprisingly, the hospital did not have a complete history of patient smoking status, including years of smoking and frequency. At the outset, only 25 percent of the structured data around smoking status was populated with binary yes/no answers. However, the analytics team was able to increase the population rate for smoking status to 85 percent by using IBM text analytics technologies. The content analytics team was also able to unlock additional information, such as smoking duration and frequency.

There were a number of reasons for this discrepancy. For example, some patients indicated that they were non-smokers, but the text analytics revealed the following from the doctors' notes:

- "Patient is restless and asked for a smoking break."
- "Patient quit smoking yesterday."
- "Quit."

Drug and Alcohol Abuse

The clinical team knew from experience that drug and alcohol abuse were significant predictors of hospital readmission rates. Only 20 percent of the patients checked off the box at admission to indicate whether they were addicted to drugs and alcohol. However, the analytics team used unstructured data sources to expand the population of the data to 76 percent of the encounters.

7.2 Use Streaming Analytics to Address Data Quality Issues In-Memory Without Landing Interim Results to Disk

IBM InfoSphere Streams can analyze high volumes of data in real-time without landing interim results to disk. As shown in Figure 7.1, streaming applications need to consider two aspects relating to the underlying data:

- *Sources*—This refers to data that is available as input to the streams application. This could be from a socket connection, database query, Java Message Service (JMS) topic/queue, or file. A schema can usually define each source, although the transport medium may vary.
- *Sinks or destinations*—This refers to data produced by a streams application and sent or written to a socket connection, database table, JMS topic/queue, file, or web service. Once again, the destination schema must be modeled.

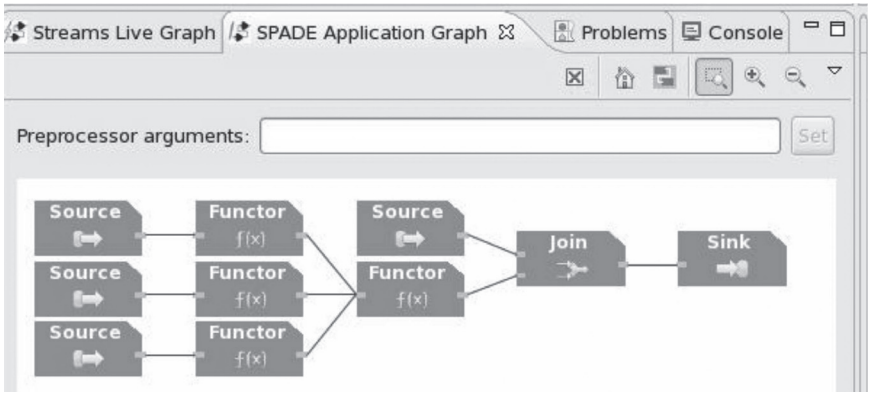


Figure 7.1: Streams application graph view from IBM InfoSphere Streams Studio[18].

Before building a streaming application, big data teams need to understand the characteristics of the data. For example, the big data team might use the Twitter API to download a sample set of Tweets for further analysis. The profiling of streaming data is similar to traditional data projects. Both types of projects need to understand the characteristics of the underlying data, such as the frequency of null values. However, profiling of streaming data also needs to consider two additional aspects of the source data:

- *Temporal alignment*—Streaming applications need to discover the temporal offset when joining, correlating, and matching data from different sources. For example, a streaming application that needs to combine data from two sensors needs to know that one sensor generates events every second, while the other generates events every three seconds.
- *Rate of arrival*—Streaming applications need to understand the rate of arrival of data:
  - Does the data arrive continuously?
  - Are there bursts in the data?
  - Are there gaps in the arrival of data?

Case Study 7.2 describes the usage of IBM InfoSphere Streams to monitor real-time network performance at a large wireless telecommunications provider.

Case Study 7.2: The use of streaming technologies to monitor real-time network performance at a wireless telecommunications provider

A large wireless telecommunications operator served customers in multiple metropolitan areas all over the country. The provider needed to analyze call detail records (CDRs), Internet protocol detail records (IPDRs) for web usage, and SMS detail records—collectively referred to as xDRs—in real-time to troubleshoot poorly performing cells.

The provider wanted to use this real-time information to accomplish the following business objectives:

1. Analyze customer call data to serve up location-dependent advertisements.
2. Identify possible network problems so that it could, for instance, initiate capital expenditure requests to upgrade poorly performing wireless towers several months before they became bottlenecks.
3. Provide customer service representatives with the latest information when a subscriber called with a service problem.

The provider's current architecture, however, made it increasingly unable to proactively address customer and network issues, due to the increased volumes of xDRs driven by 3G technologies. The team knew this problem would only get worse with the transition to 4G.

To address these issues, the technical team implemented IBM InfoSphere Streams. They had to face the following data quality issues:

- For each metropolitan market, the technical team had to merge xDRs and call quality information from different sources.
- The xDR data included a number of duplicates. For example, additional call quality data was sent when a subscriber moved between towers. The call quality information included data from records relating to signal strength, dropped calls, and calls that had to be re-routed due to poor signal strength.
- The call quality information included data from two to 20 records, and sometimes more, with the average being six records.
- The goal was to create a golden copy of an xDR with the associated call quality information.
- The team had to deal with high-volume and high-velocity data. For example, in just one mid-sized metropolitan market, they had to handle 50 gigabytes per hour of xDRs and 90 gigabytes per hour of call quality information.
- The technical team used a network switch-created field called the Universal Access Terminal Identifier (UATI) as the primary key for merge operations, but they found that this field was reused two to three times per hour for different calls by the same switch. Therefore, a combination of UATI and start/end timestamps was used to perform the merge.

To address this situation, the technical team used the concept of streaming windows (buffering based on time or count of records received) to match xDRs and call quality characteristics. The team had to make certain tradeoffs. For example, they found that if the window was too large, the system ran out of memory. On the other hand, if the window was too small, some of the call quality records would tend to fall out of it before they were matched with their corresponding xDR. By making repeated runs of a representative set of data from each market and analyzing the number of call quality records matched versus the number of "orphan" call quality records produced, the team was able to arrive at an optimum window size for each market.

Case Study 7.3 discusses the governance of time series data in a neonatal intensive care unit with IBM InfoSphere Streams.

Case Study 7.3: The governance of time series data in a neonatal intensive care unit[19]

A hospital leveraged IBM InfoSphere Streams to monitor the health of newborn babies in the neonatal intensive care unit. Using streaming technologies, the hospital was able to predict the onset of disease a full 24 hours before the onset of symptoms. From a big data governance perspective, the hospital had to establish multiple policies:

- *Data quality*—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 (a monitor attached to the baby's skin) to disengage and discontinue readings. In these situations, the streaming platform employed linear and polynomial regressions to use historical readings to fill in the gaps in the time series data.
- *Information lifecycle management*—The hospital tagged all time series data that had been modified by software algorithms. In the event of a lawsuit or medical inquiry, the hospital was able to produce both the original and the modified readings.
- *Privacy*—The hospital also established policies around safeguarding protected health information (PHI).

7.3 Cleanse Big Data Before or After Processing in Hadoop

Figure 7.2 shows an unduplicate match job in the IBM InfoSphere DataStage and QualityStage Designer client. As we discussed in chapter 4, IBM InfoSphere DataStage Version 8.7 also adds the new Big Data File stage that supports reading and writing multiple files in parallel to and from HDFS. A developer can use the IBM InfoSphere DataStage and QualityStage Designer client to design a sequence of jobs to cleanse data using IBM InfoSphere QualityStage and load data to/from HDFS with the Big Data File stage.

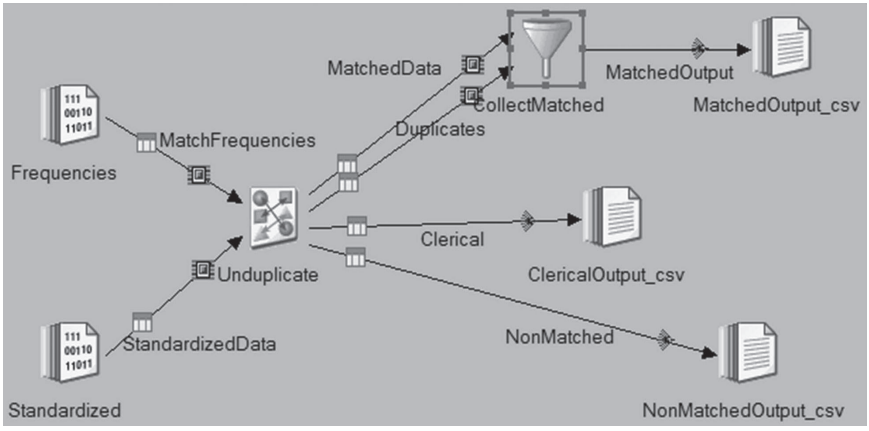


Figure 7.2: Unduplicate match job in the IBM InfoSphere DataStage and QualityStage Designer client.

Summary

Big data quality should be handled differently from traditional projects. It might need to be handled in real-time, the data is often poorly defined, confidence levels need to be established, and data stewards might be in a position to handle only a subset of the data. IBM InfoSphere Streams and IBM InfoSphere QualityStage provide robust solutions for big data quality.

CHAPTER 8

Master Data Integration

This chapter includes contributions from Nick Dimtchev (IBM), Wolfgang Nimfuehr (IBM), Jennifer Reed (IBM), Brian Vile (IBM), and Kurt Wedgwood (IBM).

To meet fundamental strategic objectives such as revenue growth, cost reduction, and risk management, organizations need to gain control over data that is often locked within silos across the business. The most valuable of this information—the business-critical data about customers, products, materials, vendors, and accounts—is commonly known as *master data*. Despite its importance, master data is often replicated and scattered across business processes, systems, and applications throughout the enterprise. Organizations are now recognizing the strategic value of master data. They are developing long-term master data management (MDM) action plans, to harness this information to drive corporate success. These include a single view of customers, materials, vendors, employees, and charts of accounts.

A *master data domain* refers to a specific category of information, such as customers, products, materials, vendors and accounts. Each data domain has specific attributes that need to be fit for purpose. For example, phone number is an important attribute of the customer data domain, because it is important for an enterprise to have valid contact information in case of need. There are relationships between master data domains that represent true understanding. For example, it is valuable for a bank to have a linked view of all the accounts and products for a given customer so that it can understand the total relationship to facilitate servicing, the sale of additional products, and profitability analysis.

Customers, accounts and products represent master data domains that have a relationship.

There should be a symbiotic relationship between MDM and big data. Big data can feed insights to MDM, and MDM can support big data with high-quality master data. There are a number of reasons why organizations should integrate big data with MDM:

- *Churn management*—Use insight from Twitter, Facebook, and voice transcripts to improve the predictability of customer churn models.
- *Risk management*—Crawl unstructured financial information such as 10-K and 10-Q reports to dynamically update changes in company ownership structures and MDM hierarchies.
- *Customer segmentation*—Use additional insight from social media and other sources to fine-tune customer segmentation and behavioral modeling.
- *Next best offer*—Identify opportunities for cross-sell and up-sell based on customer interactions with the company.
- *Reduction of operating costs in the call center*—Reduce the average handling time and frequency of calls by understanding the demographics of customers and their reasons for calling.
- *De-duplication*—It is normally the responsibility of MDM to de-duplicate master data. De-duplicated master data should be fed into the big data platform, as opposed to building another entity resolution process.
- *Preference management*—Customer preferences such as "do not call" and "do not email" are best managed within MDM. The big data analytics platform should use these preferences when determining the best customers to call as part of a marketing campaign.

The big data governance program needs to adopt the following best practices to align with MDM:

- ✓ Improve the quality of master data to support big data analytics.
- ✓ Leverage big data to improve the quality of master data.
- ✓ Improve the quality and consistency of key reference data to support the big data governance program.
- ✓ Extract meaning from unstructured text to enrich master data.
- ✓ Enrich customer master data with insights from social media to create social MDM.
- ✓ Turbo-charge MDM with Hadoop technologies.

Each best practice is discussed in detail in the rest of this chapter.

8.1 Improve the Quality of Master Data to Support Big Data Analytics

The relationship between MDM and big data is similar to MDM and data warehousing. Many organizations use MDM to cleanse data that is then fed into the data warehouse. Similarly, organizations need high-quality master data to support big data analytics. Here are some examples that justify the importance of high-quality master data to big data analytics projects:

- *Materials*—Consumer packaged goods companies use point-of-sale transaction logs from retailers to determine which products are selling at what stores. These initiatives need consistent product information from retailers so that manufacturers can compare trends across outlets.
- *Assets*—Organizations leverage real-time sensor data to build preventive maintenance models. These models can identify situations where event code A1234 (high vibration) is followed in 95 percent of the cases by event code B2345 (asset breakdown). The preventive maintenance models might be based on millions of rows of sensor event data from hundreds of pieces of equipment and thousands of sensors. Organizations can substantially improve the accuracy of their predictive maintenance models by standardizing asset hierarchies and naming conventions. Consider an example where the asset management system has two instances of the same pump with different names. The two pumps will generate event codes A1234 and A4567 for high vibration and event codes B2345 and B5689 for asset breakdown. As a result, the predictive model will be unable to pool the sensor data from these assets, which will diminish its overall ability to predict breakdowns.
- *Customers*—Big data analytics can predict the likelihood of customer churn. For example, telecommunications operators build detailed customer churn models that include big data such as calling patterns, dropped calls, and social media. However, the overall value of the churn models also depends on traditional attributes of customer master data such as date of birth, gender, location, and income.

Case Study 8.1 discusses the big data analytics program at a brand-name retailer. (The numbers in this case study have been disguised.)

Case Study 8.1: Big data analytics at a brand-name retailer

A popular brand-name global retailer was experiencing declining product profit margins due to increased promotional activity. To address this business challenge, the company decided to collect and analyze product feedback from customers in social media such as Twitter and other websites, to determine the pricing strategy for new products. If the so-called sentiment analysis was not very positive during the product launch, the company decided to update its pricing in the master product catalog and offer discounts of 30 percent. This would replace its usual practice of selling merchandise at the end of the season at a discount of 70 percent. As a result, the retailer was able to improve its profit margins, as shown in Table 8.1. It is noteworthy that the retailer used high-quality master data with standard product definitions and hierarchies to support this initiative.

The same retailer also piloted a flash event lasting just one afternoon to promote a new line of swimwear. The marketing team used only social media to attract customers to the event and anticipated that the communication would go viral. While the event was extraordinarily successful and sales exceeded projections, the marketing team uncovered some issues when analyzing clickstream data. Customers who had taken pictures of the new line could not easily find the product online. After examining the root cause, the retailer had to modify its product hierarchy so that boardshorts could be found in shorts, swimwear, and within its own subclass of boardshorts.

Table 8.1: The Hypothetical Business Benefits from Using Sentiment Analysis and Sound Product Master Data at a Global Retailer

A.	New product line value at list price	$10,000,000
B.	Deep discount at the end of the season	70%
C.	Start of the season discount	30%
D.	Difference in discounting levels (B – C)	40%
E.	**Gross profit benefit as a result of better market trends analytics (A × D)**	**$4,000,0000**

8.2 Leverage Big Data to Improve the Quality of Master Data

The master data management program can also leverage big data to improve overall data quality. Case Study 8.2 discusses how an organization used IBM InfoSphere QualityStage in conjunction with web and other data sources to improve the quality of its product master data.

Case Study 8.2: Leveraging web and other data sources to improve the quality of product master data at an information services company

An information services provider was responsible for maintaining highly granular master data about millions of consumer products. The company used a number of data sources, including product images from the web, point of sale transaction logs (POS TLogs), and store circulars, to derive a number of attributes for each item, such as the following:

- Nutritional information, including ingredients and nutrition panel information that were subject to government regulations
- Marketing claims such as "kosher," "low cholesterol," and "no trans fats" that were derived from the front of the package and were not subject to government regulations
- Product sales information from POS TLogs
- In-store promotions from store circulars
- Sentiment analysis from blogs, magazines, and social media

The company leveraged a crawling and indexing engine to pull in structured and semi-structured content from the web. It used several vendor-provided and internally generated techniques to add meaning to unstructured content using keywords, filters, and taxonomies. Because of this initiative, the company was able to achieve the following business results:

- Increase the amount of product data that it was able to manage in-house.
- Improve the quality of sparsely populated datasets such as nutritional information where it did not receive much data from manufacturers.
- Validate product information that it received from manufacturers. For example, the company used web content to validate manufacturer-provided Universal Product Codes (UPCs), which are 12-digit codes represented as bar codes on products in North America. The company also validated product attributes, such as for a shampoo that was listed as "4 oz." on the web versus "3.8 oz." in the product master.
- Provide additional analytics to manufacturers that were not possible earlier. For example, it could answer the question, "Does retailer X advertise toothpaste from our competitors at the same time as our brands?"
- Establish causality between sentiment analysis and product sales to show, for example, that a posting by an influential blogger resulted in an increase in sales of a certain product.

8.3 Improve the Quality and Consistency of Key Reference Data to Support the Big Data Governance Program

Organizations now recognize that their reference data presents a similar set of challenges as master data. Compared with master data, reference data is relatively static and may be placed in lookup tables for reference by other applications across the enterprise. Examples of reference data include codes for countries, states, provinces, currencies, and industries. In education, reference data might include codes for courses, ethnicity, and race. Here are some examples of the importance of reference data to the big data governance program:

- *ICD-9 healthcare codes*—Big claims transaction data needs a reference table of diagnoses and ICD-9 codes.
- *State codes*—Telecommunications CDRs include codes for 53 U.S. states, while the billing system includes codes for only 50 states.
- *Currencies*—One banking application uses "JPY" while another one uses "YEN" to describe Japanese yen for financial transactions.
- *Consumer products*—In Case Study 8.2, the information services provider maintained 28,000 unique values for color. For example, it maintained reference data to indicate that "RED," "RD," and "ROUGE" all referred to the same color. In addition, it used reference data to classify all the products that a particular manufacturer collectively referred to as "breakfast meals."

Ideally, big data governance needs to check key reference data against a table of agreed-upon values prior to data load. A reference data steward should flag any deviations as exceptions for subsequent review. IBM InfoSphere Master Data Management Reference Data Management Hub manages reference data such as codes for countries, states, industries, and currencies.

8.4 Extract Meaning from Unstructured Text to Enrich Master Data

MDM systems provide a 360-degree view of customers, vendors, materials, assets, and other entities. Traditional MDM systems collect structured data from a number of structured data sources. With the advent of big data, MDM projects will increasingly look to derive value from the large volumes of entity information that is hidden within unstructured text, such as social media, email, call center voice transcripts, agent logs, and scanned text. This content might reside in multiple formats, such as plain text, Microsoft Word® documents,

and Adobe® PDF documents, and in different forms of storage, such as content management repositories and file systems. In Case Study 8.3, the MDM team at a hypothetical company needs to integrate email with the customer record using IBM InfoSphere Master Data Management and IBM InfoSphere BigInsights text analytics technologies.

Case Study 8.3: Integration of email with customer MDM

The MDM program needs to adopt the following steps to enrich master data with sources of unstructured text:

1. *Define the attributes for each entity that needs to be governed.*

   Figure 8.1 describes a simplified schema for the customer entity. It includes attributes for name, company, city, country, and email address that are commonly found in a CRM system. In this theoretical example, the master data team will use these attributes to make correlations to existing IBM InfoSphere Master Data Management records with additional content from emails.

Name	Company	City	Country	Email

Figure 8.1: A simplified schema for the customer entity in IBM InfoSphere Master Data Management.

2. *Generate a dictionary file for each attribute from the MDM repository and other sources.*

   The MDM team then needs to create or reuse a dictionary containing a list of all possible values. This dictionary may be generated in multiple ways. One approach would be to create the dictionary based on all the existing values for each attribute in the IBM InfoSphere Master Data Management repository. Additional algorithms and annotation logic can then enhance these dictionaries. Figure 8.2 provides an example of a dictionary that was created by looking up the values for each customer attribute within IBM InfoSphere Master Data Management.

Name
John Doe
Paul Dean
Jane Lee
Jim Doe

Company
Acme
Akron
Alloy

City
Chicago
London

Country
USA
UK
France

Email
jdoe@acme.com
jdoe3@acme.com
pdean@alloy.com

Figure 8.2: Attribute dictionaries for the customer entity in IBM InfoSphere Master Data Management.

3. *Annotate relevant terms based on fuzzy matching and business rules.*

   Figure 8.3 shows a sample intercompany email that summarizes some business discussions. The team uses IBM text annotation techniques to locate the highlighted terms based on the dictionary. Some terms, such as "date," are not amenable to dictionary-based annotation because there are too many different ways of writing the same date. Although date is not a key attribute in this example, a date-specific annotation algorithm or rule might be more appropriate.

Jon Doe and **Janet Lee** from **Acme** met in **Chicago** with **Paul Dean** from **Akron** to discuss the materials contract. The meeting took place on 24 August 2011.

Jon from **Acme** in the **UK** is currently visiting the office in **Chicago** to help **Janet** and **Paul** with planning a joint venture between **Acme** and **Akron**. **Paul** is scheduled to spend considerable time in the **UK** later this year to oversee the execution of the project.

Please contact Jon (**jdoe@acme.com**) or Paul (**paul.d@akron.com**) for further information.

Figure 8.3: A sample email with unstructured content.

4. *Construct a query to the MDM system that consists of the annotations from the unstructured text.*

   As shown in Figure 8.4, the text analytics platform issues a single MDM query based on the annotations from the unstructured text.

Name = "Jon Doe" or "Janet Lee" or "Paul Dean" or "Jon" or "Janet" or "Paul" or "Lee" or "Doe" or "Dean"

Company = "Acme" or "Akron"

Country = "UK"

City = "Chicago" or "Akron"

Email = jdoe@acme.com or paul.d@akron.com

Figure 8.4: A single MDM query based on annotations from unstructured text.

IBM InfoSphere Master Data Management then returns the records shown in Figure 8.5 that were above the minimal matching threshold. Entity identifier 1 (EID 1) received a high matching score because MDM found a match on email, employer, country, and name, which have a high weighting in the matching algorithm.

EID	Matching Score	Name	Employer	Country	City	Email
1	High	John Doe	Acme	UK	London	jdoe@acme.com
2	Low	Jim Doe	Acme		Chicago	jdoe3@acme.com
3	Medium	Paul Dean	Akron	US	Chicago	pdean@akron.com
5	Medium	Jane Lee	Acme	US		

Figure 8.5: IBM InfoSphere Master Data Management returns records above the minimal matching threshold.

5. *Construct a record for the unstructured entity.*

   The system then constructs the matching entities based on the attributes found in the document. In Figure 8.6, the system uses the email to construct a record for matching entity identifier 1 (MEID 1) as follows:

   Name = "Jon Doe"
   Employer = "Acme"
   Country = "UK"
   Email = "jdoe@acme.com"

The record also contains the following attributes to identify the source of the information, the type of document, and the strength of that association:

DOCID = "Doc1" (identifier for the specific email)
Source = "Email"
Confidence = "High"

Name	Employer	Country	City	Email	DOCID	SOURCE	MEID	Confidence
Jon Doe	Acme	UK		jdoe@acme.com	Doc1	Email	1	High
Jon Doe	Acme				Doc1	Email	2	Low
Paul Dean	Akron		Chicago		Doc1	Email	3	Medium
Janet Lee	Acme				Doc1	Email	5	Medium

Figure 8.6: Entities constructed by text analytics intersecting the results from MDM with the extracted terms.

6. *Associate newly constructed entity records with existing MDM records.*

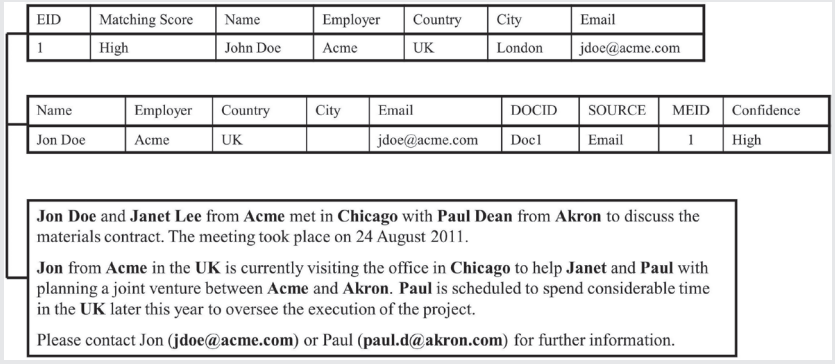
   As shown in Figure 8.7, the newly constructed entity records are automatically inserted into IBM InfoSphere Master Data Management if the confidence level is high. On the other hand, they are automatically rejected if the confidence level is low. For records with a medium confidence level, a data steward will manually review the records to determine whether they should be linked to existing MDM records.

Confidence Level	Algorithm
High	Automatically insert record into MDM
Low	Automatically reject record
Medium	Manual review of record by data steward

Figure 8.7: An algorithm to process the newly constructed entity records in IBM InfoSphere Master Data Management.

As shown in Figure 8.8, IBM InfoSphere Master Data Management will automatically link MEID 1 for Jon Doe with EID 1 for John Doe. The email in Figure 8.3 will now be associated with EID 1 as well. On the other

hand, IBM InfoSphere Master Data Management will automatically reject MEID 2 from Figure 8.6 because the confidence level is low. Finally, a data steward will manually review MEID 3 and MEID 5 because the confidence level is medium.



EID	Matching Score	Name	Employer	Country	City	Email
1	High	John Doe	Acme	UK	London	jdoe@acme.com

Name	Employer	Country	City	Email	DOCID	SOURCE	MEID	Confidence
Jon Doe	Acme	UK		jdoe@acme.com	Doc1	Email	1	High

Jon Doe and **Janet Lee** from **Acme** met in **Chicago** with **Paul Dean** from **Akron** to discuss the materials contract. The meeting took place on 24 August 2011.

Jon from **Acme** in the **UK** is currently visiting the office in **Chicago** to help **Janet** and **Paul** with planning a joint venture between **Acme** and **Akron**. **Paul** is scheduled to spend considerable time in the **UK** later this year to oversee the execution of the project.

Please contact Jon (**jdoe@acme.com**) or Paul (**paul.d@akron.com**) for further information.

Figure 8.8: Linking matching entities and emails in IBM InfoSphere Master Data Management.

By incorporating unstructured information into MDM, the master data team can build a better view of the overall customer relationship. This can also be extremely helpful during personnel changes. In the preceding example, Jon Doe and Janet Lee from Acme were working with Paul Dean from Akron. If Paul Dean left Akron before completion of the contract, how would the new representative, Lucas Alexander, get an updated status on all the work that was scheduled with Acme?

By enriching MDM with unstructured text such as this email, Lucas would be able to open the profile of Acme. He would see that his contacts were John Doe and Jane Lee. Looking further, he would see that there was an email related to the profile. Upon reading the text, he would know that he should follow up with Acme and let them know he was looking forward to continuing the work they began with Paul. This would help cement the relationship between Acme and Akron, increase Akron's retention of Acme as a client, and reduce the risk of critical leads being lost due to employee turnover or the failure of Paul Dean to enter the opportunity in the lead tracking system.

There are a number of other business applications to support the integration of text analytics with MDM. For example, bank risk departments can use the integration of text analytics with MDM to update counterparty risk. The risk department at a bank can use unstructured financial information such as U.S. Securities and Exchange Commission (SEC) filings to learn that the ownership

of a company has changed or that a large customer is also a director in three other companies. The risk department can use this information to update the customer hierarchies in MDM to establish an up-to-date picture of the overall exposure to a customer.

8.5 Enrich Customer Master Data with Insights from Social Media to Create Social MDM

Social media presents a new data type of unstructured data that is a valuable source of self-reported information. This information is useful for customer-centric efforts such as retention, cross/up-sell, and campaign management. Twitter, blogs, reviews, and Facebook all contain valuable information for decision-making. However, there are three key problems in using social media information:

- *Deriving meaning*—Mining intent, sentiment, life events, interests, and demographics from unstructured data is not always easy. The same word may be used in different contexts to mean different things.
- *Data sparsity*—Because social media information may be relatively sparse, it is not always easy to associate an entity with an existing customer record in an organization's master data.
- *Integration with structured data*—Integrating unstructured data from social media with structured data within an existing MDM record is also a major challenge. The organization may need to establish business rules to determine whether the phone number on Twitter is more trustworthy than the one in the customer's MDM record. As a first step, organizations may want to keep their internal MDM and social media entities separate but linked.

As shown in Figure 8.9, IBM InfoSphere BigInsights uses advanced text analytics techniques and business rules to derive insight from unstructured information relating to a hypothetical individual named Jane Doe. It uses the location on her Twitter account to discern that she is a football fan and lives in Tampa, Florida. Jane Doe's Tweet indicates that she is interested in yoga and that Tony C. is part of her network. Since the blog contained a reference to her Twitter handle, IBM InfoSphere BigInsights was also able to figure out that Jane Doe blogs about food-related topics.

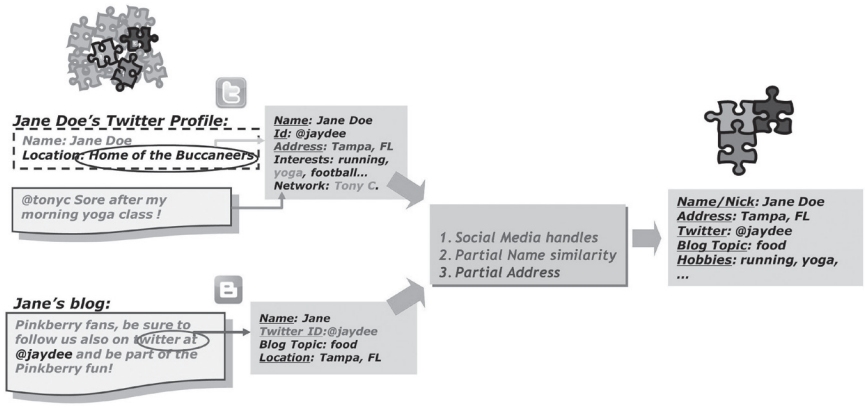


Figure 8.9: Creation of an unstructured entity from social media with IBM InfoSphere BigInsights.

As shown in Figure 8.10, IBM InfoSphere BigInsights extracts entities from social media. These entities are then fed into IBM InfoSphere Master Data Management, where they are linked with existing customer master data records using probabilistic matching techniques. This solution addresses many-to-one relationships (where a single customer may have multiple social media profiles) and one-to-many relationships (where multiple individuals such as a husband and wife share the same user ID). IBM InfoSphere Master Data Management also provides a data stewardship console for manual linkage between social media and customer entities.

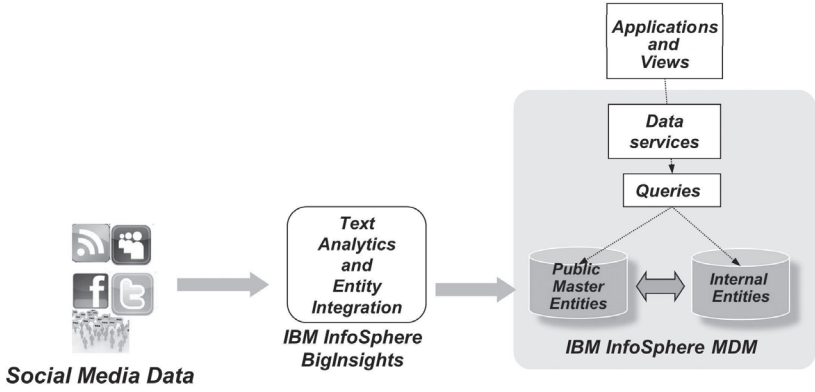


Figure 8.10: Integration of IBM InfoSphere BigInsights and IBM InfoSphere Master Data Management.

8.6 Turbo-Charge MDM with Hadoop Technologies

Marketing organizations often need to match lists of prospects against internal records to remove any customers who have made do-not-call elections. These large datasets push the limits of existing computational resources when IT needs to match 200 million prospects against a database of 100 million customers and turn around the results to marketing in 24 hours. The largest bulk entity resolution on the IBM InfoSphere Master Data Management platform was approximately one billion records. However, this exercise typically required manual partitioning of the data across several large machines, and took multiple days to process. As shown in Figure 8.11, IBM has implemented the IBM InfoSphere Master Data Management probabilistic matching engine within a MapReduce framework on an IBM InfoSphere BigInsights platform. This has helped organizations implement probabilistic matching on ultra-large datasets in hours rather than days or weeks. By combining the power of the IBM Master Data Management probabilistic matching engine and the scalability of IBM InfoSphere BigInsights, IBM has developed a scalable framework that can perform bulk entity resolution on billions of records in less than a day using a cluster of commodity hardware. Now, organizations will be able to master prospect/marketing lists in situations where this would not have been feasible before.

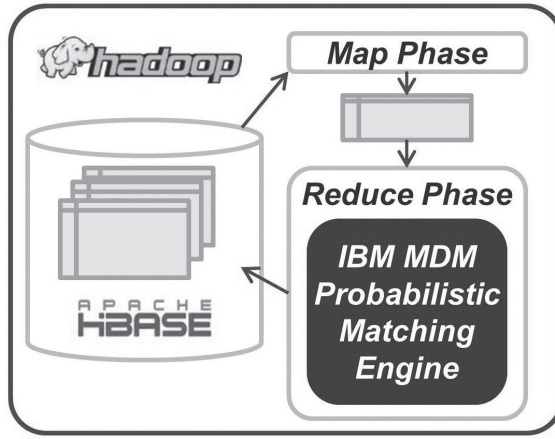


Figure 8.11: IBM Master Data Management probabilistic matching engine combined with Apache HBase inside IBM InfoSphere BigInsights.

Summary

Big data, master data, and reference data are highly synergistic. Big data should use high-quality master data on customers, materials, assets, and employees. Big data also needs high-quality reference data. In addition, master data can be enriched based on big data such as web content, call detail records, and social media. IBM InfoSphere BigInsights, IBM InfoSphere Master Data Management, IBM InfoSphere Master Data Management Reference Data Management Hub, and IBM InfoSphere QualityStage provide a strong foundation for organizations looking to embark on their big data journey.

CHAPTER 9

MANAGING THE LIFECYCLE OF BIG DATA

This chapter includes contributions from Aarti Borkar (IBM), Rani Hublou (IBM), Deidre Paknad (IBM), and Brian M. Williams (IBM).

Because of the massive increase in data volumes, organizations are being 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 has high value and ongoing obligations, or the company faces very high risks from improper disposal. With IT budgets continuing to be under pressure, over-managing information is a gross waste of capital resources.[20]

According to the 2010 Gartner study *IT Metrics: IT Spending and Staff Report*[21], IT costs are 3.5 percent of revenue and are under significant pressure, with 61 percent of these costs being a function of information volume. The big data governance program needs to establish policies that govern the lifecycle of big data to reduce legal risk and IT costs.

The best practices to manage the lifecycle of big data are as follows:

- ✓ Expand the retention schedule to include big data based on local regulations and business needs.
- ✓ Document legal holds and support eDiscovery requests.
- ✓ Compress and archive big data on Hadoop to reduce storage costs.

- ✓ Archive big data in immutable format with seamless access to Hadoop for analytics.
- ✓ Manage the lifecycle of real-time, streaming data.
- ✓ Defensibly dispose of big data no longer required based on regulations and business needs.

These best practices are discussed in detail in the remainder of this chapter.

9.1 Expand the Retention Schedule to Include Big Data Based on Local Regulations and Business Needs

Every country, state, and province has unique regulations relating to data retention. As a first step, the big data governance program needs to understand the retention requirements for each big data type by industry and jurisdiction. IBM Global Retention Policy and Schedule Management catalogs privacy laws and associates requirements and procedures to specific data sources and data categories. These tools can also reconcile conflicts of law and bring greater precision and clarity to applicable privacy requirements such as usage restrictions, transport limitations, disposition protocols, storage media, encryption requirements, and disclosure obligations.

9.2 Document Legal Holds and Support eDiscovery Requests

Most corporations and entities are subject to litigation and governmental investigations that require them to preserve potential evidence. Large entities might have hundreds or thousands of open legal matters with varying obligations for data. For example, the eDiscovery for the first phase of the trial in the BP oil spill case exceeded nine million documents and 15 terabytes of data.[22] A typical legal matter lasts three years, and many last five or more years.

The information governance program needs to control legal risk and manage costs while communicating obligations to information custodians, gathering evidence, and analyzing results. As the use of big data becomes more prevalent across the enterprise, so will its use in legal matters. For example, drilling companies that are sued for oil spills might need to produce sensor data from the rig to demonstrate that they exercised appropriate caution when dealing with the associated events.

IBM eDiscovery Solutions enable legal teams to define evidence obligations; coordinate with IT, records, and business teams; and reduce the cost of producing large volumes of evidence in legal matters. The solution includes IBM Atlas eDiscovery Process Management, which enables legal professionals to manage a legal holds workflow.

9.3 Compress and Archive Big Data on Hadoop to Reduce Storage Costs

Organizations need to compress and archive their big data at rest to reduce storage costs and to improve application performance. Big data at rest includes smart meter readings, sensor data, RFID data, and web logs that might reside in relational databases, file systems, NoSQL databases, and Hadoop. Because Hadoop avoids data loss by replicating the same data across multiple nodes in a cluster, organizations should also consider IBM InfoSphere BigInsights for fault-tolerant data archiving, as shown in Case Study 9.1.

Case Study 9.1: The economics of IBM InfoSphere BigInsights as a data archiving solution at a mid-sized company

Table 9.1 describes the economics of IBM InfoSphere BigInsights as an archiving solution for *structured and unstructured data* at a mid-sized company. The organization calculated the annual cost of existing data storage at $20,000 per terabyte, including systems administrators and software tools. The organization calculated that it could save $11,000 per terabyte despite the data replication that is inherent to Hadoop.

Table 9.1: The Economics of Hadoop as a Data Archiving Solution at a Mid-sized Company

A.	Annual cost per terabyte for existing data storage	$20,000
B.	Annual cost per terabyte for data storage within Hadoop	$3,000
C.	Number of times that data is replicated in Hadoop	3
D.	Annual cost per terabyte of Hadoop storage (B × C)	$9,000
E.	**Annual storage cost savings per terabyte (A – D)**	**$11,000**

9.4 Archive Big Data in Immutable Format with Seamless Access to Hadoop for Analytics

IBM InfoSphere Optim Data Growth solution helps organizations reduce storage costs and improve application performance by archiving *structured data*. Compared with a Hadoop solution, IBM InfoSphere Optim Data Growth offers the following benefits relative to the archiving of structured data:

- *Compliance*—Data is archived in immutable format, access is tightly controlled, and user actions are audited.
- *eDiscovery*—Data is easily retrievable during legal proceedings.
- *Legal holds*—Archived data can be subjected to legal holds.
- *Defensible disposition*—Archived data can be defensibly disposed to minimize legal impact.
- *Application and business intelligence owners*—Archived data is accessible to business intelligence and enterprise applications without code changes.
- *Data owners*—Archived data can be easily restored back to the source and is searchable.

Organizations can now leverage the power of Hadoop to perform blended analytics of archived data in Optim as well as structured and unstructured data from other sources. Organizations can store their data in immutable format in IBM InfoSphere Optim Data Growth solution, which can now create queryable data archive files for storage in HBase. As a result, organizations can combine the immutability of an Optim archive with the processing power and cost-effectiveness of the IBM InfoSphere BigInsights Hadoop platform. Finally, IBM InfoSphere Data Explorer can also search for data within Optim archive files.

9.5 Manage the Lifecycle of Real-Time, Streaming Data

Information lifecycle management (ILM) is turned on its head in the context of real-time, streaming data. When data is arriving at high velocity, big data teams need to know what data is valuable and what needs to be persisted. If the streaming analytics application can make this determination "in the moment," then it can apply ILM policies to data in motion.

For example, IBM InfoSphere Streams might analyze sensor readings every tenth of a millisecond and store the readings every second. However, when sensor readings begin to indicate anomalous behavior, the streaming analytics

application might store every reading up to and after the event. Consider Case Study 9.2, where a network monitoring system analyzes streaming data for abnormal events.

Case Study 9.2: A network monitoring system that analyzes streaming data for abnormal events

A network monitoring system analyzes NetFlow data from different routers. Each NetFlow record contains statistical information from network routers, such as the source IP address, destination IP address, and number of bytes and packets. The network monitoring application profiles the data in real-time and compares it with historical norms. It might observe an increase in traffic to a social media website at 9:00 a.m., when employees begin their workdays. However, suppose it notices an abnormally large volume of outgoing network traffic to a previously unknown destination. That might be a sign of an exfiltration (data leaving the company's network).

IBM InfoSphere Streams accomplishes real-time network analytics by keeping a portion of network history in memory. The security operations team needs to determine how much data should live in memory. For example, it might decide to keep two hours' worth of NetFlow records in memory and persist the readings to disk every minute for historical analysis.

9.6 Defensibly Dispose of Big Data No Longer Required Based on Regulations and Business Needs

Many organizations believe that keeping data forever is a good response to legal requirements. Actually, the converse is true. Any data, whether in electronic or paper format, is subject to legal discovery if it exists anywhere in the organization, whether in a storage cabinet, an employee's desk drawer, a server, or on a thumb drive. In a 2010 survey by the Compliance, Governance and Oversight Council titled *Information Governance Benchmark Report in Global 1000 Companies*, 75 percent of respondents cited the inability to defensibly dispose of data as their greatest challenge. [23] Many highlighted massive legacy data as a financial drag on the business and a compliance hazard. The big data governance program needs to establish policies that require the deletion of big data based on the retention schedule, unless it is subject to legal holds. As an example, if the retention schedule requires telecommunications CDRs to be kept for two years, then all records should be deleted after that period, except for the subset that is subject to legal holds.

Summary

The lifecycle of big data needs to be managed for data-at-rest and for data-in-motion. By managing the lifecycle of big data, organizations can reduce IT costs, improve application performance, respond to eDiscovery requests, and defensibly dispose of information. IBM's Information Lifecycle Governance Solutions including IBM InfoSphere Optim Data Growth solution and IBM InfoSphere BigInsights can help organizations manage the lifecycle of big data.

CHAPTER 10

An Introduction to Process Data Governance

According to Forrester Research, organizations kick off MDM initiatives to cleanse and integrate large volumes of customer, materials, vendor, and location data. These organizations also launch business process management (BPM) initiatives to optimize their mission-critical processes. However, Forrester notes that these initiatives remain siloed, with limited collaboration across teams.[24] Forrester's perspectives also apply to the alignment between BPM and big data governance. IBM has recognized this convergence between BPM and data governance by including IBM Business Process Manager Express within IBM InfoSphere Master Data Management V10.

Organizations need to appreciate the symbiosis between BPM and big data governance. The next three chapters provide detailed case studies regarding process data governance in retail, oil and gas, and healthcare.

CHAPTER 11

Retail Case Study: Process Data Governance of Social Media

This chapter includes contributions from Kurt Wedgwood (IBM).

Case Study 11.1 describes how a hypothetical retailer named Acme Corporation can leverage IBM InfoSphere to power a Facebook app. The marketing department wants to use master data on customers, products, employees, and store locations to enrich its Facebook app.

Case Study 11.1: Leveraging social media at a retailer

The following is a list of the key activities involved in powering the Facebook app at Acme Corporation:

- Kate (customer) receives a Facebook app request from Sally (friend).
- Kate accepts the app request based on incentives such as discounts, friend alerts, and access to premium merchandise from Acme (retailer).
- Kate opts-in and permits Acme to access her basic information and information about her friends.
- Acme matches Kate's Facebook profile to her internal record in IBM InfoSphere Master Data Management.
- Acme uses information from Kate's MDM profile and past purchases to make attractive offers for new merchandise.
- Acme also suggests the best store locations where Kate might be able to check out the promoted merchandise in person and assigns a personal shopper to be Kate's main contact.

The next step is to map the key changes in big data governance policies to the specific activities. Table 11.1 summarizes the key big data governance

policies of the retailer in Case Study 11.1 looking to use IBM InfoSphere to power a Facebook app.

Table 11.1: Key Big Data Governance Policies at a Retailer

Activity	Big Data Governance Policy
1. Kate (customer) receives a Facebook app request from Sally (friend).	Acme uses IBM InfoSphere BigInsights to identify the top influencers in social media relative to its products.
2. Kate accepts the app request based on incentives such as discounts, friend alerts, and access to premium merchandise from Acme (the retailer).	Acme uses A/B testing to fine-tune the incentives based on demographic and behavioral attributes. Acme uses IBM InfoSphere QualityStage to generate high-quality data on demographics, behaviors, and product affinities for A/B testing.
3. Kate opts-in and permits Acme to access her basic information and information about her friends.	Marketing adheres to the Facebook Platform Policies by, for example, not using data on a customer's friends outside the context of the app.
4. Acme matches Kate's Facebook profile to her internal record in IBM InfoSphere Master Data Management.	Marketing data stewards identify the attributes to link a customer's Facebook profile with his or her record in IBM InfoSphere Master Data Management. Acme adds Kate's Facebook user ID to her MDM record.
5. Acme uses information from Kate's MDM profile and past purchases to make attractive offers for new merchandise.	Merchandising data stewards establish a robust product hierarchy using multi-domain support within IBM InfoSphere Master Data Management to enable product comparisons. As a simple example, the retailer needs to know that, because Kate had purchased a "Whirlpool GX5FHDXVY," she already has a product in the "refrigerator" hierarchy.
6. Acme also suggests the best store locations where Kate might be able to check out the promoted merchandise in person and assigns a personal shopper to be Kate's main contact.	Real estate data stewards establish location master data in IBM InfoSphere Master Data Management to identify physical stores that are preferred by the customer. Store operations stewards create associate master data and define the process to assign an associate to the customer.

Summary

This chapter reviews the importance of integrating BPM and big data governance for social media in retail. The case study illustrates the importance of master data management, data quality, and business process management to the governance of social media data.

CHAPTER 12

Oil and Gas Case Study: Process Data Governance of Sensor Data

This chapter includes contributions from Arild Kristensen (IBM) and Frode Myren (IBM).

This chapter provides a case study on the process data governance of sensor data within the oil and gas industry.

Case Study 12.1: The governance of sensor data within the oil and gas industry

Figure 12.1 describes a simple process to manage oilfield sensor data, including key activities and milestones. This process is based on a discovery map that was built using IBM Blueworks Live™, which is a cloud-based version of IBM Business Process Manager Express. Table 12.1 provides an overall description of these milestones and activities.

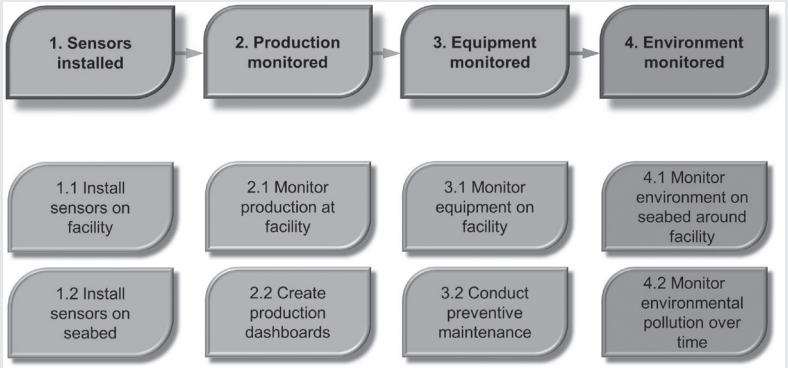


Figure 12.1: The process for monitoring oilfield sensor data built using a discovery map in IBM Business Process Manager Express.

Table 12.1: Key Milestones and Activities to Manage Oilfield Sensor Data

Seq.	Milestone/Activity	Description
1.	Sensors installed	Oil and gas companies install sensors on facilities as well as the seabed to monitor production, the state of the facility, health and safety, and adherence to environmental regulations. The sensor control systems typically support the OPC protocol, a standard that specifies the communication of real-time plant data between SCADA systems from different manufacturers.
1.1	Install sensors on facility	The modern oil facility might have more than 30,000 sensors that capture numerous types of real-time data from the exploration process, such as flows, revolutions per minute (RPM), voltage, watts, temperature, and pressure.
1.2	Install sensors on seabed	Companies might also install sensors on the seabed to monitor environmental conditions such as flow, temperature, and turbidity. Turbidity is a measurement of water quality based on the cloudiness of water caused by individual particles that might not be visible to the naked eye.
2.	Production monitored	Organizations need to monitor production of oil and gas. The oil company, acting as the operator, also calculates the production allocation to each owner of the facility.

Table 12.1: Key Milestones and Activities to Manage Oilfield Sensor Data (continued)

Seq.	Milestone/Activity	Description
2.1	Monitor production at facility	Operators install sensors to monitor oil and gas production at each facility.
2.2	Create production dashboards	Oil and gas companies also create dashboards to monitor energy production across facilities. The companies create common operations centers so that they can monitor production from a central location.
3.	Equipment monitored	Facilities use sensors to monitor equipment.
3.1	Monitor equipment on facility	Operations departments monitor equipment such as pumps and valves on each rig. Typical questions include the following:[25] • "Given a brand of turbine, what is the expected time to failure when the equipment starts to vibrate in the manner now detected?" • "Given an alarm on a well, how much time do we have to take corrective action, based on the historical behavior of the well?" • "How do we detect weather events from the observation data?" • "Which sensors have observed a blizzard within a 100-mile radius of a given location?"
3.2	Conduct preventive maintenance	Operators conduct preventive maintenance if their predictive models indicate that a particular piece of equipment is likely to fail.
4.	Environment monitored	Oil and gas companies use sensors to monitor the environment.
4.1	Monitor environment on seabed around facility	Environmental sensors may be in operation before, during, and after the operating life of the platform.
4.2	Monitor environmental pollution over time	Companies need to answer questions such as "Do the levels of salinity and turbidity in the water around the facility indicate an oil spill?"

Table 12.2 summarizes the key big data governance policies associated with managing oilfield sensor data.

Table 12.2: Key Big Data Governance Policies Relating to Oilfield Sensors

Seq.	Milestone/ Activity	Big Data Governance Policy
1.1	Install sensors on facility	The big data governance program should ensure that the SCADA systems are properly secured against the possibility of cyber attacks.
2.2	Create production dashboards	The big data governance program needs to ensure consistency of the business terms within production reports. The program needs to establish consistent definitions for key business terms such as "well," in addition to associated child terms such as "well origin," "well completion," "wellbore," and "wellbore completion." The big data governance program should leverage standard models such as the Professional Petroleum Data Management (PPDM) Association model for well data and definitions. These definitions can be stored in IBM InfoSphere Business Glossary, which also allows the program to assign data stewards to key business terms.
3.1	Monitor equipment on facility	In the past, a rig might have had only about 1,000 sensors, of which only about 10 fed databases that would be purged every two weeks due to capacity limitations. Today, oil and gas companies need to retain sensor data for a much longer period. For example, the health, safety, and environment (HSE) department might need to re-create a picture using three-month-old information to explain why a particular decision was made in the field. The big data governance program should leverage standard models such as ISO 15926 for systems and equipment on oil and gas production facilities, and associated definitions. The program also needs to play a key role in determining how much information needs to be retained, and for how long, to satisfy both internal needs and the regulators. It is important to note that the rig might generate a lot of unstructured information, such as video, pictures, and sound. As discussed in chapter 9, IBM's Information Lifecycle Governance Solutions, including IBM InfoSphere Optim Data Growth Solution, can help organizations manage the lifecycle and archival of this big data.

Table 12.2: Key Big Data Governance Policies Relating to Oilfield Sensors (continued)

Seq.	Milestone/ Activity	Big Data Governance Policy
3.2	Conduct preventive maintenance	If a specific type of equipment failed on one rig, the oil company needs to quickly pinpoint where else the same equipment has been deployed so that it can initiate the appropriate preventive maintenance. However, if the same asset has different names on different rigs, it will be difficult to locate the asset in a timely manner. As a result, big data governance has a critical role to ensure consistent naming conventions for asset data. The Institute of Asset Management and the British Standards Institute have worked together to develop strategies to help reduce risks to business-critical assets. This project resulted in the Publicly Available Specification (PAS) 55, which embodies the latest thinking in terms of best practices in asset management systems. Oil and gas companies are increasingly adopting PAS 55 as the industry standard for quality asset management. IBM InfoSphere Quality Stage can manage the standardization of asset nomenclature. In addition, IBM InfoSphere Master Data Management offers support for multiple domains including asset data.
4.1	Monitor environment on seabed around facility	As discussed earlier, oil exploration and production activities generate a lot of structured and unstructured environmental information. This information needs to be maintained well after the lifetime of the facility itself, to demonstrate adherence to environmental regulations. This information might need to be stored for 50 to 70 years, or even up to 100 years in some cases. While storage is cheap, it is not free. The big data governance program needs to establish retention schedules for specific types of information and establish the appropriate archiving policies to move information onto cheaper storage, if possible. IBM's Information Lifecycle Governance Solutions for defensible disposition as well as IBM InfoSphere Optim Data Growth Solution for archiving are applicable in this context.

Summary

This chapter reviews the importance of integrating BPM and big data governance for sensor data in the oil and gas industry. The case study illustrates the importance of metadata, master data management, data quality, information lifecycle management, and business process management to the governance of sensor data.

CHAPTER 13

HEALTHCARE CASE STUDY: PROCESS DATA GOVERNANCE OF BIG CLAIMS TRANSACTION DATA

This chapter includes contributions from Bob Leo (IBM).

A responsibility assignment (RACI) matrix can demonstrate how different organizations might be engaged in big data governance. "RACI" stands for the following:

- *Responsible*—This is the person who has delegated responsibility to manage an attribute. There may be multiple responsible parties for one attribute.
- *Accountable*—This is the person who has ultimate accountability for the data attribute. The accountable person may delegate the responsibility to manage an attribute to a responsible party. There should be only one accountable party.
- *Consulted*—This is the person or persons who are consulted via bi-directional communications.
- *Informed*—This is the person or persons who are kept informed via uni-directional communications.

Case Study 13.1 demonstrates the use of process mapping and a RACI matrix at a large health plan. To set the context for this case study, a brief primer on claim codes will be helpful.

13.1 A Primer on Claim Codes Used by Health Plans

Health plans use claim codes to reimburse providers and hospitals, to benchmark costs and quality of service, and to offer care management services that reduce medical costs.

Health plans require their providers (doctors) to include the appropriate ICD-9, ICD-10, and CPT codes when submitting claims:

- *ICD-9-CM codes*
  The International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM) is based on the World Health Organization's ICD-9. ICD-9 was originally published in 1979 and was designed to promote international comparability in the collection, processing, classification, and presentation of mortality statistics. ICD-9-CM is the official system of assigning codes to diagnoses associated with hospital utilization in the United States.
- *ICD-10-CM codes*
  The tenth revision of the ICD was published by the World Health Organization in 1999 and significantly increased the number of codes. For example, ICD-9 has just one code for angioplasty, a procedure used to widen blocked blood vessels. Under ICD-10, medical practitioners can choose among 1,170 coded descriptions that pinpoint such factors as the location and the device involved for each patient.[26] The objective of ICD-10 is to provide fine-grained analyses of the causes of diseases. The United States Department of Health and Human Services requires the implementation of ICD-10-CM by October 1, 2013.
- *CPT codes*
  Current Procedural Terminology (CPT) codes are copyrighted by the American Medical Association and describe a uniform set of medical, surgical, and diagnostic services for physicians, patients, accreditation organizations, and healthcare payers for administrative, financial, and analytical purposes.

CPT coding is similar to ICD-9-CM and ICD-10-CM coding, except that it identifies the services rendered rather than the diagnosis on the claim. The critical relationship between these codes is that the diagnosis (ICD-9-CM) should support the medical necessity of the procedure (CPT). Since both ICD-9 and CPT are numeric codes, health insurers have designed software that compares the codes for a logical relationship. For example, a claim for CPT

31256 (nasal/sinus endoscopy) would not be supported by ICD-9 826.0 (closed fracture of phalanges of the foot). Such a claim would be quickly identified and rejected.[27]

Case Study 13.1: Big claims transaction data governance at a large health plan

A large health plan processed 500 million claims per year. Each claims record contained approximately 600 attributes, in addition to unstructured text. The health plan decided to focus on claims data governance because it spent about 85 cents of every premium dollar on claims, as was the norm in the industry. The business intelligence department conducted analytics on claims data. This activity drove several downstream activities, including care management. For example, if an elderly member (patient) made a number of doctor visits for "ankle pain," a nurse from healthcare services would call the person to consider treatment for arthritis. This proactive approach would improve the quality of life for the member while also reducing medical costs for the plan (insurer).

The business intelligence department noticed that a number of entries in the diagnosis code field were not ICD-9 codes. Upon profiling the data, the business intelligence team determined that the field included both ICD-9 (diagnosis) codes and CPT codes. The business intelligence team then met with the network management team that was responsible for managing provider (doctor) relationships. After many meetings, it became clear that the network management team had allowed doctors to use either ICD-9 codes or CPT codes, despite stringent guidelines that the field was only for ICD-9 codes. As a result, the claims reports showed inconsistent data, which resulted in the health plan devoting scarce nurse resources to dealing with low-risk patients.

The business intelligence team also conducted text analytics on the free-form text fields in the claims documents. The team compared the results with the reference data for CPT codes, and found several anomalies. For example, the free-form text seemed to indicate that the procedure was "flu shot" but the CPT code was "99214" which may be used for a physical. The conclusion was that providers might have been inadvertently entering incorrect procedure codes into the claims documents. Additionally, the business intelligence team analyzed text such as "chronic congestion" and "blood sugar monitoring" to determine that members might be candidates for disease management programs for asthma and diabetes, respectively.

Figure 13.1 describes a simple process to administer big claims transaction data at the health plan. This process was developed using IBM Blueworks Live, which is a cloud-based version of IBM Business Process Manager Express.

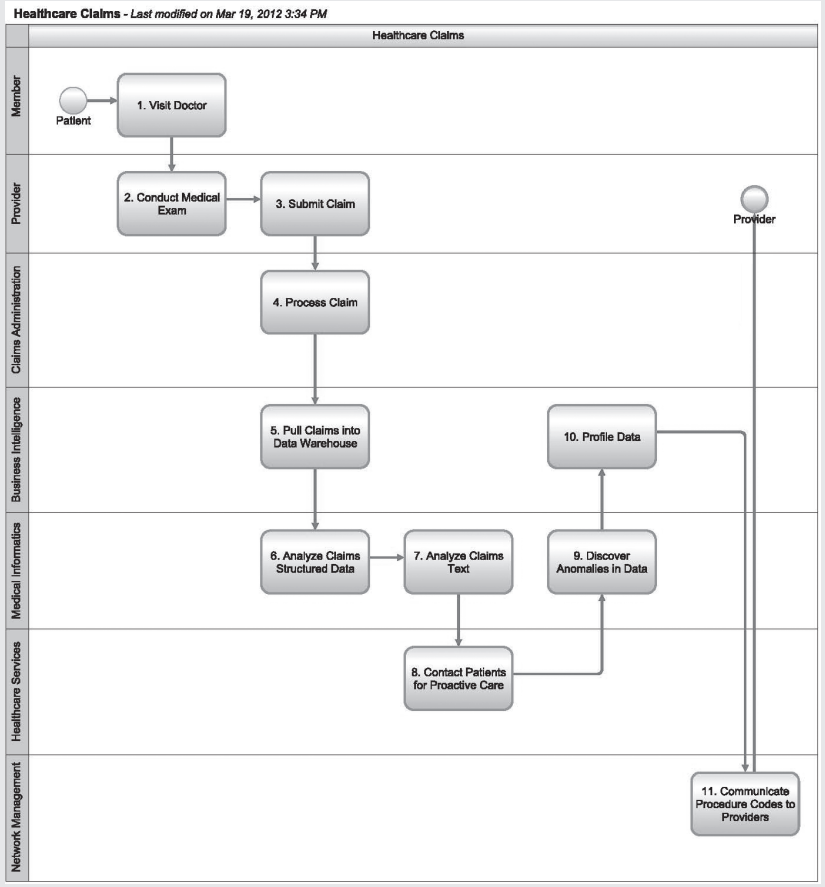


Figure 13.1: A simple claims administration process at a health plan built using IBM Business Process Manager Express.

There are a number of actors in the claims administration process:

- *Health plan*—This is the entity that pays for the medical costs of an insured party (also referred to as a healthcare payer or a health insurer).
- *Member*—This is a person who is covered under a health insurance policy, either as an enrollee or as a dependent.
- *Provider*—This refers to a doctor, therapist, practice, or hospital that provides medical services.
- *Claims administration*—This department within the health plan processes claims that are submitted by providers.

- *Business intelligence*—This department within the health plan is responsible for the data warehouse and analytics environment.
- *Medical informatics*—This department within the health plan deals with the use of information technology to solve clinical problems.
- *Healthcare services*—This department within the health plan deals with the clinical aspects of healthcare.
- *Network management*—This department within the health plan is responsible for managing the network of providers.

The health plan had to establish a number of policies to govern its big claims transaction data. A mapping of the key activities from Figure 13.1 to the big data governance policies is described in Table 13.1.

Table 13.1: Big Data Governance Policies for Big Claims Transaction Data at a Large Health Plan

Seq.	Activity	Big Data Governance Policy
4.	Process claim	The United States Health Insurance Portability and Accountability Act (HIPAA) safeguards the security and privacy of protected health information (PHI). The information security team established database monitoring using IBM InfoSphere Guardium to ensure that only authorized personnel could access claims records. For example, the insurer did not want the claims data for its senior executives and for high-profile celebrities to be randomly accessed by database administrators out of idle curiosity.
7.	Analyze claims text	As discussed earlier, the health plan used text analytics to uncover inconsistencies, such as when the procedure was "flu shot" but the CPT code was "99214," which may be used for a physical. The health plan leveraged IBM InfoSphere BigInsights for text analytics. It also used IBM InfoSphere Master Data Management Reference Data Management Hub to manage the reference data for ICD-9 and CPT codes to support these analytics.
9.	Discover anomalies in the data	The medical informatics team noticed that a number of entries in the procedure code field were not ICD-9 codes.
10.	Profile data	Upon profiling the data using IBM InfoSphere Information Analyzer, the business intelligence team determined that the diagnosis field incorrectly included both ICD-9 codes and CPT codes, despite stringent guidelines that the field was only for ICD-9 codes. As a result, the claims reports showed inconsistent data.
11.	Communicate procedure codes to providers	The big data governance program established a policy that the network management team would ask providers to only use ICD-9 codes in the relevant fields in claims documents.

As summarized in Table 13.2, the business intelligence team developed a RACI matrix to understand overall roles and responsibilities. The business intelligence team was accountable for the quality of data that drove claims analytics. However, the network management team was ultimately responsible to ensure that providers used the codes in a consistent manner. In addition, the medical informatics, healthcare services, and claims administration departments were consulted because they used the claims information on a day-to-day basis.

Table 13.2: Subset of a RACI Matrix for Healthcare Claims Data

Categories of Attributes	Responsible	Accountable	Consulted	Informed
ICD-9 and CPT codes	Network Management	Business Intelligence	Medical Informatics, Healthcare Services, Claims Administration	N/A

Summary

This chapter reviews the importance of integrating BPM and big data governance for healthcare. The big data governance program would be well-served to map policies against key activities within mission-critical business processes.

Notes

1. *The 2011 Digital Universe Study: Extracting Value from Chaos* (IDC, 2011).
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* (NASA, November 1999).
5. "An Overview of Biometric Recognition." http://biometrics.cse.msu.edu/ info.html.
6. "Biometrics in the Workplace." Data Protection Commissioner of Ireland. http://dataprotection.ie/viewdoc.asp?DocID=244.
7. Samuel Warren and Louis Brandeis. "The Right to Privacy." *Harvard Law Review*, Vol. IV, No. 5, December 15, 1890.
8. This chapter includes content from the IBM big data platform solution brief (IBM, 2012).
9. "IBM InfoSphere Streams Data Sheet" (IBM, 2011).
10. Jonny Longden. "Essential Guide to Accuracy in Web Analytics." Actionable Analytics blog, August 4, 2009. http://actionable-analytics.com/ 2009/08/essential-guide-to-data-accuracy-in-web-analytics.
11. "Velocity in Call Centers" (IBM Vivisimo, 2010). http://www.information-management.com/media/pdfs/casestudy-callcenters.pdf.
12. "Velocity in Pharmaceuticals" (IBM Vivisimo, 2010). http://vivisimo.com/ docs/casepharma.pdf.
13. http://vivisimo.com/customers/vivisimo-customers.html.
14. *The 2011 Digital Universe Study: Extracting Value from Chaos* (IDC, 2011).
15. Kai Biermann. "Betrayed by Our Own Data." *Zeit Online*, March 26, 2011. http://www.zeit.de/digital/datenschutz/2011-03/data-protection-malte-spitz.
16. Erik Larkin. "Browser Fingerprinting Can ID You Without Cookies." *PCWorld*, January 29, 2010. http://www.pcworld.com/article/188161/ browser_fingerprinting_can_id_you_without_cookies.html.
17. *IBM InfoSphere Guardium: Managing the Entire Database Security and Compliance Cycle* (IBM, 2011). See also ftp://public.dhe.ibm.com/ common/ssi/ecm/en/imd14286caen/IMD14286CAEN.PDF.
18. *IBM InfoSphere Streams: Harnessing Data in Motion* (IBM, 2010).

19. Roger Rea. "IBM InfoSphere Streams: Redefining Real Time Analytical Processing" (IBM, 2010).
20. *2010 Information Governance Benchmark Report in Global 1000 Companies* (The Compliance, Governance and Oversight Council, 2010).
21. *IT Metrics: IT Spending and Staff Report* (Gartner, 2010).
22. http://legaltalkmedia.com/LTN/TDD/TDD_051412_BPOil.mp3.
23. *2010 Information Governance Benchmark Report in Global 1000 Companies* (The Compliance, Governance and Oversight Council, 2010).
24. Rob Karel and Clay Richardson with Connie Moore, Ralph Vitti, and Charles Coit. *Avoid Process Data Headaches: Align Business Process And Data Governance Initiatives* (Forrester, November 22, 2010).
25. Emanuele Della Valle and Alessio Carenini. *Supporting Environmental Information Systems and Services Realization with the Geospatial and Streaming Dimensions of the Semantic Web.* Workshop at EnviroInfo2010. http://ceur-ws.org/Vol-679/paper9.pdf.
26. Jane Zhang. "Why We Need 1,170 Codes for Angioplasty." *The Wall Street Journal*, November 11, 2008. http://online.wsj.com/article/SB122636897819516185.html.
27. "Procedural and Diagnosis Coding Must Be Linked By Medical Necessity." University of Florida, College of Medicine Office of Compliance, September 20, 1999. http://www.med.ufl.edu/complian/q&a/cpt-codes.html.

APPENDIX

REVIEWER AND CONTRIBUTOR PROFILES

AARTI BORKAR

Aarti Borkar is the program director of product management for IBM's market-leading data governance product suite. Aarti has more than 10 years of experience in defining product direction and executing on the strategy through various product management and development roles in information management. She holds a bachelor's degree in computer engineering from Bombay University, a master's in computer science from the University of Southern California, and a master's in technology commercialization from the University of Texas at Austin.

DAVID CORRIGAN

David Corrigan is the director of product marketing for IBM's InfoSphere portfolio. Prior to his current role, he led the product management and product marketing teams for MDM and has worked in the information management space for more than 12 years. David holds an MBA from York University's Schulich School of Business and an undergraduate degree from the University of Toronto.

Tony Curcio

For the past 18 years, Tony Curcio has been focused on helping organizations maximize the impact of their data integration investments. After spending several years working in the healthcare industry as a data integration technical specialist, he joined Ardent Software, the originators of DataStage—the ETL technology at the core of IBM InfoSphere Information Server. As part of the worldwide product management team, Tony continues to work closely with customers and partners to understand how they are delivering business value and leverages this input to drive IBM's product roadmap strategy.

Nick Dimtchev

Nick Dimtchev is the worldwide information governance center of excellence and client value sales leader within IBM Software Group. Nick's organization assists clients across multiple industries and geographies with defining their information governance strategy and maximizing the business value as a result of optimizing, securing, and leveraging information as an enterprise asset. Nick is passionate about driving innovation, establishing thought leadership in the marketplace, and helping clients succeed.

Michael Eggloff

Michael Eggloff has worked for IBM since 1995. During his IBM career, he has held national and international management positions in various business organizations, such as the systems and technology group, the software group, and global business services. In his current role (since 2010), Michael is a business value consultant within the Information Agenda® team for the financial services sector in Europe, where he consults with banks in the development and realization of information strategies. Michael holds master's degrees in computer science and business administration, and teaches business informatics at the University of Applied Sciences in Mainz Germany.

Stewart Hanna

Stewart Hanna has over 15 years of experience promoting enterprise data integration and data management at IBM and other organizations; six years of technical sales leadership in Australia and New Zealand; four years in worldwide product management; and three years in worldwide sales, focused on integration, governance, and big data. Stewart is currently focused on expanding

IBM's big data business. He specializes in helping clients who are looking for innovative techniques and technologies to address the challenges of big data and capitalize on the opportunities it presents.

Jim Hare

Jim Hare is program director of product marketing for IBM's big data portfolio and is responsible for positioning and thought leadership. Prior to joining IBM in 2008, he was vice president of product marketing and business development at Celequest, a California-based operational business intelligence vendor, which was acquired by Cognos in 2007. Jim has more than 18 years of experience in enterprise software and deep expertise in business intelligence, business process management, business activity monitoring, big data, and automated software testing and monitoring. He holds an MS degree in systems management from the University of Southern California and a BS in aerospace engineering from the University of Colorado at Boulder.

Eberhard Hechler

Eberhard Hechler is an executive architect from the IBM Boeblingen R&D lab. He is currently on a two-year assignment to IBM Singapore, working as the lead architect in the communications sector for IBM's growth market unit. After two years at the IBM Kingston development lab in New York, he held several positions in software development. His main expertise includes the areas of IT architecture, industry solutions, information integration, MDM, and information architecture. He is a member of the IBM Academy of Technology and a coauthor of the books *Enterprise Master Data Management* (Pearson, 2008) and *The Art of Enterprise Information Architecture* (Pearson, 2010).

Roger Hecker

Roger Hecker is a senior product manager for IBM InfoSphere, working with clients in every industry around the world to build successful enterprise business glossary and metadata management implementations. He is a visionary product designer and key collaborator with IBM thought leaders, business partners, and customers. Working in Israel, Roger guides a team of developers who are transforming IBM InfoSphere Metadata Workbench and Business Glossary into a comprehensive information governance platform.

Rani Hublou

Rani Hublou leads IBM's Information Lifecycle Governance go-to-market team. Working closely with IBM's largest customers, Rani provides high-impact solutions for defensible disposal and value-based retention and archiving. With 25 years of IT and business strategy experience, Rani has driven step change performance improvement in the world's largest financial services, manufacturing, healthcare, media, and non-profit enterprises. Her global enterprise solution experience was gained at firms such as Oracle, SAP, McKinsey & Co., and Accenture. Rani holds both a bachelor's and a master's degree in engineering from Stanford University.

Arild Kristensen

Arild Kristensen is a business value analyst in the Information Agenda team in IBM Software Group. He has 25 years of experience across the IT industry, including experience in sales, management, and consulting. Before joining the Information Agenda team, Arild spent five years as a business solutions manager in the IBM Stavanger Centre of Excellence for oil and gas. Arild helps clients understand the business value of treating information as a corporate asset.

Bob Leo

With more than 24 years of industry experience and leadership, Bob Leo is an information technology visionary who specializes in the development of business and information solutions using an integrated customer services and engineering approach to design and development. Bob believes in a customer-driven approach to information management that includes coupling industry best practices with customer needs, creating a process to build application, infrastructure, information management, and integration solutions. As a former healthcare CIO and current healthcare information strategy consultant for IBM, Bob has, in the past six years, advised over 100 companies in the Fortune 500 on using information as a strategic asset. He is a 2004 *Computerworld* Premier 100 IT Leaders honoree.

Frode Myren

Frode Myren is an IBM Distinguished Engineer. He is the chief technical strategist for IBM Software Group in the Nordics and a senior certified executive architect leading oil and gas solutions technical strategies. Frode holds a master

of science degree in technical physics from the Norwegian University of Science and Technology (NTNU/NTH) in Trondheim, Norway.

Wolfgang Nimfuehr

Wolfgang Nimfuehr is an Information Agenda executive consultant within IBM Software Group Europe, working with clients across multiple industries (primarily in banking and insurance) to gain insights from smarter analytics by leveraging big data and Watson technologies. Previously, Wolfgang held various leading sales, technical, and consulting roles. He is an information management visionary with over 20 years experience in data warehousing, information integration, master data management, business intelligence, and data governance. Wolfgang holds a master's degree in computer science and is a regular speaker at international conferences.

Deidre Paknad

Deidre Paknad is the director of Information Lifecycle Governance Solutions at IBM. She is widely credited with having launched the first commercial applications for legal holds, collections, and retention management, and is a recognized thought leader in legal and information governance, with numerous patents in the field. Deidre is a seasoned entrepreneur and executive with 25 years of experience applying technology to inefficient business processes to reduce cost and risk. She was inducted into the Smithsonian Institution for innovation in 1999, and again in 2000.

Roger Rea

Roger Rea is responsible for driving market share growth for IBM InfoSphere Streams, leading development, marketing, sales, finance, service, and support. Previously, Roger held a variety of sales, technical, educational, marketing, and management jobs at IBM, Skill Dynamics, and Tivoli Systems. Roger earned a bachelor's degree in mathematics and computer science, *cum laude*, from the University of California at Los Angeles (UCLA). He also received a master's degree in project management from George Washington University.

Jennifer Reed

Jennifer Reed is a senior product manager within MDM, responsible for the government industry at IBM. She has worked in the government industry and specifically the public safety sector for over 17 years. With her technical expertise and background, she has been involved in designing, developing, testing, and implementing complex solutions that solve the needs of government customers. In addition to her primary role, she is also responsible for overseeing the development of MDM into big data, including unstructured data correlation and entity resolution on Hadoop/BigInsights.

Will Reilly

Will Reilly is the industry marketing and solutions director for IBM's Information Agenda, an approach he co-developed in 2008 to help organizations take a strategic approach to applying information management and analytics technologies. Since then, over 2,000 IBM customers have embarked on Information Agenda transformations. Will has over 10 years of experience in information management with Ascential Software and IBM. He was previously a management consultant specializing in procurement and supply chain transformation.

Brian Roosevelt

Brian Roosevelt is a global data governance and security sales leader for IBM's Information Management brand and a significant contributor to published materials on governance. Brian is a 17-year veteran of the security business with positions in sales, architecture, and business development, including 9+ years with IBM. He has been an active participant in data governance since the formation of the IBM Data Governance Council in 2005. He lives in his native home town of Marshfield, Massachusetts, with his wife and two sons.

Randy Schnier

Randy Schnier has had a long and varied career with IBM, including stints in VLSI circuit design, CAD/CAM 3D design software development, and the architectural team at the core of WebSphere Application Server. He is currently part of the InfoSphere Streams development team and works on product development as well as real-world usage engagements with IBM clients.

Michael Schroeck

Michael Schroeck is a partner and global leader for IBM's information management services, where he leads a team of several thousand consultants in the planning, design, implementation, and governance of information and analytics solutions. Mike is considered an industry thought leader and visionary who has authored several books and articles on topics such as business intelligence, performance management, information architecture, and analytics. He has also been a featured speaker at many industry conferences and seminars around the world and is frequently quoted in leading business and technical publications. Michael was twice named as one of the world's top "25 Most Influential Consultants" by *Consulting Magazine* and was named a Distinguished Engineer by IBM for his outstanding and sustained technical achievements and leadership.

Brian Vile

Brian Vile is IBM's program director of product marketing for the Optim, Guardium, and information governance portfolio. He is responsible for defining IBM's worldwide MDM product strategy. Brian was previously general manager of InstallShield, responsible for product strategy and development. Prior to InstallShield, he was vice president of product management at Information Resources and held high-visibility roles at Hewlett-Packard and Motorola. Brian recently served on the board of directors of the Software and Information Industry Association's Software Division.

Susan Visser

Susan Visser is the publishing program manager for IBM information management products and solutions. She helps authors publish and promote their published retail books, eBooks, and custom books. Susan also blogs and tweets about these product areas at *https://www.ibm.com/developerworks/mydeveloperworks/blogs/SusanVisser* and *https://twitter.com/#!/susvis*.

Lawrence Weber

Lawrence Weber is the program director of product marketing for IBM's Data Warehousing portfolio. From technology sales and consulting through working with business partners to embed and market IBM's wide array of products in

their own solutions, Larry has a unique perspective on the business value that technology can bring to organizations across the globe. His experience spans business development, consulting, and a number of entrepreneurial ventures.

Kurt Wedgwood

Kurt Wedgwood is an IBM information strategy consultant with ten years of consulting experience at Accenture and twelve years with IBM. He has helped retail, consumer products, travel and transportation, and industrial products companies establish and further their governance programs. Kurt resides in Seattle, Washington, and holds finance and accounting degrees from the University of Colorado and an MBA from the University of Chicago.

Brian M. Williams

Brian Williams is an executive IT architect with IBM software services federal in Bethesda, Maryland. He has over 27 years of experience with information management, information security, and relational database design and development. Most recently, he has been designing, implementing, and managing big data solutions using IBM InfoSphere Streams. In addition, he has developed and taught training courses, workshops, and conference sessions for IBM InfoSphere Streams.

IBM InfoSphere:

A Platform for Big Data Governance and Process Data Governance

In *IBM InfoSphere: A Platform for Big Data Governance and Process Data Governance*, Sunil Soares provides a big data governance framework with three dimensions:

- *Big data 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.

Use the knowledge presented in this book to understand big data support across the IBM InfoSphere portfolio, including DataStage, Streams, QualityStage, MDM, Guardium, Optim, and Data Explorer. Understand the integration between IBM Business Process Manager Express and IBM InfoSphere in terms of integrating information governance, big data, and business process management.

SUNIL SOARES is the founder and managing partner of Information Asset, LLC, a consulting firm that specializes in helping organizations build out their information governance programs. Prior to this role, Sunil was the director of information governance at IBM, where he worked with clients across six continents and multiple industries.

Sunil is the author of *The IBM Data Governance Unified Process* (MC Press, 2010), *Selling Information Governance to the Business: Best Practices by Industry and Job Function* (MC Press, 2011), and *Big Data Governance: An Emerging Imperative* (MC Press, 2012).

IBM Doc # IMM14104 -USEN-00

ISBN 978-1-58347-382-5

MC PRESS

MC Press Online, LLC
3695 W. Quail Heights Court
Boise, ID 83703-3861

Price: $16.95 US/$18.95 CAN

EAN 9 781583 473825 51695 >