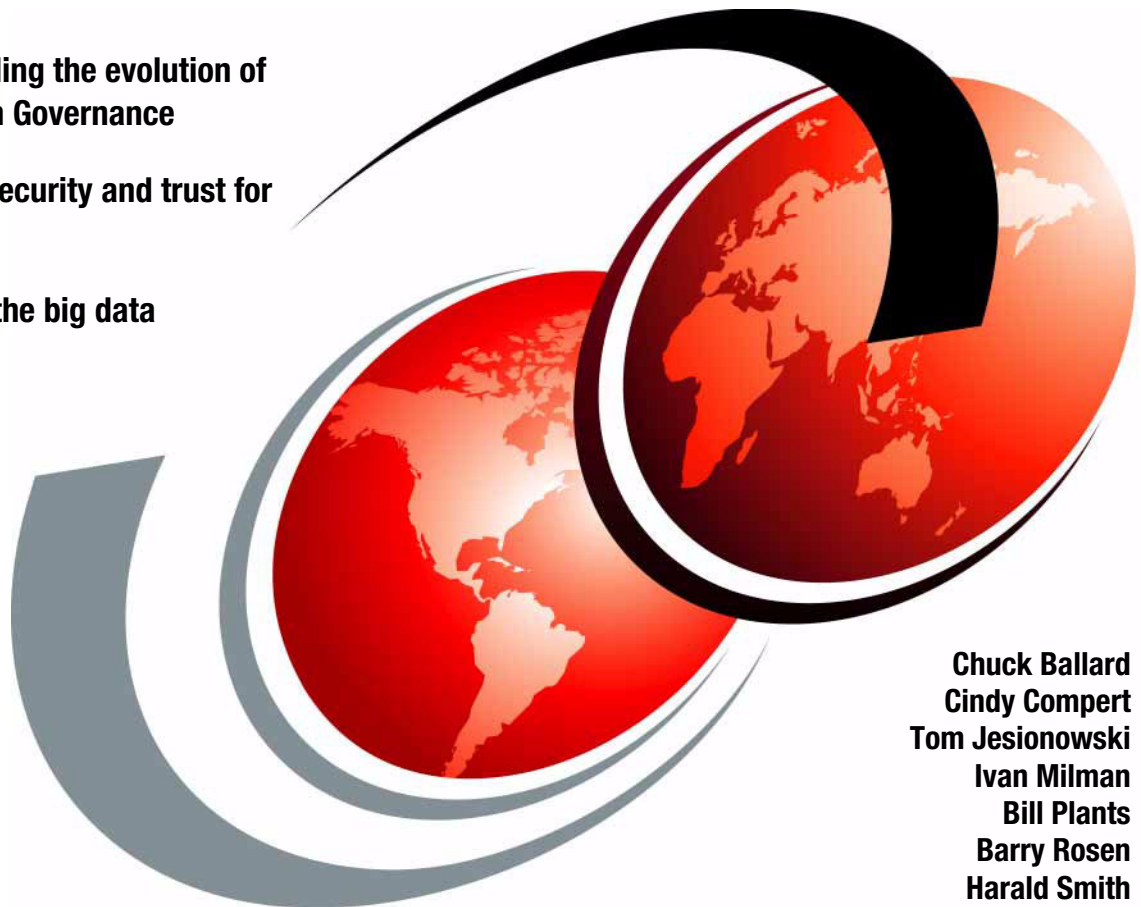


# Information Governance Principles and Practices for a Big Data Landscape

Understanding the evolution of  
Information Governance

Providing security and trust for  
big data

Governing the big data  
landscape



Chuck Ballard  
Cindy Compert  
Tom Jesionowski  
Ivan Milman  
Bill Plants  
Barry Rosen  
Harald Smith





International Technical Support Organization

**Information Governance Principles and Practices  
for a Big Data Landscape**

March 2014

**Note:** Before using this information and the product it supports, read the information in “Notices” on page vii.

**First Edition (March 2014)**

**© Copyright International Business Machines Corporation 2014. All rights reserved.**

Note to U.S. Government Users Restricted Rights -- Use, duplication or disclosure restricted by GSA ADP Schedule Contract with IBM Corp.

# Contents

<b>Notices</b> .....	vii
Trademarks .....	viii
<b>Preface</b> .....	ix
Authors .....	xi
Other contributors .....	xiv
Now you can become a published author, too! .....	xv
Comments welcome .....	xv
Stay connected to IBM Redbooks .....	xvi
<b>Chapter 1. Introducing big data</b> .....	1
1.1 What big data is .....	2
1.1.1 Origins of big data .....	2
1.2 Dimensions of big data .....	4
1.2.1 How big is big data: Volume .....	4
1.2.2 Variety .....	5
1.2.3 Velocity .....	6
1.2.4 Veracity: Can data be trusted .....	7
1.2.5 Value: The key driver .....	10
1.3 What big data looks like .....	11
1.3.1 Social media .....	11
1.3.2 Web logs .....	13
1.3.3 Machine-generated data .....	14
1.3.4 GPS and spatial data .....	15
1.3.5 Streaming data .....	16
1.4 Information Governance and big data .....	16
1.4.1 Metadata management .....	17
1.4.2 Security and privacy .....	17
1.4.3 Data integration and data quality .....	18
1.4.4 Master data management .....	19
<b>Chapter 2. Information Governance foundations for big data</b> .....	21
2.1 Evolving to Information Governance .....	22
2.2 IBM Information Governance Capability Maturity Model .....	22
2.2.1 Outcomes .....	23
2.2.2 Enablers .....	24
2.2.3 Core disciplines .....	26
2.2.4 Supporting disciplines .....	29

<b>Chapter 3. Big Data Information Governance principles</b>	33
3.1 Root principle for Big Data Information Governance	34
3.2 Leading from principle	36
3.2.1 Speed versus quality	36
3.3 Core principles for Big Data Information Governance	37
3.4 Practical application examples	41
<b>Chapter 4. Big data use cases</b>	43
4.1 Emerging big data use cases	44
4.2 Big data exploration	45
4.2.1 Identification of value	47
4.2.2 Enablement of Data Science teams	48
4.3 Enhanced 360° view of the customer	49
4.3.1 Expanding the range of customer-related data	51
4.3.2 Personalized Customer Engagements	52
4.3.3 Micro-market Campaign Management	52
4.3.4 Customer retention	53
4.3.5 Real-time demand forecasts	53
4.4 Security and Intelligence extensions	54
4.4.1 Enhancing traditional security through analytics	55
4.4.2 Network threat prediction and prevention	56
4.4.3 Enhanced surveillance insight	58
4.4.4 Crime prediction and protection	59
4.5 Operations analysis	59
4.5.1 Traffic management	62
4.5.2 Environmental monitoring and assessment	62
4.5.3 Predictive Maintenance	63
4.6 Data Warehouse modernization	64
4.6.1 Pre-processing hub	66
4.6.2 Queryable archive	67
4.6.3 Exploratory analysis	67
<b>Chapter 5. Big data reference architecture</b>	69
5.1 Traditional information landscape	70
5.2 The big data information landscape	75
5.2.1 Capabilities for the new information landscape	86
<b>Chapter 6. Introduction to the IBM Big Data Platform</b>	97
6.1 Components of the IBM Big Data Platform	101
6.2 The Data Warehouse	102
6.2.1 DB2	103
6.2.2 IBM PureData for Operational Analytics	103
6.2.3 IBM PureData System for Analytics	104
6.3 Stream computing	104

6.3.1 IBM InfoSphere Streams . . . . .	106
6.4 Apache Hadoop and related big data architectures . . . . .	107
6.4.1 IBM InfoSphere BigInsights . . . . .	108
6.5 Information Integration and Governance . . . . .	109
6.5.1 IBM InfoSphere Information Server . . . . .	110
6.5.2 IBM InfoSphere Data Replication . . . . .	113
6.5.3 IBM InfoSphere Federation Server . . . . .	115
6.5.4 IBM InfoSphere Master Data Management . . . . .	116
6.5.5 IBM InfoSphere Optim . . . . .	123
6.5.6 IBM InfoSphere Guardium . . . . .	124
6.6 Big Data Accelerators . . . . .	126
6.6.1 IBM Accelerators for Big Data . . . . .	127
6.6.2 IBM Industry Models . . . . .	128
6.7 Data visualization . . . . .	128
6.7.1 IBM InfoSphere Data Explorer . . . . .	129
6.8 The IBM Big Data Platform and the reference architecture . . . . .	130
<b>Chapter 7. Security and privacy . . . . .</b>	<b>133</b>
7.1 Why big data is different . . . . .	134
7.2 Information security defined . . . . .	135
7.2.1 Security framework . . . . .	135
7.3 Data privacy defined . . . . .	136
7.3.1 What sensitive data is . . . . .	137
7.3.2 Privacy operational structure . . . . .	137
7.4 How security and privacy intersect . . . . .	144
7.4.1 Implications and suggestions for big data . . . . .	144
7.4.2 Fit-for-purpose security and privacy . . . . .	145
7.5 Big data usage and adoption phases . . . . .	145
7.5.1 Exploration phase . . . . .	146
7.5.2 Prepare and Govern Phase (Assess and Protect) . . . . .	148
7.5.3 Inventorying and classifying sensitive data . . . . .	149
7.5.4 Consumption Phase (Sustain) . . . . .	191
7.6 Summary . . . . .	192
<b>Chapter 8. Information Quality and big data . . . . .</b>	<b>193</b>
8.1 Information quality and information governance . . . . .	194
8.2 Exploring big data content . . . . .	195
8.2.1 Knowing your data . . . . .	195
8.2.2 Call detail records . . . . .	196
8.2.3 Sensor data . . . . .	198
8.2.4 Machine data . . . . .	202
8.2.5 Social media data . . . . .	204
8.3 Understanding big data . . . . .	209

8.3.1 Big Data Exploration . . . . .	209
8.4 Standardizing, measuring, and monitoring quality in big data. . . . .	218
8.4.1 Fit for purpose. . . . .	219
8.4.2 Techniques for Information Quality Management . . . . .	220
8.4.3 Governance and trust in big data . . . . .	233
<b>Chapter 9. Enhanced 360° view of the customer . . . . .</b>	<b>235</b>
9.1 Master data management: An overview . . . . .	236
9.1.1 Getting a handle on enterprise master data . . . . .	237
9.1.2 InfoSphere Data Explorer and MDM . . . . .	242
9.2 Governing master data in a big data environment . . . . .	252
<b>Related publications . . . . .</b>	<b>255</b>
IBM Redbooks . . . . .	255
Other publications . . . . .	255
Online resources . . . . .	257
Help from IBM . . . . .	258



# Notices

This information was developed for products and services offered in the U.S.A.

IBM may not offer the products, services, or features discussed in this document in other countries. Consult your local IBM representative for information on the products and services currently available in your area. Any reference to an IBM product, program, or service is not intended to state or imply that only that IBM product, program, or service may be used. Any functionally equivalent product, program, or service that does not infringe any IBM intellectual property right may be used instead. However, it is the user's responsibility to evaluate and verify the operation of any non-IBM product, program, or service.

IBM may have patents or pending patent applications covering subject matter described in this document. The furnishing of this document does not grant you any license to these patents. You can send license inquiries, in writing, to:

*IBM Director of Licensing, IBM Corporation, North Castle Drive, Armonk, NY 10504-1785 U.S.A.*

**The following paragraph does not apply to the United Kingdom or any other country where such provisions are inconsistent with local law:** INTERNATIONAL BUSINESS MACHINES CORPORATION PROVIDES THIS PUBLICATION "AS IS" WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESS OR IMPLIED, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF NON-INFRINGEMENT, MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. Some states do not allow disclaimer of express or implied warranties in certain transactions; therefore, this statement may not apply to you.

This information could include technical inaccuracies or typographical errors. Changes are periodically made to the information herein; these changes will be incorporated in new editions of the publication. IBM may make improvements and/or changes in the product(s) and/or the program(s) described in this publication at any time without notice.

Any references in this information to non-IBM websites are provided for convenience only and do not in any manner serve as an endorsement of those websites. The materials at those websites are not part of the materials for this IBM product and use of those websites is at your own risk.

IBM may use or distribute any of the information you supply in any way it believes appropriate without incurring any obligation to you.

Any performance data contained herein was determined in a controlled environment. Therefore, the results obtained in other operating environments may vary significantly. Some measurements may have been made on development-level systems and there is no guarantee that these measurements will be the same on generally available systems. Furthermore, some measurements may have been estimated through extrapolation. Actual results may vary. Users of this document should verify the applicable data for their specific environment.

Information concerning non-IBM products was obtained from the suppliers of those products, their published announcements or other publicly available sources. IBM has not tested those products and cannot confirm the accuracy of performance, compatibility or any other claims related to non-IBM products. Questions on the capabilities of non-IBM products should be addressed to the suppliers of those products.

This information contains examples of data and reports used in daily business operations. To illustrate them as completely as possible, the examples include the names of individuals, companies, brands, and products. All of these names are fictitious and any similarity to the names and addresses used by an actual business enterprise is entirely coincidental.


## **COPYRIGHT LICENSE:**

This information contains sample application programs in source language, which illustrate programming techniques on various operating platforms. You may copy, modify, and distribute these sample programs in any form without payment to IBM, for the purposes of developing, using, marketing or distributing application programs conforming to the application programming interface for the operating platform for which the sample programs are written. These examples have not been thoroughly tested under all conditions. IBM, therefore, cannot guarantee or imply reliability, serviceability, or function of these programs. You may copy, modify, and distribute these sample programs in any form without payment to IBM for the purposes of developing, using, marketing, or distributing application programs conforming to IBM's application programming interfaces.

# Trademarks

IBM, the IBM logo, and ibm.com are trademarks or registered trademarks of International Business Machines Corporation in the United States, other countries, or both. These and other IBM trademarked terms are marked on their first occurrence in this information with the appropriate symbol (® or ™), indicating US registered or common law trademarks owned by IBM at the time this information was published. Such trademarks may also be registered or common law trademarks in other countries. A current list of IBM trademarks is available on the Web at <http://www.ibm.com/legal/copytrade.shtml>

The following terms are trademarks of the International Business Machines Corporation in the United States, other countries, or both:

BigInsights™	IMST™	S-TAP®
DataStage®	Informix®	Smarter Cities®
DB2®	InfoSphere®	System i®
developerWorks®	Optim™	System z®
Global Business Services®	PureData™	Tivoli®
Guardium®	pureScale®	WebSphere®
IBM PureData™	QualityStage®	z/OS®
IBM Watson™	Redbooks®	
IBM®	Redbooks (logo)  ®	

The following terms are trademarks of other companies:

Netezza, and N logo are trademarks or registered trademarks of IBM International Group B.V., an IBM Company.

Evolution, and Kenexa device are trademarks or registered trademarks of Kenexa, an IBM Company.

Linux is a trademark of Linus Torvalds in the United States, other countries, or both.

Microsoft, Windows, and the Windows logo are trademarks of Microsoft Corporation in the United States, other countries, or both.

Java, and all Java-based trademarks and logos are trademarks or registered trademarks of Oracle and/or its affiliates.

UNIX is a registered trademark of The Open Group in the United States and other countries.

Other company, product, or service names may be trademarks or service marks of others.

# Preface

The growth of data over the past five years has been staggering, yet it will pale beside the data volumes that are predicted for just the next five years alone. The variety of data continues to grow far beyond traditional databases and spreadsheets to encompass a huge array of sensor feeds, social media blogs and tweets, video recordings, call records, operational logs, and many more forms of data. This volume and variety of data comes at a faster and faster velocity. Volume, variety, and velocity are the original three “Vs” of big data, which is data that is expected to transform individual organizations and entire industries through new insights.

All of this big data underpins the analysis of multiple sources along multiple dimensions to find unexpected correlations. In turn, these new analytical insights change business models, whether supporting a next best action or an offer to a customer, rapidly identifying fraud, spotting urban crime patterns or system network threats, managing traffic and logistics, or predicting maintenance needs in advance to avoid downtime. Big data is seen by organizations to be a natural resource that drives competitive advantage and generates revenue and value.

At the same time, all this big data poses expanding risks to any organization. The speed and volume of arrival preclude human review of most content. The variety of data opens the possibility to find correlations that can be used for detrimental purposes, such as large-scale criminal theft of private customer information. The sources for big data are often outside the control of the organization and often complex enough to raise questions about the veracity of the information, such as the origin or bias of the data, or the overall quality of the data by itself or in combination with other data.

Information Governance is intended to help organizations address these risks and attain the wanted value of all data, whether it is traditional structured data or the emerging sources of unstructured big data. By using a framework that focuses on outcomes and core principles that support the organization's strategies, Information Governance maps out the people, processes, and disciplines, with reference architecture and technology, that are necessary to address these challenges.

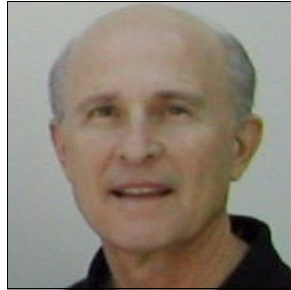
This IBM® Redbooks® publication assumes that the reader has a basic working knowledge of information strategy, architecture, and requirements. It provides an introduction to big data and Information Governance, including key principles for the latter. It delves further into the driving use cases for big data across industries, describes an emerging reference architecture for big data, and then describes the IBM Big Data Platform in that context. Most importantly, it explores the usage of the IBM Big Data Platform technology to support the core Information Governance disciplines for Information Security and Privacy, Information Quality, and Master Data Management.

This book describes and depicts the usage of many of the product modules and components of the IBM Big Data Platform, including the ones in the following list:

- ▶ IBM InfoSphere® BigInsights™
- ▶ IBM InfoSphere Data Explorer
- ▶ IBM InfoSphere Data Replication
- ▶ IBM InfoSphere Federation Server
- ▶ IBM InfoSphere Guardium®
- ▶ IBM InfoSphere Information Server
- ▶ IBM InfoSphere Master Data Management
- ▶ IBM InfoSphere Optim™
- ▶ IBM InfoSphere Streams
- ▶ IBM InfoSphere Warehouse
- ▶ IBM PureData™ System for Operational Analytics
- ▶ IBM Accelerators for Big Data
- ▶ IBM Industry Models
- ▶ IBM DB2®

## Authors

This book was produced by a team of specialists from around the world working with the International Technical Support Organization, in San Jose, CA.



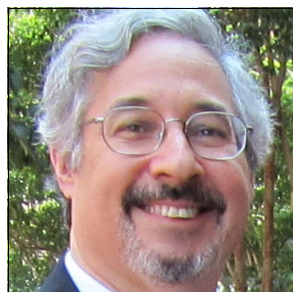
**Chuck Ballard** is a Project Manager at the International Technical Support organization, in San Jose, California. He has over 35 years experience, holding positions in the areas of Product Engineering, Sales, Marketing, Technical Support, and Management. His expertise is in the areas of database, data management, data warehousing, business intelligence, and process re-engineering. He has written extensively on these subjects, taught classes, and presented at conferences and seminars worldwide. Chuck has both a Bachelors degree and a Masters degree in Industrial Engineering from Purdue University.



**Cindy Compert** has 30 years of experience in IT, including previous positions in Technical Sales, Support, Channels, and Corporate Strategy. Cindy has been with IBM 12 years, having come to IBM with the Informix® acquisition. She advises a number of enterprise clients on Data Privacy and Data Security and is especially keen on facilitating synergies among IT, Security, and Compliance offices. She is an active member of the International Association of Privacy Professionals (IAPP) and blogs and tweets regularly about Governance and big data privacy topics. Cindy holds a CIPM Certification through the IAPP (Information Privacy Program Manager) and has a BA and MBA from New York University.



**Tom Jesionowski** is a Senior Managing Consultant at the IBM Global Business Services® division Center of Competency for Strategy and Analytics. A seasoned data professional with an engineering career spanning 38 years, his most recent 18 years were invested working within IT and business intelligence across numerous industries. An effective communicator, he has published trade journal articles on business intelligence and data governance. He co-invented a patented approach for data governance communication and collaboration. Tom is a former nuclear-trained US Navy submariner, and holds a Bachelor's degree in Computer Information Systems.



**Ivan Milman** is a Senior Technical Staff Member at IBM working as a security and governance architect for IBM Master Data Management (MDM) and InfoSphere product groups. Ivan co-authored the leading book on MDM, *Enterprise Master Data Management: SOA Approach to Managing Core Information*. He is working on IBM managed data offerings for the cloud. Over the course of his career, Ivan has worked on various distributed systems and security technology, including OS/2 Networking, DCE, IBM Global Sign-On, and IBM Tivoli® Access Manager. He has also represented IBM to standards bodies, including The Open Group and IETF. Before his current position, Ivan was the lead architect for the IBM Security Access Management Family family of security products. He is a member of the IBM Academy of Technology and the IBM Data Governance Council. Ivan is a Certified Information Systems Security Professional and a Master Inventor at IBM, and has been granted 14 US patents.



**Bill Plants** is a Guardium Technical Sales Engineer with the Guardium North America sales team. He has over 12 years experience as an UNIX Administrator and Database Administrator before working for IBM, with the last 11 years in an Information Management technical sales role. Bill joined IBM as part of the Informix acquisition in 2001, and has an Associate Sciences degree in Marine Technology and a Bachelor of Science degree in Computer Information Systems.



**Barry Rosen** is an IBM senior certified Global Executive Architect and Chief Field Advocate for the IBM Software Group Big Data Industry team with over 25 years of experience in Banking/Financial Services, Financial Markets, and related government entities. His responsibility is to drive customer success through the advancement of industry-leading big data architectures and deployment roadmaps in Data Quality, Data Integration, Data Warehousing, Data Governance, Metadata Management, MDM/RDM, BI, and Analytics. Barry has written extensively on these subjects, taught classes, and presented at conferences and seminars worldwide. He holds a Master's degree in Engineering Management and Computer Information Systems from Northeastern University and a Chemical Engineering degree from Worcester Polytechnic Institute.



**Harald Smith** is a Software Architect in IBM Software Group in Littleton, MA, specializing in information quality, integration, and governance products. He has over 30 years experience working in product and project management, software and application development, technical services, and business processes, and is IBM certified in delivering IBM Information Management solutions. Harald is the co-author of *Patterns of Information Management* by IBM Press, and is an IBM developerWorks® Contributing Author, for which he has written extensively, and has been issued four US patents. He has a Bachelor's degree in History and a Master's degree in Social Science from the University of Chicago.

## Other contributors

We want to thank others who contributed to this book, in the form of written content, subject expertise, and support.

### From IBM locations worldwide

- ▶ Tuvia Alon: Senior Technical Consultant and Solution Architect, IBM Software Group, Information Management, IBM Israel Software Labs, IL, US
- ▶ Anju Bansal: Manager - Big Data Accelerators, IBM Software Group, Information Management, Silicon Valley Lab, San Jose, CA, US
- ▶ Mandy Chessell: FEng, CEng, FBCS, Distinguished Engineer, Master Inventor, Chief Architect for InfoSphere Solutions, IBM Software Group, Information Management, Hursley Park, Winchester, UK
- ▶ Anshul Dawra: Software Architect, IBM Software Group, Information Management, Big Data Development, Silicon Valley Lab, San Jose, CA, US
- ▶ Vineet Goel: Sr. Product Manager, InfoSphere Optim for Big Data and Data Warehouse, IBM Software Group, Information Management, San Jose, CA, US
- ▶ Scott Kimbleton: Sr. Managing Consultant, Lead, Information Governance North America, Lead, Global Information Governance Center of Excellence, Jacksonville, FL, US
- ▶ Peter T Nguyen: IT Consultant, IBM Australia LTD



- ▶ Cindy Saracco: Solution Architect, Big Data Software Developer, IBM Software Group, Information Management, Silicon Valley Lab, San Jose, CA, US
- ▶ Sonali Surange: Software Architect and Developer, IBM Software Group, Information Management, Silicon Valley Lab, San Jose, CA, US
- ▶ Robert Uleman: Streams expert on the Worldwide Information Management Technical Sales Acceleration Team, IBM Software Group, which is based in California, US

#### **From the International Technical Support Organization**

- ▶ Mary Comianos: Publications Management, San Jose, CA
- ▶ Ann Lund: Residency Administration, Poughkeepsie, NY
- ▶ Wade Wallace: Editor, Austin, TX

## **Now you can become a published author, too!**

Here's an opportunity to spotlight your skills, grow your career, and become a published author—all at the same time! Join an ITSO residency project and help write a book in your area of expertise, while honing your experience using leading-edge technologies. Your efforts will help to increase product acceptance and customer satisfaction, as you expand your network of technical contacts and relationships. Residencies run from two to six weeks in length, and you can participate either in person or as a remote resident working from your home base.

Find out more about the residency program, browse the residency index, and apply online at:

[ibm.com/redbooks/residencies.html](http://ibm.com/redbooks/residencies.html)

## **Comments welcome**

Your comments are important to us!

We want our books to be as helpful as possible. Send us your comments about this book or other IBM Redbooks publications in one of the following ways:

- ▶ Use the online **Contact us** review Redbooks form found at:

[ibm.com/redbooks](http://ibm.com/redbooks)

- ▶ Send your comments in an email to:  
[redbooks@us.ibm.com](mailto:redbooks@us.ibm.com)
- ▶ Mail your comments to:  
IBM Corporation, International Technical Support Organization  
Dept. HYTD Mail Station P099  
2455 South Road  
Poughkeepsie, NY 12601-5400

## Stay connected to IBM Redbooks

- ▶ Find us on Facebook:  
<http://www.facebook.com/IBMRedbooks>
- ▶ Follow us on Twitter:  
<http://twitter.com/ibmredbooks>
- ▶ Look for us on LinkedIn:  
<http://www.linkedin.com/groups?home=&gid=2130806>
- ▶ Explore new Redbooks publications, residencies, and workshops with the IBM Redbooks weekly newsletter:  
<https://www.redbooks.ibm.com/Redbooks.nsf/subscribe?OpenForm>
- ▶ Stay current on recent Redbooks publications with RSS Feeds:  
<http://www.redbooks.ibm.com/rss.html>



# Introducing big data

Having a group of IT professionals agree on the definition for *big data* is like getting a group of surfers to agree on the perfect wave. Both are highly dependent on perception, skills, experience, and goals. Until early 2013, no surfer thought it was possible to surf a 90-foot wave. Yet with Jet Ski technology, which provided a new method of approaching a mega-wave, a vision of the possible, and a highly skilled, determined, fearless surfer, the art of the possible was achieved.

Just like the achievement of surfing the largest wave, big data can be the perfect computing mega-wave: untamed, potentially dangerous, exhilarating, yet rewarding in solving challenges previously thought impossible. Defining big data equally represents a challenge for the two following reasons:

1. The name is inaccurate; data does not have to be voluminous to be classified as big data.
2. There is a large perception/misconception of what it is, based largely on vendor hype, and sensationalist news stories.

This chapter defines big data, reveals its origins, articulates its value, and introduces Information Governance in the context of big data.

# 1.1 What big data is

This book uses a singular definition of big data. Big data is defined as “Extracting insight from an immense volume, variety, and velocity of data, in context, beyond what was previously possible.” (This is the IBM definition). Because data has become so voluminous, complex, and accelerated in nature, traditional computing methods no longer suffice. For example, how perform you do the following tasks:

- ▶ Analyze 500 million daily call detail records in real time to predict customer churn?
- ▶ Mine tweets to prevent crime by determining the location and time for a planned “bash-mob” (a flash mob that is dedicated to criminal activities) crime spree and deploying resources to that location?
- ▶ Protect national security by analyzing acoustical data to detect, classify, locate, and track potential threats and secure perimeters and border areas?

All of these are revolutionary applications that require revolutionary technology approaches.

big data represents the confluence of having new types of available information, new technical capability, and processing capacity, and the desire and belief that it can be used to solve problems that were previously impossible. At the same time, many of the concepts of big data are not new. Where did big data originate?

## 1.1.1 Origins of big data

Although it appears that big data originated in the last five years, the term itself was first used back in October 1997 in an IEEE paper on Visualization. You can find this paper at the following website:

[http://www.evl.uic.edu/cavern/rg/20040525\\_renambot/Viz/parallel\\_volviz/paging\\_outofcore\\_viz97.pdf](http://www.evl.uic.edu/cavern/rg/20040525_renambot/Viz/parallel_volviz/paging_outofcore_viz97.pdf)

In that paper, authors Michael Cox and David Ellsworth defined “the problem of big data” and how working with large volumes of data presented a challenge for visualization. The economics of digital processing also crossed the financial divide just a year before. According to IBM Systems Journal, in 1996, digital storage became more cost-effective than paper, enabling increasing analytic processing efficiencies.<sup>1</sup>

---

<sup>1</sup> *The Evolution of Storage Systems*, found at:  
<http://www.research.ibm.com/journal/sj/422/morris.pdf>

The origins of the three key dimensions of “big data” (volume, variety, and velocity) were first described by Gartner Group’s Doug Laney in 2001 in a research paper, where he conveyed the impact of e-commerce on data volumes, increased collaboration, and the desire to more effectively use information as a resource.<sup>2</sup>

Since that time, large-scale information processing continued to evolve without much fanfare until 2004, when Jeffrey Dean and Sanjay Ghemawat at Google published the ground-breaking paper that introduced the MapReduce programming construct for parallel processing on massive numbers of clustered commodity hardware (presented at the Sixth OSDI 2004 conference Symposium on Operating System Design and Implementation).<sup>3</sup> Although the concept of MapReduce functional programming was not new, what was revolutionary was the method of providing grid-level reliability and processing at massive scale without needing to know the details about how to distribute or process the data. Also revolutionary was the concept of not needing a schema to describe the underlying data, and that the schema could dynamically change without recoding, a definite shift in data design principles. Largely because many big data use cases require exploration and refinement, a schema-less design lends itself to nimble problem solving. It is no surprise that Google originated this paradigm in response to the need to create search indexes to support its search engine’s exponentially growing popularity.

Google’s version of MapReduce evolved one step further to a critical inflection point just one year later. In 2005, Doug Cutting and Mike Cafarella at Yahoo created Hadoop, an open source framework that included not just MapReduce, but also a highly available file system (HDFS) along with tools to support this new style of massive parallel processing. Not surprisingly, Hadoop evolved out of work that was done at Yahoo on another search engine project called Nutch.<sup>4</sup> Interestingly, it took a further six years until Hadoop and the little yellow elephant (named for Doug Cutting’s son’s stuffed toy elephant) moved into mainstream IT.<sup>5</sup> By then, both processing models and scale technology had evolved sufficiently to recognize that big data encompassed not just Hadoop, but also capabilities such as VLDB (very large databases) or analytic databases, and Stream (real-time) computing.

---

<sup>2</sup> <http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf>

<sup>3</sup> <http://research.google.com/archive/mapreduce.html>

<sup>4</sup> [http://en.wikipedia.org/wiki/Apache\\_Hadoop#cite\\_note-11](http://en.wikipedia.org/wiki/Apache_Hadoop#cite_note-11)

<sup>5</sup> [http://www.computerworld.com/s/article/358518/Big\\_data\\_goes\\_mainstream](http://www.computerworld.com/s/article/358518/Big_data_goes_mainstream)

In June 2013, big data took center stage in the public eye through news stories about US Government surveillance programs that used advanced analytics, and through big data's debut entry into the Oxford Dictionary along with 'tweet' (v) and 'stream' (v).<sup>6</sup> Indeed, big data now has been woven into the fabric of the public consciousness.

## 1.2 Dimensions of big data

big data is most frequently described according to the three dimensions or characteristics that Doug Laney first described: *volume*, *variety*, and *velocity*. More recently, experienced big data practitioners have come to appreciate that *veracity* is also a key element, that is, the "trust factor" behind the data. Although technologists do not always acknowledge the final "V" (*value*), gaining insight needs to translate into business value in order for any big data effort to be truly successful.

### 1.2.1 How big is big data: Volume

The first characteristic of big data, volume, is frequently cited as the most compelling and in many cases is thought of as "the definition." In other words, having a large volume of data to process constitutes big data. According to a widely quoted IBM CMO study,<sup>7</sup> each day, 2.5 quintillion bytes of data are created. In 1999, large data volumes for real-time analytics were touted as "very large" at 300 GB, a number that today could fit into a single system's memory.

Because the internet facilitated a massive data explosion, you can appreciate the massive volumes that major websites alone generate daily. Google indexes 20 billion pages *per day*.<sup>8</sup> Twitter has more than 500 million users and 400 million tweets per day, with a recent record peak rate of 143,199 tweets per second on 8/3/13.<sup>9</sup> Similarly, Facebook generates equally staggering statistics: 2.7 million 'Likes', 500 TB processed, and 300 million photos that are uploaded *per day*.<sup>10</sup>

---

<sup>6</sup> <http://public.oed.com/the-oed-today/recent-updates-to-the-oed/june-2013-update/a-heads-up-for-the-june-2013-oed-release/>

<sup>7</sup> <http://www-935.ibm.com/services/us/cmo/cmstudy2011/cmo-registration.html>

<sup>8</sup> <http://www.verticalmeasures.com/search-optimization/how-google-understands-you-infographic/>

<sup>9</sup> <http://expandedramblings.com/index.php/march-2013-by-the-numbers-a-few-amazing-twitter-stats/>

<sup>10</sup> <http://www.businessweek.com/articles/2012-08-23/facebooks-is-bigger-than-yours>

Clearly, these numbers are impressive. Size is indeed one important dimension that has radically influenced processing techniques such as the usage of MapReduce. However, big data is much more about analytics and deriving insight from information.

As an example, telecom providers have been processing massive volumes of information in the form of Call Detail Records (CDRs) for decades, long before the advent of big data. Typically, CDRs were used for billing purposes, using a monthly processing cycle. Outside of an audit or legal action, they were rarely needed. Since that time, deregulation and the explosion in mobile device usage have radically increased CDR volumes. Telecoms now typically process from 100 million to half a billion CDRs per day. Not only have CDR volumes increased, there is now a need to analyze CDRs in real time and near real time to uncover such things as the following items:

- ▶ Consumer usage patterns that can be used to drive marketing and retention programs and reduce churn.
- ▶ Revenue leakage because of billing errors.
- ▶ Fraud. Telecom fraud alone accounts now for \$40 billion in lost revenues worldwide.<sup>11</sup>

Providers also need to make real-time usage information available to consumers and track usage because billing has evolved from monthly to pay-as-you-go and usage-based costing. Using traditional processing techniques, it was previously impossible from a cost and time perspective to perform the type of analytics that was needed to provide these services. You can see why many telecom providers have become big data early adopters.

## 1.2.2 Variety

Each day, we now grapple with an astounding variety of information in many forms. Consider a typical day in the life of a digitally connected family: father, mother, and two children. The high school son might send a series of texts to his friends, with embedded photos and videos from a concert he attended the previous night. His middle-school age sister might log in to the school website to use her digital textbook to complete her homework, and then collaborate online to edit a book report. After work, Mom searches on the internet for a healthy dinner recipe while filling out her diet and fitness log in preparation for a marathon. Dad tweets his favorite team's football scores while making plans to attend a photography meet-up through a social media site.

---

<sup>11</sup> [http://www.heavyreading.com/document.asp?doc\\_id=223398](http://www.heavyreading.com/document.asp?doc_id=223398)

Using this example, you observe the usage of text, video, images, tweets, apps, online collaboration, and social media. Each one of these activities generates a large volume of information about click stream behavior, preferences, and interests, all which can be leveraged through big data analytics to derive new insights. For example, a Data Scientist might use text analytics from tweets to understand product sentiment for a new product introduction, and combine that with product sales to determine where to invest advertising dollars.

IBM estimates that unstructured data represents 90% of all the real-time data being created today (2011 CMO Study, IBM Institute of Business Value). It is easy to see through this example that unstructured information is estimated to be growing at 15 times the rate of structured information.<sup>12</sup> In addition, having new types of data pushes the boundaries of Information Governance as you combine data in new ways to reveal new patterns. Along with new patterns arise new risks, such as when analyzing smart meter information; there is an opportunity to also uncover personal habits, such as individuals who stay up late, and therefore compromise privacy.

### 1.2.3 Velocity

Webster's Dictionary defines velocity as "quickness of motion" and "rate of occurrence or action". When we describe the velocity of big data, we mean the rate of generated and consumed information, that is, the ability to stream or process information flowing at a rapid rate, frequently in real time. Smart electric meter data is a good example. With traditional electric meters, meter usage was read monthly by a meter-reader, an individual who walked the neighborhood and noted down the electrical usage for each home and business. With the advent of smart meters, not only has the process been automated, the meter read rate has accelerated to 15-minute intervals, 24 hours a day. The velocity of meter reads for a customer just went from once per month to four times per hour, for a total of 2880 reads per month. Assuming a large electric utility with 5 million customers, those reads quickly accumulate to 14.4 billion per month. Not only is the utility using this information for demand-based energy pricing, consumers are also accessing online portals to examine energy usage and find ways to economize, such as reducing consumption at peak periods.

---

<sup>12</sup> *Understanding Big Data - Analytics for Enterprise Class Hadoop and Streaming Data*, found at: <http://public.dhe.ibm.com/common/ssi/ecm/en/iml14296usen/IML14296USEN.PDF>



Velocity can also be construed as the difference between real time versus *rear-view mirror* analytics, which is analyzing what happened in the past. For this reason, integrating information from machine-generated data, such as sensors or devices, presents the biggest velocity challenge as real-time automation that drives predictive and preventive analytics also replaces traditional manual processes. Velocity is driving process change and generating more effective outcomes.

For example, hospital patient monitoring and care are undergoing a transformation from manual charting of vital signs on a periodic basis, to automating the collection and analysis of patient vital signs that are generated from electronic monitoring devices in real time. How does this level of velocity look? How is it used? Sample data streams could contain data about ECG, blood pressure, brain pressure, brain electrical activities, and so on. Electrocardiography (ECG) samples alone amount up to 1,000 readings a second to construct a waveform signal of heart function, for a total of 86.4 million readings a day per patient. For more information, go to the following website:

<http://lifesciences.ieee.org/images/pdf/06513228.pdf>

In a leading hospital ICU, big data analytics on streaming data is being used to predict rising brain pressure in patients with traumatic brain injuries. The goal is to provide a warning to physicians and nurses of pending changes in the patient's condition, allowing them to take preventive action to keep patients safe from rising brain pressure and further complications. Data velocity in this case could mean the difference between life and death. For a typical patient, collected real-time data monitoring adds up to 300 MB per day from monitors, up to 2.7 GB per day for EEG, and up to 700 alarms per day.<sup>13</sup> Analyzing real-time streams of vital signs that are collected continuously from the bedside monitor allows caregivers to spot subtle changes in the patient's pulse, blood and intracranial pressure, heart activity, and respiration. As a result, medical teams can proactively assess changes in the patient's condition to detect and react to complications, heal patients faster, and reduce readmittance rates. For more information about this topic, go to the following website:

<http://www-03.ibm.com/press/us/en/pressrelease/40624.wss>

### 1.2.4 Veracity: Can data be trusted

Veracity or “data trustworthiness” has the following three characteristics:

- ▶ The quality or cleanliness/consistency/accuracy of the data
- ▶ The provenance or source of the data, along with its lineage over time
- ▶ The intended usage because the usage can dramatically affect what is considered an acceptable level of trust or quality

---

<sup>13</sup> <http://www.ibmdatahub.com/blog/how-big-data-changing-patient-care>

The following questions highlight considerations for veracity:

- ▶ Where did the data come from?
- ▶ Did the data originate internally within the organization or externally?
- ▶ Is it publicly available data, such as phone numbers, or is it behavioral data from a data aggregator?
- ▶ Is the data from a transaction that can be audited and proven?
- ▶ Is the data truth or opinion?
- ▶ Is the data an intentional fabrication?
- ▶ Is the raw data usable as is, such as in the case of fraud detection, where identifying the aberrations are the focus, or does it require standardization and cleansing?
- ▶ What governance methods does an organization use to vet and rank veracity and classify its dimensions? As more data sources move from internal to acquired externally, this issue has become more pressing.
- ▶ How do you classify the trust factor? Organizations seriously must consider classification as part of the governance process.

For organizations that transact business externally online, the degree to which you trust an online user's identity, and how you use that trust for analytic purposes, is also a veracity and governance decision. Do guest users have a higher propensity to buy versus browse? Is the profile information supplied by a registered user reliable? Does it correlate with purchasing behavior?

Publicly available data is another area that warrants further examination. Public data does not necessarily translate into accurate and reliable data, especially when multiple sources are combined. For example, a man who died in 1979 is listed by a website that advertises itself as a "People Search Engine". The website used consolidated public data to determine that the man is 92 years old and living in Florida, although he never lived there and his wife had remarried and later was widowed again. He also gained an incorrect middle initial with his move, which just happens to be the initial of his wife. Had that website managed its master data effectively and cross-referenced it with the Social Security Death Index (SSDI), it would have concluded that he died in 1979.

The same holds true for social data, whether from social media or product or service reviews. Extracting sentiment through text analytics is a useful activity in terms of discerning reactions to, for example, products and events. In doing so, analysts must recognize there is a good likelihood that some or all of the postings or reviews could be false, either paid for by companies or individuals seeking to improve their ratings, or to disparage competitors. Gartner Group estimates that by 2014, false reviews constitute 10 - 15% of all reviews.<sup>14</sup> Case in point: A travel website with over 200 million unique visitors per month recently removed over 100 reviews that were created by a hotel chain executive. The executive created positive reviews of his properties while creating negative reviews of his rivals.<sup>15</sup>

Having more data also does not necessarily imply a higher level of trust or confidence. Recently, Aberdeen Group research identified the “Big Data Trust Paradox”, where the trust in data declined as the number of data sources increased. Specifically, 70% of organizations with fewer than 20 unique data sources had a high level of trust in their data versus only 43% for those organizations that had more than 20 sources. It is understandable that the more data sources, the more work is required to match, standardize, and master the data.<sup>16</sup>

Veracity is based on the data, algorithms, and assumptions. For example, in 2008, Google began tracking flu infection rates. Google Flu Trends<sup>17</sup> derives estimates of flu infection rates by using tweets and flu-related search terms. When the data is compared with CDC manual reporting rates, the numbers are shown to be accurate and providing delivery several days earlier, providing the necessary lead time in fighting a potential epidemic. During the 2013 season, those estimates spiked to nearly double the CDC reports. Experts asserted that increased news coverage resulted in increased searches by individuals who were not ill. For more information about this topic, see the following website:

<http://www.nature.com/news/when-google-got-flu-wrong-1.12413>

---

<sup>14</sup> <http://www.gartner.com/newsroom/id/2161315>

<sup>15</sup> <http://www.travelweekly.com/Travel-News/Hotel-News/Accor-suspends-exec-who-posted-fake-reviews/>

<sup>16</sup> *The Big Data Imperative: Why Information Governance Must Be Addressed Now*, found at: <http://public.dhe.ibm.com/common/ssi/ecm/en/im14352usen/IML14352USEN.PDF>

<sup>17</sup> <http://www.google.org/flutrends/>

## 1.2.5 Value: The key driver

Big data projects frequently turn out to be only experiments when the delivered value to the business is ignored. In the 2012 IBM Institute for Business Value Study “Analytics: The real-world use of big data - How innovative enterprises extract value from uncertain data”,<sup>18</sup> nearly half of the respondents identified customer-centric objectives as their organization’s top priority when they were asked to rank their top three objectives for big data. Companies clearly see big data as providing the ability to better understand and predict customer behaviors, and by doing so, improve the customer experience. Research has also shown that companies who invest in data as an asset for decision making are more successful. The following key data points illustrate this assertion.

In a recent Harvard study of public North American companies, organizations in the top third of their industry regarding the usage of data-driven decision making were, on average, 5% more profitable than their competitors.<sup>19</sup> Another study by IBM and MIT of 3,000 executives, managers, and analysts working across more than 30 industries and 100 countries highlighted even more compelling differences between high and low performing organizations. Top-performing organizations were found to use analytics five times greater than low-performing organizations. Additionally, organizations who strongly agreed that using analytics differentiated them within their industry were twice as likely to be top performers as lower performers.<sup>20</sup>

In the online world, retailers that use data analytics showed a revenue growth of 24% and an earnings growth of 22% over 10 years versus non-analytic users that had -1% and -15% growth.<sup>21</sup> In the area of social media and listening, outperforming enterprises are, on average, 1.7 times as likely to capture data at every customer interaction.<sup>22</sup>

In this context, value is any application of big data that drives revenue increases (such as customer loyalty analytics), identifies new revenue opportunities, improves quality and customer satisfaction (such as Predictive Maintenance), saves costs (such as fraud analytics), or drives better outcomes (for example, patient care). Certainly, it is critical that big data implementers gain experience using the technology in a ‘sandbox’ mode, where experimentation is encouraged. At the same time, identifying a test case and business sponsor ensures that early experimentation can lead to business success.

<sup>18</sup> <http://www-935.ibm.com/services/us/gbs/thoughtleadership/ibv-big-data-at-work.html>

<sup>19</sup> *Big Data: The management revolution*, found at:

<http://blogs.hbr.org/2012/09/big-datas-management-revolution/>

<sup>20</sup> *Analytics: The New Path to Value*, a joint MIT Sloan Management Review and IBM Institute, found at: <http://public.dhe.ibm.com/common/ssi/ecm/en/gbe03371usen/GBE03371USEN.PDF>

<sup>21</sup> “Big Data - Turning Bytes Into Insights”, by CIBC, December 12, 2012. To get a full copy of the report, send a request to <mailto:micons1t@us.ibm.com>.

<sup>22</sup> <http://public.dhe.ibm.com/common/ssi/ecm/en/yte03002usen/YTE03002USEN.PDF>

## 1.3 What big data looks like

Most data management professionals are comfortable with the concept of structured data, which is data that corresponds to a repeatable pattern, such as the familiar traditional database RDBMS with its row and column format. You can also relate to the more familiar semi-structured and unstructured formats, such as clear text, images, videos, audio, and XML. When you open the lens of information to encompass big data, you see many types of structure. Some examples are shown in the following sections.

### 1.3.1 Social media

Blogs, tweets, social networking sites (such as LinkedIn and Facebook), blogs, news feeds, discussion boards, and video sites all fall under this category. How one posts and views these elements on a web page versus the internal formatting and metadata that is provided for API access and processing are two different things. For example, tweets are posted as 140 character text strings. Reading a tweet through an API for purposes of text analytics means accessing a JavaScript Object Notation (JSON) formatted object. JSON is a language-independent lightweight data-interchange format that is based on a subset of the JavaScript Programming Language and is heavily used for exchanging data from tweets and blogs.

A tweet example in JSON format is shown in Example 1-1.

*Example 1-1 JSON format Tweet example*

---

```
{
  "coordinates": null,
  "created_at": "Fri Jan 17 16:02:46 +0000 2010",
  "favorited": false,
  "truncated": false,
  "id_str": "28039232140",
  "entities": {
    "urls": [
      {
        "expanded_url": null,
        "url":
        "http://www.research.ibm.com/cognitive-computing/machine-learning-applications/identify-theft-protection.shtml",
        "indices": [
          69,
          100
        ]
      }
    ]
  }
}
```

```

    },
    ],
    "hashtags": [

    ],
    "user_mentions": [
    {
        "name": "IBM, Inc.",
        "id_str": "16191875",
        "id": 11106875,
        "indices": [
        25,
        30
        ],
        "screen_name": "ibm"
    }
    ],
    },
    "in_reply_to_user_id_str": null,
    "text": "the future of identity protection from our leading
researchers",
    "contributors": null,
    "id": 28034030140,
    "retweet_count": null,
    "in_reply_to_status_id_str": null,
    "geo": null,
    "retweeted": false,
    "in_reply_to_user_id": null,
    "user": {
        "profile_sidebar_border_color": "CODEED",
        "name": "IBM, Inc.",
        .....

```

---

More information about the JSON API is available at the following website:

<https://developers.google.com/blogger/docs/2.0/json/using>

Tools such as IBM InfoSphere BigInsights hide the complexities of incorporating this type of data so that developers do not need to understand the complexities of writing to the respective APIs.

## 1.3.2 Web logs

Web logs exist in various semi-structured formats. Typically, they contain environmental information, and entries that document web application server activities from start to shutdown, that is, essentially everything that happens on your server. Logs contain transactional information about each connection, from the initial request to connection close, including any errors that might have occurred. Web logs traditionally are used for diagnosing errors and technical issues. More recently, they are being combined with other data sources to understand user website activity patterns, and to identify potential security breaches.

A sample of an IBM WebSphere® Application Server Community Edition server.log file is shown in Example 1-2.

*Example 1-2 WebSphere server.log file example*

---

```
2013-12-16 18:09:39,112 INFO [SinglePoolConnectionInterceptor]
Removing
ManagedConnectionInfo:org.apache.geronimo.connector.outbound.ManagedCon
nectionInfo@3e393e39. mc:
org.tranql.connector.jdbc.ManagedXAConnection@7660766] from pool
org.apache.geronimo.connector.outbound.SinglePoolConnectionInterceptor@
78c478c4
2013-12-16 18:09:41,336 INFO [KernelContextGBean] bound gbean
org.apache.geronimo.configs/system-database/2.1.8-wasce/car?J2EEApplica
tion=null,JCAConnectionFactory=SystemDatasource,JCAResource=org.apache.
geronimo.configs/system-database/2.1.8-wasce/car,ResourceAdapter=org.ap
ache.geronimo.configs/system-database/2.1.8-wasce/car,ResourceAdapterMo
dule=org.apache.geronimo.configs/system-database/2.1.8-wasce/car,j2eeTy
pe=JCAManagedConnectionFactory,name=SystemDatasource at name
org.apache.geronimo.configs/system-database/JCAManagedConnectionFactory
/SystemDatasource
2013-12-16 18:09:41,666 INFO [KernelContextGBean] bound gbean
org.apache.geronimo.configs/system-database/2.1.8-wasce/car?J2EEApplica
tion=null,JCAConnectionFactory=NoTxDatasource,JCAResource=org.apache.ge
ronimo.configs/system-database/2.1.8-wasce/car,ResourceAdapter=org.apac
he.geronimo.configs/system-database/2.1.8-wasce/car,ResourceAdapterModu
le=org.apache.geronimo.configs/system-database/2.1.8-wasce/car,j2eeType
=JCAManagedConnectionFactory,name=NoTxDatasource at name
org.apache.geronimo.configs/system-database/JCAManagedConnectionFactory
/NoTxDatasource
2013-12-16 18:09:44,788 INFO [SystemProperties] Setting
Property=openejb.vendor.config to Value=GERONIMO
```

```
2013-12-16 18:09:45,569 INFO [service] Creating
TransactionManager(id=Default Transaction Manager)
2013-12-16 18:09:45,616 INFO [service] Creating
SecurityService(id=Default Security Service)
2013-12-16 18:09:45,616 INFO [service] Creating
ProxyFactory(id=Default JDK 1.3 ProxyFactory)
2013-12-16 18:09:45,803 INFO [config] Configuring Service(id=Default
Stateless Container, type=Container, provider-id=Default Stateless
Container)
2013-12-16 18:09:45,803 INFO [service] Creating Container(id=Default
Stateless Container)
2013-12-16 18:09:45,913 INFO [config] Configuring Service(id=Default
Stateful Container, type=Container, provider-id=Default Stateful
Container)
2013-12-16 18:09:45,913 INFO [service] Creating Container(id=Default
Stateful Container)
2013-12-16 18:09:45,975 INFO [OpenEJB] Using directory
C:\IBM\InfoSphere\Optim\shared\WebSphere\AppServerCommunityEdition\var\
temp for stateful session passivation
2013-12-16 18:09:45,975 INFO [config] Configuring Service(id=Default
BMP Container, type=Container, provider-id=Default BMP Container)
2013-12-16 18:09:57,991 INFO [SinglePoolConnectionInterceptor]
Removing ManagedConnectionInfo:
org.apache.geronimo.connector.outbound.ManagedConnectionInfo@6cac6cac.
mc: org.apache.activemq.ra.ActiveMQManagedConnection@38193819] from
pool
org.apache.geronimo.connector.outbound.SinglePoolConnectionInterceptor@
7ef07ef
```

---

### 1.3.3 Machine-generated data

Machine-generated data constitutes a wide variety of devices, from RFIDs to sensors, such as optical, acoustic, seismic, thermal, chemical, scientific, and medical devices, and even the weather. As an example, the US National Weather Service publishes weather data on its public website in various formats. Although we can view the current weather for Islip, NY in an easy-to-understand format and understand that on September 19, 2013 at 5:56 p.m. that it is 68 degrees Fahrenheit with a few clouds, our big data application may ingest this reading as a string of raw data in METAR (weather standard) format.

Other examples of machine-generated data include road sensors, TV set top boxes, fitness sensors, and video cameras.



### 1.3.4 GPS and spatial data

Geospatial data has become ubiquitous. From the GPS systems in our cars, in planes, and ships, to GPS apps on smartphones, we use GPS to guide our movements. In turn, GPS is used to track our movements, such as emergency beacons, and retailers who use in-store WiFi networks to access shoppers' smartphones and track their shopping habits.

Another use for geospatial information in the big data world is Location Based Services (LBS), which is the ability to deliver services or perform tailored advertising that is based on the location of moving objects such as cars or people with mobile phones. For example, a coffee shop with three branches in a city might want to know which cars are close to each of its stores at this moment. To solve this problem, you assess the vehicle location data and filter the locations by distance to a set of given locations. The data that describes the location of the coffee shops might look like the data that is shown in Example 1-3.

#### *Example 1-3 Sample location data*

---

Four fields: location ID, latitude, longitude, radius

File Contents:

```
loc1,37.786216,-122.409074,500
```

```
loc2,37.791134,-122.398774,250
```

```
loc3,37.787776,-122.40122,300
```

The code snippet to perform the calculation is as follows (using the InfoSphere Streams Geospatial Toolkit):

```
type LocationType = rstring id, rstring time, PointT location, float64  
speed, float64 heading ;
```

```
type FixedLocType = rstring id, PointT center, float64 radius ;
```

```
composite GeoFilter
```

```
{
```

```
    graph
```

```
    ...
```

```
    stream<LocationType, tuple<rstring fixedLocId, float64 distance>>
```

```
NearbyVehicles
```

```
    = Join(VehicleLocations as V ; FixedLocations
```

```
as F)
```

```
{
```

```
    window
```

```
        V: sliding, count(0) ;
```

```
        F: sliding, count(3) ;
```

```
    param
```

```
        match: distance(V.location, F.center) < F.radius;
```

```
    output NearbyVehicles;
```

```
        fixedLocId = F.id,  
        distance   = distance(V.location, F.center);  
    }  
    ...  
}
```

---

### 1.3.5 Streaming data

Stream data is a special category of big data. Rather than being a format, it is a processing style. Streaming data is taking virtually any kind of data and making it available in a continuous near real-time feed. The advantage of this approach is analytics at the “speed of thought.” Some examples where streaming data is useful include fraud detection, security, such as detecting sounds from unknown vehicles near a secure site to prevent an incident, traffic monitoring, connected car, and medical monitoring. Frequently, streaming applications use messaging protocols to send and receive information. Streams receive messages, process them, then might send out response message or send data elsewhere. The geospatial processing example in 1.3.4, “GPS and spatial data” on page 15 is also an example of using streaming data to detect vehicle locations within a given radius.

## 1.4 Information Governance and big data

Information Governance is the glue that drives value and mitigates risk. There are several key areas where Information Governance for big data is critical, such as metadata management, security and privacy, data integration and data quality, and master data management. It is interesting to note that big data innovators recognize the importance of governance to the success of their projects. According to a recent study, 58% of the organizations who report having active big data efforts included security and governance processes in their efforts.<sup>23</sup> This book illustrates how IBM information governance solutions can rapidly improve confidence in your big data and accelerate the value gained while reducing risk and driving innovation.

---

<sup>23</sup> *Analytics: The real-world use of big data - How innovative enterprises extract value from uncertain data*, found at:  
<http://www-935.ibm.com/services/us/gbs/thoughtleadership/ibv-big-data-at-work.html>

## 1.4.1 Metadata management

Although there is a virtual avalanche of available information, less than 0.5% of it is used in analytics. IDC recently estimated that 23% of content could be useful if tagged.<sup>24</sup> Providing context around content through tagging, using the wealth of available metadata, and implementing governance processes can rapidly accelerate progress in this area.

## 1.4.2 Security and privacy

Security and maintaining privacy is a growing concern for the implementation of big data projects. The proportion of data in the digital universe that requires protection is growing faster than the digital universe itself, from less than a third in 2010 to an estimate of more than 40% in 2020.<sup>25</sup> There are many considerations about how big data impacts security and privacy, such as legal, competitive, moral and ethical, safety, and political reasons. In many of these areas, the implications are only beginning to emerge. Legally, organizations must comply with security and privacy laws. However, there are few if any regulations governing the protection and disclosure of new data types, such as geolocation. In addition, there are many inconsistent regulations between countries and even states that must be considered.

Combining data sources can also cause more, unanticipated exposure, such as combining IP addresses with demographic information, or in medical research to uncover disease causation, where an individual's identity could be accidentally revealed. Competitively, organizations who secure sensitive data and protect the privacy of their customers through disclosure and opt-out choices will excel in the marketplace. Witness recent high-profile privacy breaches that greatly impacted an organization's reputation, especially in banking and healthcare.

Morally and ethically, big data exposes unanticipated issues, and decisions that must be made about the type of analysis and decision-making it provides. Is it moral or ethical for a company to do the following tasks:

- ▶ Make credit decisions that are based on your Facebook friends?
- ▶ Share medical data that also exposes information about your family's disease risk?
- ▶ Send coupons for baby products that are based on buying behavior to a minor child who was pregnant although her family did not know it?

<sup>24</sup> *Digital Universe study*, found at:

[https://www-950.ibm.com/events/ww/grp/grp037.nsf/vLookupPDFs/ED\\_Accel\\_Analytics\\_Big\\_Data\\_02-20-2013%20v2/\\$file/ED\\_Accel\\_Analytics\\_Big\\_Data\\_02-20-2013%20v2.pdf](https://www-950.ibm.com/events/ww/grp/grp037.nsf/vLookupPDFs/ED_Accel_Analytics_Big_Data_02-20-2013%20v2/$file/ED_Accel_Analytics_Big_Data_02-20-2013%20v2.pdf)

<sup>25</sup> *The Digital Universe in 2020: Big Data, Bigger Digital Shadows, and Biggest Growth in the Far East*, found at: <http://idcdocserv.com/1414>

- ▶ Monitor your personal Facebook account as a method of employee screening?
- ▶ Charge users who log in with Apple devices for travel services more because research showed that they are willing to pay more for travel?

Machine data is also a source of concern when it exposes information about individual behavior and habits. For example, smart electrical meter data can reveal household activity patterns. Again, policy and legislation has not kept pace with individual privacy rights.

Safety is increasingly an issue regarding stalking and fraud. Geolocation and GPS devices in particular can reveal the locations of individuals who are at risk for violent crimes, such as domestic violence or potential terrorism targets.

Lastly, there are political implications for privacy and security as they relate to international relations and government policy, such as recently publicized NSA activities that conflict with privacy laws abroad.

There is a growing recognition that governance is a key aspect of managing and getting value out of big data. According to Gartner, by 2018, 25% of progressive organizations will manage all their unstructured data by using information governance. That number stands at less than 1% today.<sup>26</sup>

### 1.4.3 Data integration and data quality

Integration and ensuring that data is trusted are key enablers of successful big data projects. Quality is relative to the requirement, that is, drawing from governance principals to provide data that is “fit to purpose.” In the case of customer or product information, it is critical to standardize and identify the lineage of all data elements to ensure consistency and correctness and to monitor and maintain your data quality. As with traditional data, establishing data quality metrics that are aligned with business objectives enables you to quickly uncover data quality issues and establish remediation plans.

---

<sup>26</sup> <http://searchcompliance.techtarget.com/tip/Big-data-era-requires-new-approach-to-information-governance-strategy>

### 1.4.4 Master data management

Master data is the high-value, core information that is used to support critical business processes across the enterprise (information about customers, suppliers, partners, products, materials, employees, accounts, and so on), and is at the heart of every business transaction, application, and decision. Master data management (MDM) can provide a compelling starting point for big data analysis. MDM, by definition, focuses on the highest value entities within an organization. Many organizations imagine that they want to analyze social media to determine customer sentiment, but do they know who their customers are? Do they know their best customers?

Integration points between MDM and big data include ingesting and analyzing unstructured data, creating master data entities, loading new profile information into the MDM system, sharing master data records or entities with the IBM Big Data Platform as the basis for big data analysis, and reusing the MDM matching capabilities in the IBM Big Data Platform, such as customer matching. Together, big data and MDM can help extract insight from the increasing volume, velocity, variety, and decreasing veracity of data in context beyond what was previously possible.





# Information Governance foundations for big data

Welcome to the evolution!

In the course of doing your daily job, you are continuously enmeshed in the now. You struggle with the tactical details and issues that consume the work day, continuously focused on some small part of a much bigger machine that keeps getting bigger and more challenging to manage. You deal in the vast collection of micro-tasks that make up part of a much bigger workflow.

This chapter is going to pull you away from that setting and have you look at governance from a much broader perspective. Broaden your viewpoint and look at how all the moving parts interact in the era of big data!

## 2.1 Evolving to Information Governance

During a kick-off meeting for a governance program, a CFO got up to address a room full of IT and business professional. He stated that if we designed our systems today from scratch, they would look nothing like the environment we own. He went on to elaborate that we arrived here by layering thousands of good and valid decisions on top of one another. Our systems evolved.

Big data and Big Data Information Governance also evolved out of the good work that was done by those who preceded us. These items evolve into something that only a few of us can envision today. Along the way, we evolved the technology and the way we manage our daily tasks working with data. What started as good engineering practices for mainframes became data management.

Then, with new technological advances, we encountered new problems, introduced new tasks and disciplines, and created Information Governance in the process. We were standing on the shoulders of data management, armed with new solutions to new problems. Now we face the four Vs of big data and each of those new data system characteristics have introduced a new set of challenges driving the need for Big Data Information Governance as a response to changing velocity, volume, veracity, and variety.

## 2.2 IBM Information Governance Capability Maturity Model

So, do you need another framework for the new challenges? The answer to that question depends upon how comprehensive the last framework was and how well it scales to address the new challenges. First, there are several frameworks out in the marketplace to choose from, but this IBM Redbooks publication relies upon the IBM Information Governance Capability Maturity Model or IBM Data Governance Capability Maturity Model (CMM), as shown in Figure 2-1 on page 23. The model is described in greater detail in other publications from IBM and that detail is not revisited here. This chapter describes the impact of big data on that model and builds upon that prior work.



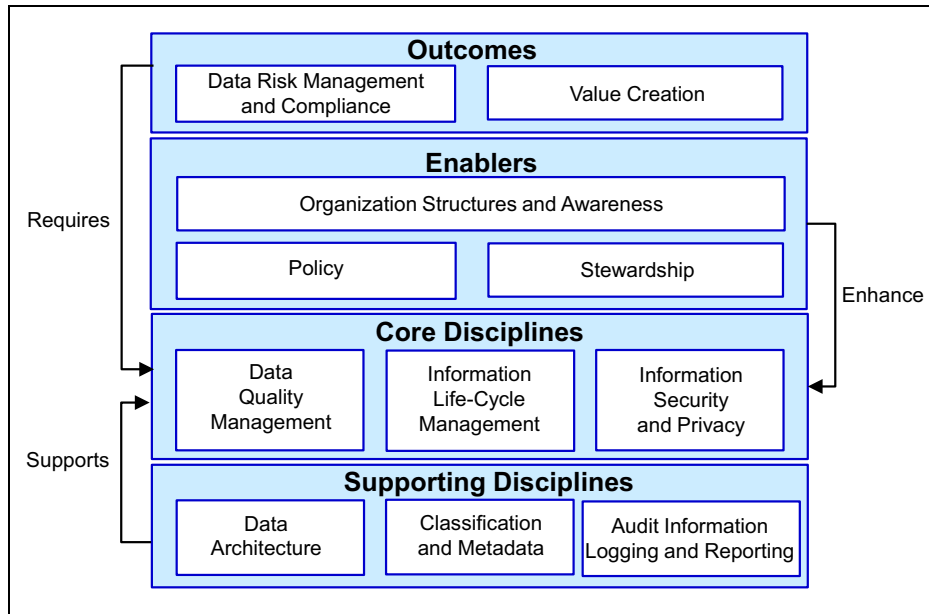


Figure 2-1 IBM Information Governance Capability Maturity Model

## 2.2.1 Outcomes

In the IBM Data Governance CMM, the *outcomes* are used to establish why governance activities are needed. The outcomes provide the drivers for the data governance efforts. The key drivers of risk management, compliance, and value creation continue to be influencers for Big Data Information Governance. What has changed is the scale of their impact. Consider that as the value or risks that are associated with data scale upward with big data, so do the effect of the drivers. big data means bigger risks and bigger rewards depending upon the effectiveness of governance programs.

### Value creation and risk management

Many of the use cases that are presented later in this book exist to fulfill the value creation domain of the CMM. For that reason, value creation is not explored now, but risk management is covered. Perhaps the biggest risk that is associated with any advancement is to *not* consider the new risk exposures that occur with the adoption of the new technology. The remainder of this chapter covers the activities that are needed to both create value while concurrently mitigating the new risks that are presented with the advent of big data.

## 2.2.2 Enablers

*Enablers* are increasingly relevant to big data primarily because of the effect of the scaling risk and reward factors. The increased importance of the outcomes adds to the imperative to define a managed structure for the enablers. Fortunately, regarding big data, the makeup of policy, organization, and stewardship are largely unchanged. Policies that are based upon sound guiding principles are still an essential component of the program. The content of the policy requires updates to incorporate the effects of increased volume, velocity, variety, and veracity.

### Policy impacts

As an example, consider how a policy might evolve because of the increased velocity of data. In the era of 24-hour batch processing cycles, the time-to-action (TTA) starting with the acquisition of information was measured in hours or even days. This time frame was manageable for humans to react to, so few policies were written that consider TTA. As the velocity of data increases, and more real-time systems emerge, the TTA action is often reduced to milliseconds, well beyond the ability for a human to react.

As a result of this fundamental shift, the increased velocity of data necessitates an addition to policies. A policy defining how velocity must be managed and accommodate must be introduced. For example, for each system making an automated response to data flowing into the environment, a policy is needed that requires controls to prevent runaway processes and limits the operational risk that is posed by the automated decision process. The policy requires the control without specifying how the control can be achieved.

A practical example of this type of control resulted from flash crashes that were observed in various markets. The crashes resulted from automated selling that drove prices lower, which triggered automated trades driving prices even lower until the cycle concluded at the bottom. The automated decision and market feedback set a downward spiral in motion. Today, these crashes are inhibited by automated responses to either slow trading or stop all trading outright. Those controls are now a matter of policy.

Likewise, ballooning volumes require revisiting policy statements for Information Lifecycle Management. Various data might require the expansion of the policy for data classification and quality to accommodate data types beyond structured data. The value of retaining the data longer must be weighed against the risk that the data might pose to electronic discovery during litigation.

Veracity, or the truthfulness of the data, impacts the policies and standards surrounding the quality of data that is acquired from sources outside your direct control. This data includes third-party feeds, and social media. Data consumers that are unable to validate the veracity of information from these sources can inadvertently represent them as factual when the ability to verify that the data source does not exist. This situation can create an operational risk for data that is used for internal decisions, or a legal risk if the data is released to the public space and presented as factual.

You have seen this in the news, where organizations have encountered embarrassing press when names on mailing addresses were replaced with inappropriate content harvested from big data sources. The poor veracity was not caught because of gaps in the process controls for the data. Although the incidents might have been fleeting, they serve as the canary in the coal mine of potentially more disastrous results ahead. The question remains, how can veracity be governed to reduce the risks?

## **Organizational impacts**

By the time any organization is making the foray into the big data arena, they should have a data governance program; if not, then they should definitely start! The added complexities being introduced to the enterprise requires a higher degree of coordination. As described in “Policy impacts” on page 24, the leadership of an organization must have a means to retain control over value creation and risk management, which is how the organization component is brought to bear.

No matter the organization format that is in play, the organization’s leadership must retain the means of directing the actions that they deem critical to the business. The organizational layout should be articulated through a policy to avoid ambiguity in the conduct of the core and supporting disciplines. This situation has not changed with big data, but has become a stronger imperative because of the speed of information flowing into the corporate decision engine and the faster pace of actions that are tied to the information.

## **Data stewardship impacts**

Data stewards, already strained by the mounting volumes of structured data, are further taxed by the burden of new unstructured data, the volumes of data, and the demand for changes to systems in response to the insight that is gained. Then, add the challenge of getting the business to acknowledge their ongoing responsibility for the data. Too often, the business forgoes quality, security, and analytics support by deferring to the deployment of a new business capability. This self-destructive behavior, when left unchecked by policy and supporting organizational controls, exposes the enterprise to greater operational risks and erodes the value of the data.

Data stewardship in the era of big data requires a clear understanding of obligations and decision rights. Earlier, the need for feedback rules to prevent runaway processes was put forward. The organization must determine who within the organization decides when and how the feedback rules are applied. This is an illustration of the changes that are needed for the various stewardship roles. Each organization must examine for themselves how the stewardship roles evolve based on the manner in which velocity, veracity, volume, and variety has changed the information landscape. The roles cannot, and must not, remain unchanged from their current state.

### **2.2.3 Core disciplines**

The core disciplines of data governance cover data quality management, Information Lifecycle Management, and information security and privacy. The need for these capabilities has not diminished in the era of big data. Quite the contrary, the disciplines become more challenging to administer because of the increased complexity of the data landscape, and volume, veracity, and variety each play an increasingly significant role in the direction the core disciplines take over time.

#### **Information security and privacy**

Perhaps the most frequently encountered situation that is encountered during big data projects is information privacy. Who owns the data? Although many businesses might feel that they have addressed information privacy through their existing practices, those policies and practices might require revisiting to ensure that they remain current with the changing technology.

As increasing volumes of social media, which might have been published publicly, are harvested for marketing analytics, answering the question of who owns the data and managing the data becomes trickier. This situation complicates the data security landscape and data classification tasks. For example, how do you classify comments on Twitter feeds? These comments are put into the public sphere with little expectation of privacy, so it is not unreasonable to assign that big data streaming data as public data in your data classification scheme. Compare that to a comments page on the company website, or on a corporate Facebook page, where distribution might entail a smaller audience or a semi-public expectation of privacy. At this stage of the assessment, it is time to invite the lawyers into the governance policy discussion. The outcome of their decision trickles down into the data.

There is increasing reliance upon third parties to oversee the custody of data in clouds, or as services. In these cases, the data physically is outside the direct control of the business and is governed through various contracts and service agreements. These agreements must evolve from the templates they were built upon. They must address and incorporate the policies and rules that are needed to protect the data as though it is still in-house. Simply put, custody over the data might have transferred outside of the direct control of an organization, but responsibility over the content of the data remains within the organization.

Begin a dialog between the legal department, IT, and the business to refine the language of these contracts. The new requirements are used to modify future contracts, and are infused into existing contracts as they come up for renewal. This review and update process must be continuous to evolve the contracts as regulations, threats, and technology transforms over time.

## **Information Lifecycle Management**

The classic view of Information Lifecycle Management (ILM) is that it is a method for managing the process of archiving data and managing backups. Over time, the discipline evolved and adopted many of the long-held practices from records management. One of those practices, Defensible Disposal, can run against the big volumes aspect of big data.

Part of the promise of big data is that with reduced constraints on storage volumes, more data can be kept and mined for additional value and insight. That value proposition must always be balanced by the risk that is posed by possessing and retaining the data. Under the premise of Defensible Disposal, data that is no longer of value and not subject to retention because of regulatory controls or legal holds should be destroyed. Because there is new value in the increasing volumes, it can be argued that the data should be retained longer until the value is overshadowed by the risk. This simple shift in balance changes the dynamics driving Defensible Disposal discussions.

Unfortunately, there is no simple answer as to how to address this shift in data retention practices. Each organization must determine how to manage this dynamic based upon their unique business model and strategic objectives. Retention rules written a decade ago might no longer reflect the reality of today. The rules might have attempted to bind the growth of data stores to manageable levels and that logic might no longer hold true. Do you keep the rules as written, or extend the time frames to account for the new capacity levels and the newly discovered value in the historical data? Who makes that decision?

Legal departments often weigh in heavily on these discussions. Often, they steer an organization to keep what is required by regulation and law, and then only for the least amount of time possible. Meanwhile, the business wants to keep access to the rich volumes of data they have diligently captured while interacting with their customers. The insight into customer actions and behaviors can be derived and used to create a competitive advantage.

One example of a compromise between these camps balances the value with the risk of data being retained, which might be achieved by obfuscating the historical data after a set period. This enables the immediate and long-term business needs to be met while protecting the organization from the risk that is associated with keeping too much data. The one exposure that does remain is the possibility that the obfuscated data might be combined with other data to reverse the obfuscation process.

**Data quality management**

Data quality is always a factor, but the task of managing and reporting on quality is harder than before. This situation is particularly true with regard to unstructured data and data from sources outside the direct control of the organization. There is an increased dependency upon data from public sources that might not have the same rigorous validation and verification as internally managed processes. As a result, big data introduces a higher degree of variance and uncertainty.

To assess the impact on your environment, use the matrix that is shown in Table 2-1 as a guide. The eight measures of quality<sup>1</sup> in the left column are impacted by the changes to the four Vs of big data.

*Table 2-1 The Four Vs of big data*

Quality characteristics	Volume	Veracity	Variety	Velocity
Completeness	?	?	?	?
Auditability (Provenance and Lineage)	?	?	?	?
Timeliness	?	?	?	?
Precision	?	?	?	?
Accuracy	?	?	?	?
Relevance	?	?	?	?

<sup>1</sup> *Improving Data Warehouse and Business Information Quality* by English, and *Journey to Data Quality* by Lee, et al.

Quality characteristics	Volume	Veracity	Variety	Velocity
Integrity	?	?	?	?
Validity	?	?	?	?

You and your team should investigate the impact the change each dimension of big data has on the quality measures, and how your business reacts to the changes and triage changes in your quality measures. This impact is of critical importance to data that you capture that is subject to regulatory controls and compliance mandates. Engaging the stakeholders that are responsible for compliance helps in this task. It is important to balance the value that is promised by the four Vs with the changes to existing risks and the new risks that are introduced.

As an example, because of the increased velocity of data, it might be necessary to add streaming analytics processes that are focused on the evaluation of the precision, accuracy, and integrity of the data while it is in motion. Because of the high velocity of the systems, the window of availability to respond to issues is reduced. Waiting for traditional batch processing cycle times exposes the business to heightened risk and any delay protracts the exposure, such as in cases of fraud or public safety threats.

## 2.2.4 Supporting disciplines

The supporting disciplines from the Information Governance CMM include established data management domains: Classification and Metadata, Audit Information Logging and Reporting, and Data Architecture. As with the other domains of the Information Governance CMM, these areas are each impacted in their own unique way by each of the four Vs of big data. The intent of this section is to approach the domains systematically and ensure full coverage of the governance factors influencing the implementation of a big data solution. Although this section explores some of those impacts, you will undoubtedly encounter impacts that are unique to your business.

### Classification and metadata

Because big data relies heavily upon large volumes of unstructured data, it becomes easy to assume that data classification takes on a lesser role. The reality is different. At the end of the information pipeline are the data scientists, report analysts, and others who convert data into knowledge. The users of the information rely upon segmentation to create order out of chaos and to refine the insight that the data provides.

This situation requires improved capability to dissect the unstructured content, then regroup and compare it in a new and meaningful manner. This has been a core process of analytics since the creation of the first database and continues to be true for advanced technologies, such as IBM Watson™, which is one the most advanced deconstruction, classification, and analytics processing engines in existence.

However, you do not need to have the technology of IBM Watson to successfully classify data to suit your business needs. You must build up the capability beginning with your master data management or data modeling practices. This capability does require practice and discipline to master.

Practice is the operative word here, as an organization does not learn to perform these tasks overnight. There is a reliance upon the core capabilities being in place, such as organization and stewardship. If your organization does not already consider MDM and Enterprise Data Modeling as required activities, it is incumbent upon the architects, business sponsors, and program managers to build this capability as part of the big data initiative.

### **Audit information logging and reporting**

Interestingly, most organizations consider this capability last or after they discover a problem and realize they are unable to trace back to the root cause or point of origin of the errant data. Often, there are issues with no audit trail for the data because of an absence of time stamps that are associated with data streaming into a big data environment, or the time stamps are not captured and retained. As a result, the slowly changing nature of data cannot be tracked.

The problems from the lack of appropriate logging and auditability of data can become compounded with streaming high velocity big data from multiple sources into a single integrated store. If not correctly synchronized during the integration process, “old” data that arrived late, perhaps over a batch process pathway, can overwrite the most recent data, which arrived through a high velocity streaming pathway.

This situation also impacts the veracity of your big data solution. After users realize that the data cannot be trusted because of the lack of controls over the synchronization of the data, they choose to obtain their information elsewhere in the data environment. Veracity also impacts the risk assessments that are associated with various regulatory controls. The ability to trace data that has been included into reports going to various regulatory bodies, or to prove compliance, remains a core business requirement regardless of the underlying technology.



Given the need for ongoing controls, it is advisable to establish control points throughout the architecture. After control points are established, the identification of who among the various stakeholders is responsible for the content crossing the control point is critical. This stakeholder is a business role and not part of an Information Technology group. The parties that are responsible must be engaged in the definition of various quality metrics and facts that are related to the control point location, including the requirements for auditability. This can include the ability to understand and measure the core dimensions of quality that were presented earlier.

## **Data architecture**

Eventually, all of the other factors converge to influence the design of your architecture. The Enterprise, Solution, and Information Architects must factor in the requirements necessary to design, deploy, and operate the environment. Some of these factors and considerations include integration of big data sources, placement of appropriate governance functions, such as information security and privacy requirements and information quality validations, storage and segmentation of big data to optimize its subsequent use, and provision of capabilities for downstream users, such as Data Scientists to access, analyze, and report on the contents.

In summary, the Information Governance CMM provides a framework for examining the impact of evolving technology such as big data on business operations, governance requirements, and the information technology architecture.

The following chapters explore the direct business value and opportunities that are associated with big data and the technical architecture and landscapes that are needed to deliver on the value proposition. Concurrently, it is incumbent upon the reader to factor in how human resources, the controls for compliance and risk management, and ongoing operational processes are governed. You must consider the greater question: How will your organization respond to the challenges of this evolution?





## Big Data Information Governance principles

When business and IT professionals first examine the subject of Information Governance, many are overwhelmed at the complexity of the subject. The number of systems and the competing business agendas throw so many variables into the mix that the task at hand appears unmanageable. Successfully solving the problem requires that you focus on the fundamentals, decide what your core values for information are, and then tie your decision-making processes to those principles.

The principles that are needed for governing big data are rooted in fundamental data management practices. They have evolved as technology advances introduced new opportunities to create value and new risks to be managed and controlled. The principles that are covered here serve as a foundation for the decisions that are embedded in the use cases that are explored throughout this book.

At first glance, you might be tempted to dismiss these principles as common sense, or good data management. Although this is true, the principles often become poorly practiced in many organizations. What is common sense is seldom common practice. Principles are brushed aside in favor of short cuts and compromises. As part of your Information Governance journey, you must ask yourself, what is the state of the fundamentals in your organization and why does that situation exist?

It is up to the leadership of an organization to drive the principles into action. This task is accomplished by ensuring that the principles are communicated widely and broadly to every worker handling data. The workforce needs them to serve as their guidance as they make the thousands of small decisions each day that your organization requires to succeed. Without them, the organization sets off in a thousand random directions. The principles must become part of an organization's foundation and “DNA”.

Second, the principles serve as the basis for the creation of a policy that embodies how the principles are routinely applied for your organization. The policy is secondary to communicating and embodying the principles into the organization. As the second level of control, it codifies those business requirements that you want to have endure over more than one year.

### **3.1 Root principle for Big Data Information Governance**

The principles that are presented here are based on a simple core precept: *The objective of governing information for an organization is to move information as quickly as is practical while keeping the quality as high as is practical and as secure as is practical.*

This is the master principle from which the others are derived. A simple concept but one that has wide-reaching implications. It sets up a delicate balancing act pitting productivity and quality as companions that at times work at cross purposes, as shown in Figure 3-1 on page 35.

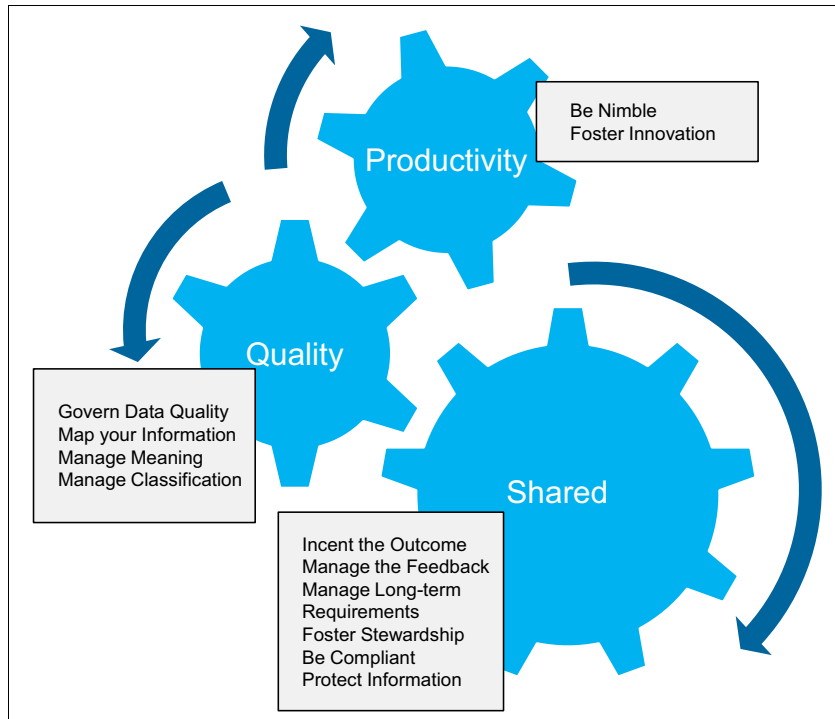


Figure 3-1 Root principle

The usage of the term *practical* in the core principle is intentional. Perfection is simply not an option for most organizations. In reality, productivity and quality need only to remain ahead of your competitors and improve steadily. *That* is the standard by which your customers and shareholders judge you. For a short period, you might need to favor one or the other, but staying out of balance for too long can lead to the demise of any organization.

An organization must sustain a proper balance between ensuring that the data is fit for purpose and having a healthy production tempo. Only your organization can identify when you have reached the tipping point where any additional governance and oversight delivers only modest value or minimally reduced risk. That decision belongs with senior leadership and cannot be delegated.

## 3.2 Leading from principle

The role of senior leadership is to balance productivity and quality by providing their guidance after they recognize when one should be favored over the other, as shown in Figure 3-2. This function is the primary responsibility of any Enterprise Data Governance Council. They are the action arm of the executive suite regarding the balancing of these competing forces.



*Figure 3-2 Speed versus quality*

### 3.2.1 Speed versus quality

The speed at which the aggregated impact and results of a customer interaction reaches line managers, sales and marketing, finance teams, and the leadership of an organization has a material impact on the business. As a practical example, market trading organizations locate data centers and trading functions as close as practical to the market systems to leverage a microsecond advantage over their competitors.

Likewise, the quality of the information influences whether those same information consumers trust the data more than their own intuition. In some cases, the confidence in the data has declined so much as to render the information unusable, which undermines the investment in the data.

The investment in the people and infrastructure becomes eroded simply because the business practices governing the operation and maintenance of the systems that capture, integrate, and distribute information never kept up with the increasing volume, variety, veracity, and velocity.

It is because of this balancing act that leadership over governance must occur high up within the organization to steer the actions lower in the organization.

### 3.3 Core principles for Big Data Information Governance

The core principles serve the leadership by providing a basis for their decision, and a means to communicate their intent to the lower reaches of the organization. With an understanding of these principles, decisions that are required during periods of high productivity remain likely to stay in alignment with the intent of the Enterprise Data Governance Council.

Many of these principles are common sense and common practice in the realm of structured data. Their origins can be traced to good data management practices that were developed decades earlier. They are an extension of those practices, which often provide the foundation for many regulatory requirements, such as SOX, HIPPA, Dodd-Frank, and others. They are extended here to include the principles that are necessary for big data. If your organization has not yet put in place a governance program/structure based on principles, then consider the following items to be a good starting point:

- ▶ Be nimble: The focus of data governance practices must allow for nimble responses to changes in technology, customer needs, and internal processes.
  - The organization must be able to respond to emergent technology (everyday, not just that day).
  - Consider how rules and bureaucratic controls might reduce productivity.
  - Standardize where doing so does not significantly impact the work flow.
- ▶ Be compliant: You must be able to apply social compacts to data brought into the organization by a person or process in association with policies and regulations.
  - Public company: Meet the obligation to protect the investment of the shareholders and manage risk while creating value.
  - Private company: Meet privacy laws even if financial regulations are not applicable.
  - Both types of companies: Fulfill the obligations of external regulations from international, national, regional, and local governments.
- ▶ Manage quality: Because information and data is the core of your business, the quality of the content of that information is paramount to your continued success.
  - For big data, the data must be fit for purpose; context might need to be hypothesized for evaluation.
  - Quality does not imply cleansing activities, which might mask the results.

- ▶ Encourage the outcome: Your people are the means to quality, security, and governance/compliance in your data.
  - Continuously generate awareness as a critical task for the governance program. As an example, many companies offer periodic training about the protection of personal information.
  - Audit and measure the organization against these principles routinely and include the results in performance monitoring for individuals and departments.
- ▶ Map your information: Understanding your complete business and the flow of information across all processes enables you to succeed with your first two principles. This requires the capture and recording of data at rest and data in motion throughout the organization.
  - Provenance and lineage are still, if not more important, in big data.
  - What is the source? This is a crucial question to validate analytics results as fit for purpose.
  - Where is it going or has gone? This is a crucial question for ongoing system maintenance and support efforts.
  - Enables audit reporting, which is considered an essential regulatory compliance function.
- ▶ Manage meaning: Data is the language of your business. To that end, understanding the language that you use and managing it actively reduces ambiguity, redundancy, and inconsistency, which relates directly to the quality of information.
  - Big data might not have a logical data model provided, so, as early as practical, any structured data should be mapped to the enterprise model.
  - Big data still has context and thus modeling becomes increasingly important to creating knowledge and understanding.
  - Determine the degree of ambiguity that you can tolerate in reporting and analytics. Measure and monitor this ambiguity as a metric.
  - Plan for meaning to change over time. The definitions evolve over time and the enterprise must plan to manage the shifting meaning.
- ▶ Manage classification: It is critical for the business/steward to classify the overall source and the contents within as soon as it is brought in by its owner to support of information lifecycle management, access control, and regulatory compliance
  - Public versus private.
  - Retention period/schedule.
  - Security level.



- Applicable regulatory controls (for example, PII, PCI, and illegal content).
- Non-destructive testing to make assessment/classification in a staging zone (for data at rest) or threat assessment/function in stream (for data in motion).
- If you do not have positive control over the incoming stream (for example, a Twitter feed is uncontrolled), then you must have components to monitor and classify what crosses the boundary. This is needed to be able to exclude any data that does not meet the acceptance criteria.
- ▶ Protect information: Protection of data quality and access is essential to the ability to maintain the trust of your customers and their customers.
  - Your information protection must not be compromised for the sake of expediency, convenience, or deadlines (for example, simple data exploration). If conditions warrant them, exceptions should be decided by the appropriate management team and documented.
  - Accept as a fact the difficulty in predicting how new risks manifest themselves.
  - Protect not just what you bring in, but what you join/link it to, and what you derive. Your customers will fault you for failing to protect them from malicious links.
  - The enterprise must formulate the strategy to deal with more data, longer retention periods, more data subject to experimentation, and less process around it, all while trying to derive more value over longer periods.
- ▶ Foster stewardship: Ensuring the appropriate use and reuse of data requires the action of an employee. This role cannot be automated and requires the active involvement of a member of the business organization to serve as the steward over the data element or source.
  - For the source data set/content (including structured data elements, metadata, and unstructured file/message contents) that is used in a critical business process, report, or analytical process, a data steward from the business should be assigned responsibility for defining the content, validation rules, security level, and quality standards for those data elements. The record of a data steward must remain current always.
  - A steward is not from IT (unless they own the business process). A steward is from the business; they oversee the data that is acquired or created for a legitimate business purpose.
  - Stewards exert control over the type and subject area of data from a business perspective. Information technology is *only* an enabler unless the data is from their own business process.

- ▶ Manage long-term requirements: Policies and standards are the mechanism by which management communicates their long-range business requirements. They are essential to an effective governance program.
  - Projects that impact the manner in which data is captured, stored, or moved must refer to all applicable policies and standards. Those policies and standards must be treated as equals to the business requirements.
  - If you are initiating a big data project, you might need a big data policy (or revisit existing policies) to see what applies and what unique exposures/value opportunities are created. Revisit the rules because big data is changing the game.
  - Factor in that big data implies less capacity for human intervention and identify new approaches for responding.
  - You must build in interruption processes to avoid runaway processes and monitor for such events. Teams must plan for feedback mechanisms to control the rate of change to prevent becoming overwhelmed by a runaway process or decision engine.
- ▶ Manage feedback: As a companion to policies and standards, an escalation and exception process enables communication throughout the organization when policies and standards conflict with new business requirements. It forms the core process to drive improvements to the policy and standard documents.
  - Every policy or standard must have a channel to facilitate changes that are identified by the people that are required to comply with the policy or standard. This is in keeping with the first guiding principle.
  - In general, with policies and standards, an escalation process must exist, be known, and followed.
  - You must consider the ability to create and manage exceptions to policy and ensure that exceptions are cleared over time. Exceptions should never remain a permanent solution.
- ▶ Foster innovation: Governance must not squelch innovation. Governance can and should make accommodations for new ideas and growth. This is managed through management of the infrastructure environments as part of the architecture.
  - Constant innovation must function with limited constraints from governance during development with an exception of designing for change and protection of data.
  - As part of any technology development effort, the project must report on what changes are required to policies, standards, existing processes, data models, or business terms. The intent is to instill the drive governance core principles into the design.

- ▶ Control third-party content: Third-party data plays an expanding role in big data. There are three types (defined below) and governance controls must be adequate for the circumstances. They must consider applicable regulations for the operating geographic regions; therefore, you must understand and manage those obligations.
  - Outsourced delivery (proxy for you): Contracts need to reflect your policies, as you are still responsible for the content of the data (for example, traceability).
  - Providers of data (publishing or selling to you): What is your responsibility (for example, a bad action that is based on bad data) versus that of the third-party provider (wrong emergency response)? The responsibility belongs with the third party, depending on the terms and conditions.
  - Subscribers of data (buying from you): The responsibility belongs with the deliverer, depending on the terms and conditions.

### 3.4 Practical application examples

Throughout the use cases in the following chapters, the principles here are referenced as a demonstration of the application of the principles. In witnessing the citing of the principle being applied, readers should gain a greater appreciation for the value of the principles. It is hoped you will better understand how they can add value to your governance and oversight role.





## Big data use cases

Chapter 1, “Introducing big data” on page 1 describes the phenomenon of big data, with the tremendous explosion in data volume, velocity, and variety (three Vs). Most discussions of big data start with these three Vs, and the numbers that are involved can be staggering. “According to Twitter’s own research in early 2012, it sees roughly 175 million tweets every day...”<sup>1</sup> That pales next to the volume of sensor data now produced. As an example, a single “Boeing jet generates 10 terabytes of information per engine every 30 minutes of flight... There are about 28,537 commercial flights in the sky in the United States on any given day. Using only commercial flights, a day’s worth of sensor data quickly climbs into the petabyte scale.”<sup>2</sup> That is only sensors on jet engines in the United States! There are sensors on other parts of the planes, sensors in automobiles, sensors for weather and water quality, sensors in mobile phones, and so on.

These three Vs plus a fourth V, veracity, represent the dimensions of the information governance challenge that every large organization deals with daily as they struggle to extract value from their information resources and make better decisions, improve operations, and reduce risk.

---

<sup>1</sup> <http://www.mediapost.com/publications/article/173109/a-conversation-on-the-role-of-big-data-in-marketing.html#ixzz2dSe0F5xt>

<sup>2</sup> <http://gigaom.com/2010/09/13/sensor-networks-top-social-networks-for-big-data-2/>

## 4.1 Emerging big data use cases

For any organization, only some segment of this data volume is of interest and use. But with the variety and velocity of this data, often coming in unstructured formats, it can be difficult for organizations to use these new data sources effectively to gain new insights and improve decision making. Organizations are challenged with a number of considerations, such as what other data sources are available and how they can be used with what is known. Here are examples:

- ▶ How to manage, use, and analyze the data when it is not in a common, familiar, or readily usable format (such as a relational database).
- ▶ How to use non-traditional sources of customer data (such as call center data or social media comments) to better understand and predict customer behavior.
- ▶ How to process time-sensitive data for real-time decision making, such as identifying security risks.
- ▶ How to use the range of log and sensor data to respond to machine-based events, predict downtime, or ensure that service-level agreements are maintained.
- ▶ How to find and derive high value data from the new data sources that can connect to traditional sources for improved insights.
- ▶ How to take advantage of new technologies to lower the total cost of ownership (TCO) while still deriving maximum value from their data.

Although there are certainly differences and variations across industries, five primary use cases have emerged around these challenges:

- ▶ Big data exploration
- ▶ Enhanced 360° view of the customer
- ▶ Security and intelligence extensions
- ▶ Operations analysis
- ▶ Data Warehouse augmentation

These use cases are not mutually exclusive. Identification and usage of sensor-based information may become summarized and correlated with other traditional information, such as sales or pricing in an augmented or extended Data Warehouse, and then lead to analytics providing greater insight into better marketing campaigns to customers or better identification of fraud or security risks.

## 4.2 Big data exploration

A common phrase now is that data is “the new oil”<sup>3</sup>, that is, a resource that offers tremendous potential but requires exploration and refinement to extract that value. Just as the oil and gas industry relies on exploration to identify the most productive resources for drilling, the first step in using big data is to discover what you have and to establish the ability to access it and use it to support decision-making and day-to-day operations. This aspect is not new. For many years now, a key step of any information integration effort has been to understand the data. This remains a critical and imperative step for big data, yet there is so much of it that the task quickly appears overwhelming. Big Data Exploration is the way that you get started.

Why is this understanding and exploration so imperative? If you do not know what is available and what you have, then there is no meaningful way to derive new value from it in any other use case. If you do not understand the volume of the data, then you cannot understand how to appropriately manage the environment for it. If you do not understand how often the data arrives or how quickly it becomes useless, then you cannot make timely, and often real-time, decisions. If you do not know the variety of information, then you do not know how to integrate what you have or how to relate it to other sources of information, and then there is still no meaningful way to derive new value. If you do not understand what the data contains, then you do not know what risks that data might present to your organization, whether poor decision-making or potential exposure to regulatory compliance and the associated penalties and fines.

Organizations have multiple applications that manage information, from their CRM systems to enterprise content management, Data Warehouses, and rapidly growing corporate intranets. The challenge is that any important decision, customer interaction, or analysis inevitably requires information from multiple different sources. As the world has changed, you have seen the addition of many other sources of data that people need to perform their day-to-day work and make important decisions. The growth of so-called “raw” data—collected from, as examples, sensors, machine logs, and clickstreams from websites, presents yet another challenge. How do organizations add context to this data to fuel better analytics and decision making?

---

<sup>3</sup> <http://www.marketplace.org/topics/tech/data-economys-new-oil>

Big data exploration addresses the challenge that every organization faces: Information from a vast array of new and traditional sources is stored in many different systems and silos in many different forms, as shown in Figure 4-1.

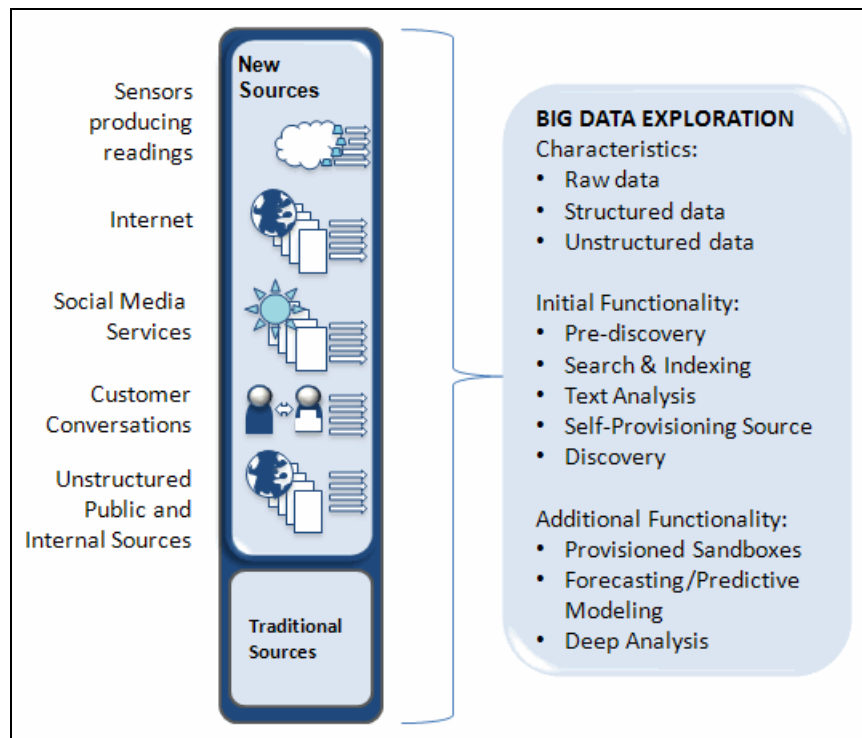


Figure 4-1 New sources of information drive Big Data Exploration



## 4.2.1 Identification of value

Big data exploration becomes a method to identify the data from which you can derive value or profit versus that data that simply is stored and becomes waste. The following examples are described in the IBM white paper *Data Profit vs. Data Waste*, IMW14664USEN:

- ▶ A leading healthcare provider substantially enhanced its compliance position by automating the capture of new compliance-related data from all sources, including internal memos, web page updates, and regulatory bulletins, and the selective distribution of that data to targeted staff members. This reduced the company's exposure to fines, injunctions, and other compliance-related risks. More effective and automated usage of its existing unstructured data assets also enabled the company to reduce its knowledge worker headcount by 140 full-time employees (FTEs), resulting in labor savings of \$11.2 million annually.
- ▶ A leading aircraft manufacturer implemented a portal that gave its employees vastly improved visibility into all unstructured data relating to the supply chain for a major new model. The speed with which supply-chain managers were able to retrieve critical data from spec sheets, bills of lading, and other documents allowed them to repeatedly avoid parts delivery delays and errors, which saves the company millions. The company also estimated that it was able to reduce knowledge worker headcount on the project by 9% and realize another \$1 million savings on the printing, distribution, and storage of paper documents alone.

Clearly, finding and identifying data of interest, particularly in and across diverse repositories of unstructured and structured content, is one key aspect of Big Data Exploration. This improves information access and knowledge sharing, reduces potentially duplicated and extraneous work, and ultimately enhances decision making. Basic capabilities such as indexing, search, discovery, and text analytics are needed to support the incorporation of many of these sources into the organization.

## 4.2.2 Enablement of Data Science teams

Another key aspect is the enablement of teams around the emerging concept of Data Science and the corresponding role of the data scientist. Data Science encompasses a range of skills such as statistical and mathematical analysis and modeling, pattern recognition and learning, and data visualization techniques going beyond basic data analysis. Data scientists establish hypotheses, find potentially relevant data sources, and then apply analytical and modeling techniques to test the data against the hypotheses. The work of these teams is also exploratory work. Data sources must be evaluated for usefulness and bias that might skew results. Data sources may require filtering or summarizing to be correlated against other standard information. Different models, algorithms, and approaches might be needed to perform tasks such as time series analysis, audio and visual analysis, or the application of optimization techniques, such as linear or quadratic programming.

For example, significant decreases in store sales on weekends before Christmas over several years might suggest that weather is a likely factor. The data scientist might start with this as his hypothesis, gathers historical sales transaction data over a 5-year time window for selected cities, and then incorporates weather data from the National Weather Service for the same time frame and cities into the analysis, and other data that might help find potential correlations, such as airline delays and traffic records. After analysis, where the hypothesis is validated, the data scientist might identify that specific data sources (for example, real-time airline delays) have the highest likelihood of predicting sales impacts in real time. The recommendation to the marketing department might be that this source is used to identify emerging sales trends and adjust sales offers from store channels to online channels.

Big data exploration is thus foundational to subsequent work by providing insight into this varied content and where the volume and variety of data might be effectively applied. This effort, often undertaken by emerging teams of data scientists, data analysts, and domain experts, focuses on key tasks such as the following ones:

- ▶ Identifying promising sources of information
- ▶ Finding characteristics of possible use (such as detail and statistical) and the relationship to existing sources of information
- ▶ Validating hypotheses or finding correlations to generate hypotheses (for example, snow storms significantly decrease store sales but increase online sales)

- ▶ Finding characteristics and risks from a governance perspective (does this source contain sensitive information or potentially create sensitive information when merged with existing sources?)
- ▶ Determining whether to operationalize (and retain) specific data sources

The outcome is the transformation of raw big data into refined and usable information that can support subsequent use cases.

## 4.3 Enhanced 360° view of the customer

Gaining a full understanding of customer behavior, such as why they buy, how they prefer to shop, why they switch, what they buy next, what factors lead them to recommend a company to others, and what they dislike, is strategic for every company. The IBM Institute for Business Value report “Real-world use of big data”<sup>4</sup> cites as its number one recommendation that organizations should focus their big data efforts first on customer analytics that enable them “to truly understand customer needs and anticipate future behaviors.”<sup>5</sup>

Organizations have been working with systems and applications over the last 10 - 15 years that focus on their customers through Customer Relationship Management (CRM) and Master Data Management (MDM) systems. However, storing this level of customer data treats the customer as an information object; the customer has such things as a name, an address, and a phone number. Although the information is associated and correlated with events and transactions such as orders or calls, there is often little insight into those aspects that engage the customer with the organization.

---

<sup>4</sup> “Analytics: The real-world use of big data”, by IBM Institute of Business Value in collaboration with Saïd Business School at the University of Oxford, 2012. Use the following URL to request a copy of the article.

[https://www14.software.ibm.com/webapp/iwm/web/signup.do?source=csuite-NA&S\\_PKG=Q412IBVB](https://www14.software.ibm.com/webapp/iwm/web/signup.do?source=csuite-NA&S_PKG=Q412IBVB)  
igData

<sup>5</sup> Ibid.

Organizations see big data as providing the opportunity to better understand and predict customer behaviors, and by doing so, engage more effectively with their existing and potential customers and improve the customer experience. Data sources such as transactions, multi-channel interactions, social media, syndicated data through sources like loyalty cards, and other customer-related information, as broadly shown in Figure 4-2, have increased the ability of organizations to create a complete picture of customers' preferences and demands, which has been a goal of marketing, sales and customer service for decades.

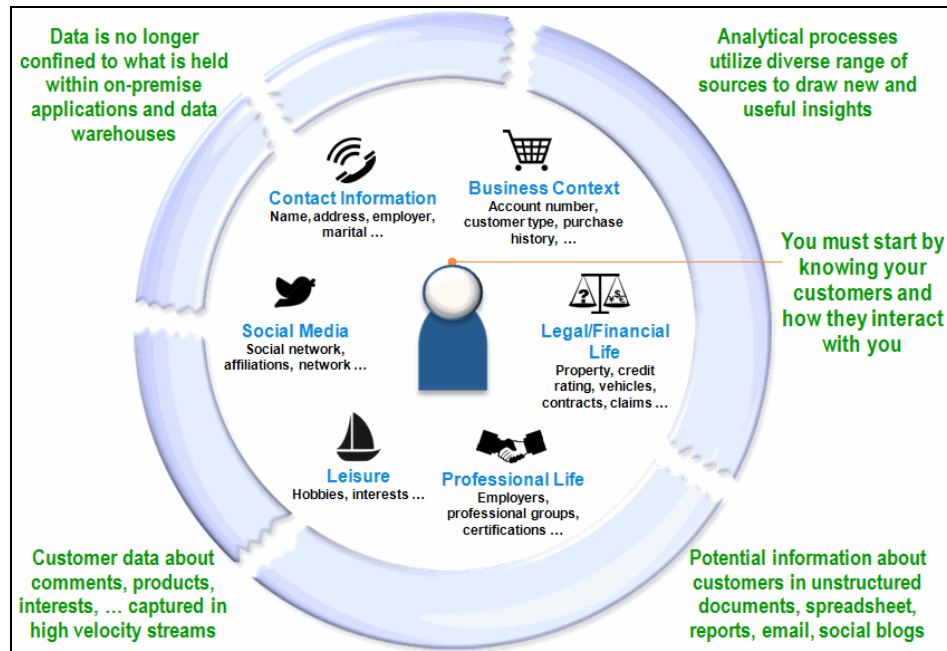


Figure 4-2 The expanded range of customer data

That principle applies not only in the retail sector, but equally in the telecommunications, healthcare, government, banking and finance, and consumer products sectors where consumers and citizens are involved, and in business-to-business interactions with partners and suppliers.

### 4.3.1 Expanding the range of customer-related data

Social media is one example of a new big data source where collaborative channels such as Facebook or Twitter have enabled customers to be more vocal about expectations and outcomes. Issues with an organization's products or services can be immediately expressed and exposed to a large range of individuals, which can have a potentially negative impact on sales or customer support. Conversely, the relationships and “likes” of given individuals within the organization's target market can generate significant interest in what the organization offers.

By combining such data with other internal and external sources, organizations create more than just a “mailing label” view of customers. This 360° view supports analytical results such as the following:

- ▶ Fine-grained customer segmentation
- ▶ Lifetime Value Score
- ▶ Next best offer
- ▶ Recent interaction history
- ▶ Products owned
- ▶ Existing relationships to organizational employees

Armed with this type of information, marketing can target distinct customer groups with timely, mobile offers, agents can identify the next best offer or action to keep a customer loyal, and customer support can respond quickly to emerging issues.

In addition to these analytics that give strategic insights into customer behavior, the importance of the 360° view extends to the front-line employees. Forward-thinking organizations recognize the need to equip their customer-facing professionals with the correct information to engage customers, develop trusted relationships, and achieve positive outcomes such as solving customer problems and up-selling and cross-selling products. To do these tasks, they need to be able to navigate large amounts of information quickly to discover what is needed for a particular customer.

### 4.3.2 Personalized Customer Engagements

One focal point for this enhanced view of the customer is Personalized Customer Engagements. A more comprehensive understanding of each customer leads to greater revenue and lower churn (also called *churn rate* or *attrition rate*, this is a measure of loss of customers in a period). Organizations expect this direct one-on-one personal experience to accomplish the following goals:

- ▶ Increase revenue by building comprehensive views of customer behaviors and preferences by segment across all channels
- ▶ Increase customer intimacy by adding social analytics as an additional source of valuable insight into their interaction with the organization and its products
- ▶ Increase customer satisfaction with faster and complete insight into the individual customer's experience and service levels
- ▶ Improve customer retention by predicting the likelihood of a customer's departure and delivering customer retention offers in near real time
- ▶ Increase offer-acceptance by analyzing customer usage and response to offers in real time

### 4.3.3 Micro-market Campaign Management

The analysis of customer response to offers touches on another focal point of the broad use case: Micro-market Campaign Management. Organizations increasingly find that targeted messaging to micro-markets (that is, small segments of their customer base) achieve higher returns than broad generalized marketing campaigns. To do this effectively, though, organizations must accomplish the following tasks:

- ▶ Capture and analyze consumer sentiment for brand and product affinities
- ▶ Identify and drive localized brand and product affinity
- ▶ Crowd source and test new product ideas
- ▶ Enrich client customer data with email content to differentiate email marketing service offerings
- ▶ Improve clients' email campaign results through analytics on customer data that is enriched with email content
- ▶ Improve return on marketing spending with tailored offers and promotions; reasonable increases in response generally correlate with an increase in purchases

- ▶ Optimize offers and cross-sell to increase average products per customer; even small increases, say from an average of 1.5 to 2.0 products per customer, can have significant revenue uplift
- ▶ Track social media sentiment analysis to measure the impact of targeted campaigns

Ultimately, this focus can facilitate the goals of campaign management, whether in the success of new product launches or the targeted increase in sales in selected channels or market segments.

#### **4.3.4 Customer retention**

Customer retention activities are also enhanced with big data. Not only can the organization become more directly and personally engaged with the customer, but it can also improve its opportunities for customer retention based on behaviors that are seen across correlated pieces of data instead of only transactions. Organizations can accomplish this task through multiple sources of big data, such as the following:

- ▶ Use branch teller notes, call center notes, and client emails to identify changing client behaviors or issues
- ▶ Track social media sentiment analysis to measure dissatisfaction with the organization or its products and services
- ▶ Identify relationships between customers through social media that are not visible in traditional customer master data, which might impact the organization's brand
- ▶ Improve the ability to identify customer attrition risk and take preventive measures to improve customer retention

Tied to Personalized Customer Engagements and Micro-market Campaigns, active retention practices offer more opportunities to continue existing revenue streams than trying to attract new customers in competitive markets.

#### **4.3.5 Real-time demand forecasts**

With demand forecasting, the target is to help optimize inventory in manufacturing or in retail. Often, this is part of an effort to improve the efficiency of an entire supply chain. By tying in to big data such as social media or even upcoming weather forecasts, emerging trends can be identified and demands satisfied when needed.

Demand forecasting can be facilitated by bringing together the following:

- ▶ Historical sales patterns and price points by store, chain, product, category, sales channel, or region
- ▶ Local sentiment information to understand localized trends
- ▶ Weather or news information that might correlate to or trigger higher demand for specific goods (for example, shovels for approaching snowstorms, and plywood boards for impending hurricanes)

The goal becomes a smarter, timely response with improved promotion tactics and execution. By timely positioning to meet demand, positive feedback can be generated in support of subsequent customer or supply chain requirements.

## 4.4 Security and Intelligence extensions

At the same time that organizations are trying to enhance their understanding of their customers, they are also engaged at identifying and preventing threats to their security.<sup>6</sup>

Maintaining data security is critical. In this era of big data, the average cost of security-related events is estimated to cost American companies USD 40,000,000 per year.<sup>7</sup> Worldwide, cybercrime costs people and businesses USD 388,000,000,000 in lost time and money, which is USD 100,000,000,000 more than the cost of the world's illegal drug market.<sup>8</sup> Identity theft is at a three-year high, affecting 12,600,000 US consumers and costing more than USD 21,000,000,000 each year.<sup>9</sup>

Data breaches not only cost money; they can also drive down stock prices and cause irreparable brand damage. Protecting brand reputation and preserving customer trust are two of the top three organizational goals that depend on good data protection.<sup>10</sup>

---

<sup>6</sup> *Data protection for big data environments*, IMM14127USEN

<sup>7</sup> *The Big Data Imperative: Why Information Governance Must Be Addressed Now*, found at: [http://ibm.com/common/ssi/cgi-bin/ssialias?subtype=WH&infotype=SA&appname=SWGE\\_IM\\_DD\\_US&htmlfid=IML14352USEN&attachment=IML14352USEN.PDF](http://ibm.com/common/ssi/cgi-bin/ssialias?subtype=WH&infotype=SA&appname=SWGE_IM_DD_US&htmlfid=IML14352USEN&attachment=IML14352USEN.PDF)

<sup>8</sup> *Norton Study Calculates Cost of Global Cybercrime: \$114 Billion Annually*, found at: [http://www.symantec.com/about/news/release/article.jsp?prid=20110907\\_02](http://www.symantec.com/about/news/release/article.jsp?prid=20110907_02)

<sup>9</sup> *Identity Fraud Hits Three-Year High; More Than \$21 Billion Lost*, found at: <http://www.darkreading.com/privacy/identity-fraud-hits-three-year-high-more/240149005>

<sup>10</sup> *The Business Case for Data Protection: What Senior Executives Think about Data Protection*, found at: <http://www.ponemon.org/library/the-business-case-for-data-protection-what-senior-executives-think-about-dataprotection>



Data protection is also required by law. More than 50 international laws, such as Canada's Privacy Act, Germany's Federal Data Protection Act, Argentina's Personal Data Protection Act, and Korea's Act on Personal Information Protection, mandate data protection.<sup>11</sup>

#### 4.4.1 Enhancing traditional security through analytics

To identify and protect against threats, organizations must enhance traditional security solutions to prevent crime by analyzing all types and sources of big data (unstructured and streaming) and sources of under-used data. In addition, using the associated big data technologies to augment and enhance the security solutions through the range of analytical techniques improve intelligence, security, and law enforcement insight.

To combat these sophisticated threats, organizations must adopt approaches that help spot anomalies and subtle indicators of attack. Doing so requires collecting and analyzing data from the security infrastructure and beyond, including traditional log and event data and network flow data, vulnerability and configuration information, identity context, threat intelligence, and more. In short, security has become a big data problem.

Forward-leaning organizations are exploring custom analytics that use additional big data technologies on various unstructured data sources, including email, social media feeds, business transactions, and full network packet payloads. Forward-leaning organizations are turning to big data platforms, such as those based on an enterprise-class Apache Hadoop (Hadoop) system to help solve advanced security challenges.

The types of analysis big data platforms provide typically uses historical baselines, statistics, and visualizations to uncover evidence of past fraud or security breaches. Here are some examples:

- ▶ Correlating millions of global DNS requests, HTTP transactions, and full packet information to identify malicious communications that are associated with botnets
- ▶ Uncovering fraud by correlating real-time and historical account activity, and by using baselines to spot abnormal user behavior, unlikely application paths, and suspicious transactions
- ▶ Linguistic and predictive analytics to profile email and social networking communications and to identify suspicious activity, triggering proactive measures before an incident takes place

---

<sup>11</sup> *International Privacy Laws*, found at: <http://www.informationshield.com/intprivacylaws.html>

Big data analytics must store, process, and analyze a wide variety and volume of structured, semi-structured, and unstructured data that is not currently analyzed by today's security solutions.

The Security and Intelligence Extension use case has several key drivers, including the needs to accomplish the following tasks:

- ▶ Analyze data from existing and new sources (data in motion and at rest) to find patterns and associations
- ▶ Analyze data that is continuously generated, such as video, audio, and smart devices
- ▶ Incorporate up-to-date and often real-time intelligence information (currency)
- ▶ Predict, detect, act, and respond to network and computer security threats sooner
- ▶ Use telecommunications and social data to track criminal/terrorist activity over time

Based on these drivers, the following three primary scenarios have emerged within this use case.

#### 4.4.2 Network threat prediction and prevention

All organizations are vulnerable to threats over their computer networks, both external and internal, whether the organization is a government, a financial institution, a utility, or even a manufacturer or retailer.

Cyber attacks might be the most reported threat, particularly in light of publicly announced data breaches or where networked systems are overwhelmed, but they are just one of the external threats. These attacks may range from state-sponsored or state-aided efforts to organized cyber criminals or individual hackers. Often, these attacks are persistent and sophisticated attempts to compromise an organization. Advanced persistent threats may include malware (infected software or computer viruses), denial-of-service attacks, or other attacks that are intended to disable, disrupt, or compromise an organization's operations.

Another emerging external threat is *hacktivism* (social activism through “hacking”), which is computer-based efforts to protest against a given organization's business or business practices. Such attacks may come from groups or individuals, usually with the intent to present a point-of-view that is visible to other outsiders, such as customers. Such an attack may come simply from an association with a country or other organizations who are the primary targets of the hacktivist.

Insider threats remain a third category to predict and prevent. There are many possible triggers for these threats, such as organizational decisions, an employee's economic condition, changes in an employee's performance or productivity, or the employee's relationship with managers or co-workers. Some situations might be innocuous and accidental, and others might be more malicious or based on economic rewards. Users might steal or sell intellectual property, customer lists, or even employee data.

An example of how these threats can come together with high cost to an organization was the denial-of-service attacks that were launched by the hacker group Anonymous against Sony and its PlayStation Network after Sony removed support for Linux. Personal user data and credit card information was stolen *en masse* and the network was down for nearly one month. The cost to Sony was approximately USD 171,000,000, not to mention the damage to its brand reputation.<sup>12</sup>

The impact of a cyber incident within an organization can be large, affecting productivity (lost man hours) and financial performance (lost sales, credit rating, and stock price), incurring incremental costs (overtime and regulatory fines); and hidden costs (liabilities and lost opportunities), and compromising an organization's reputation.

Predictive ability carries the potential for discovering emerging network threats and mitigating the risks of compromise and exposure. To support predictive ability, organizations require a broader mix of data:

- ▶ High volume network and DNS events.
- ▶ Rapidly changing identifiers.
- ▶ User access and historical usage patterns.
- ▶ Historical records of how malware infections have spread or the network paths that were used in prior attacks.
- ▶ Email, document repositories, and the movement of data across the network.
- ▶ Social networks, job tasks, access privileges, and other data that creates high-risk points.
- ▶ Human resources reports, news feeds, and employee reviews for potential insider threats.
- ▶ Social media feeds, news feeds, and government reports might point to risks of hacktivism.

---

<sup>12</sup> *One Year Later: Reflecting on the Great PSN Outage*, found at:  
<http://www.ign.com/articles/2012/04/21/one-year-later-reflecting-on-the-great-psn-outage>

By bringing together a broad and diverse mix of big data, an organization increases its opportunity to identify potential attack vectors and inside risks and put mitigating strategies in place.

### 4.4.3 Enhanced surveillance insight

Some security risks take a physical form. In cities or utilities, this risk might be from criminal activity targeting individuals or plant operations. For retailers, this risk can be theft either at stores or in their own warehouses. For insurance companies, understanding the physical situations that are associated with accidents, such as data taken from street-based cameras or pictures that are uploaded to the internet, help reduce fraud. Achieving greater insight into these physical risks comes in the form of surveillance media, typically video and audio data from internal or external sources.

Video- and audio-based equipment produces a huge volume of constant, streaming information. A recent example of such a solution requires the processing of 275 MB of data from over 1000 sensors in one-fourteenth of a second.<sup>13</sup>

Surveillance insight can be improved by using broader information sources and types:

- ▶ Collection and processing of machine data (internet, satellite, video, and audio)
- ▶ Unstructured data analysis, correlation, and pattern matching
- ▶ Integration with surveillance platform activity workflow
- ▶ Deployed security surveillance systems to detect, classify, locate, and track potential threats at highly sensitive locations

By putting this type of data with streaming analytics functions, organizations can process high volumes of information rapidly and yield significant information about physical conditions that might pose threats.

---

<sup>13</sup> TerraEchos: Streaming data technology supports covert intelligence and surveillance sensor systems, found at:  
<http://public.dhe.ibm.com/common/ssi/ecm/en/imc14726usen/IMC14726USEN.PDF>

#### 4.4.4 Crime prediction and protection

Predicting criminal activity is a focal point for law enforcement organizations globally. This prevention can be improved efforts to identify criminals or locations of likely crimes to better staff such areas and reduce crime in those locations. Other organizations, such as financial or municipal services, are looking to address fraud. Fraudulent activity may increase an organization's risk, decrease its revenue, create issues of compliance with regulatory agencies (potentially including fines), and might impact the relationship of the organization with customers, employees, or partners.

Besides traditional information such as known parties or customers, organizations are looking to bring significant new sources of information to bear to track criminals in real time to predict and prevent crime. These sources include the following ones:

- ▶ Mobile and social media communication channels
- ▶ Linguistic and identity analysis from text and voice sources
- ▶ Consolidated information about offenders, arrests, calls for service (911), human resources, and geographic information from multiple channels
- ▶ Historical information about member benefits and entitlements

By analyzing this big data, investigators can gain insight from billions of records and reach the appropriate agents within minutes, not days or weeks, enabling them to respond faster and with appropriate resources. This insight may include visualizations of patterns of crime or fraud, detection of complex threats or hidden facts, and active monitoring of known suspects through multiple aliases and watch lists.

### 4.5 Operations analysis

Operations analysis focuses on better usage of machine data. Machine data is often equated with logs or sensor data, and that data is certainly part of the realm of machine data, but it is much broader area. Machine data is any information that was automatically created from a computer process, application, or other machine without the intervention of a human. For example:

- ▶ System or application logs
- ▶ Sensor readings
- ▶ Meters
- ▶ GPS signals (for example, from mobile phones and bus or truck location devices)

- ▶ Call data records
- ▶ Configuration data
- ▶ Data from APIs and message transactions
- ▶ Website clickstreams
- ▶ Network packet data

Each of these types of machine data contains a wide variety of distinct data. Consider the range of data coming from computer system generated logs. The chart in Figure 4-3 highlights some of the types of log messages that can be produced on an ongoing basis, in some cases after changes (for example, a database change log) and in other cases in a constant stream. Even on this segment of machine data, there is a tremendous range of incoming data and analytical techniques that can be applied.

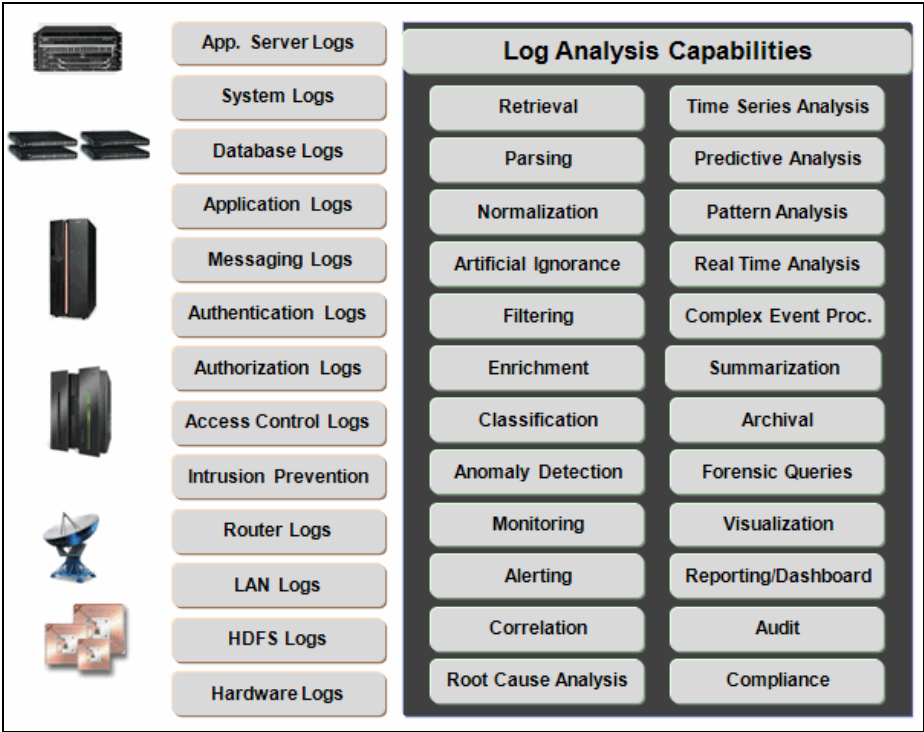


Figure 4-3 Log data and the associated log analysis capabilities

With machine data in its raw format, organizations cannot use the wide variety of machine data that is generated. Consider that machine data has the following attributes:

- ▶ Grows at exponential rates
- ▶ Arrives in large volumes in a wide variety of formats, often without standards
- ▶ Arrives at differing time intervals, often in-motion
- ▶ Needs to be combined with existing enterprise data
- ▶ Requires complex analysis and correlation across different types of data sets
- ▶ Requires unique visualization capabilities that are based on data type and industry/application

A good example is an airline. One airline flight generates ~5 GB of data from hundreds of sensors and indicators every second. Every 30 minutes of a flight creates ~10 TB of data.<sup>14</sup> If a given plane has 10 hours of travel time in a day, it generates 200 TB in that time and maybe 57 PB over the course of a year. Multiply that figure times all the airplanes in a given fleet, and there is much data to work with.

By analyzing that data from an operational perspective, an airline company can start to be predictive by identifying what parts of the plane are not performing optimally and that are going to need repair. Repairs can be scheduled and problems resolved before they fail and cause delays. By detecting anomalies and patterns from the machine data, organizations can avoid the high cost of emergency repairs or delays that upset customers. This situation is not isolated to airlines or other organizations in the transportation or utility industries; it applies to any organization manufacturing goods, maintaining buildings, or operating systems and networks.

Operations analysis takes advantage of big data technologies to analyze large volumes of these multi-structured, often in-motion machine data and gain insight from it using a broad array of analytical capabilities to yield better business results. Organizations may perform the following actions:

- ▶ Analyze machine data to identify events of interest
- ▶ Apply predictive models to identify potential anomalies
- ▶ Combine information to understand service levels
- ▶ Monitor systems to avoid service degradation or outages

There are several key scenarios within this use case, which are described in the following sections.

---

<sup>14</sup> *Data on Big Data*, found at: <http://marciaconner.com/blog/data-on-big-data/> and *Sensor Networks Top Social Networks for Big Data*, found at: <http://gigaom.com/2010/09/13/sensor-networks-top-social-networks-for-big-data-2/>

## 4.5.1 Traffic management

Traffic occurs in both the physical world and the world of computer systems and networks. By looking at the range of data that is associated with either systems or networks, operations analysis can help predict and improve traffic flow.

In the physical world, streaming real-time data that is gathered from cameras at entries and exits to cities, GPS data from taxis, buses, and trucks, airline flight information, and weather information can be combined to create a better understanding of where actual bottlenecks exist and where alternative routes are. For example:

- ▶ Transit authorities can identify the location of buses along their respective routes, or identify buses that do not follow their assigned route, and identify opportunities to improve the riders' experience.
- ▶ Riders can in real time estimate the time of arrivals of the buses at their next stops along their route, or the probability of delays at stops or travel times at stops at different times of the day or different days of the week from historical data.
- ▶ Travelers can predict the best time and method to travel to an airport, such as when to leave to catch a flight at the airport based on current traffic conditions and possible airline delays.

In the world of data, real-time data is gathered from, as examples, network switches and packets, from application and system logs, from monitors of processor and I/O channels, from response times in user interfaces, and from message queues. Some of this information might help analyze security threats. Some of this data might identify key bottlenecks causing delays in processing or potential downtime for customers or employees. Again, prediction of potential failures or consistent spikes in network traffic allows operators and administrators to take corrective action before failures impact business operations.

## 4.5.2 Environmental monitoring and assessment

A broad array of meters and sensors are now deployed for environmental monitoring. This monitoring can range from the multiple streams of physical, chemical, and biological data from sensors that are deployed in rivers and bays to assess water supplies to the sensors that are deployed in buildings for improved energy consumption.



The data from such sensors and meters can be analyzed against larger meteorological data and aggregated to identify weather-based factors and dynamics. Social media feeds can be coupled with this data to identify additional factors that are reported through human interactions. Such data can be visualized and delivered in standard format to scientists, engineers, policy makers, educators, or facility managers for real-time assessment. The data can potentially be shared with outside systems, researchers, policy makers, or operational units to increase collaboration and improve resource management.

Some benefits, depending on industry, may include the following ones:

- ▶ Help resource management respond more effectively to changes to local water resources. Users can access, aggregate, analyze, and set up automated alerts using a web portal.
- ▶ Create faster and more accurate flood prediction.
- ▶ Identify and track pollution and location-based debris to increase public safety.
- ▶ Streamline seafood, shipping, and monitoring operations.
- ▶ Improve building energy usage by identifying hotspots or alternative approaches to cooling.

### **4.5.3 Predictive Maintenance**

Most industries have equipment of some type that requires maintenance. The equipment might be part of complex manufacturing processing, might be heating or cooling systems for buildings, or even might be the sensors and meters producing data. Identifying issues and resolving them before systems shut down saves time, money, and resources.

As an example, in the Chemical and Petroleum industry, organizations apply new technologies and processes that capture and transform raw data into information to accelerate the discovery of petroleum, improve production efficiency in both upstream and downstream channels, maximize asset usage, reduce unplanned shutdowns, and extend the life of their assets. It is vital to safety in the oil and gas industry to ensure safe and reliable operations of wells and rigs and avoid catastrophic failures.

Analysis of machine data helps accomplish the following tasks:

- ▶ Optimize the well production yield and lower the production cost to match customer demand
- ▶ Reduce the cost that is related to downtime, labor, maintenance, repair frequency, catastrophic failures frequency, and problem ID to solution time, resulting in the following items:
  - Increasing additional hours of availability per year per site
  - Increasing the mean time between shutdowns per site
- ▶ Mitigate risk and operate in an environmentally aware and responsible manner
- ▶ Predict which equipment is most likely to fail and evaluate the potential cost of each failure
- ▶ Develop failure forecasting methods that “learn” as conditions change

On an oil rig, big data is analyzed to detect events leading to non-productive time in the operating environment and provide the information that operations personnel need to initiate action to either optimize operations or take corrective action to mitigate non-productive time.

Through operational modeling and predictive analytics, optimizing oil field and refining assets can produce significant positive economic impact. The producer can optimize assets by sharing information across functions, visualizing interactive data and collaborating both inside and outside the organization.

Although this example is of a large production environment operating with many environmental variables, Predictive Maintenance is relevant across industries, whether in retail grocery chains ensuring that their freezers and coolers are functioning adequately to protect food or in financial service organizations maintaining large system networks delivering banking and trade transactions in a timely fashion.

## 4.6 Data Warehouse modernization

Traditionally, the Data Warehouse is the central repository for an organization's data, allowing the collection and correlation from operational, transactional, and master data sources. The data is structured as it enters the Data Warehouse to support a common understanding. From the Data Warehouse, the data can be dispersed to the varied data marts and business intelligence and reporting solutions, which answer common and repeated business questions and drive business decisions.

For example, daily order and sales transactions are fed from operational systems into the Data Warehouse. These transactions might record the customer who ordered and purchased the items, so they are linked in the warehouse to customer master data that is fed from customer data hubs. The structure of the information in the Data Warehouse supports the generation of, for example, summarized order and sales data by date, channel, geography, and demographic segment. Business users receive reports or views of this information and use it to focus on particular geographies that are struggling for sales or identify emerging channels that are preferred by certain classes of buyers.

However, the advent of big data is modifying the traditional view of the Data Warehouse. The need or desire to use a wider variety of data alters the approaches to ingesting or incorporating data for subsequent analytics. Organizations must address a mix of structured, unstructured, and streaming data that often has low latency requirements (analysis in minutes or hours not weeks or months) and still support queries and reports across these diverse types of data.

For example, raw sensor data (such as from an electronic car, military sensor, or smart meter) may be ingested, cleansed and transformed, and then analyzed from an operational perspective. But, the data may be required for further analysis or query support, so after being transformed into a structured or semi-structured form, the data can then be placed into the Data Warehouse for decision support, business intelligence, or other further analysis.

Similarly, organizations are recognizing that there is a wealth of untapped information in open social media data. However, the Data Warehouse and standard relational databases are not suited for analyzing social media data. Social media data is unstructured and requires different technologies to process and extract useful information.

It is likely impractical to store all this new data in the Data Warehouse itself. This data is changing rapidly and has both a great variety and high volume. The potential time that is required to model to integrate and to deploy this data within the standard warehouse structure, let alone acquire the necessary storage to contain it, means that organizations cannot afford to store it all in the classic Data Warehouse environment. This is both a technical issue and financial issue for enterprises attempting to use big data with the Data Warehouse alone.

The Data Warehouse also stores low-touch “cold” data that is infrequently accessed (for example, large volumes of old historical transaction data). What is needed is a secondary compute and storage platform that is cost-effective for storing big data and that can work in tandem with the existing Data Warehouse system.

There are three different approaches for Data Warehouse modernization, as shown in Figure 4-4.

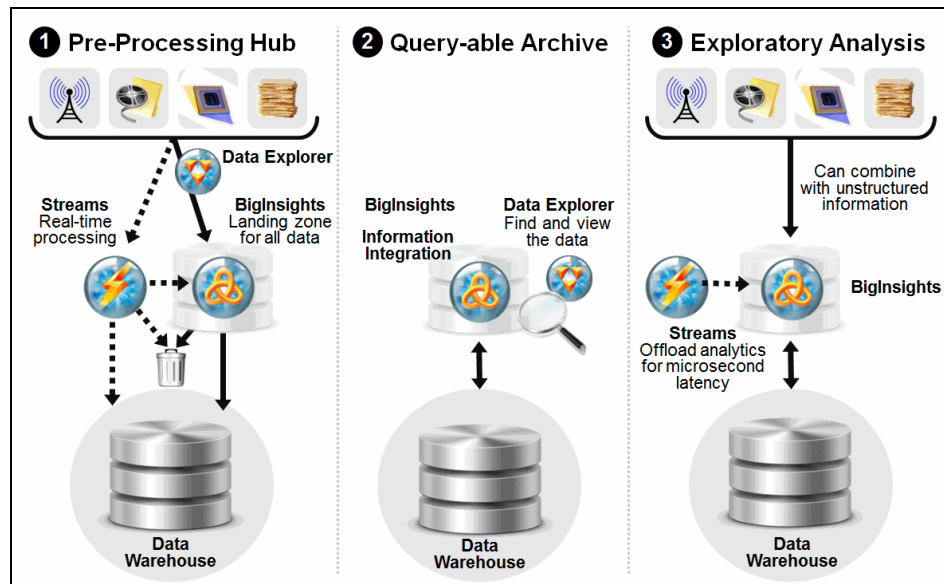


Figure 4-4 Data Warehouse modernization scenarios

The approaches that are adopted are not simply about acquiring and storing new data sources; they are also about organizations offloading rarely used or historical data to adjunct technologies where they can still readily access, query, explore, and analyze the data but keep the storage costs lower.

#### 4.6.1 Pre-processing hub

A pre-processing hub positions an enterprise-grade Hadoop capability as a staging area or landing zone for data before determining what data should be moved to the Data Warehouse. Early exploration might be needed to determine what data you want to move to run deeper analytics or store the data with lower cost. This is not a required step, but can be used in areas where organizations want to leave some of their data at rest.

Stream computing can also be used a real-time component by processing and analyzing streaming data, without having to store it first, and determining what data should be saved (either in Hadoop Distributed File System (HDFS) or the Data Warehouse). In some cases, data does need to be saved; being able to process and act on information as it is happening can also reduce storage in the warehouse. With this landing zone approach, data can be cleansed and transformed before loading it in to the Data Warehouse.

### 4.6.2 Queryable archive

In the queryable archive approach, infrequently accessed or aged data can be offloaded from the warehouse and application databases by using information integration software and tools. The offloaded data can then be federated with data still in the Data Warehouse for queries and reports.

Data federation involves virtually consolidating data from multiple sources, making them appear as a single data source to the user. Data federation enables the user to access data anywhere in the enterprise, regardless of its format or vendor. The complexities that are typically associated with querying data from multiple sources (such as database type, schema, or structural differences) are hidden from the user.

### 4.6.3 Exploratory analysis

In the exploratory analysis approach, stream computing enables analytics on data in motion, often giving organizations the ability to perform analytics that might have previously been done in the warehouse, therefore optimizing the warehouse and enabling new types of analysis. Different data types can be combined (structured, unstructured, and streaming) with warehouse data, enabling deep analytics to provide insights that were not previously possible. Stream computing can act as an analysis filter to find the high-value nuggets of data that then can be stored in BigInsights or the Data Warehouse.

Data Warehouse modernization builds on an existing Data Warehouse infrastructure, using big data technologies to increase its value. This use case allows organizations to do the following tasks:

- ▶ Add new sources to existing Data Warehouse investments
- ▶ Optimize storage and provide queryable archives
- ▶ Rationalize data for greater simplicity and lower cost
- ▶ Increase analytics usage across global operations
- ▶ Enable complex analytical applications with faster queries
- ▶ Scale predictive analytics and business intelligence

- ▶ Improve Data Warehouse and reporting capabilities and performance
- ▶ Improve scalability of the Data Warehouse while lowering costs and simplifying the structure

Organizations are using this approach to support global growth initiatives with improved business performance visibility.

Overall, this approach is about an organization optimizing and using their warehouse infrastructure to support the volumes of newly emerging and existing historical data to handle the range of data formats without requiring transformation of every source, and to still allow the range of ad hoc queries, reporting, and analysis that can drive new insights and value.



## Big data reference architecture

As organizations take advantage of emerging big data use cases, they need optimal architectures to handle the data volume, velocity, and variety, and information governance. To understand how organizational architectures must evolve, it is useful to look at the impact to the traditional information reference architectures that are in place and see how these reference architectures are changing with both new technologies and new concepts.

This chapter reviews the traditional information reference architecture and the emerging big data reference architecture.

## 5.1 Traditional information landscape

The traditional information architecture was designed originally to address transaction processing. It then expanded to enhance reporting around those historical transactions, and then grew to provide a broader view of core master data (people or organizations, such as customers, employees, patients, vendors, and products), and throughout its history has enriched reporting with more analytical insights. In this traditional information landscape, operational data is collected into centralized or distributed Data Warehouses, and then provided downstream to data marts, and analytical and reporting engines.

A traditional information landscape typically focuses on delivering application independent, trusted insight to the business, as shown in Figure 5-1. Information analysis is performed on the data at-hand from the operational processes.

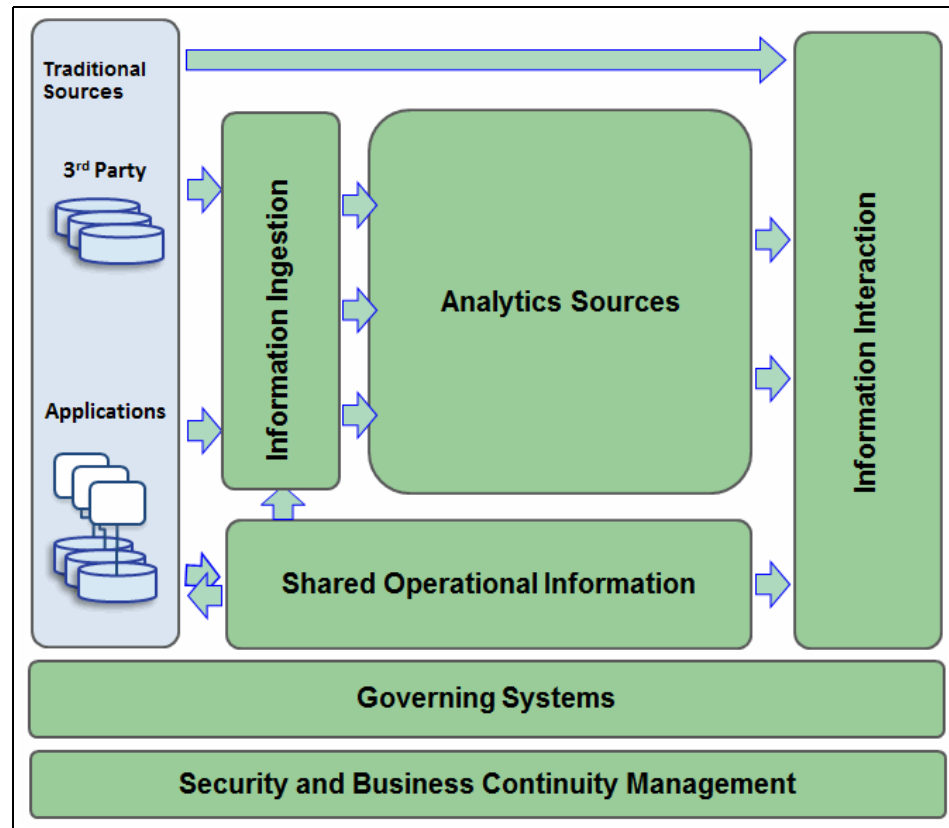


Figure 5-1 Traditional information reference architecture



In this view, you see the following items:

- ▶ **Functional Areas:** Collections of related subsystems delivering a major IT function (represented by shaded boxes, such as Information Ingestion and Analytics Sources).
- ▶ **Information Provisioning:** Various provisioning mechanisms for locating, retrieving, transforming, and aggregating information from all types of sources and repositories (represented as arrows between the functional areas).

Characteristics of this traditional information landscape include the following functional areas:

- ▶ **Information Ingestion:** Mechanisms for collection and performing initial processing on raw data from many traditional sources. This area includes staging areas and transformation engines that are performing enrichment and restructuring operations on the raw data.
- ▶ **Shared Operational Information:** Application independent operational data in managed repositories, which includes master data, reference data, and content management repositories.
- ▶ **Analytics Sources:** Collections of information from many sources, stretching back in time and managed for analytical operations.
- ▶ **Information Interaction:** Interaction with authoritative, insightful information to improve the operation of the organization.
- ▶ **Governing Systems:** Protecting, managing, and improving information while demonstrating compliance to preferred practices, policy, and regulations.
- ▶ **Security and Business Continuity Management:** Ensuring access to information is only for authorized purposes and that there is the ability to recover from unforeseen circumstances.

At the center of this reference architecture are the Analytics Sources. These sources typically include the following items:

- ▶ *A landing area:* Refers only to the most raw data capture, transformation, and analytics:
  - Captures most of the raw data for filtering and transformations; not usually modeled relationally at this point.
  - Supports analytics and highly cost-effective and highly parallel data transformations.
  - Provides for analytics through SQL and non-SQL ad hoc and static analysis.

- ▶ *An operational data store (ODS)*: Refers to integrated, subject-oriented data that is current but volatile and used for tactical analysis:
  - Serves as an interim area for a Data Warehouse with current and recent data only. Almost always on well-modeled relational tables.
  - Stores time-sensitive operational data that needs to be accessed efficiently.
  - Provides support for both simple queries along with complex reporting.
- ▶ *A Data Warehouse (DW)*: Refers to a modeled, integrated data environment crossing and linking multiple subject-oriented content over time. Less volatile than an ODS.
  - Depending on organizational size and maturity, might exist across enterprise (Enterprise Data Warehouse (EDW)) or divisions (lines of business) or might be dimensional data marts.
  - Contains both current and historical data.
  - Supports complex queries over large volumes of data.
  - Provides for ad hoc queries, reporting, and mining. OLAP analysis is common.

To reflect these components in the reference architecture, add the following components to the view of the Functional Areas and Information Provisioning:

- ▶ **Technical Capabilities**: Specialized types of technology performing specific roles (white boxes, such as Transformation Engines, Master Data Hubs, User Reports, and Dashboards).
- ▶ **Zones**: The scope of concern for a particular cross-cutting service (boxes with dashed borders). A zone has associated requirements and governance that any system placed in the zone must adhere to.

The Analytic Sources functional area can now be described or organized from the perspective of these defined zones for specialized processing, as shown in Figure 5-2. An information zone defines a collection of systems where information is used and managed in a specific way.

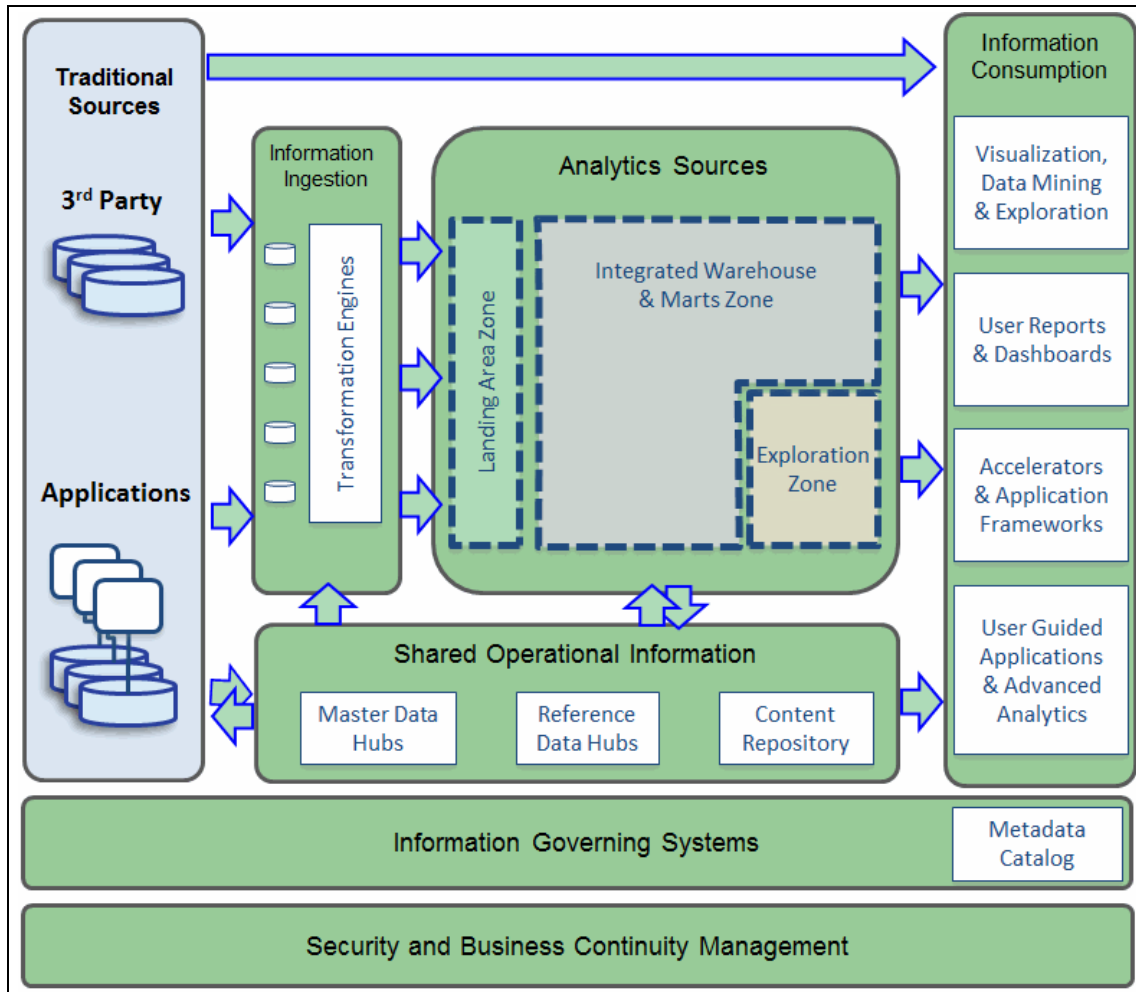


Figure 5-2 Information Zones in a traditional information reference architecture

At least at a general level, data enters the information landscape, is processed, transformed, and integrated along with shared operational information, and is brought into the Analytics environment. At this point, information might go into one or more zones; further mapping and aggregation to suit your analytical needs is possible. In the traditional information landscape, these zones include the following ones:

- ▶ Landing and Analytical Zone:
  - An area for raw data for querying, exploration, data transformations, and pseudo-archival (also known as the online queryable archive).
  - An area that integrates and modernizes data with traditional Integrated Warehouses, Discovery, MDM 360° View, and Content Management.
- ▶ Integrated Data Warehouse and Data Mart Zone (also known as the “Guided Area”):
  - An area for detailed, aggregate, and dimensional data.
  - Includes data models, so there are “guided” analytics because the analytics are performed on known and mapped data schemas.
- ▶ Discovery Zone (also known as the “Exploratory Area”). This is an area for discovery and prediction, including capabilities such as data mining, classification building, and association mapping.

The information in these zones variously supports downstream information consumption, such as dashboards, data mining, or advanced analytics. The zones overlap with one another when the same information is being used for multiple purposes. This approach is necessary when there is a high volume of information, making it uneconomical for each team to have its own private copy of the information. However, in the traditional information landscape, there is often little or no overlap between these information zones.

The traditional information landscape incorporates a logical and fairly linear approach to information creation and delivery:

1. Receive and process operational data.
2. Extract operational data to a staging area or landing zone.
3. Normalize data for enterprise consumability in the Data Warehouse.
4. Provide guided and interactive access through data marts.
5. Deliver data for deeper analysis and modeling in exploratory zones.
6. Archive “cold” data to reduce costs (old, historical data is removed from the Data Warehouse and becomes unavailable in the information landscape).

Enterprise environments adopt tools to support the elements in the above list. In some cases, staging areas and landing zones might be brought into the Data Warehouse to provide faster transformations for downstream analysis, although usually at greater storage cost.

## 5.2 The big data information landscape

As described in Chapter 4, “Big data use cases” on page 43, there is an ever-widening array of data sources entering the organizational information landscape, as shown in Figure 5-3.

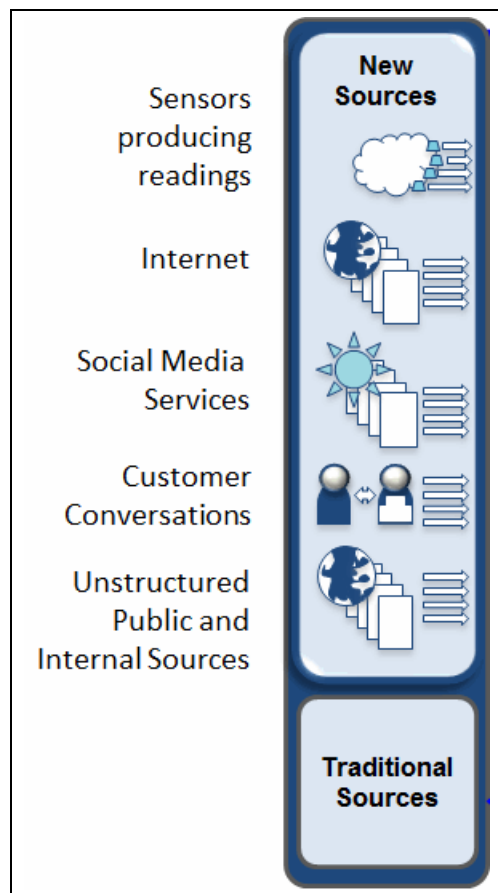


Figure 5-3 Expanding range of data sources

The use cases requesting these new sources require analytical capabilities that are distinctly different from those capabilities that are provided in the traditional information landscape. The traditional information landscape has worked with structured sources in a fairly repeatable and linear fashion. The new, big data paradigm needs to work with far more than the traditional internal data subsets. This paradigm incorporates a large volume of unstructured information, looks for non-obvious correlations that might drive new hypotheses rather than working from expected correlations, and must work with data flowing into the organization in real time that requires real-time analysis and response.

The distinctions between the traditional approach and the new requirements are shown in Figure 5-4.

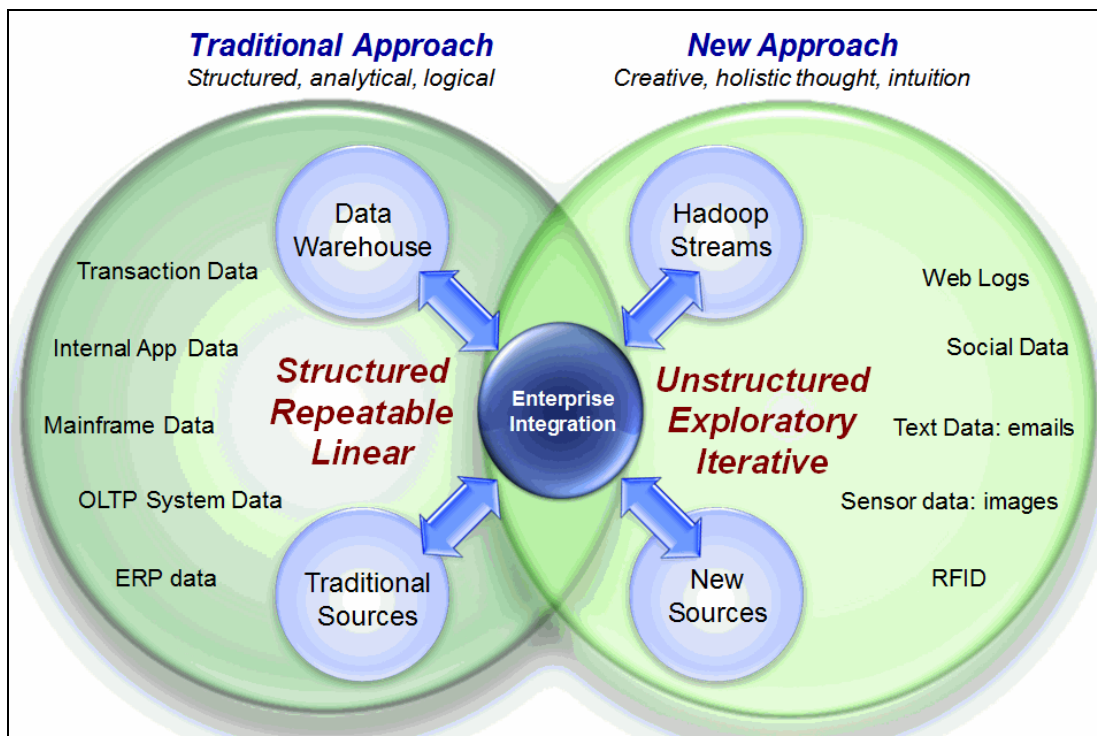


Figure 5-4 New analytical requirements for big data

Information processing in the traditional information landscape focuses on high-value and relatively stable structured data that is brought in through repeated operations and processes (such as transactions, reports, and Business Intelligence) that is optimized for fast access and analysis. Data is cleansed, transformed, and structured before going in to the Data Warehouse to maximize its utility.

However, the new data sources, as you have seen in the big data use cases, are high volume, often unstructured, or highly variable. They might require iterative or non-linear exploration and use. They might not lend themselves to structured models. Raw data may be used or stored as-is without any modification to rapidly respond to content or to preserve fidelity and lineage. Data must be optimized for flexibility where the questions and requirements are ad hoc and ever-changing. The new analytical approach is one of intuitive and exploratory scientific research, which enables a data scientist to contribute to the enterprise in new ways.

These analytical approaches and information landscapes do not compete with each other. In essence, there are two complementary analytical approaches, but they are all big data approaches, and the information landscape must be able to incorporate and integrate both to achieve new business insights.

An organization that wants to augment its existing information with new sources of information needs a focused approach to ensure these outcomes:

New and additional information is relevant to the business requirements when the following items are true:

- ▶ The best sources for the new information are being used.
- ▶ The new information is made available to the business systems and people that need to access the new information.
- ▶ The new information is made available while it is still relevant and useful.

Investigation and exploration of possible data sources is necessary to ensure the optimal use of resources and delivery of information that satisfies the business requirements. Through this investigation, your technical team can build a model of the information that might improve the operations of your organization. Mapping this information to your organization's systems identifies three types of information gaps:

- ▶ Information that is missing from the organization's systems
- ▶ Information that is not of sufficient quality for the decision makers
- ▶ Information that is available in a system other than the one that is used by the decision-makers

Filling these information gaps requires a focus on three activities:

1. Agreeing on the information strategy for the organization

This activity describes how to provide the type and quality of information that is needed to support the organization's business strategy. The information strategy must cover three types of work:

- a. Exploratory investigation of patterns in data to understand the potential cause and effect of proposed activity.
- b. Projects that change IT infrastructure or business processes.
- c. Everyday business operations that run continuously.

2. Developing an evolving information landscape

This activity explains how the IT infrastructure can be transformed over several iterations to deliver the information strategy:

- a. Incorporating new sources of information into your organization's systems. Often, this involves deploying new systems to store this information.
- b. Using techniques such as master data management (MDM) and content management to create consolidated sources of information for operational use.
- c. Connecting systems together, so information can flow to the places where it is needed.

3. Operating an information governance program

This activity actively controls and monitors how information is being used and managed.

The information strategy must take a holistic approach to the management of information. It requires a “system of systems” style of architecture where each system has a well-defined role, both in serving its direct users and also being a part of the management and flow of information across the organization.

The information landscape must evolve to address both these new and rapidly changing data sources and the business needs, requirements, and decisions for which an organization wants to use these data sources.

Figure 5-5 on page 79 shows an architectural approach to augmenting existing systems with new data sources to support new information requirements that are needed by the organization. This information architectural approach has three goals:

1. Combine the best information from across the organization for use by both operational and analytical systems.



2. Augment this consolidated information with relevant facts and event triggers by analyzing previously untapped sources of information.
3. Help more easily locate information and deliver it to where the business most needs it.

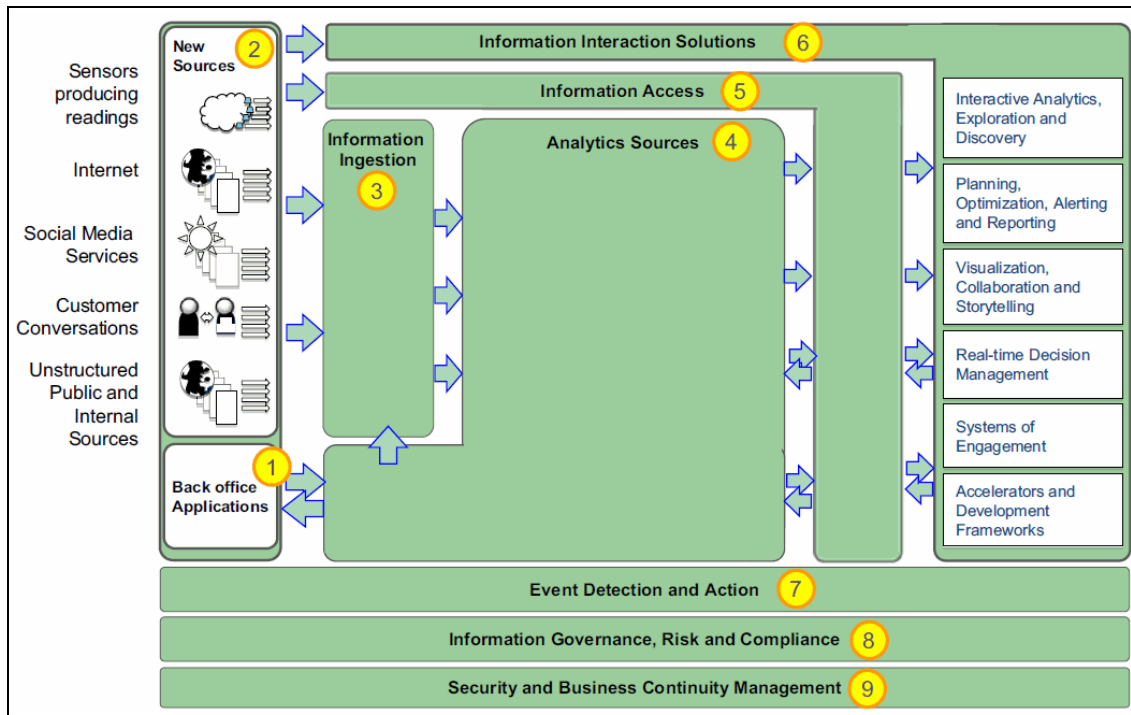


Figure 5-5 The new big data information landscape

Here are the description of the functional areas that are shown and numbered in Figure 5-5:

► Back-office Applications (Area 1)

Back-office applications run and process the business transactions. These systems include the order processing, billing, marketing, product development, and sales types of systems. In general, they operate independently using their own data, but they can exchange information with the operational systems in the analytics sources.

► New Sources (Area 2)

New sources are systems that supply information to augment the information that is generated by the back-office applications. These sources might come from third parties, or might be internal sources, such as logs, emails, and other text-based sources that were too expensive to process in the past.

► Information Ingestion (Area 3)

Information ingestion collects together raw information from the back-office applications and new sources for validation, correlation, cleansing, and transformation.

► Analytic Sources (Area 4)

Analytics sources provide the information for different types of analytics processing:

- Some of this analytics processing occurs inside of the systems hosting the analytics sources.
- Some analytics processing occurs in the provisioning engines as information is moved between the analytics sources.
- Some analytics processing occurs in the information interaction systems.

The analytics sources in the new information landscape include shared operational systems, such as master data hubs, reference data hubs, activity data hubs, and content management hubs, and systems such as Data Warehouses, MapReduce (Hadoop), files, databases, and data marts that host historical information that is harvested from many sources.

► Information Access (Area 5)

Information access enables the information interaction solutions to locate and consume information without needing to understand exactly where it is physically stored and maintained. This capability is part of an approach that is called *information virtualization* and is a key addition to the new information landscape.

► Information Interaction (Area 6)

Information interaction solutions include the systems of engagement and advanced analytics capability. They predominantly work with information drawn from the analytics sources. Information interaction solutions can produce some information that is fed back into the analytics sources.

► Event Detection and Action (Area 7)

Event detection and action triggers business processes and other activity when significant events are detected in the applications, analytic sources, and information interaction solutions. This is another key addition to the new information landscape.

► Information Governance, Risk, and Compliance (Area 8)

Information governance, risk, and compliance provide the capability to ensure that information is appropriately managed throughout its lifecycle with the level of quality and protection that is consistent with the information's sensitivity and use.

► Security and Business Continuity (Area 9)

Security and business continuity management ensures that all systems have the security and availability that are appropriate to their importance to the business.

The shape of the new information landscape suggests some changes are taking place in the architecture. If you drill further into the zones that comprise the central areas (Information Ingestion and Analytic Sources), you see more of the detail of this evolution, as shown in Figure 5-6.

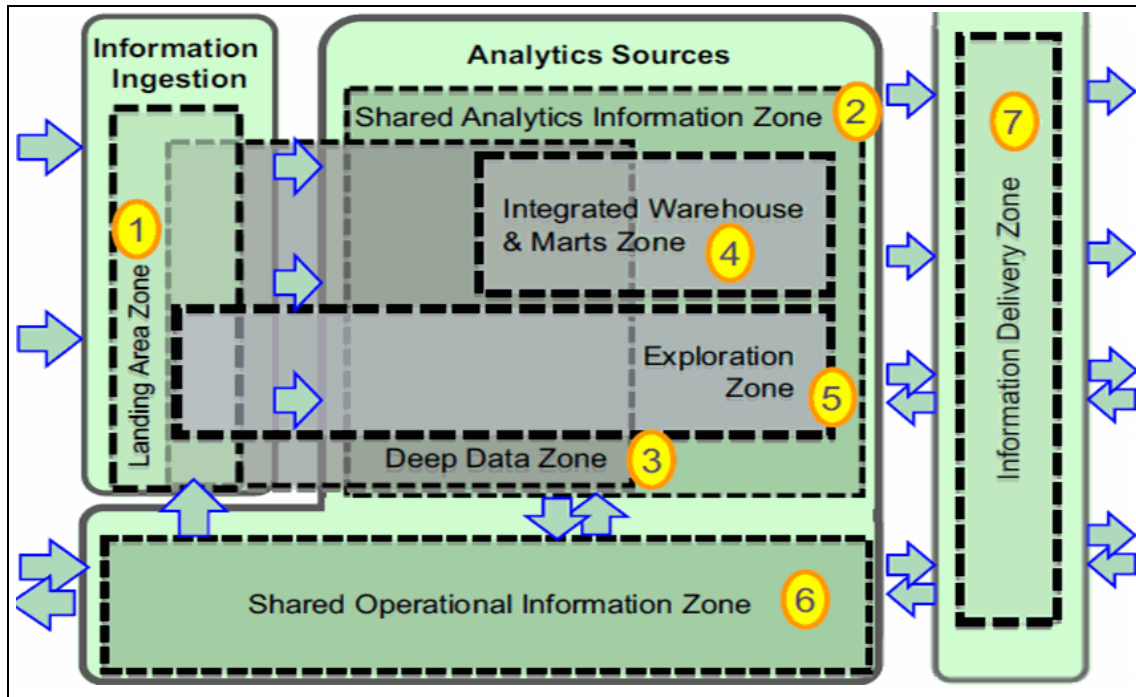


Figure 5-6 Information Zones in the new big data information reference architecture

Here are the descriptions of the information zones that are shown in Figure 5-6:

► Landing Area Zone (Zone 1)

The Landing Area Zone manages raw data that is received from the applications and the new sources. This data had minimal verification and reformatting that is performed on it.

► Shared Analytics Information Zone (Zone 2)

The Shared Analytics Information Zone contains information that was harvested for reporting and analytics.

► Deep Data Zone (Zone 3)

The Deep Data Zone contains detailed information that is used by analytics to create new insight and summaries for the business. This data is kept for some time after the analytics processing completes to enable detailed investigation of the original facts if the analytics processing discovers unexpected values.

► Integrated Warehouse and Marts Zone (Zone 4)

The Integrated Warehouse and Marts Zone contains consolidated and summarized historical information that is managed for reporting and analytics.

► Exploration Zone (Zone 5)

The Exploration Zone provides the data that is used for exploratory analytics. Exploratory analytics uses a wide variety of raw data and managed information.

► The Shared Operational Information Zone (Zone 6)

The Shared Operational Information Zone has systems that contain consolidated operational information that is being shared by multiple systems. This zone includes the master data hubs, content hubs, reference data hubs, and activity data hubs.

► Information Delivery Zone (Zone 7)

The Information Delivery Zone contains information that was prepared for use by the information interaction solutions. The Information Delivery Zone typically contains read only information that is regularly refreshed to support the needs of the systems using it. It provides some of the authoritative information sources that are used in information virtualization where the original source of information is not suitable for direct access.

Figure 5-7 on page 83 summarizes some of the distinct characteristics of each of these information zones.

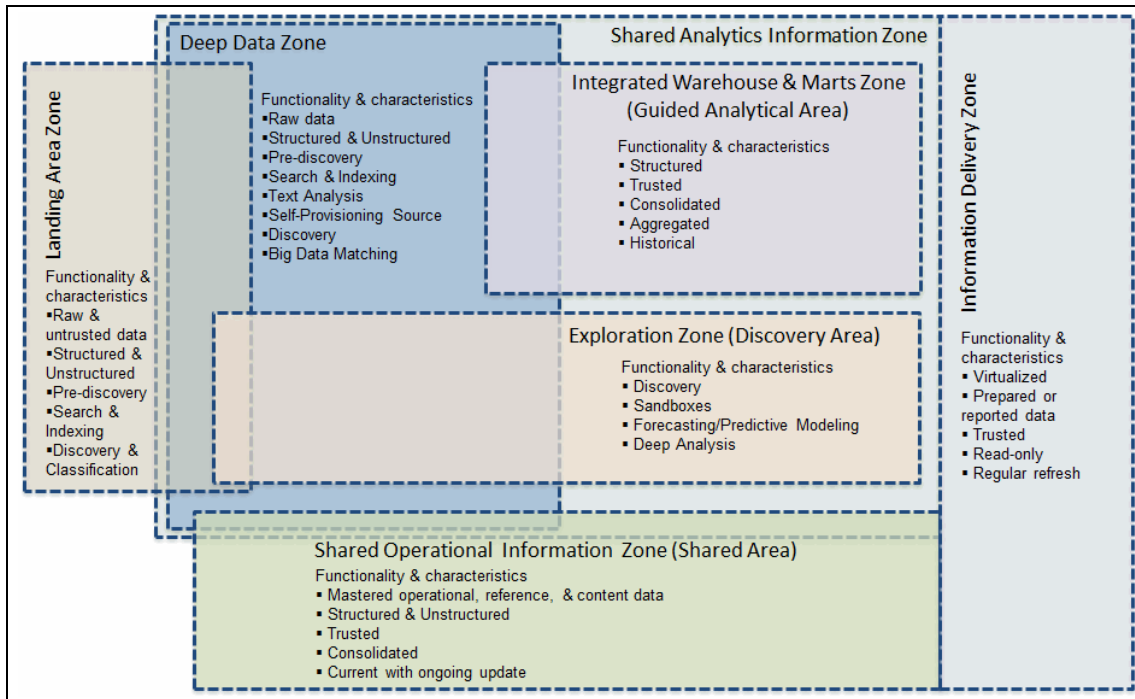


Figure 5-7 Characteristics of Big Data Information Zones

The Big Data Information Zones tie in to the use cases that you examined in Chapter 4, “Big data use cases” on page 43. Big Data Exploration, for example, brings new sources of data in to the organization, typically into the Landing Zone. There, the data can be initially examined and inspected to ensure that it does not violate information governance policies, or contain harmful code or simply useless data. After the data is past this initial evaluation, the data may be moved into the Exploration Zone for further evaluation, such as sandbox testing against business hypotheses. Sources that prove to be of ongoing use are brought into the information supply chain as part of production processes. Depending on their data content and use, they might flow into the Deep Data Zone or into the Shared Operational Information Zone, or be aggregated into the Integrated Warehouse and Marts Zone.

To move information in to and through these information zones, you must also consider the systems that support the zones as part of the new information landscape. In some cases, these systems fall distinctly within one zone, but in others these systems and their capabilities span across multiple zones, as shown in Figure 5-8.

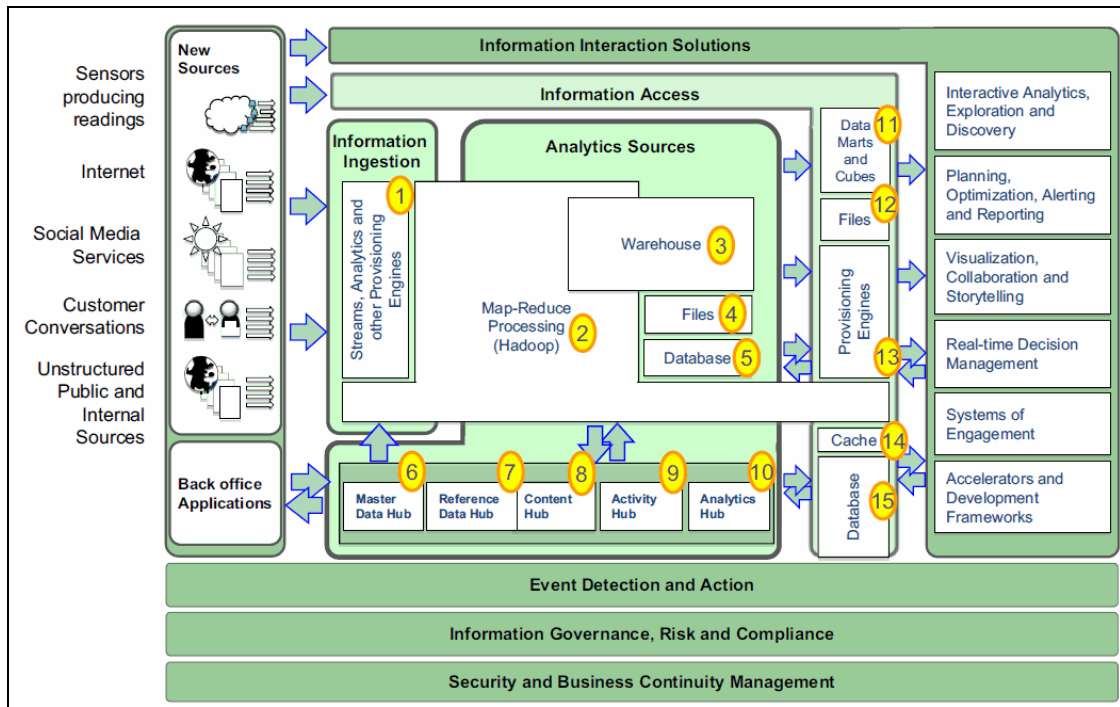


Figure 5-8 Systems in the new big data information landscape

Here are the descriptions of the systems that are shown in Figure 5-8:

► Streams, Analytics and Other Provisioning Engines (System 1)

Streams, Analytics and Other Provisioning Engines take raw data and perform initial verification, correlation, consolidation, and transformation to make the information consistent, which simplifies subsequent processing. The streams provisioning engine is important when receiving information from high-speed feeds that arrives too quickly to store. A streams engine can extract important facts to pass on to the other systems.

- ▶ MapReduce Processing (Hadoop) (System 2)

MapReduce Processing (Hadoop) provides a flexible storage system that can hold data in many formats. Schemas and other forms of annotations can be mapped on to the data after it is stored, allowing it to be used for multiple purposes.
- ▶ Warehouses (System 3)

Warehouses are still needed to provide efficient access to consolidated, aggregated, and reconciled information for analytics, reports, and dashboards.
- ▶ Files (System 4)

Files are used for many purposes, particularly for moving large amounts of information.
- ▶ Databases (System 5)

Databases offer structured storage for real-time access.
- ▶ Master Data Hubs (System 6)

Master Data Hubs consolidate and manage master data such as customers, suppliers, products, accounts, and assets, as part of a master data management (MDM) program.
- ▶ Reference Data Hubs (System 7)

Reference Data Hubs manage code tables and hierarchies of values that are used to transform and correlate information from different sources.
- ▶ Content Hubs (System 8)

Content Hubs manage documents and other media, such as image, video, and audio files that must be controlled and managed through formal processes.
- ▶ Activity Data Hubs (System 9)

Activity Data Hubs manage consolidated information about recent activity, including analytical decisions that are related to the entities in the Master Data and Content Hubs. This type of data is normally managed by applications. However, because the organization is processing information about activity from outside of the scope of its applications, a new type of hub is needed to manage the dynamic nature of this type of information.
- ▶ Analytics Hubs (System 10)

Analytics Hubs appear among the operational hubs to create additional insight in real time. These hubs are supporting advanced analytics, such as predictive analytics and optimization, along with business rules.

- ▶ **Data Marts and Cubes (System 11)**  
Data Marts and Cubes still provide subsets of information that are formatted for a particular team or style of processing. They are read-only copies of information that are regularly refreshed from the analytical sources.
- ▶ **Files (System 12)**  
Files are still used as input to certain analytics processing, particularly data mining.
- ▶ **Provisioning Engines (System 13)**  
Provisioning Engines are either serving information in real time to the information interaction solutions, or creating and maintaining specialized stores of information in the Information Delivery Zone, which are used by the information interaction solutions.
- ▶ **Cache (System 14)**  
Cache represents in-memory stores providing fast access to read-only information for the information interaction solutions.
- ▶ **Databases (System 15)**  
Databases (in this context) can provide real-time storage for information that is created by the information interaction solutions.

Collectively, these systems supply authoritative information to the information interaction solutions whether they are analytics, reporting, decision management, or other downstream capabilities. Authoritative sources are maintained by using a deterministic and governed information supply chain that needs to synchronize the authoritative sources with other copies of the information. Each system that is shown in Figure 5-8 on page 84 is responsible for implementing part of the information supply chain.

## 5.2.1 Capabilities for the new information landscape

The big data information landscape not only must expand to incorporate both traditional and big data sources, but it also requires a broader set of capabilities to support the consumption, processing, and delivery of information. Capabilities describe a function, which is something like a set of useful parts that are needed together to create solutions. An information landscape for big data requires capabilities in seven broad categories, as shown in Figure 5-9 on page 87.



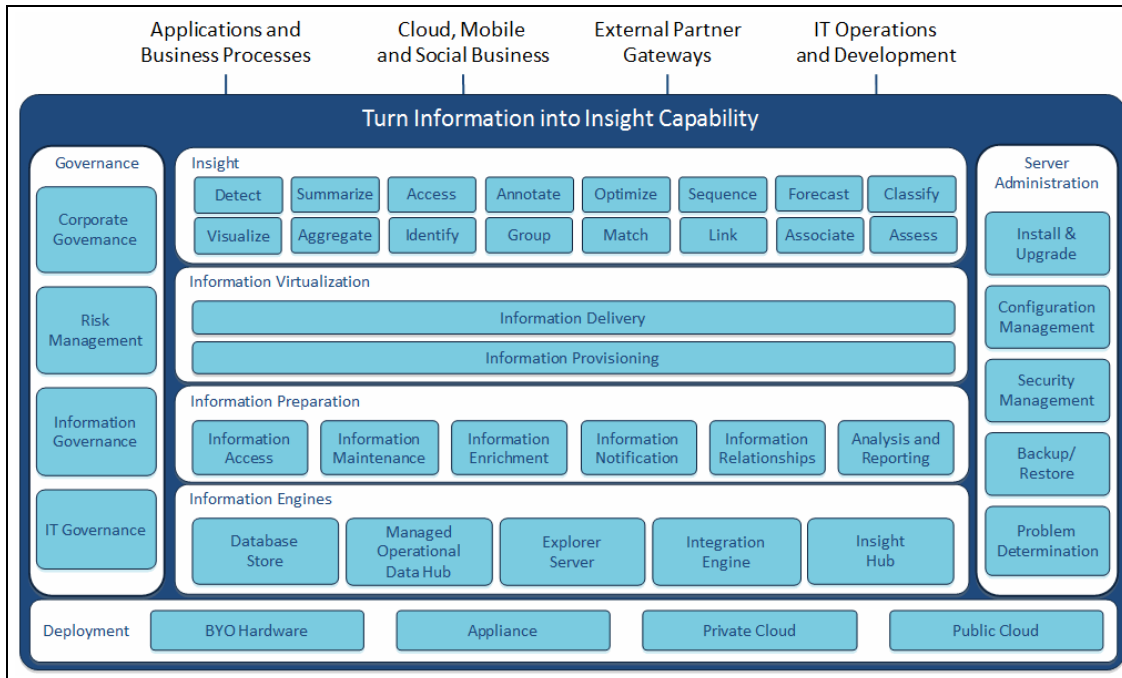


Figure 5-9 Capabilities for the big data information architecture

Here are the seven capability categories and their components:

- ▶ Insight
- ▶ Information Virtualization
- ▶ Information Preparation
- ▶ Information Engines
- ▶ Deployment
- ▶ Governance
- ▶ Server Administration

Here are descriptions of these categories and their components:

- Insight: Added value services that provide additional insight over the raw data.

The Insight capabilities, which are shown in Figure 5-10, are provided by a big data platform to extract value from the information that is stored.

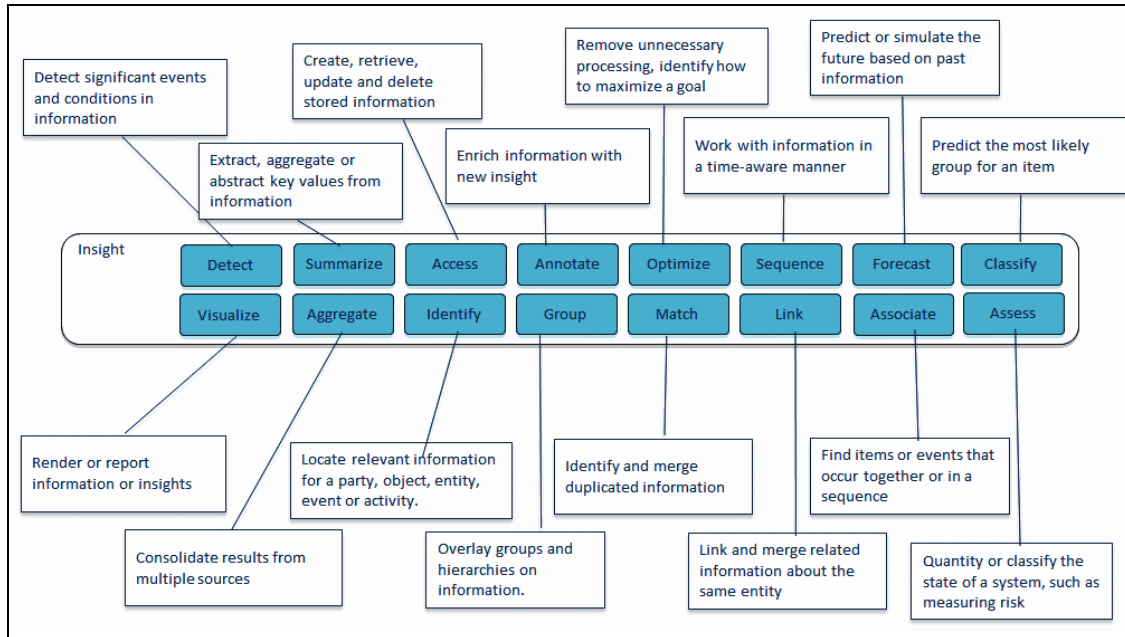


Figure 5-10 Insight capabilities for the big data information architecture

Here are descriptions of the capabilities:

- Detect: Detect significant events and conditions in information.
- Summarize: Extract key values from information.
- Access: Create, retrieve, update, and delete stored information.
- Annotate: Enrich information with new insight.
- Optimize: Remove unnecessary processing.
- Sequence: Work with information in a time-aware manner.
- Forecast: Predict or simulate the future based on past information.
- Classify: Predict the most likely group for an item.
- Render: Graphically display information to a user.
- Aggregate: Consolidate results from multiple sources.
- Identify: Locate relevant information for a party, object, event, or activity.

- Group: Overlay groups and hierarchies on information.
  - Match: Identify and merge duplicated information.
  - Link: Link together related information.
  - Associate: Find items or events that occur together or in a sequence.
  - Assess: Quantify or classify the state of a system, such as measuring risk.
- Information Virtualization: Mechanisms that access, reformat, collate, and copy information to the location where it is needed by the organization.

The Information Virtualization capabilities, which are shown in Figure 5-11, provide access to information through well-defined interfaces. The information is typically dispersed among heterogeneous data stores. Information Virtualization provides two capabilities: It delivers this information in a coherent view as though it were a single data source, and it provisions information to the consumer as though the information is local.

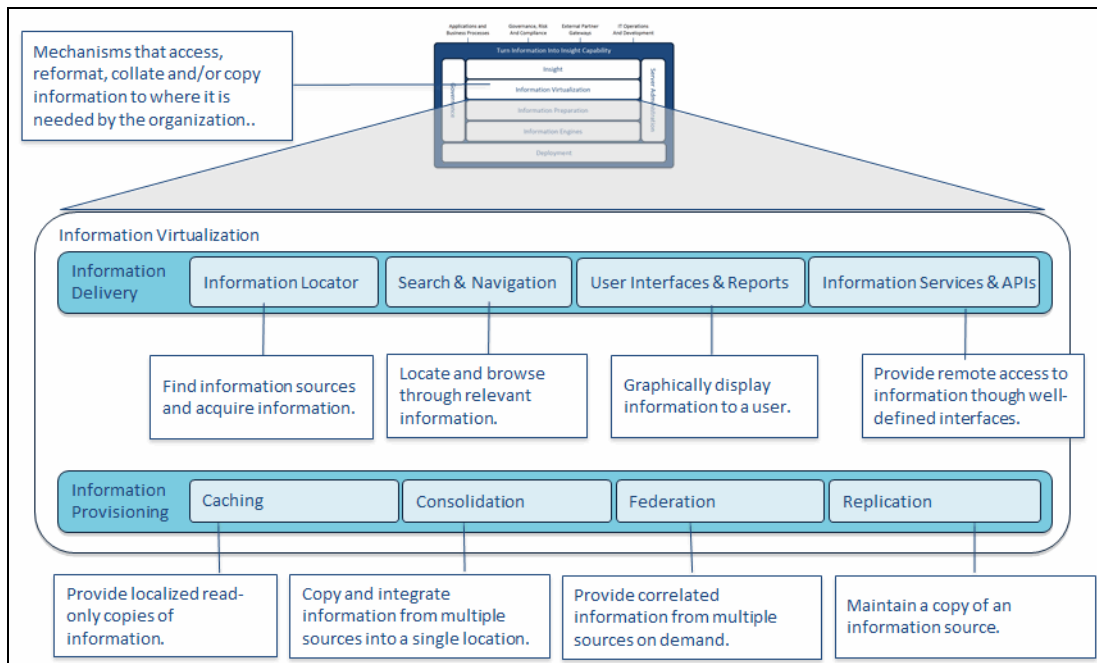


Figure 5-11 Information Virtualization capabilities for the big data information architecture

Here are descriptions of the capabilities:

- Information Locator: Find information sources and acquire information.
- Search and Navigation: Locate and browse through relevant information.
- User Interfaces and Reports: Graphically display information to a user.

- Information Services and APIs: Provide remote access to information through well-defined interfaces.

Provisioning capabilities include the following items:

- Caching: Provide localized read-only copies of information.
- Consolidation: Copy and integrate information from multiple sources in to a single location.
- Federation: Provide correlated information from multiple sources on demand.
- Replication: Maintain a copy of an information source.
- Information Preparation: Mechanisms to maintain and prepare information for effective consumption.

The Information Preparation capabilities, which are shown in Figure 5-12, describe how information is transformed and improved within a big data platform.

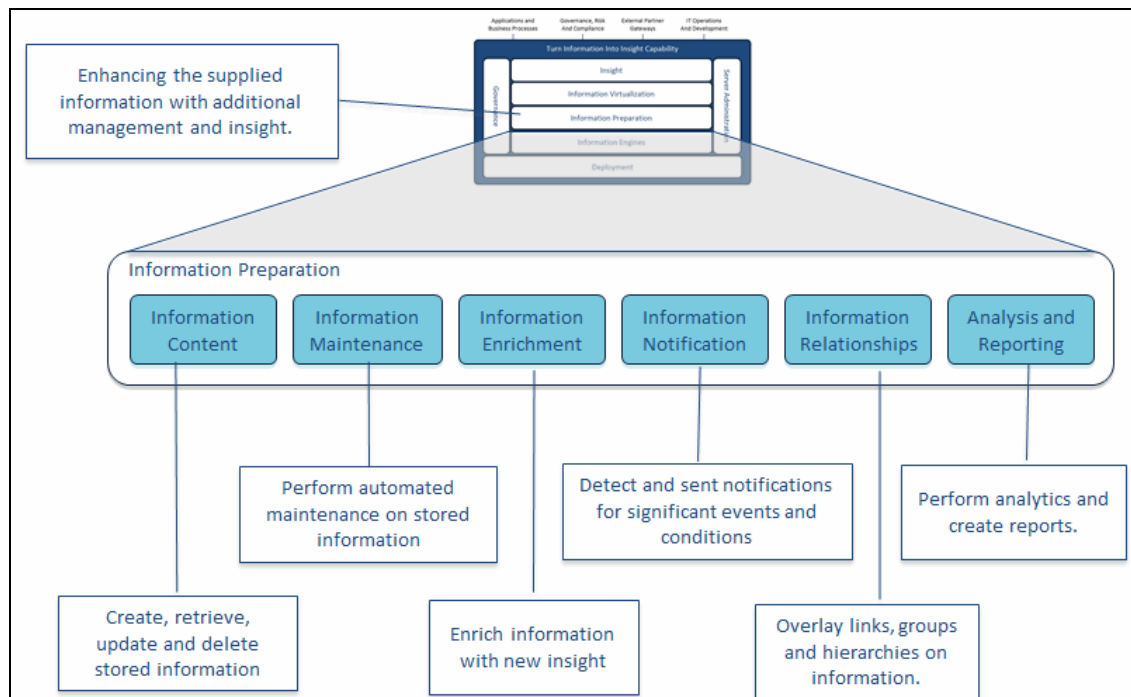


Figure 5-12 Information Preparation capabilities for the big data information architecture

Here are descriptions of the capabilities:

- Information Rationalization: Reorganize stored information for future processing.
- Information Maintenance: Perform automated maintenance on stored information.
- Information Enrichment: Enrich information with new insight.
- Information Notification: Detect and send notifications for significant events and conditions.
- Information Relationships: Overlay links, groups, and hierarchies on information.
- Analysis and Reporting: Perform analytics and create reports.

- Information Engines: Specialist servers and engines for processing and delivering information.

The Information Engines capabilities, shown in Figure 5-13, are the mechanisms that manage information. The Information Engines, working together, enable the Information Virtualization and Information Preparation capabilities. They are how information is managed, whereas the Information Virtualization and Information Preparation capabilities reflect the direct business value that the information enables.

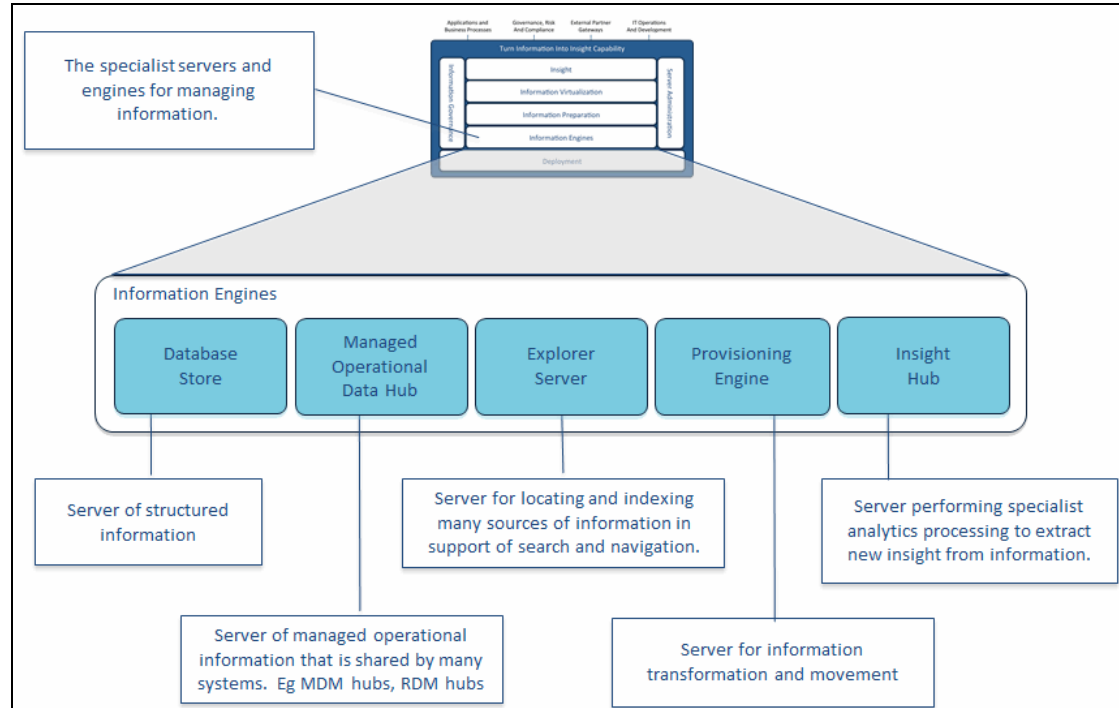


Figure 5-13 Information Engines capabilities for the big data information architecture

Here are descriptions of the capabilities:

- Database Store: A server for managing collections of structured information, such as an Online Transactional Processing database, Operational Data Store, Data Warehouse, and Online Analytics Processing Cubes.
- Managed Operational Data Hub: A server of managed operational information that is shared by many systems, such as Master and Reference data hubs and the Metadata catalog.

- Explorer Server: A server for locating and indexing many sources of information in support of search and navigation.
- Provisioning Engine: A server for information transformation and movement that includes Consolidation, Federation, Messaging, Replication, and Streaming engines.
- Insight Server: A server performing specialist analytics processing to extract new insight from information, such as Analytics, MapReduce, and Matching engines.
- Deployment: Approaches for managing the infrastructure that supports the big data capabilities.

The Deployment capabilities, which are shown in Figure 5-14, describe the approaches to hosting Information Management capability.

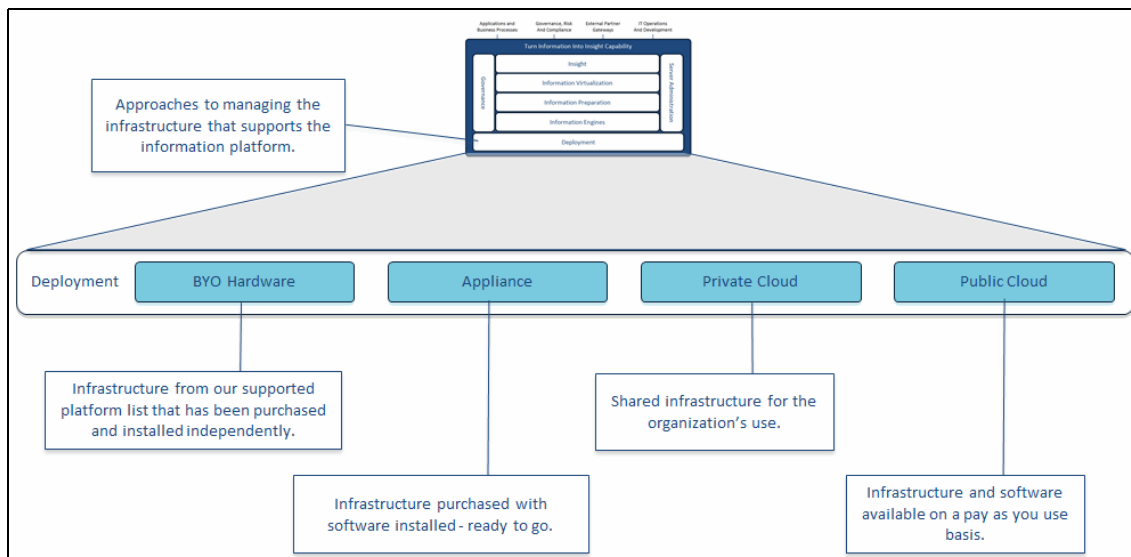


Figure 5-14 Deployment capabilities for the big data information architecture

Here are descriptions of the capabilities:

- BYO Hardware: Infrastructure from your supported platform list that was purchased and installed independently.
- Appliance: Infrastructure that is purchased with software installed and ready to deploy.
- Private Cloud: Shared infrastructure for the organization's use.
- Public Cloud: Infrastructure and software that is available on a pay-as-you-use basis.

- Governance: Capabilities for protecting, managing, and improving the operation of the organization.

Governance capabilities, which are shown in Figure 5-15, provides the means by which an organization ensures that it meets its obligations. These obligations might be to legal authorities, shareholders, customers, suppliers, and employees.

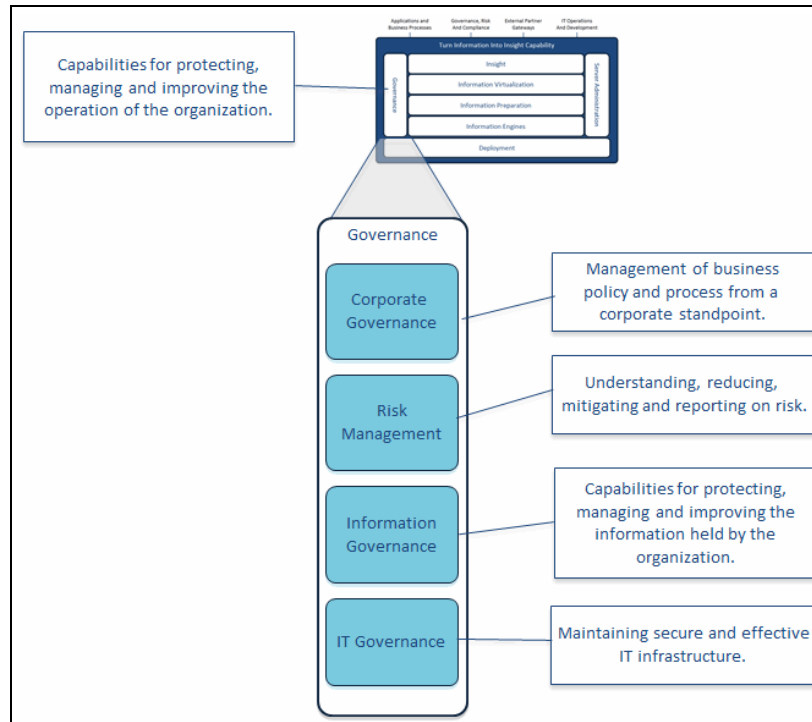


Figure 5-15 Governance capabilities for the big data information architecture

There are four subdisciplines of governance:

- Corporate Governance: Management of business policy and processes from a corporate standpoint.
- Risk Management: Understanding, reducing, mitigating, and reporting on risk.
- Information Governance: Capabilities for protecting, managing, and improving the information that held by the organization.
- IT Governance: Maintaining secure and effective IT Infrastructure.



- **Server Administration:** Services for managing a big data platform infrastructure (hardware and software resources).

The Server Administration capabilities, which are shown in Figure 5-16, are found in every Information Management product that stores data. Additional products exist that aggregate the management of servers (nodes) to reduce the total cost of ownership (TCO) for large enterprises.

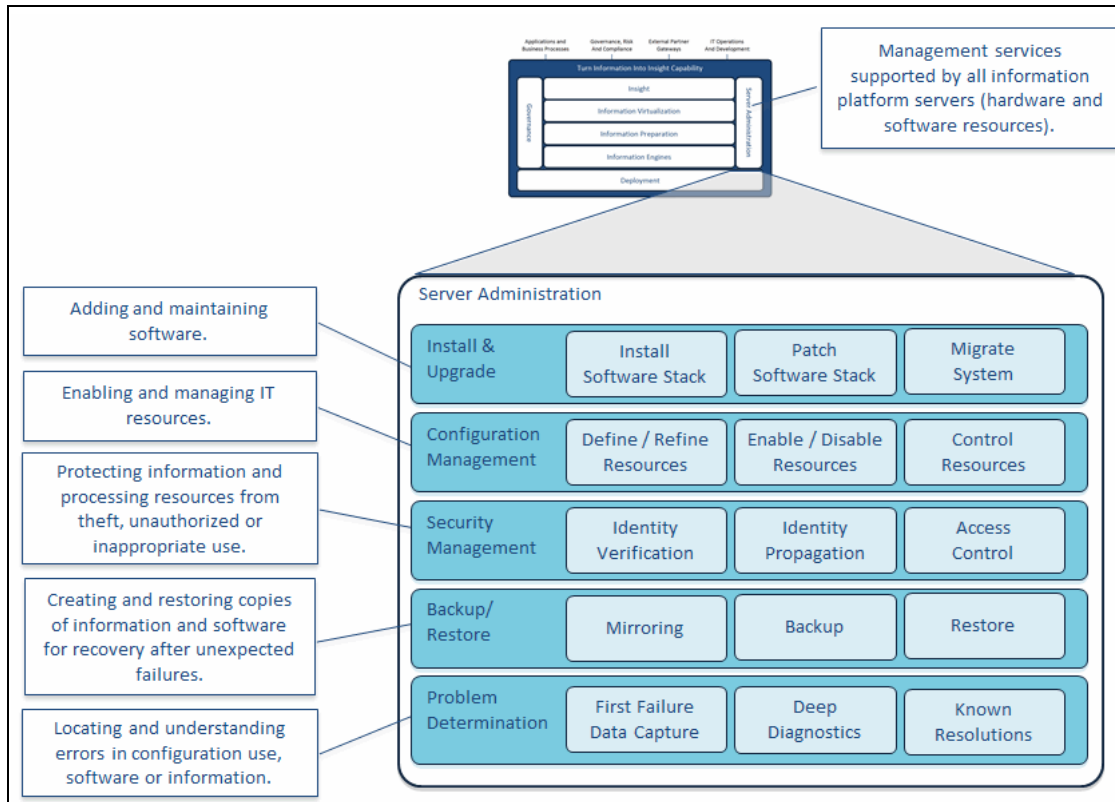


Figure 5-16 Server Administration capabilities for the big data information architecture

Here are descriptions of the capabilities:

- **Install and Upgrade:** Adding and maintaining software.
- **Configuration Management:** Enabling and managing IT resources.
- **Security Management:** Protecting information and processing resources from theft, or unauthorized or inappropriate use.

- Backup / Restore: Creating and restoring copies of information and software for recovery after unexpected failures.
- Problem Determination: Locating and understanding errors in configuration use, software, or information.

When these systems and capabilities are brought together and connected across the new big data information landscape, there is an increased obligation to manage and govern the way that information is used in the organization. For example, bad information added in one system has the opportunity to contaminate other systems and has a greater impact than ever before.

Chapter 6, “Introduction to the IBM Big Data Platform” on page 97 looks at the IBM Big Data Platform that supports the new information landscape. Subsequent chapters consider some of the critical information governance dimensions regarding the big data sources and how the IBM Big Data Platform can be used to address these dimensions.



# Introduction to the IBM Big Data Platform

Looking at the range of big data and the emerging use cases that are associated with it, organizations are using the wealth of new information to drive new analytical applications and solutions. It is these targeted applications and solutions that drive the requirements for the IBM Big Data Platform. At a general level, as shown in Figure 6-1 on page 98, a platform to support big data should be able to perform the following functions:

- ▶ Analyze various information
- ▶ Analyze information in motion
- ▶ Analyze extreme volumes of information
- ▶ Discover and experiment
- ▶ Manage, plan, and govern the information

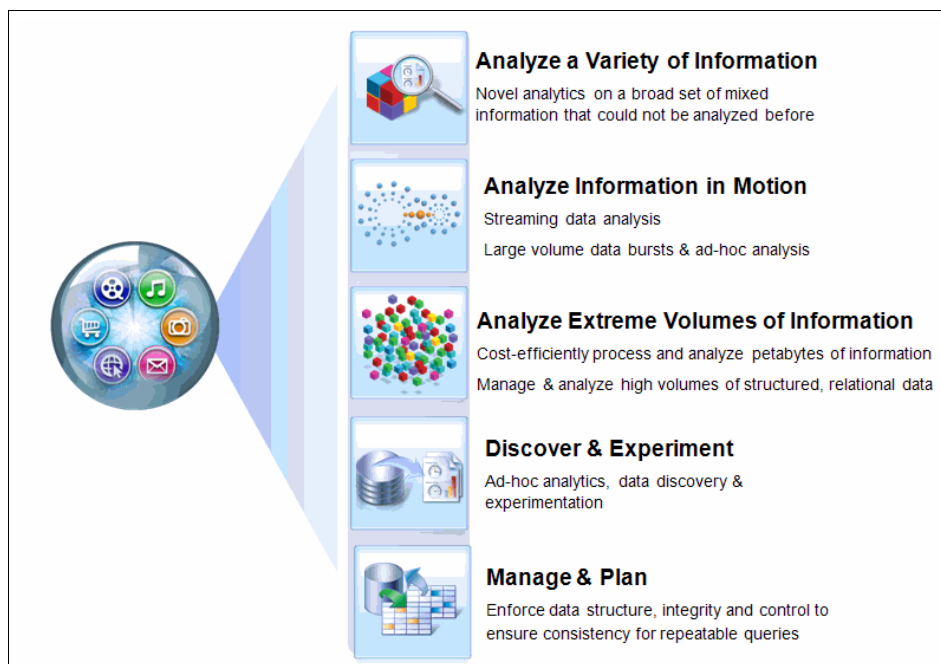


Figure 6-1 Big data platform - wanted capabilities

## Big data platform: Wanted capabilities

A platform for big data does not stand in isolation. As shown in the use cases that are described in Chapter 4, “Big data use cases” on page 43, the variety, volume, and constant delivery of big data serves business needs, whether to increase revenue, optimize processes, reduce cost, or reduce risk. At this broader level, a big data platform fits into an architecture that performs the following functions:

- ▶ Integrate and manage the full variety, velocity, and volume of data.
- ▶ Assess and explore data for potential and effective use.
- ▶ Apply advanced analytics to various information in its native form. Often, novel analytics is used on a broad set of mixed information that could not previously be analyzed.
- ▶ Build new analytic applications to consume the data at multiple points.
- ▶ Visualize all available data for ad hoc or operational analysis.
- ▶ Schedule and optimize the necessary data processing.
- ▶ Secure and govern the overall big data environment.
- ▶ Enforce integrity and control to ensure consistency for repeatable queries.

With these factors in mind, a big data platform, as shown in Figure 6-2, incorporates an underlying infrastructure, a processing platform allowing multiple styles of information computing, integration, and storage, an analytics layer supporting multiple types of analysis and visualization, and solutions that can build upon the entire platform to transform a business around the identified and emerging use cases.



Figure 6-2 A Big data platform in its broader context

IBM provides capabilities throughout these layers. This publication focuses on the big data platform components that process, store, integrate, deliver, accelerate, visualize, and govern information. These focal points are shown in Figure 6-3.

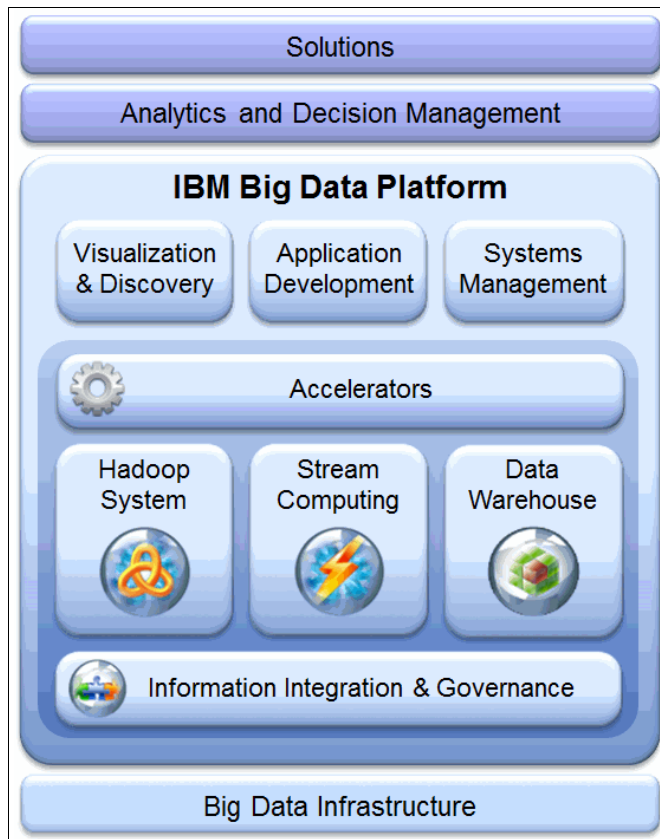


Figure 6-3 The IBM Big Data Platform

## 6.1 Components of the IBM Big Data Platform

The remainder of this chapter introduces you to the IBM products that comprise the IBM Big Data Platform. Specifically, this chapter describes the following items:

- ▶ The Data Warehouse
  - IBM DB2
  - IBM PureData System for Operational Analytics
  - IBM PureData System for Analytics
- ▶ Stream computing: IBM InfoSphere Streams
- ▶ Apache Hadoop and related big data architectures: IBM InfoSphere BigInsights
- ▶ Information Integration and Governance
  - IBM InfoSphere Information Server
  - IBM InfoSphere Data Replication
  - IBM InfoSphere Federation Server
  - IBM InfoSphere Master Data Management
  - IBM InfoSphere Optim
  - IBM InfoSphere Guardium
- ▶ Accelerators
  - IBM Industry Models
  - IBM Accelerators for Big Data
- ▶ Visualization and discovery: IBM InfoSphere Data Explorer

## 6.2 The Data Warehouse

The Data Warehouse is the center point for an organization's core data, whether this data is customers, products, orders, sales, or other historical information. The Data Warehouse supports high-performance analytics on a large volume of data that is structured for control, consistency, and integrity. As shown in Figure 6-4, the Data Warehouse is a primary starting point for big data because it forms the foundation for business decision making and exists in most organizations.

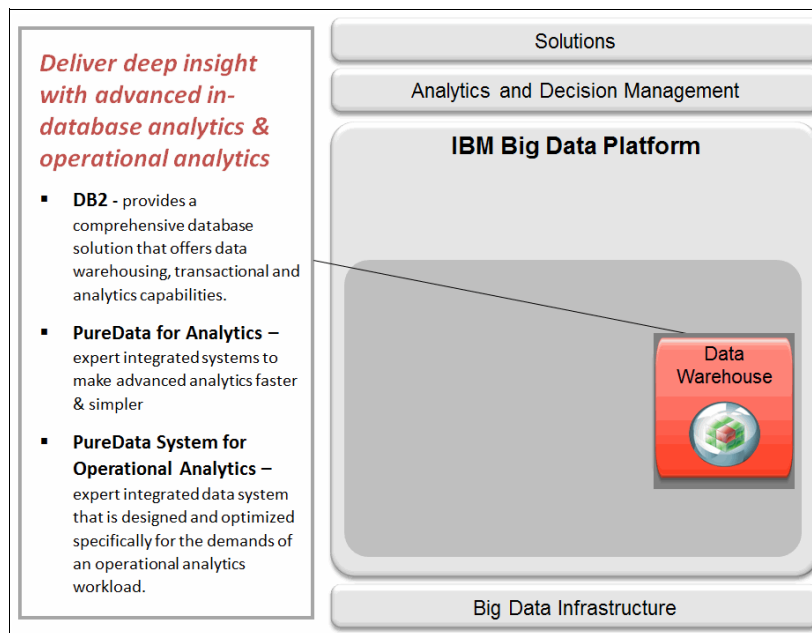


Figure 6-4 Data Warehouse in the IBM Big Data Platform

IBM delivers complementary capabilities that support the range of organizational requirements to store, report, and analyze their data. DB2 provides the foundation for the information warehouse environment. IBM PureData for Operational Analytics scales to address the heavy-duty requirements of an enterprise Data Warehouse while IBM PureData for Analytics provides fast up-and-running environments to offload intensive analytics work.



## 6.2.1 DB2

DB2 provides a comprehensive database solution for the enterprise. It is the ideal multi-workload database solution that offers data warehousing, transactional, and analytics capabilities in one package. DB2 provides storage optimization, highly reliable system availability, workload management, and performance to help reduce overall database costs.

DB2 has the following capabilities:

- ▶ Optimized database storage and performance using integrated technologies, such as columnar storage
- ▶ Speed of thought analytics with BLU Acceleration and sophisticated data warehousing capabilities
- ▶ Faster, more cost-effective queries using advanced data handling features
- ▶ High availability and workload management capabilities for improved throughput and reduced database costs
- ▶ Simplified database administration and development tools for easier data design, modeling, movement, and transformation
- ▶ Data access at your fingertips with mobile synchronization

## 6.2.2 IBM PureData for Operational Analytics

IBM PureData for Operational Analytics provides a powerful range of capabilities that go beyond traditional Data Warehouses. IBM PureData System for Operation Analytics is a Data Warehouse system for delivering insights to business operations for real-time decision making. It is optimized to handle complex analytics and designed to handle 1000+ concurrent operational queries.

Using multidimensional cubing services, IBM PureData System for Operational Analytics delivers rapid insight into high volumes of fast-moving data. Users can create, edit, import, export, and deploy cube models over the relational warehouse schema to analyze multiple business variables. Cubing services help optimize performance for online analytical processing (OLAP) queries, providing more power for users to analyze data and generate business insight that can enhance both profitability and customer satisfaction.

With in-database analytics, you run analytics on your data where it is, that is, the warehouse. This eliminates the time, cost, and risk that is associated with copying data out of the warehouse to analyze it. Using multidimensional cubing services, the IBM PureData System delivers rapid insight into high volumes of fast-moving data. Users can create, edit, import, export, and deploy cube models over the relational warehouse schema to analyze multiple business variables.

### 6.2.3 IBM PureData System for Analytics

With simple deployment and optimization, no tuning, minimal on-going maintenance, and powered by Netezza® technology, IBM PureData System for Analytics is a simple data appliance for serious analytics with fast time-to-value and low total cost of ownership (TCO). It simplifies and optimizes the performance of data services for analytic applications, enabling complex algorithms to run in minutes instead of hours.

The system is running in hours, requires minimal upfront design and tuning, and minimal ongoing administration. It provides standard interfaces to best of breed analytics, business intelligence, and data integration tools, and easy connectivity to other IBM Big Data Platform components. IBM PureData System for Analytics also eliminates the need for complex database management tasks, such as defining and optimizing indexes and manually administering storage.

IBM PureData System for Analytics dramatically simplifies analytics by consolidating all analytic activity in one place, that is, where the data is. Moving analytical functions to IBM PureData System for Analytics is straightforward with an embedded analytic platform that is delivered with every appliance.

## 6.3 Stream computing

Unlike traditional computing, stream computing focuses on the following items:

- ▶ Current fact finding
- ▶ Analysis of data in motion before it is stored
- ▶ A low-latency paradigm through a push model
- ▶ A data-driven approach that brings the data to the query

As shown in Figure 6-5 on page 105, stream computing sits next to and complements an existing Data Warehouse.

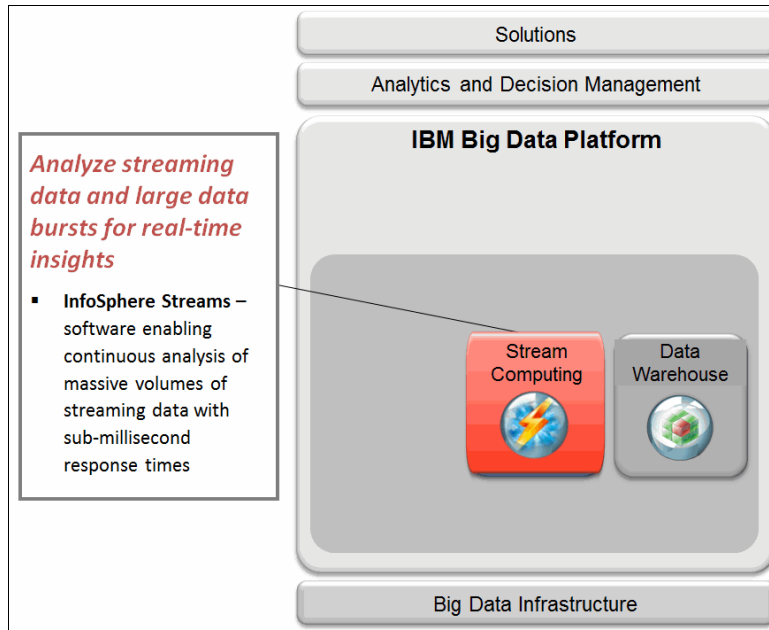


Figure 6-5 Stream computing in the IBM Big Data Platform

With certain kinds of data, there is no time to store it before acting on it because it is constantly changing. This data is often (but not always) generated by a sensor or other instrument. Speed matters for use cases such as fraud detection, emergent healthcare, and public safety where insight into data must occur in real time and decisions to act can be life-saving.

High-velocity, high-volume data calls for in-motion analytics. Streaming data might arrive linked to a master data identifier (a phone number, for example). In a neonatal ICU, sensor data is clearly matched to a particular infant. Financial transactions are unambiguously matched to a credit card or Social Security number. However, not every piece of data that streams in is valuable, which is why the data should be analyzed by a tool such as IBM InfoSphere Streams before being joined with the master data.

### 6.3.1 IBM InfoSphere Streams

IBM InfoSphere Streams analyzes large data volumes with Micro-Latency. Rather than accumulating and storing data first, the software analyzes data as it flows in and identifies conditions that trigger alerts (such as outlier transactions that a bank flags as potentially fraudulent during the credit card authorization process). When this situation occurs, the data is passed out of the stream and matched with the master data for better business outcomes. InfoSphere Streams generates a summary of the insights that are derived from the stream analysis and matches it with trusted information, such as customer records, to augment the master data.

IBM InfoSphere Streams is based on nearly 6 years of effort by IBM Research to extend computing technology to rapidly do advanced analysis of high volumes of data, such as:

- ▶ Analysis of video camera data for specific faces in law enforcement applications
- ▶ Processing of 5 million stock market messages per second, including trade execution with an average latency of 150 microseconds
- ▶ Analysis of test results from chip manufacturing wafer testers to determine in real time if there are failing chips and if there are patterns to the failures

InfoSphere Streams provides a development platform and runtime environment where you can develop applications that ingest, filter, analyze, and correlate potentially massive volumes of continuous data streams that are based on defined, proven, and analytical rules that alert you to take appropriate action, all within an appropriate time frame for your organization.

The InfoSphere Streams product goes further by allowing the applications to be modified dynamically. Although there are other systems that embrace the stream computing paradigm, InfoSphere Streams takes a fundamentally different approach to how it performs continuous processing and therefore differentiates itself from the rest with its distributed runtime platform, programming model, and tools for developing continuously processing applications. The data streams that are consumable by InfoSphere Streams can originate from sensors, cameras, news feeds, stock tickers, or various other sources, including traditional databases. The streams of input sources are defined and can be numeric, text, or non-relational types of information, such as video, audio, sonar, or radar inputs. Analytic operators are specified to perform their actions on the streams.

## 6.4 Apache Hadoop and related big data architectures

Apache Hadoop<sup>1</sup> is a set of open source capabilities that cost-effectively store and analyze a wide variety and large volume of information to gain insights that were not previously possible. It can quickly analyze data in its native format, without imposing a schema or structure. The IBM Big Data Platform builds on the Apache Hadoop project and incorporates the latest technology from IBM Research through the IBM InfoSphere BigInsights product, as shown in Figure 6-6.

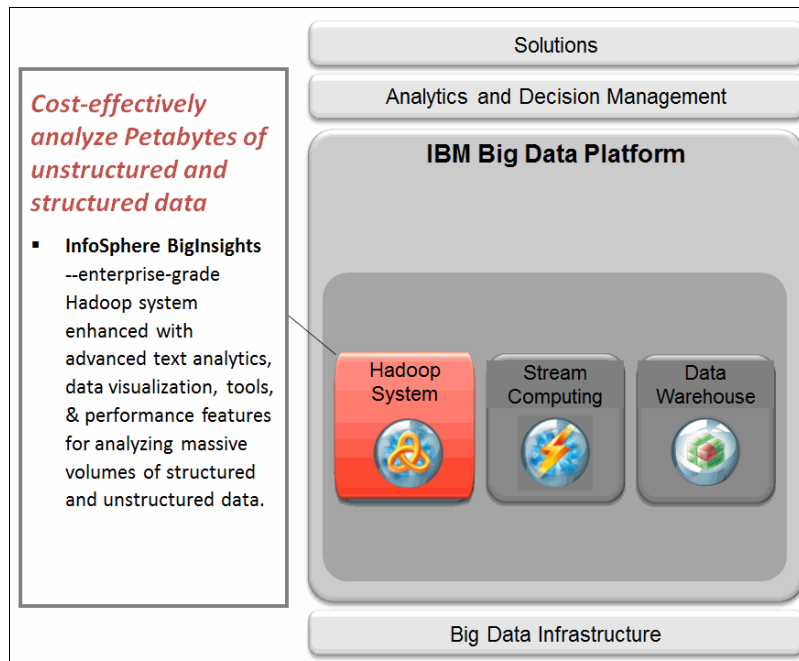


Figure 6-6 Hadoop system in the IBM Big Data Platform

<sup>1</sup> For more information about Apache Hadoop, see <http://hadoop.apache.org/index.html>.

## 6.4.1 IBM InfoSphere BigInsights

InfoSphere BigInsights allows enterprises of all sizes to cost-effectively manage and analyze the massive volume, variety, and velocity of data that consumers and businesses create every day. InfoSphere BigInsights can help you increase operational efficiency by augmenting your Data Warehouse environment:

- ▶ It can be used as a queryable archive, allowing you to store and analyze large volumes of multi-structured data without straining the Data Warehouse.
- ▶ It can be used as a pre-processing hub to help you to explore your data, determine what is the most valuable, and extract that data cost-effectively.
- ▶ It can also allow for ad hoc analysis, giving you the ability to perform analysis on all of your data.

InfoSphere BigInsights manages a wide variety and huge volume of data. The diverse sources and types of big data, which is combined with the explosive growth in data volume, might overwhelm organizations, making it difficult to uncover nuggets of high-value information. InfoSphere BigInsights helps build an environment that is suited to exploring and discovering data relationships and correlations that can lead to new insights and improved business results. Data scientists can analyze raw data from big data sources alongside sample data from the enterprise warehouse in a sandbox-like environment. Then, they can move any newly discovered high-value data in to the enterprise Data Warehouse and combine it with other trusted data to help improve operational and strategic insights and decision making.

Developers have access to professional-grade application development tools that makes it simple to build applications without being a Hadoop expert. The unified tools enable developers to sample data and define, test, and deploy analytics applications from the Eclipse tools, and to administer, run, and monitor deployed applications from the web console.

The enterprise-level management capabilities that InfoSphere BigInsights adds include the following capabilities:

- ▶ Performance optimization
- ▶ Development tools
- ▶ Enterprise integration
- ▶ Analytic accelerators
- ▶ Application and industry accelerators
- ▶ Visualization
- ▶ Security

InfoSphere BigInsights includes a broad palette of analytics tools and capabilities to help organizations build powerful, custom analytic applications that deliver results and insights that are tailored to specific business needs. One of the available tools is IBM BigSheets. BigSheets is a browser-based, spreadsheet-like tool that enables data scientists and business users to explore data that is stored in InfoSphere BigInsights applications and create analytic queries without writing any code. Built-in analytic macros address common data exploration requirements, further improving data accessibility.

## 6.5 Information Integration and Governance

The value of information is enhanced when diverse sources can be brought together and integrated, including the great variety of big data sources that is coupled with traditional organizational sources. At the same time, as described in Chapter 2, “Information Governance foundations for big data” on page 21 and Chapter 3, “Big Data Information Governance principles” on page 33, these data sources must be appropriately governed. The capabilities to integrate and govern big data are core components of and the foundation for the IBM Big Data Platform, as shown in Figure 6-7.

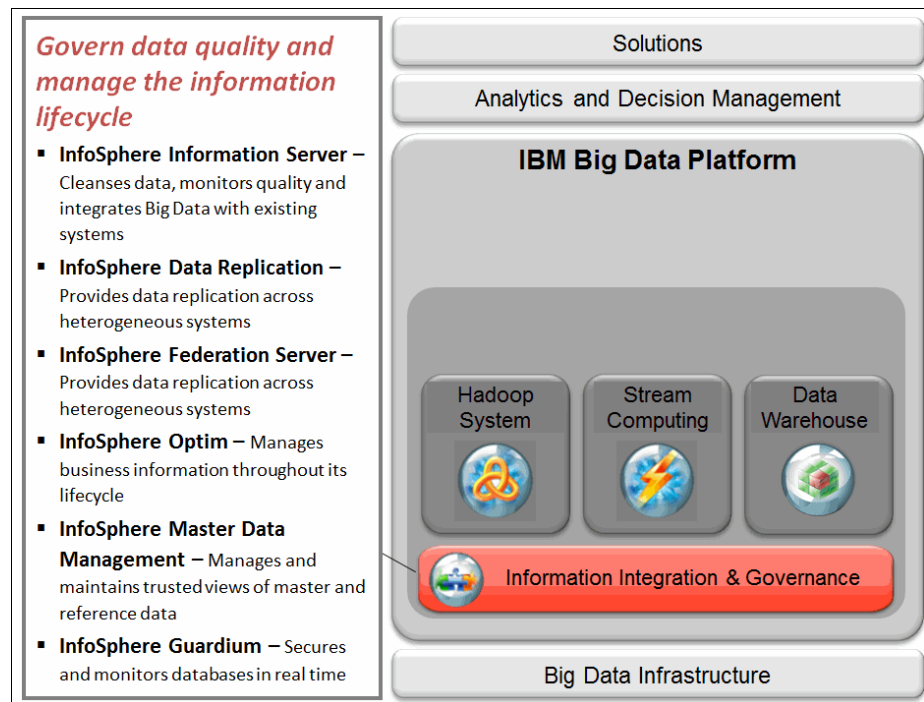


Figure 6-7 Integration and governance in the IBM Big Data Platform

These capabilities help you build confidence in big data, make well-founded decisions, and take decisive actions to accelerate your information-intensive projects.

The primary products for information integration and governance are IBM InfoSphere Information Server, IBM InfoSphere Data Replication, IBM InfoSphere Federation Server, IBM InfoSphere Master Data Management, IBM InfoSphere Optim, and IBM InfoSphere Guardium.

### **6.5.1 IBM InfoSphere Information Server**

IBM InfoSphere Information Server provides a broad set of capabilities, all built on a common platform to support cross-team productivity and effectiveness. These end-to-end capabilities help organizations understand and govern data, create and maintain data quality, and transform and deliver data. At the core of these capabilities is a common metadata repository that stores imported metadata, project configurations, reports, and results for all components of InfoSphere Information Server.

InfoSphere Information Server enables you to transform data in any style and deliver it to any system. InfoSphere Information Server provides multiple options for retrieving and delivering data, whether through bulk (extract, transform, load (ETL) and extract, load, transform (ELT)), virtual (federated), incremental (data replication), or real-time (web service or message queue) delivery. The product supports high performance integration on large volumes of information for Data Warehouses, Hadoop systems, streaming information sources, and transactional systems. It also incorporates a broad range data cleansing and transformation functions to ensure the appropriate integration of data with the correct level of quality. InfoSphere Information Server also benefits from balanced optimization capabilities that enable organizations to choose the deployment option (such as ETL or ELT) that works best for their architecture.

InfoSphere Information Server Enterprise Edition consists of three offerings, each of which is also available independently:

- ▶ IBM InfoSphere Business Information Exchange provides comprehensive capabilities to help understand and govern an organization's information. This package encourages a standardized approach to discovering IT assets and defining a common business language improving alignment of business and IT goals.



- ▶ IBM InfoSphere Information Server for Data Integration helps transform data in any style and deliver it to any system through connectivity to database management systems, big data sources, messaging queues, enterprise resource planning (ERP) and other packaged applications, industry formats, and mainframe systems, all with unique native API connectivity and parallelism.
- ▶ IBM InfoSphere Information Server for Data Quality has rich capabilities to help create and monitor data quality, turning data in to trusted information. By analyzing, cleansing, monitoring, and managing data, organizations can add significant value, make better business decisions, and improve business process execution.

## **IBM InfoSphere Business Information Exchange**

IBM InfoSphere Business Information Exchange encourages a standardized approach to defining a common business language and discovering your IT assets, and helps you achieve the following goals:

- ▶ Create a well-documented, end-to-end information blueprint to ensure that you have aligned your business requirements with your enterprise and reference architectures before starting your strategic project.
- ▶ Establish a common business language and manage business perspectives about information, aligning those views with the IT perspective.
- ▶ Automate the discovery of relationships within and across data sources to accelerate project deployment and also manage and explore data lineage to create trusted information that supports data governance and compliance efforts.
- ▶ Provide a solid foundation for different types of information integration and governance projects, including information integration, lifecycle management, and security and privacy initiatives.

## **IBM InfoSphere Information Server for Data Quality**

IBM InfoSphere Information Server for Data Quality provides rich capabilities that enable you to cleanse data and monitor data quality, turning data into trusted information. By analyzing, cleansing, monitoring, and managing data, you can make better decisions and improve business process execution. Use this product for the following tasks:

- ▶ Automate source data investigation, information standardization, and record matching, all based on business rules that you define.
- ▶ Use comprehensive and customizable data cleansing capabilities in batch and real time.
- ▶ Enrich data and make sure that the best data across sources survives.

- ▶ Match records to eliminate duplicates, householding, and many other operations.
- ▶ Monitor and maintain your data quality. Establish data quality metrics that are aligned with business objectives that enable you to quickly uncover data quality issues and establish a remediation plan
- ▶ Correctly manage and support a data governance program.

### **IBM InfoSphere Information Server for Data Integration**

IBM InfoSphere Information Server for Data Integration enables you to transform data in any style and deliver it to any system, ensuring faster time to value. Built-in transformation functions and a common metadata framework help you save time and expense. InfoSphere Information Server also provides multiple options for delivering data, whether through bulk (extract, transform, and load (ETL) and extract, load, and transform (ELT)), virtual (federated), or incremental (data replication) delivery. Here are some examples:

- ▶ Use unique project blueprinting capabilities to map out your data integration project to improve visibility and reduce risk.
- ▶ Use unique discovery capabilities to understand the relationships within and across data sources before moving forward.
- ▶ Deliver faster time to value by deploying an easy-to-use graphical interface to help you transform information from across your enterprise.
- ▶ Integrate data on demand across multiple sources and targets, while satisfying the most complex requirements with the most scalable run time that is available.
- ▶ Save time and expense by using hundreds of built-in transformation functions, and promote collaboration through a common metadata framework.
- ▶ Select from multiple data delivery options, whether through bulk data delivery (ETL and ELT) or incremental data delivery (Change Data Delivery).
- ▶ Benefit from balanced optimization capabilities and choose the deployment options (such as ETL and ELT) that work best for you.
- ▶ Take advantage of connectivity to DBMS, big data sources, messaging queues, ERP, and other packaged applications, industry formats, and mainframe systems, all with unique native API connectivity and parallelism.

### **IBM InfoSphere Information Server Enterprise Edition**

IBM InfoSphere Information Server Enterprise Edition provides all of the capabilities of the other three packaged editions, providing your entire team with the full set of capabilities that are required to support trusted and governed information.

## 6.5.2 IBM InfoSphere Data Replication

IBM InfoSphere Data Replication provides a highly flexible one-stop shop for high-volume, robust, secure information replication across heterogeneous data stores. It uses real-time data replication to support high availability, database migration, application consolidation, dynamic warehousing, master data management (MDM), service-oriented architecture (SOA), Business Analytics, and extract-transform-load (ETL) or data quality processes. It also delivers outstanding capabilities for loading real-time information into a Data Warehouse or operational data store, which can help organizations enhance business agility and visibility in to key processes.

The downtime costs for the inaccessibility of mission-critical data and applications range from contractual fines and lost productivity to a loss of credibility; all of which might result in lost customers. With InfoSphere Data Replication software, data can be synchronized between two systems to provide continuous availability. If the primary system is impacted by a planned or unplanned outage, a secondary system is available to serve customers and keep the business up and running.

InfoSphere Data Replication also enables continuous delivery of data to support critical business operations. Transaction consistency is maintained throughout the process to preserve units of work and referential integrity. The software supports full transaction granularity with before-and-after images of all transactional changes. The platform is data event-aware, so it can be used to trigger specific business processes. In addition, fault-tolerance capabilities allow organizations to recover to the last committed transaction.

Instead of using triggers or performing queries against the database, InfoSphere Data Replication software reads the native database log to capture changes. For businesses faced with shrinking batch windows or overutilized applications, this log-based change data capture (CDC) approach helps ensure that the performance of even the most demanding mission-critical applications running on the source system is not adversely affected. In addition, InfoSphere Data Replication supports near-zero latency for pervasive integration projects.

### Replicating your data

IBM InfoSphere Data Replication copies information between your heterogeneous data stores in near real time. It provides replication with transactional integrity to support database migration, application consolidation, warehousing, master data management, Business Analytics, and data quality processes.

Here are the capabilities that are supported:

- ▶ Faster, simpler data replication that maintains transactional integrity and consistency for enterprise data volumes:
  - Delivers scalable, low-latency replication with transactional integrity.
  - Offers data replication for active-active databases and high availability.
  - Provides automatic replication of the DB2 data definition language (DDL), such as create and drop table, alter data type, and add column.
  - Offers conflict detection and resolution to support data changes made to multiple databases.
  - Reduces batch windows by providing continuous delivery of changed data into target systems.
- ▶ Centralized, easier-to-use platform that helps simplify deployment and data integration processes:
  - Provides a GUI to help enable faster deployment of data integration processes.
  - Enables zero-download database migrations and application upgrades.
  - Offers full monitoring capabilities to increase visibility into the health and performance of the replication environment.
  - Supports flexible implementation to enable one-way, two-way, many-to-one, and one-to-many delivery of data.
- ▶ Support for heterogeneous data that moves information, at a lower cost, between a wide range of systems and data sources:
  - Supports various heterogeneous source and target databases, including DB2 for Linux, UNIX, and Windows.
  - Integrates with IBM InfoSphere DataStage® to provide changed data feeds to extract, transform, and load (ETL) processes.
  - Supports service-oriented architectures (SOAs) by packaging data transactions into extended markup language (XML) documents or delimited formats for use with messaging middleware, such as IBM WebSphere MQ.
  - Combines with IBM InfoSphere Data Replication for DB2 for z/OS® to replicate heterogeneous data to and from IBM DB2 for z/OS.

### 6.5.3 IBM InfoSphere Federation Server

IBM InfoSphere Federation Server allows organizations to virtualize their data and provide information in a form that applications and users need while hiding the complexity of the underlying sources. Data federation allows information to be accessed through a common interface that centralizes the control of data access.

Federation is also known as enterprise information integration (EII). It provides an optimized and transparent data access and transformation layer with a single relational interface across all enterprise data.

A federated system has the following capabilities:

- ▶ Correlates data from local tables and remote data sources as though all the data is stored locally in the federated database.
- ▶ Updates data in relational data sources as though the data is stored in the federated database.
- ▶ Moves data to and from relational data sources.
- ▶ Uses data source processing strengths by sending requests to the data sources for processing.
- ▶ Compensates for SQL limitations at the data source by processing parts of a distributed request at the federated server.
- ▶ Accesses data anywhere in your enterprise, regardless of what format it is in or what vendor you use, without creating databases and without disruptive changes to existing ones, by using standard SQL and any tool that supports Java Database Connectivity (JDBC) or Open Database Connectivity (ODBC).

With a federated system, you can send distributed requests to multiple data sources within a single SQL statement. For example, you can join data that is in a DB2 table, an Oracle table, a web service, and an XML tagged file in a single SQL statement.

#### **A consolidated view of your data**

IBM InfoSphere Federation Server creates a consolidated view of your data to support key business processes and decisions. It incorporates the capability to access and integrate diverse data and content sources as though they were a single resource, regardless of where the information is.

InfoSphere Federation Server allows you to perform the following tasks:

- ▶ Combine data from disparate sources, such as DB2, Oracle, and SQL Server into a single virtual view.
- ▶ Access, join, and manipulate data of different formats, such as relational, mainframe, and XML from the same interface.
- ▶ Extend your data visualization solution by adding InfoSphere Information Services Director to provide reusable services for Information as a Service (IaaS) and SOA.
- ▶ Interface with standard SQL tools, development environments, portals, and other standard IT infrastructure components.
- ▶ Incorporate data from multiple sources, such as up-to-the-minute inventory, sales, and customer information into reports and analytics with a single query.
- ▶ Insert, delete, or update multiple data sources simultaneously no matter where the information is.
- ▶ Implement security policies throughout the unified view of the enterprise.
- ▶ Enable hybrid data warehousing by combining data replication and data federation. Frequently used data is replicated into the warehouse. The warehouse's reach is then extended with data federation.
- ▶ Extend IBM DB2 pureScale® to be a federated database server, which lets all data sharing members process federated SQL.

## 6.5.4 IBM InfoSphere Master Data Management

Any enterprise that accumulates large volumes of fast-moving structured and unstructured data needs a well-defined approach to store, analyze, and use it, and a way to gauge its trustworthiness.

Information about customers and products is concealed within big data, and it must be linked to existing customer information that is stored in traditional structured applications; otherwise, the new insights that are extracted from big data risk becoming another information silo.

Master Data Management (MDM) plays a crucial role in translating data into meaningful and useful intelligence. Together, big data and MDM help CIOs and CMOs understand customers at an individual level and become more responsive to their needs. MDM creates context for big data by providing trusted information about how incoming unstructured data fits into the business environment. MDM can also help identify and structure big data that is used in controlled environments. Conversely, big data creates context for MDM by providing new insights from social media and other sources, which helps companies build richer customer profiles.

InfoSphere Master Data Management is a complete and flexible MDM solution that creates trusted views of party, product, or other domains to improve operational business processes, big data, and analytics. It supports all domains, architectural styles, and use cases across all industries, and offers quick time-to-value through pre-built and customizable data models and business services.

### **InfoSphere Master Data Management Standard Edition**

InfoSphere MDM Standard Edition is a flexible and proven best-in-class MDM solution that delivers complete, accurate, and real-time views of master data to users throughout your organization. It delivers business value for MDM projects with a quick time to value.

This edition incorporates a virtual MDM engine, which allows organizations to quickly establish an MDM registry for creating a single trusted view across existing systems. It also allows you to dynamically assemble a composite view of the master information that is tailored to each consuming system's needs.

At the same time, it also includes the following key features:

- ▶ **Master Data Matching:** Highly accurate probabilistic matching and search that is optimized for MDM use cases. Maximizes the correct resolution of fragmented or duplicate records, which typically requires manual intervention.
- ▶ **Relationships, Groupings, and Hierarchies:** Provide significant value in managing households, associations, business-to-business relationships, and so on, to better understand and manage relationships between master data entities.
- ▶ **Real-time and Batch Integration:** Use multiple options for integration with applications, such as customer relationship management (CRM), enterprise resource planning (ERP), and web portals to best suit your business requirements.
- ▶ **Data Stewardship and Data Governance:** Use intuitive user interfaces to inspect and resolve matches, data quality issues, and more, in real time, including relationships and hierarchies. YOu can author and edit the golden record.
- ▶ **MDM Workflow:** Implement multi-step and multi-role policies for correct ongoing stewardship and master data governance to ensure that the quality, integrity, and security of your master data are in line with your organization's processes.
- ▶ **Policy Management:** Administer, monitor, and enforce governance policies to ensure data quality through a probabilistic approach.

- ▶ **Integration:** Designed to run in a heterogeneous environment, but is also integrated with a range of IBM products, including pre-built integrations with IBM InfoSphere Information Server for Data Quality and shared metadata. Design and pre-built integrations help accelerate implementation.
- ▶ **Support for big data:** Designed to handle big data volumes, link various data, including social media to existing master records, and is optimized for real time.
- ▶ **Scalability, Performance, and Availability:** Process high volumes of data with low latency, and scale for rapid growth in transaction and data volumes to address your current and future needs.

### **InfoSphere Master Data Management Collaborative Edition**

InfoSphere Master Data Management Collaborative Edition provides best-in-class capabilities for Product Information Management (PIM) and supports any domain through the collaborative authoring style of MDM. It can help you create a single, up-to-date repository of product and other related information that can be used to accomplish the following tasks:

- ▶ You can aggregate information from any upstream or legacy system.
- ▶ You can collaborate and enforce business processes to ensure data accuracy and drive time to market.
- ▶ You can synchronize the trusted information to all downstream systems.

This edition includes integrated, highly scalable, and configurable workflows to streamline business processes such as New Product Introduction (NPI) and drive time to market. The sophisticated workflow dashboard helps govern workflow tasks and activities.

It also includes Multi-Channel Hierarchy Management to organize and browse products and services through different hierarchies, such as product type, organizational hierarchy, departmental hierarchy, web hierarchy, enterprise resource planning (ERP) hierarchies, and other system hierarchies. Organizations can use relationships and linkages to manage bundles, groups, offers, and so on, that are targeted for specific channels.

Other important capabilities include the following ones:

- ▶ **Content Management:** Manage unstructured content through an integrated user interface for working with images, files, and other digital content/media, bringing together a complete picture of product or service components.
- ▶ **Commerce Initiatives:** Purpose-built solution with a data model, workflows, and integration tools to complement IBM WebSphere Commerce.



- ▶ **Supplier Portal:** Establish a supplier portal for data submission with included IBM WebSphere Portal components. Manage vendor onboarding and streamline supplier interactions.
- ▶ **Global Data Synchronization (GDS) Module:** Manage data subscriptions through GDSN network standards to maximize the effectiveness of trading partner relationships, improve overall supply chain efficiency, and increase revenue.
- ▶ **Organizational Modeling with Role and Attribute Level Security:** Segment various attributes of product data by business responsibility (such as logistics, finance, pricing, marketing, and supplier), and control user actions (view only, edit, and approve) based on their roles. Ensure that users can access only the data that is relevant for their role or business function.
- ▶ **Enterprise Business Rules:** Use a single interface for product creation, rules authoring, and rules association through an integrated solution with IBM Operational Decision Manager.
- ▶ **Intuitive Out-of-the-Box User Interfaces (UIs):** The native user interface (UI) is dynamic, adapting to the configurations in the data model, access controls, and workflow, and incorporates user settings and preferences to help users reduce implementation time. Extend the UI through custom windows for enhanced capabilities
- ▶ **High Performance, Scalability, and Reliability:** Grow to manage millions of products with translated and localized product data across geographies.

### **InfoSphere Master Data Management Advanced Edition**

InfoSphere MDM Advanced Edition is a flexible and proven best-in-class MDM solution that accommodates multiple MDM styles and domains to provide a comprehensive set of MDM capabilities. It can help you strategically transform your organization through improved business processes and applications. One of the key aspects in this edition is the combination of Virtual and Physical MDM engines with unique hybrid capabilities, which allow for a broad range of use cases.

Here are some key features of this edition:

- ▶ **Business Services:** Provide pre-built, coarse-grained, and fine-grained services that accelerate the integration with multiple applications, such as CRM, ERP, and web portals.
- ▶ **Pre-built and Extensible Data Models for any Domain:** Support cross-industry and multi-domain (party, product, account, and so on) requirements and is enhanced by industry-specific templates that provide you with a comprehensive set of capabilities that can be tailored to your environment.

- ▶ **Master Data Matching:** Highly accurate probabilistic matching and search that is optimized for MDM use cases. Maximizes the correct resolution of fragmented or duplicate records, which typically requires manual intervention. Deterministic matching can also be used.
- ▶ **Relationships, Groupings, and Hierarchies:** Provide significant value in managing households, associations, business-to-business relationships, and so on, to better understand and manage relationships between master data entities.
- ▶ **Data Stewardship and Data Governance:** Use intuitive user interfaces to inspect and resolve matches, suspected matches, data quality issues, and more, in real time, including relationships and hierarchies. Author and edit the golden record.
- ▶ **MDM Workflow** lets you implement policies for correct ongoing stewardship and master data governance to ensure that the quality, integrity, and security of your master data are in line with your organization's processes.
- ▶ **Integration:** Designed to run in a heterogeneous environment, but also integrated with a range of IBM products, including IBM InfoSphere Information Server for Data Quality and shared metadata. Pre-built integrations help accelerate implementation.
- ▶ **Support for big data:** Designed to handle big data volumes and link various data, including social media to existing master records. Optimized for real time.
- ▶ **Performance and Scalability:** Designed to support large transaction loads and high-availability operational environments.

### **InfoSphere Master Data Management Enterprise Edition**

InfoSphere Master Data Management Enterprise Edition is a complete, flexible, and proven MDM solution that creates trusted views of master data to improve operational business processes and analytics. InfoSphere MDM supports all domains, architectural styles, and use cases across all industries. IBM offers quick time to value through pre-built and customizable data models and business services. The solution is optimized for real-time integration and supports the volume, variety, and veracity of big data initiatives.

This edition incorporates the broad set of capabilities and features that are contained in the other editions in one comprehensive product. Of particular note are the following capabilities:

- ▶ **Support for all MDM Implementation Styles:** Virtual master registry, hybrid, centralized physical master repository, and collaborative authoring allow you to grow with your project and adapt to changing demands of your business.

- ▶ Support for all Master Data Domains: Party, product, account, and custom domains address all of your organization's needs.
- ▶ Support for big data: Designed to handle big data volumes, link various data, including social media, to existing master records, and is optimized for real time.
- ▶ Link Unstructured Content to Master Data: Use potential information about customers and their behavior from documents, reports, emails, social blogs, and more.
- ▶ Governance and Stewardship User Interfaces: Inspect and resolve data quality issues in real time, including relationships and hierarchies, author and edit the golden record, manage complex relationships and hierarchies. Ease of access and manipulation of the golden record.
- ▶ Data Quality: Matching, linking, and semantic reconciliation of master data from different data sources to create and maintain the trusted golden record to drive accuracy in operations and analytics.
- ▶ Data Governance: Integrated workflow provides the capabilities to implement and coordinate multi-step and multi-role workflow for data stewardship and data governance in-line with your organization's policies. Actively prevent poor quality master data from entering your environment.
- ▶ Loading: Multiple flexible options for loading master data. Data can be loaded through business services, in batch or real time, for ease of integration with your existing IT applications: CRM, ERP, web portals, and so on

Overall, this edition can address an organization's needs across multiple master data domains, use cases, and implementation styles by creating trusted views of data assets and elevating the effectiveness of most important business processes and applications.

## **InfoSphere Identity Insight**

InfoSphere Identity Insight is an advanced entity analytics solution with sophisticated recognition algorithms that are optimized to help predict and preempt criminal activity faster by recognizing the true identity of someone or something ("who is who") and determining the potential value or danger of relationships ("who knows who") among customers, employees, vendors, and other external forces.

In many organizations, the raw data that represents identities and relationships already exists. The problem with most systems is that there is no simple way to manage, analyze, and resolve the volume of data that you need to gain the maximum insight from it.

With IBM InfoSphere Identity Insight, organizations can manage, analyze, and integrate data in real time from any source, such as customer databases, vendor lists, employee databases, regulatory compliance lists, and streaming data feeds. IBM InfoSphere Identity Insight can also help identify the network value of customers or their market segments, based on a comprehensive view of the customer.

InfoSphere Identity Insight uses advanced algorithms that are optimized to recognize nefarious individuals and organizations despite their sophisticated attempts to mask their identity, their unscrupulous relationships, and their activities. By examining relevant individuals, their relationships, and their actions, InfoSphere Identity Insight can enable a comprehensive picture of who a person is, who the person knows, and what the person does:

- ▶ **Who is Who – Identity Resolution:** By accumulating identity context over time, InfoSphere Identity Insight uses various enterprise sources of information to determine whether individuals really are who they say they are. The IBM proprietary Entity Resolution technology looks at these attributes across time to help ensure the most accurate identity, and determines whether a previous assumption should be corrected based on new facts.
- ▶ **Who knows who – Relationship Resolution:** After an accurate identity is established, complex relationships can be uncovered. InfoSphere Identity Insight processes resolve identity data to determine whether people are, or ever have been, related in any way.
- ▶ **Who does what – Complex Event Processing:** With “who is who” and “who knows who” established, InfoSphere Identity Insight then applies complex event processing to evaluate all transactions of the entity, and, optionally, of associated entities. The capability to determine all occurrences of the same person across the information landscape is required to have a clear picture of how an individual is interacting with the organization. This sophistication allows the IBM solution to discover well-concealed fraudulent and criminal activities.

This process mines the overall information landscape in real time, bringing to light all of the associations and occurrences that involve the same person or a group. If a questionable situation or pattern is detected, the system proactively generates and sends real-time alerts to analysts, security, or other personnel for further investigation.

## 6.5.5 IBM InfoSphere Optim

InfoSphere Optim Solutions help organizations meet the requirements for information governance and address challenges that are exacerbated by the increasing volume, variety, and velocity of data. InfoSphere Optim Solutions focus on three core governance areas:

- ▶ Information lifecycle management
- ▶ Test data management
- ▶ Data privacy

### InfoSphere Optim Archive

IBM InfoSphere Optim Archive offers a single, scalable solution to implement consistent data lifecycle management strategies throughout the enterprise. InfoSphere Optim Archive helps organizations to accomplish the following goals:

- ▶ Manage the growth of application data and Data Warehouses for better data control and reduced storage costs.
- ▶ Apply business rules and criteria to segregate and archive historical data more safely and improve compliance.
- ▶ Improve data lifecycle management for more efficient storage processes.
- ▶ Use a single, scalable solution across applications, databases, Data Warehouses, operating systems, and platforms.

By archiving old data from huge Data Warehouse environments, businesses can improve response times and reduce costs by reclaiming valuable storage capacity. As a result, organizations gain more control of their IT budget while simultaneously helping their big data and Data Warehouse environments run more efficiently and reducing the risk of exposure of sensitive data.

### IBM InfoSphere Optim Test Data Management

IBM InfoSphere Optim Test Data Management optimizes and automates the test data management process. Prebuilt workflows and services on demand facilitate continuous testing and agile software development. With IBM InfoSphere Optim Test Data Management, organizations can accomplish the following goals:

- ▶ Streamline test data management processes to help reduce costs and speed application delivery.
- ▶ Analyze and refresh test data on demand for developers and testers.
- ▶ Create production-like environments to shorten iterative testing cycles, support continuous testing, and accelerate time to market.
- ▶ Protect sensitive data and help reduce risk in testing, training, and development environments.

By creating realistic, right-sized data sources for testing, development, and testing teams can enhance the accuracy of testing and identify problems early in the testing cycle. By implementing data masking capabilities, they can protect sensitive data and help ensure compliance with privacy regulations.

### **InfoSphere Optim Data Privacy**

IBM InfoSphere Optim Data Privacy de-identifies confidential data on demand throughout the enterprise, including big data platforms. It masks data statically or dynamically in applications, databases, and reports across production and non-production environments to improve data protection and support compliance initiatives. InfoSphere Optim Data Privacy helps organizations with the following tasks:

- ▶ Mask confidential data on demand in applications, databases, and reports that are based on business policies to protect data privacy.
- ▶ Support compliance initiatives throughout the enterprise.
- ▶ Use a single, scalable enterprise data masking solution to mask data everywhere across applications, databases, and operating systems.

## **6.5.6 IBM InfoSphere Guardium**

InfoSphere Guardium solutions help organizations implement the data security and privacy capabilities that they need for big data environments. With these solutions, organizations can protect against a complex threat landscape, including insider fraud, unauthorized changes, and external attacks, while remaining focused on business goals and automating compliance.

### **IBM InfoSphere Guardium Data Activity Monitor**

IBM InfoSphere Guardium Data Activity Monitor prevents unauthorized data access, issues alerts on changes or leaks to help ensure data integrity, automates compliance controls, and protects against internal and external threats. Continuous monitoring and real-time security policies protect data across the enterprise without changes to databases or applications or performance impact. InfoSphere Guardium Data Activity Monitor can do the following tasks:

- ▶ Monitor and audit all data activity for all platforms and protocols.
- ▶ Enforce security policies in real time for all data access, change control, and user activities.

- ▶ Create a centralized repository of audit data for enterprise compliance, reporting, and forensics.
- ▶ Support heterogeneous environments for all leading platforms, file shares, and operating systems, including big data environments.

### **IBM InfoSphere Guardium Data Encryption**

IBM InfoSphere Guardium Data Encryption provides encryption capabilities to help you safeguard structured and unstructured data and comply with industry and regulatory requirements. This software performs encryption and decryption operations with minimal performance impact and requires no changes to databases, applications, or networks. InfoSphere Guardium Data Encryption offers the following capabilities:

- ▶ Transparent, rapid implementation: Requires no changes to applications, the underlying database, or hardware infrastructure.
- ▶ Centralized key and policy management: Delivers a unified management system to help simplify data security management.
- ▶ Compliance-ready capabilities: Provides granular auditing and reporting to help you meet data governance requirements, such as the Health Insurance Portability and Accountability Act (HIPAA) and PCI Data Security Standard (PCI DSS).

### **IBM InfoSphere Guardium Data Redaction**

IBM InfoSphere Guardium Data Redaction protects sensitive data in documents, forms, and files from unintentional disclosure by detecting and removing the data from the document version that is openly shared. It supports many document formats, including scanned documents, PDF, TIFF, XML, and Microsoft Word and allows organizations to do the following tasks:

- ▶ Protect sensitive unstructured data in documents, forms, and files.
- ▶ Transform tedious, manual redaction alternatives into automated processes for speed, accuracy, and efficiency.
- ▶ Safeguard proprietary or personal information from internal/external misuse and fraud and support regulatory compliance by applying data governance controls.
- ▶ Integrate with records management systems.
- ▶ Secure need-to-know, web-based document viewing.
- ▶ Capture all sensitive data in documents or document repositories for reporting or discovery processes.

Together, InfoSphere Guardium and InfoSphere Optim Solutions provide a complete data protection approach to secure diverse data types (structured, unstructured, online, and offline) across different locations, including production and non-production (development, test, and training) environments and big data platforms. Their data security and privacy capabilities help organizations streamline compliance by implementing unified controls and consistent enforcement of security policies across the entire enterprise.

InfoSphere Guardium and InfoSphere Optim Solutions for data security and privacy have the following characteristics:

- ▶ Business-driven: Set policies that are based on business requirements.
- ▶ Simple to use: Use providers for protection policies.
- ▶ Ready for big data: Support leading databases, operating systems, and applications across mainframe, distributed, and big data environments.

By focusing on data security and privacy within big data implementations, organizations can accomplish the following goals:

- ▶ Prevent data breaches: Avoid disclosing or leaking sensitive data.
- ▶ Help ensure data integrity: Prevent unauthorized changes to data, data structures, configuration files, and logs.
- ▶ Reduce the cost of compliance: Automate and centralize controls and simplify the audit review process.
- ▶ Protect privacy: Prevent disclosure of sensitive information by masking or de-identifying data in databases, applications, and reports, on demand, across the enterprise.

## 6.6 Big Data Accelerators

IBM accelerators for big data focus on faster time-to-value, whether supporting specific types of big data, comprehensive data models, or industry specialized packs for the Data Warehouse, or providing added analytical capabilities, as shown in Figure 6-8 on page 127.



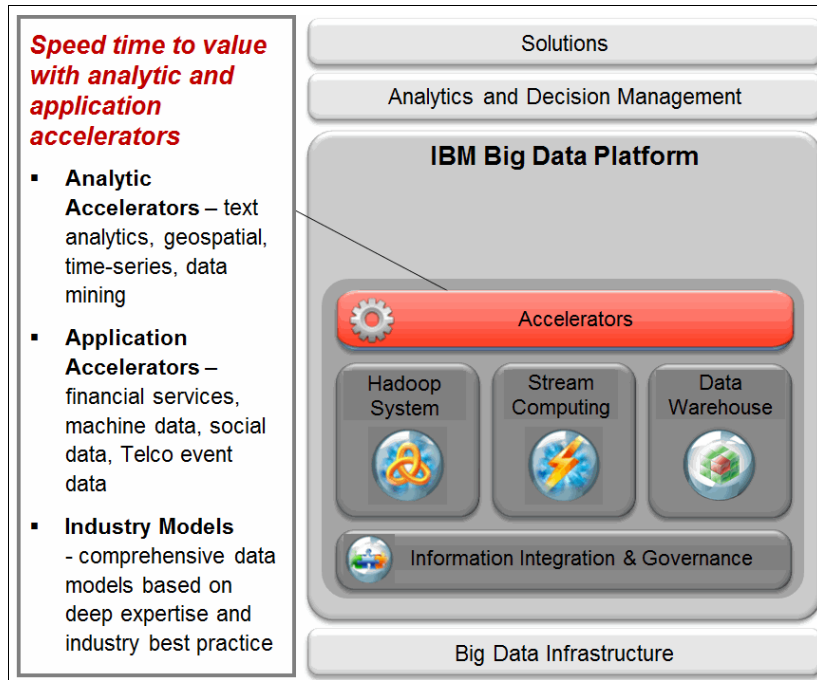


Figure 6-8 Accelerating value with the IBM Big Data Platform

### 6.6.1 IBM Accelerators for Big Data

IBM offers three analytic components that accelerate the development and implementation of big data solutions. These components, or accelerators, address specific data types or operations, such as analytics for machine log data, social media data, and telecommunications event data.

Accelerators provide business logic, data processing capabilities, and visualization for specific use cases. By using accelerators, you can apply advanced analytics to help integrate and manage the variety, velocity, and volume of data that constantly enters your organization. Accelerators also provide a development environment for building new analytic applications that are tailored to the specific needs of your organization.

Accelerators enable data scientists to iterate your data, refine your analysis, and produce results that are meaningful to your specific business questions. Developers use accelerators to speed solution implementation by providing customization points to suit the context of your use case and industry.

### **IBM Accelerator for Social Data Analytics**

The social data analytics accelerator can be used to efficiently analyze customer information from various types of social media data. This analysis supports enhanced insights for effective, targeted marketing campaigns and timely product marketing decisions, which helps organizations gain competitive intelligence and build customer retention and new customer acquisition programs.

### **IBM Accelerator for Machine Data Analytics**

The machine data analytics accelerator can ingest, parse, and extract various machine data from sources, such as log files, smart devices and telemetry, and help process that data in minutes instead of days and weeks. By using the machine data analytics accelerator, organizations gain insights into operations, customer experience, transactions, and behavior. The resulting information can be used to proactively boost operational efficiency, troubleshoot problems and investigate security incidents, and monitor an end-to-end infrastructure to avoid service degradation or outages.

## **6.6.2 IBM Industry Models**

IBM Industry Models combine deep expertise and industry preferred practice in a usable form (a “blueprint”) for both business and IT communities to accelerate industry solutions. Part of the IBM InfoSphere portfolio, the IBM Industry Models are based on the experience of more than 500 clients, and more than ten years of development.

The IBM Industry Models include comprehensive data models containing Data Warehouse design models, a business terminology model, and analysis templates to accelerate the development of business intelligence applications. They also include preferred practice business process models with supportive service definitions for development of a service-oriented architecture (SOA). The result is acceleration of projects and also reduced risk.

## **6.7 Data visualization**

With big data, it is not simply enough to provide integration solutions; it is also critical to provide Data Science teams working with the data the ability to discover, understand, and visualize the data and subsequent information results across the diverse data sources that are shown in Figure 6-9 on page 129.

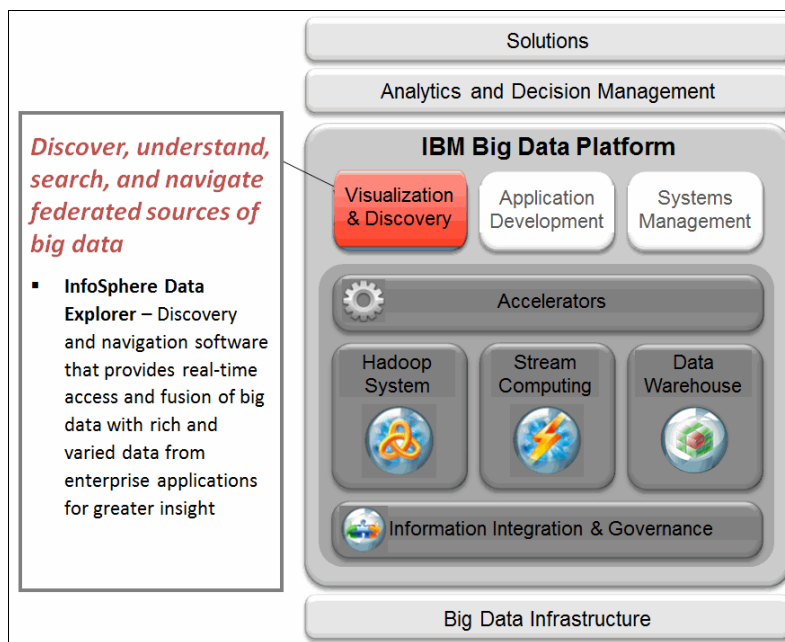


Figure 6-9 Exploring and visualizing data within the IBM Big Data Platform

### 6.7.1 IBM InfoSphere Data Explorer

IBM InfoSphere Data Explorer is a platform for navigating and filtering large amounts of data in nearly any format. It offers a visual dashboard that joins master data about an individual with that person's detailed transaction history from emails, social media, packaged applications, and more.

This data can help sales staff, particularly inbound and outbound marketers, connect to prospects on a more individual level and achieve better sales rates. Designed to handle both social and sensor generated data, InfoSphere Data Explorer lets users specify an “axis” and then searches for matching data that it can fuse onto existing master data. With master data as the starting point, InfoSphere Data Explorer can also place real-time activity streams into context for analysis.

Intelligence, law enforcement, and fraud analysts need to use all relevant pieces of information to connect the dots and make well-informed decisions instantly. InfoSphere Data Explorer works from known master data to give intelligence agencies a unified view of all their information sources to deliver visibility, transparency, and insight.

## 6.8 The IBM Big Data Platform and the reference architecture

Overall, the capabilities of the IBM Big Data Platform allow organizations to incorporate and address the following items:

- ▶ The accelerating variety of information:
  - Social media/sentiment analysis
  - Geospatial analysis
  - Brand strategy
  - Scientific research
  - Epidemic early warning system
  - Market analysis
  - Video analysis
  - Audio analysis
- ▶ Information in motion at high velocity:
  - Smart Grid management
  - Multimodal surveillance
  - Real-time promotions
  - Cyber security
  - ICU monitoring
  - Options trading
  - Click-stream analysis
  - CDR processing
  - IT log analysis
  - RFID tracking and analysis
- ▶ Extreme volumes of information:
  - Transaction analysis to create insight-based product/service offerings
  - Fraud modeling and detection
  - Risk modeling and management
  - Social media/sentiment analysis
  - Environmental analysis
- ▶ Discovery, exploration, and experimentation with new and existing information:
  - Sentiment analysis
  - Brand strategy
  - Scientific research
  - Ad hoc analysis
  - Model development
  - Hypothesis testing
  - Transaction analysis to create insight-based product/service offerings
- ▶ Management planning and governance to support business policies and decision making:
  - Operational analytics through BI reporting

- Planning and forecasting analysis
- Predictive analysis

Chapter 5, “Big data reference architecture” on page 69 reviewed the core capabilities and landscape for the big data reference architecture. The capabilities that are provided by the IBM Big Data Platform support that reference architecture to address the broad analytical needs of an organization. Figure 6-10 highlights the overlay of the IBM Big Data Platform on the reference architecture. This is a broad view and the specifics of any organization's information landscape, particularly the configuration of specific information processing zones, might require adjustments to optimize processing and ensure alignment to the information governance principles.

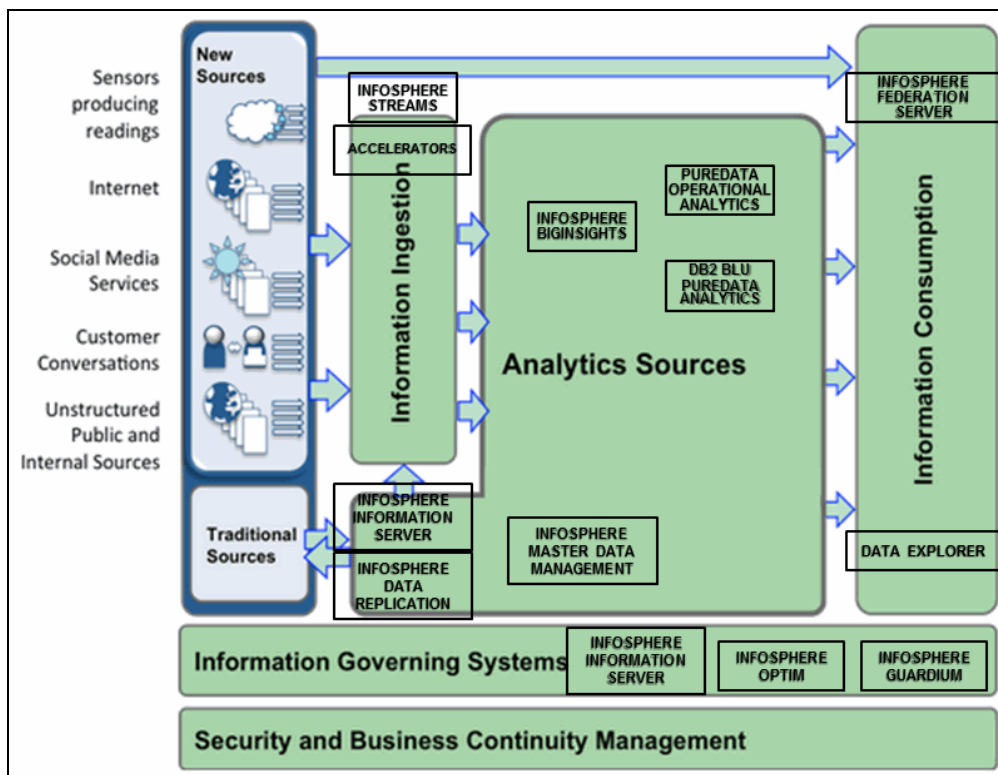


Figure 6-10 IBM Big Data Platform

In summary, this chapter has highlighted the range of capabilities for working with big data within the IBM Big Data Platform. The subsequent chapters highlight Information Governance challenges and risks within the key big data use cases and where the IBM Big Data Platform can be used to address those organizational requirements.





## Security and privacy

From a security and privacy perspective, big data is different from other traditional data, and requires a different approach. Many of the existing methodologies and preferred practices can easily be extended to support the big data paradigm. This chapter describes the data security and privacy landscape and operational models, and how they can be applied to big data. This chapter also highlights usage examples by showing how IBM Software solutions, such as IBM InfoSphere Optim and IBM InfoSphere Guardium, can be used to secure and protect your critical information.

## 7.1 Why big data is different

On the surface, big data appears to have similar risks and exposures to traditional data. However, there are several key areas where it is dramatically different:

- ▶ More data translates into higher risk of exposure in the event of a breach.
- ▶ More experimental usage means that the organization's governance and security discipline is less likely to be in place, especially in the initial stages of testing and deployment.
- ▶ New types of data are uncovering new privacy implications, with few privacy laws or guidelines to protect that information. Examples include the connected home and digital meters for monitoring electricity usage (eMeters), cell phone beacons that broadcast physical location, health devices such as medical, fitness and lifestyle trackers, and telematics data that tracks automobile locations.
- ▶ Exposure of other data. There is an increased security risk of revealing not just privacy data, but also data under compliance regulations such as Basel II, Health Insurance Portability and Accountability Act (HIPAA), Payment Card Industry Data Security Standard (PCI DSS), and the Sarbanes-Oxley Act (SOX), and critical internal company information, such as intellectual capital (software programs, algorithms, and so on).
- ▶ Immature markets. Many areas of Hadoop are still evolving.
- ▶ Data linkage and combined sensitive data. The act of combining multiple data sources can create unanticipated sensitive data exposure, often with a lack of awareness. As an example, a study by Dr. Latanya Sweeney in 1997 was able to identify individuals 87% of the time by matching them against the US Census, just by using date of birth, gender, and postal code.<sup>1</sup>
- ▶ Anonymizing production data. Much of the value of big data lies in uncovering patterns that do not require identification of the individual. As a result, organizations are increasingly using anonymization and de-identification to remove individual identifiers while still gaining utility from the data.

---

<sup>1</sup> *Simple Demographics Often Identify People Uniquely*, found at:  
<http://dataprivacylab.org/projects/identifiability/index.html>



## 7.2 Information security defined

According to the US National Institute of Standards and Technology (NIST), Information Systems Security, also known as INFOSEC or ISS, is defined as “The protection of information and information systems from unauthorized access, use, disclosure, disruption, modification, or destruction in order to provide confidentiality, integrity, and availability.”<sup>2</sup>

This chapter focuses on security as it pertains to the protection of data. Other areas of security that should be considered in the big data environment include user and application level security. The two drivers for information security include compliance mandates and guarding against intrusions and breaches.

### 7.2.1 Security framework

Although it is beyond the scope of this book to describe security frameworks in detail, this book can touch on the usage of frameworks in planning your security implementation for big data. A security framework provides a set of preferred practices, operational standards, and controls to guide organizations in assessing vulnerabilities, planning and implementing secure computing, and reducing the risk of data and systems intrusion.

These frameworks specify the basis for an Information Security Management System (ISMS). There are several security frameworks in use today, such as Control Objectives for Information and Related Technology (COBIT), NIST, and International Standards Organization (ISO 27000). Of the three, ISO 27000 is the only framework that contains a standard for which companies can be certified. It is also multinational in nature. It is interesting to note that although ISO 27000 specifies standards for Security Policy, Asset Management, Cryptography, and Access Control, it does not specifically address data security itself. Rather, a security framework supports data security through controls such as user authentication and separation of duties. Check with your organization to better understand which framework it uses, and existing standards, policies, and procedures for data security. Also, ensure that you understand your organization's major compliance requirements, such as Basel II, PCI, HIPAA, and SOX. For more information about ISO27000, go to the following website:

<http://www.27000.org/>

---

<sup>2</sup> <http://nvlpubs.nist.gov/nistpubs/ir/2013/NIST.IR.7298r2.pdf>

## 7.3 Data privacy defined

The International Association of Privacy Professionals (IAPP) defines data or information privacy as “The claim of individuals, groups or institutions to determine for themselves when, how and to what extent, information about them is communicated to others.”<sup>3</sup>

Privacy laws throughout the world are a patchwork of laws and regulations. Most contain a shared set of principles that are known as Fair Information Practice Principles (FIPPs) that govern notice, choice and consent, collection, use, access, disposal, and program management of information. For more information about FIPPs, go to the following website:

[http://itlaw.wikia.com/wiki/Fair\\_Information\\_Practice\\_Principles](http://itlaw.wikia.com/wiki/Fair_Information_Practice_Principles)

Here are three key principles that big data practitioners of which need to be aware:

1. Choice and Consent means that individuals have a right to opt-in or opt-out of data collection.
2. Collection and Use confines the collection of personal information only for the purposes that are specified.
3. Retention and Disposal (Data minimization) specify that data can be retained only while it is needed, with stipulations for disposal afterward.

You can see the potential impact of these three principles for big data in that the information you collect and store should be used only for the purposes that are specified in your privacy notice and retained only while it is needed. Consult with your privacy office or legal counsel to understand your organization's privacy policies because they form the foundation of the decisions that you make about how to manage and secure sensitive data.

---

<sup>3</sup> [https://www.privacyassociation.org/resource\\_center/privacy\\_glossary#I](https://www.privacyassociation.org/resource_center/privacy_glossary#I)

### 7.3.1 What sensitive data is

In general, sensitive data is any information that identifies an individual. This data can range from your name, address, physical characteristics, location, email address, to your cell phone, which has a unique identifier. The definition of what constitutes sensitive data varies by country and even by state. In the European Union (EU) for example, the definition of sensitive data is much broader, and extends to identifiers such as trade union membership and even political beliefs. Essentially any data about an identified or identifiable individual is considered to be personal data. Two major categories of sensitive data are Personally Identifiable Information (PII), also known as Personal Information outside the US, and Personal Health Information (PHI). In addition to the PII examples described above, Personal Health Information is typically associated with HIPAA, and includes any information that is related to an individual's health, condition, and treatment. One of the first steps in protecting data is to understand your organization's privacy policies and to then identify what is considered sensitive within your organization.

### 7.3.2 Privacy operational structure

As you undertake technical measures to protect information, it is critical to understand where these activities fit within the business practices of privacy that are typically managed by a privacy office. Here are brief descriptions of the operational structure, frameworks, maturity model, and privacy lifecycle. Within those practices, we highlight areas that are relevant to technical big data practitioners.

#### Privacy frameworks

A privacy framework is an extension of the FIPPS, and consists of a set of guidelines that govern privacy policies for the creation, usage, sharing, disposal, and program management of sensitive information. Frameworks provide the structure and implementation roadmap for a privacy program. Major frameworks include American Institute of Certified Public Accountants/Chartered Accountants of Canada (AICPA/CICA) Generally Accepted Privacy Principles (GAPP), Organization for Economic Co-operation and Development Privacy Guidelines (OECD (EU))<sup>4</sup>, Privacy by Design (PbD), Asia Pacific Economic Cooperation (APEC), ISO/IEC 29100:2013, and Canadian Personal information Protection and Electronic Documents Act (PIPEDA). Consult with your privacy or compliance office to get a better understanding of the framework that your organization uses.

---

<sup>4</sup> <http://oecdprivacy.org/>

## Privacy Maturity Model

A useful tool for assessing your organization's readiness in terms of protecting big data is the AICPA/CICA's Privacy Maturity Model, which aligns closely to the AICPA/CICA privacy framework. The maturity model specifies 73 criteria and five maturity levels, which are organized by each of the 10 framework principles. The five maturity levels range from ad hoc to optimized, with regular review and feedback. The following categories are of particular note and should be considered when implementing your big data privacy controls:

- ▶ 1.2.3: Personal Information Identification and classification
- ▶ 1.2.4: Risk Assessment
- ▶ 1.2.6: Infrastructure and Systems Management
- ▶ 3.2.2: Consent for new Purposes and uses
- ▶ 4.2.4: Information developed About Individuals
- ▶ 8.2.1: Information security Program.

For more information about the model, see the following website:

<http://www.cica.ca/resources-and-member-benefits/privacy-resources-for-firms-and-organizations/docs/item48094.pdf>

## Privacy Operational Lifecycle

IAPP has defined a Privacy Operational Lifecycle that describes the four major steps for managing data privacy along with the activities under each step.<sup>5</sup>

This lifecycle, which is shown in Figure 7-1 on page 139, is a useful framework for aligning operational activities to privacy objectives. There are four phases in the lifecycle: Assess, Protect, Sustain, and Respond. Later in the chapter, you see how elements of each phase can be used to guide data protection of critical information. Although some of the steps and activities are implemented at a business and organizational level, it is important for the technical practitioner to understand their role in the process.

---

<sup>5</sup> *Privacy Program Management: Tools for Managing Privacy Within Your Organization*, by Densmore

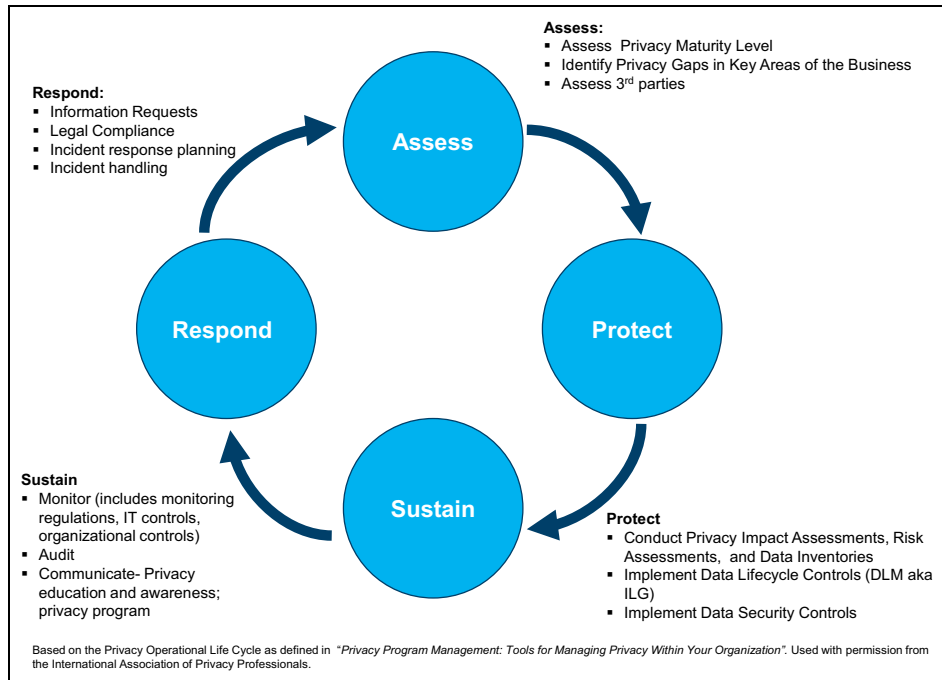


Figure 7-1 Privacy Operational Lifecycle

- **Assess:** In this first phase, you assess the organization's privacy maturity level, identify privacy gaps in key areas of the business, and assess third-party partners regarding their privacy practices. For more information about these actions, see the *IAPP Privacy Program Management* book at the following website:

<https://www.privacyassociation.org/store/merchandise/533d4496-3cff-4073-96e8-e5e191339d78>

For purposes of this book, assume that the organization has already conducted a maturity assessment both within the business and assessed its third parties.

- Protect: The Protect phase is the heart of the Privacy and Security program and includes the following activities:

- Conduct Privacy Impact Assessments.

A Privacy Impact Assessment (PIA) is a tool that identifies the type of personal or sensitive information to be used, determines the level of security risk, and identifies the solutions that are used to mitigate the risk. A sample Privacy Impact Assessment template is available on the FDIC US Government website:

<http://www.fdic.gov/about/privacy/assessments.html>

A PIA typically takes the form of a high-level questionnaire and is performed on an application or data subject area basis and covers all aspects of privacy, such as what data is held, who has access, data sharing, data quality, maintenance, and administrative controls. It may also include a data inventory. Typically, an application or business data owner is responsible for leading the assessment, with the assistance of the privacy or security office.

- Conduct Risk Assessments.

A Risk Assessment is a process that assesses and documents the risks that are associated with retaining sensitive data. It typically follows the PIA. Risk Assessments are also typically conducted for security and are more comprehensive.

- Create Data Inventories.

A Data Inventory documents the information that is held, including specific details such as data element names, locations, and specifically calls out sensitive information. Because the organization has already defined policies on sensitive data, much of the work in this phase is in understanding where that data is, its format, and presenting the information in a structured format, such as a report that can then be used to implement privacy and security technical controls.

In the big data environment, there are several tools that are available to assist you in this process. For structured data, such as a Data Warehouse, both IBM InfoSphere Discovery and IBM InfoSphere Guardium can identify and classify sensitive information, such as PII and PHI. In semi-structured and unstructured Hadoop files, InfoSphere Data Explorer can be used to search the file system to identify embedded sensitive information. For example, machine data logs may contain email addresses, IP addresses, and host names. Later in this chapter, we look at examples of how these tools help you to create a data inventory.

– Implement Data Security Controls.

In this activity, you take steps to mitigate exposure and can choose from various technical solutions that are covered in further detail in this chapter. Data Security Controls can take several forms. For example, user and role level security may specify access level controls to specific files and file systems. Application level security can be used to fence information and display specific only window, reports, or results to those who have authorized credentials.

- **Masking:** Sensitive data can be obfuscated and masked in test environments to minimize the risk of exposure. Masking techniques substitute real data with realistic looking fictitious data so that the data utility is largely preserved. Masking can also be applied to de-identify individuals while preserving the statistical integrity of the data for analytics purposes. IBM InfoSphere Optim Data Privacy for Hadoop provides a call-level embeddable API that can be used to mask sensitive data (check product name) and substitute realistic yet fictitious values without disrupting application integrity or utility. Later in this chapter, we described several usage scenarios for masking data going in to Hadoop, on Hadoop, and within big data databases, such as Netezza.
- **Encryption:** Encryption is an additional solution that can be used to scramble sensitive data so that only authorized users can see the clear text information while others see only a string of numbers and letters that obscure the original source. Encryption is a key element in a data protection strategy in that it can provide a PII “safe harbor” in the event of a security breach. Many US states and countries have Encryption Safe Harbor laws that specify that breach disclosure is not required if the PII data is encrypted and the encryption keys are compromised along with the data. Consult with your legal counsel to understand how Encryption Safe Harbor regulations might apply for your jurisdiction.

Several jurisdictions also specify encryption as a protective technology in their data protection laws, such as Nevada and Massachusetts, and countries that have National Data Encryption Laws, such as the UK Data Protection Act, EU Data Protection Directive, South Korean PIPA, and South Africa IPA. IBM Guardium Data Encryption provides a policy-based solution to encrypt and secure static data at rest, including data in traditional databases and file systems, and Hadoop. It uses a Data Encryption Security Server that centrally administers, manages, and controls policies and keys and access to unencrypted data.

- Data/Database Activity Monitoring: Monitors the database and Hadoop file system data, prevents unauthorized data access, provides alerts on changes or leaks to help ensure data integrity, automates compliance controls, and protects against internal and external threats. Continuous monitoring and real-time security policies should protect data across the enterprise without changes to databases or applications or performance impact.
- Implement Data Lifecycle Controls (Data Lifecycle Management, also known as Information Lifecycle Governance).

Information Lifecycle Governance, also known as Data Lifecycle Management (DLM) or Information Lifecycle Governance (ILG), is a framework for managing data from requirements to retirement. Retirement specifically means disposing of data in a manner that is defensible in a court of law, which includes documenting policies for data retention and disposition, proving that the data was retained in a manner that preserved the information in its original format, and that the data was destroyed in such a way that a verifiable data disposition audit trail exists. There are three reasons to dispose of data:

- Policy: Legal, regulatory, and corporate policies.  
Understanding that regulations and corporate policies dictate the disposition of data, you must ensure that policies are equally applied to big data.
- Privacy laws.  
Privacy laws typically stipulate data minimization, only holding data while it is needed and for the purposes that are specified. The retirement or data disposal aspect of ILG is an interesting case as it relates to big data. Because the economics of big data are such that it is more economical and feasible to store massive amounts of information, there is a tendency to just “park” data in case it is ever needed. Using the data dump approach to big data means that the organization could be violating privacy regulations and subjecting itself to further legal or regulatory action.
- Breach risk.  
There is an increased risk of retaining excessive amounts of data. In the event of a breach, the potential for massive volumes of data to be exposed dramatically increases. Finally, if an organization is retaining data past its legal retention period, if a legal matter were to arise, the holder of this information is legally bound to produce it, putting it at increased financial risk such as class-action lawsuits or fines.



There are two solutions to the data retention conundrum:

- Dispose of it.
- In the case of PII, de-identify or anonymize it.

InfoSphere Optim for Hadoop provides both Archiving and De-identification capabilities (masking) to help you implement effective ILG practices.

- **Sustain:** In this phase, you support ongoing activities of the privacy program, such as monitoring and auditing. The Sustain phase also includes non-technical aspects, such as communicating the privacy program and policies, and conducting education and awareness. Monitoring typically refers to IT controls. However, in this case, the privacy organization also is engaged in monitoring regulations and monitoring organizational and business process controls.

IT controls that support the Sustain phase for big data include data usage monitoring and auditing of key events, such as logins, object creation, who is running MapReduce jobs, and what data jobs are accessing. InfoSphere Guardium provides capabilities to monitor and audit Hadoop data usage and data usage in more traditional data stores. With InfoSphere Guardium, you can implement Hadoop Data Activity Monitoring and create a secure, detailed, and verifiable audit trail of users and activities, including privileged users and file or object creation and manipulation. By doing so, you gain visibility in to cluster activity involving sensitive data, that is, the who, what, when, and how. Monitoring also gives you real-time alerts and anomaly detection for suspicious activity, which is critical in preventing breaches. Guardium also integrates with business processes for audit compliance, including reports dissemination to appropriate personnel for sign-off and review, plus report retention and sign-offs per audit requirements.

- **Respond:** In this final phase of the Privacy Operational Lifecycle, you respond to both information and compliance requests and plan incident response and incident handling. Although much of this phase is aligned to business processes, you can see that IT and IT Security play a role as well. For example, compliance requests could include producing audit reports.

InfoSphere Guardium has a wealth of compliance reporting capabilities, including modules that are designed specifically for key regulations, such as SOX, HIPAA, PCI, and Data Privacy.

## 7.4 How security and privacy intersect

You can have security without privacy, but you cannot have privacy without security. Security is the broad and comprehensive foundation that governs people, process, applications, and users, and privacy is specific to safeguarding the information of individuals. Good security practices support effective privacy practices.

### 7.4.1 Implications and suggestions for big data

In Information Governance for big data projects, it is useful to have a broad vision and objectives while starting small with a single project to gain experience. Incremental progress is useful to demonstrate the value of information governance and collaboration to ensure that you meet all compliance and security/privacy objectives. The key to success in a big data environment is managing governance at the point of impact, and using multiple complementary approaches to secure critical data. Different types of data have different protection requirements, so organizations must take a holistic approach to safeguarding information no matter where it is. This approach includes the following items:

- ▶ Understanding where the data exists. Organizations cannot protect sensitive data unless they know where it is and how it is related across the enterprise.
- ▶ Safeguarding sensitive data, both structured and unstructured. Structured data that is contained in databases must be protected from unauthorized access. Unstructured data in documents and forms requires privacy policies to redact (remove) sensitive information while still allowing required business data to be shared.
- ▶ Protecting nonproduction environments. Data in nonproduction, development, training, and quality-assurance (QA) environments must be protected while remaining usable during application development, testing, and training processes.
- ▶ Securing and continuously monitoring access to data. Enterprise databases, Data Warehouses, and file shares require real-time insight to ensure data access is protected and audited. Policy-based controls are required to rapidly detect unauthorized or suspicious activity and alert key personnel. In addition, databases and file shares must be protected against new threats or other malicious activity and continually monitored for weaknesses.
- ▶ Demonstrating compliance to pass audits. It is not enough to develop a holistic approach to data security and privacy; organizations must also demonstrate compliance and prove it to third-party auditors.

IBM InfoSphere solutions for data security and privacy support this holistic approach to data protection. They incorporate intelligence that enables organizations to proactively address IT threats and enterprise risks.

### **7.4.2 Fit-for-purpose security and privacy**

Information flows through enterprise systems and divisions, much like goods through a physical supply chain. Raw materials are turned into products; oil is refined. Data must be refined as well. Technology and architectures must become more dynamic and allow information to flow at the speed of the business. Because of technology limitations, placing data can present a bottleneck to enterprise information flow. The big data pattern breaks the limits of old technologies and methods to allow companies to react more swiftly to become more competitive, and drive higher business value. As described in Chapter 4, “Big data use cases” on page 43, Big Data Exploration addresses the challenge that every organization faces, which is how to make sense of the plethora of information from a vast array of new and traditional sources to derive new value.

## **7.5 Big data usage and adoption phases**

There are three phases to big data usage and adoption. Each of these phases corresponds to a phase in the governance process and also aligns to the privacy lifecycle:

1. Exploration Phase (Assess): Identify raw data of interest to present together or for further downstream processing.
2. Prepare and Govern Phase (Assess and Protect): This phase takes place in the Integration and Information Governance Zone: profile, identify risk, determine required integration and processing, and generate metadata.
3. Consumption phase (Sustain/Respond): Whether an application is providing a unified view of federated data for decision making, a 360 degree view of the customer, or perhaps streaming analytics, when an application is deemed to be delivering value and used for business decisions, governance methods must be in place, along with the appropriate measurement criteria. Defense in depth is a key principle of securing data, be it big data or traditional data. Your security implementation is only as strong as its weakest link.

## 7.5.1 Exploration phase

As described by Privacy by Design principles, the security and privacy approach must fit the business usage and stage of the development lifecycle. In the initial Big Data Exploration phase, which is shown on the left of Figure 7-2, your primary goal is to identify useful data for the purpose at hand and to make all important and interesting data available. By keeping raw data in the Discovery Zone, you can provide initial exploration and analytics to key users and meet Governance Principle #1: Move information as quickly as practical while keeping the quality as high as practical and as secure as practical.

What is key with Big Data Exploration in the Exploratory phase is its goal of identifying value and use before further processing, such as standardization, matching, refinement, movement, and downstream ingestion or landing. One method of performing big data Exploration is with IBM InfoSphere Data Explorer, which is a powerful tool that helps you profile and classify this information, visualizing and navigating federated data and content.

Fit-for-Purpose Governance	
Initial or Exploratory Use Cases	Used for Business Decisions
Little security concerns	Protect, Secure, Encrypt
Sporadic change management	Audit trail tracking access & changes
No data retention requirements	Preserve data for N years
Little to no regulation	Legislated requirements
No / isolated data quality concerns	Data quality imperatives
Sources of information are “interesting”	Sources must be trusted
No difference in data governance requirements once the data is used for making operational business decisions	

Figure 7-2 Fit-for-Purpose Governance

This type of initial work must be performed in a quarantined landing zone, where only a small number of authorized individuals have access to the data. Because you have not yet evaluated the usefulness of this raw data, you also have not yet classified the information to determine the appropriate security and privacy controls for the Consumption phase.

### **Exploration phase: Security and privacy risks**

The initial Exploration phase presents the most risk by virtue of allowing any and all raw data in the initial assessment. There might be cases where raw data yields the most insights, such as in the case of fraud detection and entity analytics, tying seemingly unrelated individuals together based on a common data element. Raw data may also be brought in from other sources and placed in the Information Ingestion and Operational Information Zone.

In this ungoverned zone, several security and privacy risks are apparent:

- ▶ Risk of exposing sensitive data to a breach: Because you have not yet identified where sensitive data is, it is easy to see that exploratory processes that use customer or human resources data could contain PII, such as names and addresses. Because big data typically means massively greater volumes of information, the potential fines and penalties could be dramatically increased.
- ▶ Additional exposures are also possible in the sense that combining or linking data sources might expose sensitive information when data elements are combined, such as combining telematics location information with name and address.
- ▶ Plain old “bad” data: There is a risk in using unvetted external data sources, such as product reviews, and tweets. With a new data source, it is critical to validate the source and lineage. Validation may include classification with “sniff tests”, such as value ranges and transaction size. Assigning a “trust” or validation ranking also might be useful in weighting the veracity of the source. It might be useful to also compare it with another data source. “Bad” data presents a privacy risk in that it might result in inaccurate decisions that affect individuals, such as credit-worthiness.

Although the chart specifies “little security concerns”, that does not mean “no security concern”, for the reasons that are mentioned above. There is a general acceptance that there are unknown risks in raw data. A simple “sanity” test must be applied here so that if the data sources concern customer, employee, financial data, or intellectual capital, then basic security and privacy controls should be used, such as user IDs and secure passwords and limiting access to a select group of individuals. Additional controls, such as file system encryption and activity monitoring (described in “Privacy Operational Lifecycle” on page 138) can also be added if the subject area is of high risk. A preferred practice recommendation is to apply basic controls to the quarantine zone.

Applying an agreed upon set of basic controls to the quarantine zone by default ensures that no matter what the data source, the risk is diminished. You might consider performing rudimentary Privacy Impact and Security Risk Assessments when the subject areas are known.

## **7.5.2 Prepare and Govern Phase (Assess and Protect)**

In this phase, you finalize your data sources and prepare them for consumption. First, you must understand the data source, its “trust factor”, the data context and meaning, and how it maps to other enterprise data sources. You also must determine whether to operationalize (and retain) specific data sources, and which zone to land the data, that is, Hadoop, Data Warehouse, leave in place, and so on. As part of this process, it is critical to conduct both a PIA and a Security Risk Assessment. After you determine that the subject area contains information that is either private or has specific security requirements, you must complete the following steps:

1. Inventory and classify sensitive data. InfoSphere Data Explorer contains capabilities to assist you with this effort, as described in 7.5.3, “Inventorying and classifying sensitive data” on page 149.
2. Identify and match against the legal, contractual, and organizational data protection requirements with assistance from your security, privacy, and compliance organization.
3. Identify protection standards for each classification. Again, your organization should have defined the level of protection that is needed. For example, all credit card numbers must be encrypted in accordance with PCI DSS.
4. Identify the gaps and set up remediation plans. In this step, you identify complementary IBM solutions and how they can help you secure and protect your big data investment. The solutions that are described in the following sections are InfoSphere Data Explorer, InfoSphere Guardium for Hadoop, InfoSphere Guardium Data Encryption, and InfoSphere Optim for Data Privacy.

These four steps are specific to IT implementation and are described in detail in the following sections.

### 7.5.3 Inventorying and classifying sensitive data

In this step, you identify any data elements and defining metadata around sensitive information that pertain to both security and privacy. Effective data privacy begins with an agreement that outlines the purpose, accountabilities, and participants in the data privacy strategy. Not all data must be protected in the same way; some data might be considered low risk and not worth the time and effort that is required to secure it. The first step is to define sensitive data, but defining it is not solely an IT function. A cross-functional team, including marketing, sales, line-of-business (LOB), operations, and IT, should work together to create the definition. High-value data such as design specifications or corporate secrets might not require protection under legal mandates, but organizations most certainly want to protect it with stringent controls. The team should decide what data to protect based on business priorities. The goal is to create a clear definition of what must be protected regarding regulatory mandates. The questions that follow provide an example of the considerations that should guide this part of the definition phase.

Define sensitive data by answering the following questions:

- ▶ What data is considered sensitive?
- ▶ What are the components of PII?
- ▶ What is the definition of high-risk data or corporate secret data?
- ▶ What data is affected by legal mandates and what is not?
- ▶ Where is sensitive data copied across the enterprise to and from the big data environment?
- ▶ Is sensitive data being shared with third parties or outsourced for testing, development, or QA work?
- ▶ Who has a valid business need to know sensitive data?

The outcome of your efforts should be a business glossary, which is an authoritative dictionary of data privacy terms and relationships that are employed across the enterprise. Designed to be accessible across the enterprise, the business glossary defines the terms that are used to build enterprise data privacy policies. All employees can use this central source for standard definitions of sensitive data, which helps remove reactionary processes and guesswork.

A good rule of thumb is any customer, human resources, financial data, or intellectual capital all require information inventories. For the Data Warehouse environment, both InfoSphere Guardium and InfoSphere Discovery can identify and classify sensitive data, such as Social Security Numbers or credit card numbers. Additionally, Guardium can scan the network for unregistered databases, such as developer databases on workstations. For big data, more sophisticated text analytics might be required.

### **Identifying sensitive data using InfoSphere Data Explorer**

Knowledge and location of sensitive information (individual and combined data elements) that might be included in a big data repository is critical. It can be performed at various stages of processing to identify and understand privacy and compliance implications as the variety of data sources are identified and included in to the gathering of data. Consider these factors of the data source identification and impact of loading into the big data environment:

- ▶ Understand the generation of the data and the nature of the business/enterprise value and risk to the organization. If the data source is documented and sourced, then the value of the information and risk can be factored into the inventory.
- ▶ The individual data source is not understood or the nature of the business/enterprise value and risk to the organization is not understood. This situation must be investigated so that the value of the information and risk can be factored into the inventory.
- ▶ Combination of multiple data sources with links to associate the disparate sources to provide business/enterprise value and risk to the organization.

To illustrate this situation, here is a fictional financial company that recently progressed from a big data sandbox to a pilot and realized the need to investigate and verify whether any sensitive information is dispersed on the data/database nodes. If found, the data must be identified and classified, and the results made available for analysis and review. Collaboration in this process is critical in an enterprise environment because there are many application data owners and the data sources can cross these application boundaries. After access is provided, it is not a far leap to the data being included in local repositories or being passed internally or externally (it is shared).

The following “use cases” were agreed on for a proof of concept using InfoSphere Data Explorer:

- ▶ Ability to connect to the data stores (relational databases, SharePoint repositories, HTML, email, and XML-based content)
- ▶ Ability to discover data in different file formats (text, log, compressed, PDF, and Microsoft Office files)



- ▶ Ability to identify and classify data (PII data, such as credit card numbers, Social Security numbers, telephone numbers, and email addresses)
- ▶ Ability to provide a common discovery dashboard/reporting capabilities and support for user collaboration

In the results, InfoSphere Data Explorer identified and classified PII data in the big data sandbox. Here are the pertinent comments:

- ▶ Not all data sources were defined or understood for the content. This POC provided the company the opportunity to reassess the process of loading the data and where to include a task to verify the data for sensitivity and compliance in the process flow before placing it in to the big data lake.
- ▶ The discovery of PII information in the big data sandbox during the POC also underlined the need to better understand the source data systems, review the extraction and load methods of the data, and include steps and processes to address any sensitive data and compliance issues (that is, assess the risk).
- ▶ Improved compliance, reduced risk, and better metadata management. The big data initiative can provide such a robust environment that sometimes there is a rush to include content before its security risk is considered.

### ***Architecture***

The key to a successful information optimization solution is a rich platform that can automate the optimization process while being quickly deployed and configured to meet each individual organization's need. InfoSphere Data Explorer is built on a modern architecture and takes advantage of XML and XSL standards. By using a web-based interface, administrators can perform complex configurations through a simple point-and-click administrative interface. Configuration can also be done through an extensible set of REST/SOAP APIs. Through web services and SOA protocols, InfoSphere Data Explorer can easily pull data from existing applications, creating a universal gateway to securely access information among disparate systems within an organization.

InfoSphere Data Explorer provides various templates that simplify solution configuration and deployment. These templates include pre-built connectors, predefined reports, display templates and more. For each feature, InfoSphere Data Explorer provides an easily adaptable template to create custom configurations. This gives the administrator the power to configure features and functions that are tailored to their own information environment.

InfoSphere Data Explorer, which is shown in Figure 7-3, is configured to scale broadly across your organization and support redundancy and reliability. InfoSphere Data Explorer can support 99.99 percent up time and ensure high performance across a distributed environment. In addition, InfoSphere Data Explorer supports the ability to handle terabytes of information with minimal infrastructure costs compared to other solutions on the market.

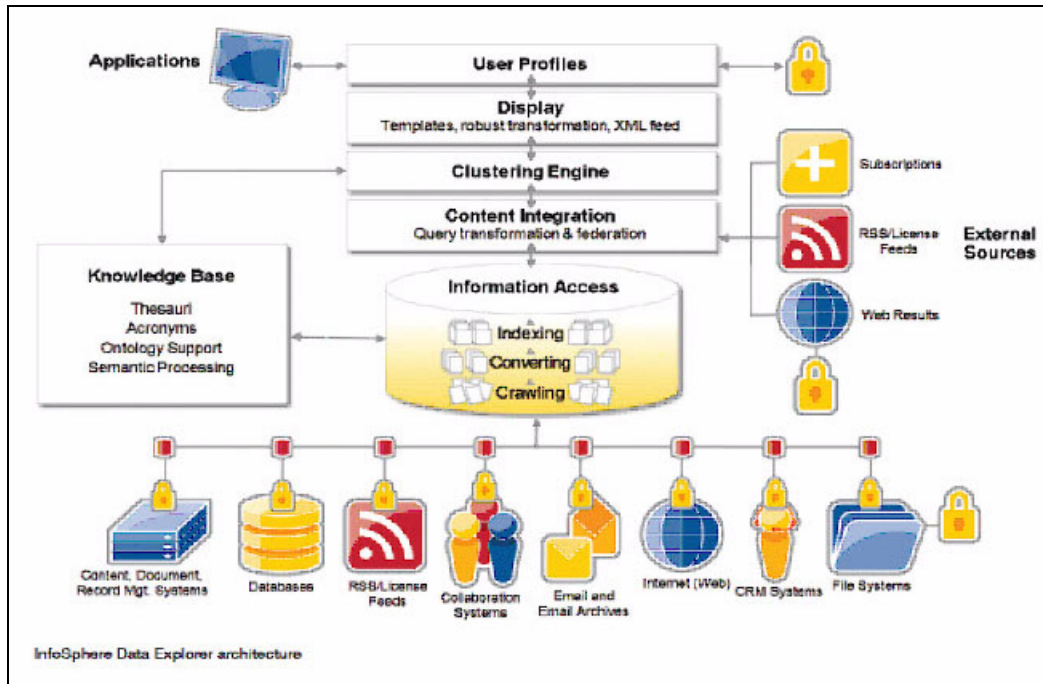


Figure 7-3 IBM Data Explorer architecture

InfoSphere Data Explorer provides these features and benefits:

- ▶ Secure, federated discovery, navigation, and search:
  - Provides access to data stored by a wide variety of applications and data sources, both inside and outside the enterprise, including content management, customer relationship management, supply chain management, email, relational database management systems, web pages, networked file systems, Data Warehouses, Hadoop-based data stores, columnar databases, and cloud and external web services.
  - Includes federated access to non-indexed systems, such as premium information services, supplier or partner portals, and legacy applications through the InfoSphere Data Explorer Query Routing feature.

- A relevance model accommodates diverse document sizes and formats while delivering more consistent search and navigation results. Relevance parameters can be tuned by the system administrator.
- A security framework provides user authentication and observes and enforces the access permissions of each item at the document, section, row, and field level to ensure that users can view only information that they are authorized to view in the source systems.
- Provides rich analytics and natural language processing capabilities, such as clustering, categorization, entity and metadata extraction, faceted navigation, conceptual search, name matching, and document de-duplication.
- ▶ Rapid development and deployment framework:
  - InfoSphere Data Explorer Application Builder enables rapid deployment of information-centric applications that combine information and analytics from multiple sources for a comprehensive, contextually relevant view of any topic, such as a customer, product, or physical asset.
  - A widget-based framework enables users to select the information sources and create the personalized view of information that is needed to perform their jobs.
  - Entity pages enable presentation of information and analytics about people, customers, products, and any other topic or entity from multiple sources in a single view.
  - An Activity Feed enables users to “follow” any topics, such as a person, company, or subject and receive the most current information, and post comments and view comments that are posted by other users.
  - A comprehensive set of application programming interfaces (APIs) enables programmatic access to key capabilities and rapid application development and deployment options.
- ▶ Distributed, highly scalable architecture:
  - A compact, position-based index structure includes features such as rapid refresh, real-time searching, and field-level updates.
  - Updates can be written to indexes without taking them offline or rewriting the entire index, and are instantly available for searching.
  - Updates can be written to indexes without taking them offline or rewriting the entire index, and are instantly available for “shared nothing” deployment.

- ▶ Flexible data fusion capabilities
  - Information from multiple sources can be combined into “virtual documents”, which contain information from multiple sources.
  - Large documents can be automatically divided into separate objects or subdocuments that remain related to a master document for easier navigation and comprehension by users.
  - Enables creation of dynamic “entity pages” that allow users to browse a comprehensive, 360 degree view of a customer, product, or other item.
- ▶ Collaboration features:
  - Users can tag, rate, and comment on information.
  - Tags, comments, and ratings can be used in searching, navigation, and relevance ranking to help users find the most relevant and important information.
  - Users can create virtual folders to organize content for future use and optionally share folders with other users.
  - Navigation and search results can return pointers to people to enable the location of expertise within an organization and encourage collaboration.
  - Shared Spaces allow users to collaborate about items and topics that appear in their individualized views.

## **Setting up remediation plans**

Now that you have a good idea of what information must be protected, it is time to decide on how each of the described methods can assist in your data protection strategy. The methods include file and directory permissions, monitoring data usage, encryption, masking (anonymization), and archiving.

### ***File and directory permissions***

Authorization is the process of providing access level control (ACL) of data. An entitlement that is granted to an account on a target system enables the account owner (user) to perform a specific task or function. An entitlement can be a role, responsibility, or group membership. For example, if user Joe is granted the 'Budget Account' role on a target system, then Joe can use that entitlement to access and generate budget-related reports from the target system. Relational database sources have a comprehensive and granular set of controls for access and manipulation of the data; unstructured data sources have a less comprehensive set of controls for directory and file level access.

Some Hadoop systems offer methods of setting ACL permissions (InfoSphere BigInsights has this feature). The BigInsights System Administrator can change permissions on directories and files to allow or restrict the access of users, and assign those users to roles for easier management. BigInsights follows the standard Access Control Lists standards that are used within POSIX file systems.

These unstructured data sources do not have the granularity of enterprise relational database systems that provide ACL to the column level of a table, view, or synonym.

### ***Monitoring data usage activity: Guardium data activity monitoring***

The proliferation of data from endpoint devices, growing user volumes, and new computing models such as cloud, social business, and big data, have created demands for data access and analytics that can effectively handle staggering amounts of data. Hadoop-based systems (a set of open source components) have emerged to address the challenges of analyzing data and turning it into actionable insight. Hadoop consists of a distributed file system and a MapReduce application framework and additional optional components, such as Hive for Data Warehouse queries and HBase, a NoSQL database, for fast record lookups. From a data governance perspective, this environment of mixed technologies and data formats across a large volume of storage represents a significant potential risk of sensitive data exposure, misuse of the information, and data breaches. Any company building such an information platform looks at all existing data sources as potential value and at least needs to be considered for inclusion. How does security and governance fit onto this new platform to meet compliance?

For this reason, organizations using any combination of relational databases, Hadoop, or NoSQL databases still need to implement preferred practices for auditing and compliance:

- ▶ Continuous real-time monitoring to ensure that data access is protected and audited.
- ▶ Policy-based controls that are based on access patterns to rapidly detect unauthorized or suspicious activity and alert key personnel.
- ▶ Protection of sensitive data repositories against new threats or other malicious activity.

- ▶ Demonstrate compliance to pass audits: It is not enough to develop a holistic approach to data security and privacy; organizations must also demonstrate and prove compliance to internal and external auditors.
- ▶ Provide patterns of data access (who, what, where, when, and how) to which data and what location. This profiling can be of added value for analysis of identifying detailed data that is of value, summarized and made available in a warehouse zone.

### ***Benefits***

InfoSphere Guardium can be the solution to your monitoring and compliance of all data access to your company's heterogeneous data repositories needs. InfoSphere Guardium provides full visibility into the Hadoop data environment, making it possible to identify unauthorized activities, such as data tampering or hacking, in real time. Automation of the entire security and compliance lifecycle reduces labor costs, facilitates communication throughout the organization, and streamlines audit preparation. This solution has the following benefits:

- ▶ Single integrated appliance
- ▶ Definition of granular security policies, using indicators of possible risk (appropriate for the particular environment), including the file or data object, type of access (reading, updating, deleting), user ID, MapReduce job name, and more
- ▶ Definition of actions in response to policy violations, such as generating alerts and logging full incident details for investigation
- ▶ Automating compliance workflow for routine activities, such as remediation or approval of new MapReduce jobs, including steps such as sign-offs, commenting, and escalation
- ▶ Ready-to-use reports that are tailored for Hadoop and a full and customizable reporting capability
- ▶ Dynamically scalable for large organizations and growth

### ***Architecture***

A nonintrusive architecture (see Figure 7-4 on page 157) that requires no configuration changes to the Hadoop servers, InfoSphere Guardium provides full visibility into data activity through the Hadoop software stack, and provides full separation of duties.

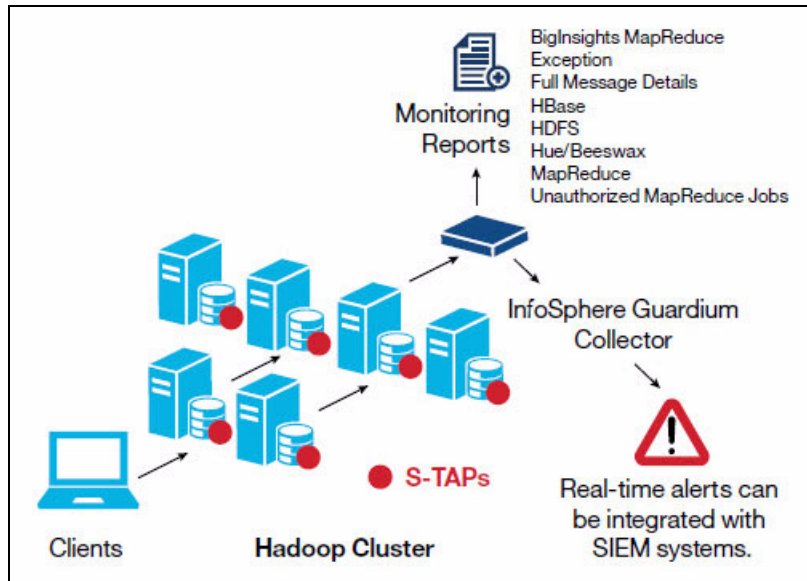


Figure 7-4 Guardium and Hadoop

InfoSphere Guardium components can be grouped into appliances and agents:

► Appliances:

- Collectors: The collector is the appliance that is used for real-time capture and analysis of the database activity.
- Aggregators: The aggregator appliance is used to offload reporting activity from the collectors, and to provide consolidated reporting from multiple collectors.
- Central Managers: The central manager (CM) is a specialized function that is enabled on an aggregator appliance. The central manager function is used to manage and control multiple Guardium appliances.

► Agents:

- Software TAP agent (IBM S-TAP®): The S-TAP agent is installed on the database server and is used to monitor and relay the observed activity to the Guardium collector appliance.
- Guardium Installation Manager agent (GIM): The GIM agent is installed on the database server and is used to facilitate the agent installation, updating, and configuration modification of agents.

- Change Audit System agent (CAS): The CAS agent is installed on the database server and is used to capture change audit information of configuration files and more on the database server.
- Instance Discovery agent: The instance discovery agent is installed on the database server and is used to obtain database, listener, and port information.

S-TAPs are installed on key servers in the Hadoop cluster, such as the NameNode. These software taps rapidly stream network packets over to a separate, tamper-proof collector. Because processing of the network traffic occurs on the collector, the impact on the Hadoop cluster is minimal.

The InfoSphere Guardium deployment for a NoSQL environment follows the same architecture as Hadoop. In Figure 7-5, in a MongoDB cluster, S-TAPs are deployed on the individual MongoDB shards and report to a collector.

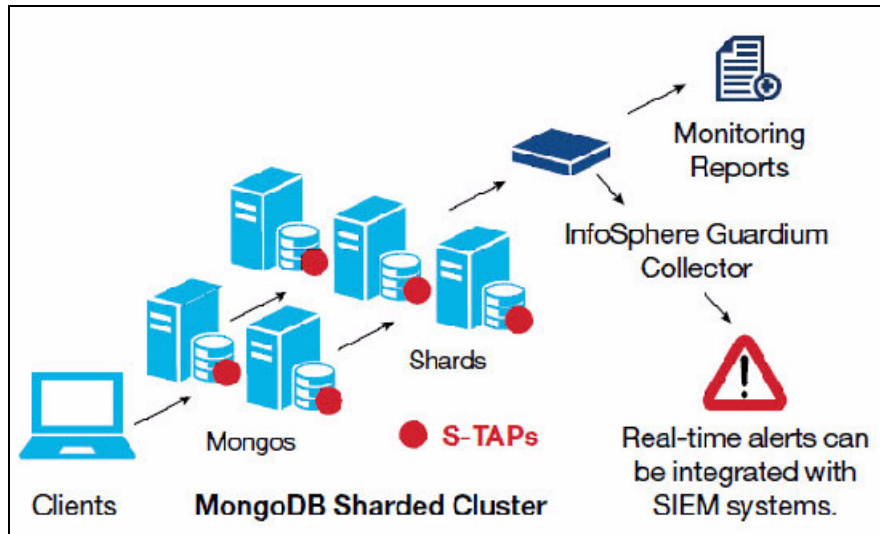


Figure 7-5 Guardium and MongoDB



The InfoSphere Guardium web console provides centralized management of alerts, report definitions, compliance workflow processes, and settings (such as archiving schedules) without the involvement of Hadoop administrators, thus providing the statement of direction (SOD) that is required by auditors and streamlining compliance activities, as shown in Figure 7-6.

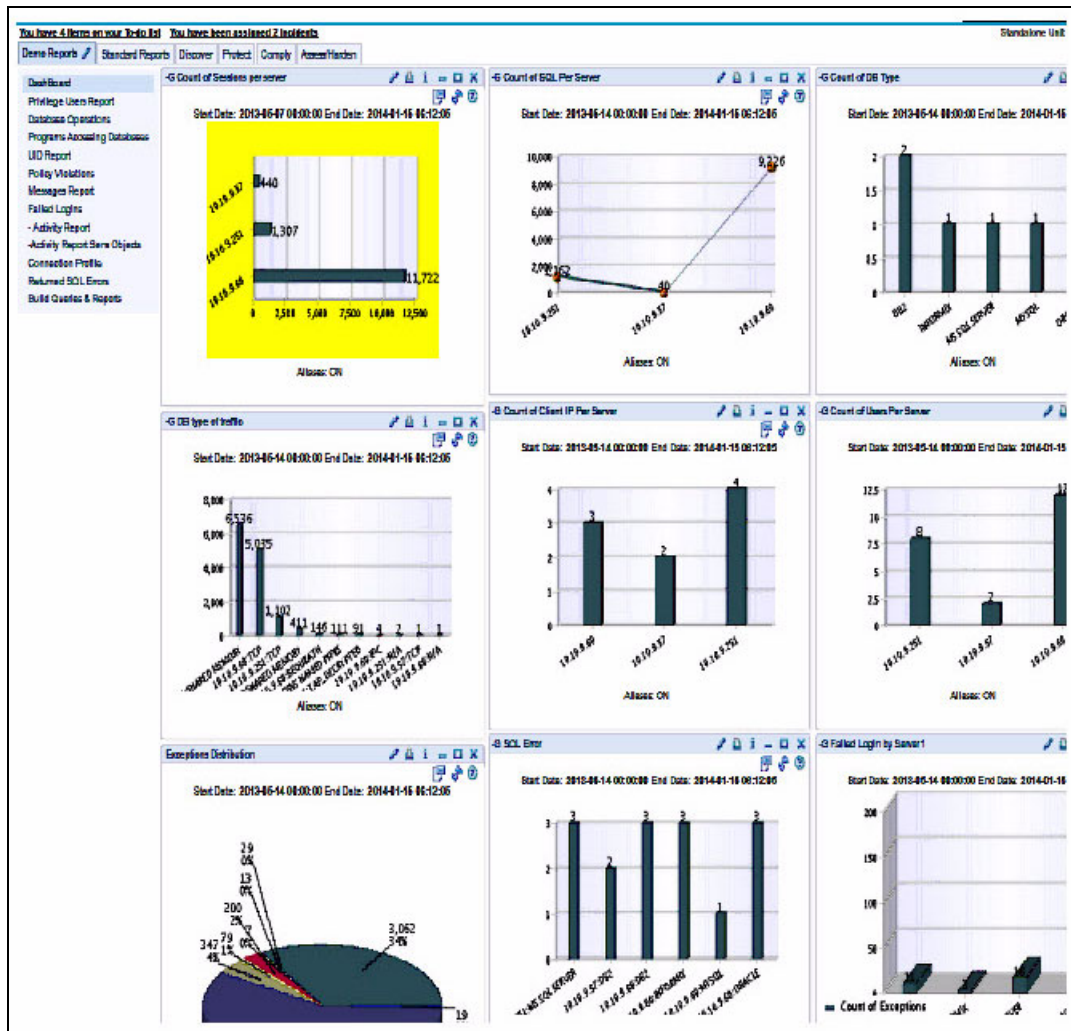


Figure 7-6 InfoSphere Guardium Web Console examples

## Enterprise

The InfoSphere Guardium is focused on being enterprise ready; that is, InfoSphere Guardium can scale in an efficient and cost-effective manner. Every release of InfoSphere Guardium introduces significant improvement in scalability, integration, and automation-related features, with one goal in mind: streamline the administration, configuration, and usage of the solution in large environments.

InfoSphere Guardium can scale horizontally by adding more collectors as more data repositories are added to the big data environment, and vertically to centralized management and aggregation of the audit data. Real-time alerts occur at the collector level, and reporting and workflow occur at the central management and aggregation level. The entire InfoSphere Guardium deployment can be managed from one console, from configuration of alerts at the collectors, to the management of compliance at central management. The InfoSphere Guardium architecture is shown in Figure 7-7.

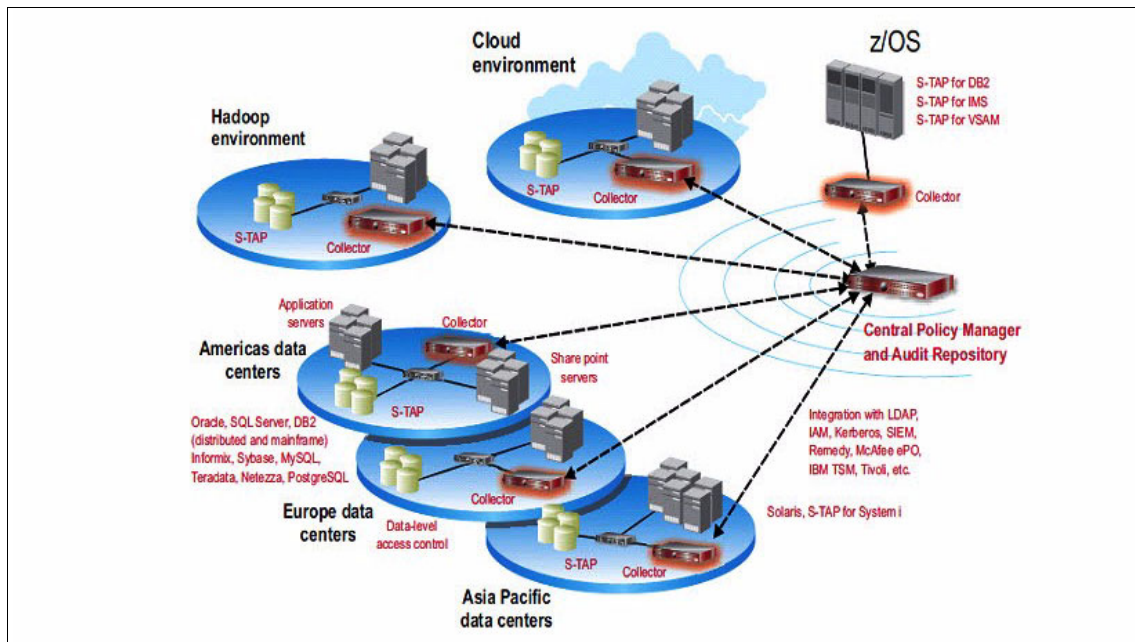


Figure 7-7 InfoSphere Guardium architecture

InfoSphere Guardium has built-in load balancing, failover, and redundancy for all appliances, and collector monitoring at the central manager for an enterprise view of the health of the system. Additionally, InfoSphere Guardium has self-monitoring alerts that can notify the InfoSphere Guardium administrator if any issue arises. All these capabilities are enabled through the InfoSphere Guardium console.

### ***Encrypting sensitive data using Guardium Data Encryption***

The most proven line of defense against data breaches is to protect the data itself with encryption. A comprehensive encryption strategy includes the ability to protect both structured and unstructured data stores, which might include database and file server files, documents, image scans, voice recordings, and logs, across a heterogeneous IT infrastructure. To gain the maximum benefit from their encryption programs, security-conscious enterprises are using intelligent encryption that uses access controls to ensure that data is decrypted only for authorized requests. This encryption scheme has the following characteristics:

- ▶ Security for your structured and unstructured data:
  - High performance encryption, access control, and auditing
  - Data privacy for both online and backup environments
  - Unified policy and key management for centralized administration across multiple data servers
- ▶ Transparency to users, databases, applications, and storage:
  - No coding or changes to existing IT infrastructure
  - Protect data in any storage environment
  - User access to data
- ▶ Centralized Administration
  - Policy and key management
  - Strong separation of duties
  - Audit logs
  - High availability

## Benefits

Simple to deploy and manage, InfoSphere Guardium Encryption Expert provides powerful encryption of data across enterprise IT infrastructures transparently and with minimal performance impact. With high-speed encryption, integrated key management, and context-aware access controls, it helps protect valuable information to help companies address industry and government regulations and lower the risk of devastating data breaches. A business does not want to be limited from new IT innovations and choices, but still is required to protect and meet compliance:

- ▶ As organizations shift from physical to virtual environments, and subsequently use infrastructure in the cloud, the security of their data becomes an even greater concern.
- ▶ PCI DSS, HIPPA, HIPPA-HITECH compliance is often necessary.
- ▶ U.S-EU Safe Harbor is a provision of a statute that specifies that certain conduct is deemed not to violate a given rule. This provides protection to a business if there is a data breach and the breached data was encrypted: this greatly reduces the requirement for disclosure of the breach and notification.

## Architecture

The flexibility and scalability of the solution's design stems from its separation of the manager from the agents. InfoSphere Guardium Encryption Expert Manager provides centralized administration of encryption keys and data security policies, while InfoSphere Guardium Encryption Expert Agents help protect structured and unstructured data stores.

Figure 7-8 on page 163 shows an overview of the InfoSphere Guardium Encryption Expert.

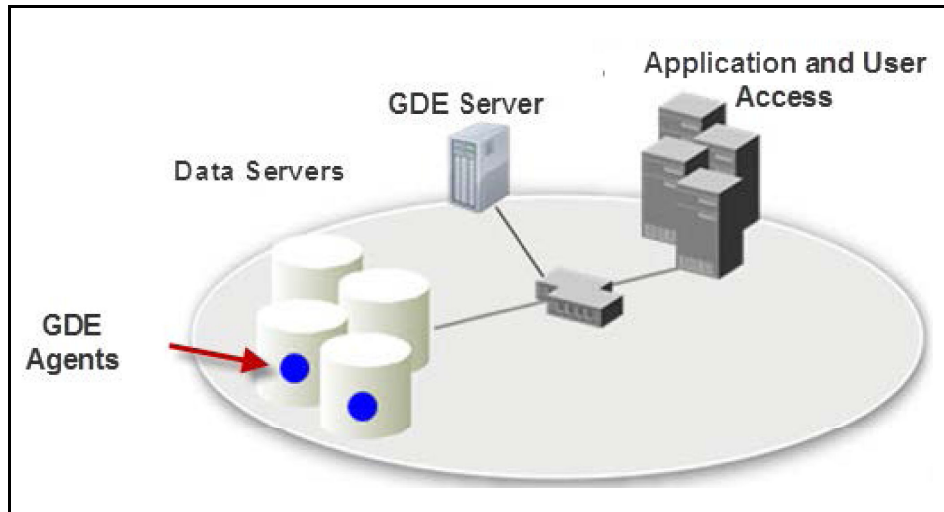


Figure 7-8 Guardium Data Encryption overview

InfoSphere Guardium Encryption Expert (GDE) consists of the following two major components:

- InfoSphere Guardium Encryption Expert Manager

The InfoSphere Guardium Encryption Expert Manager integrates key management, data security policy management, and event log collection in to a centrally managed cluster that provides high availability and scalability to thousands of InfoSphere Guardium Encryption Expert Agents. This enables data security administrators to easily manage standards-based encryption across Linux, UNIX, and Windows operating systems in both centralized and geographically distributed environments. The InfoSphere Guardium Encryption Expert Manager stores the data security policies, encryption keys, and audit logs in a hardened appliance that is physically separated from the InfoSphere Guardium Encryption Expert Agents.

- InfoSphere Guardium Encryption Expert Agents

InfoSphere Guardium Encryption Expert Agents are software agents that sit above the file system logical volume layers. The agents evaluate any attempt to access the protected data and apply predetermined policies to either grant or deny such attempts. They maintain a strong separation of duties on the server by encrypting files and leaving their metadata in the clear so IT administrators can perform their jobs without directly accessing the information.

The agents perform the encryption and decryption, and access control works locally on the system that is accessing the data at rest, which enables encryption to be distributed within the data center and out to remote sites while being centrally managed through the InfoSphere Guardium Encryption Expert Manager cluster.

The two GDE components are shown in Figure 7-9.

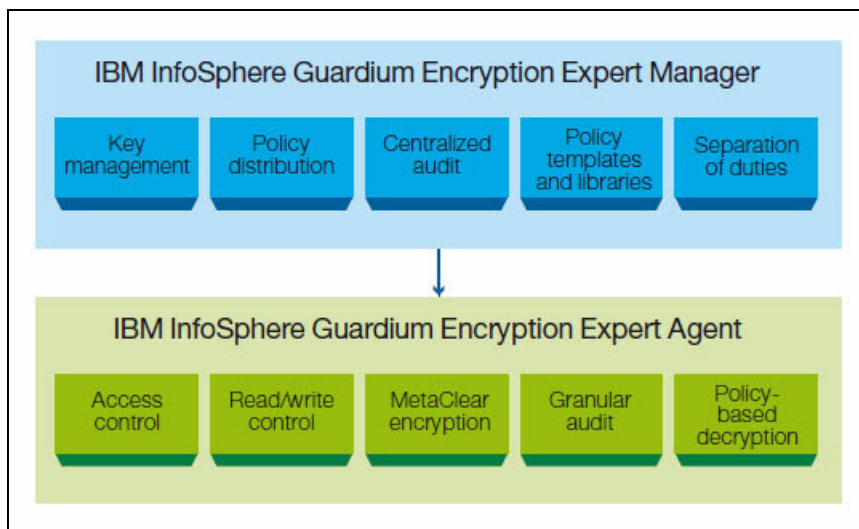


Figure 7-9 GDE Server and Agent

Security management domains combine role-based administration of IBM InfoSphere Guardium Encryption Expert with the ability to compartmentalize the management for policies, data encryption keys, agent configurations, and audit logs, which provides specific management within a business enterprise for various business units and their privacy/compliance requirements and ownership through role-based administration and domains, as shown in Figure 7-10.

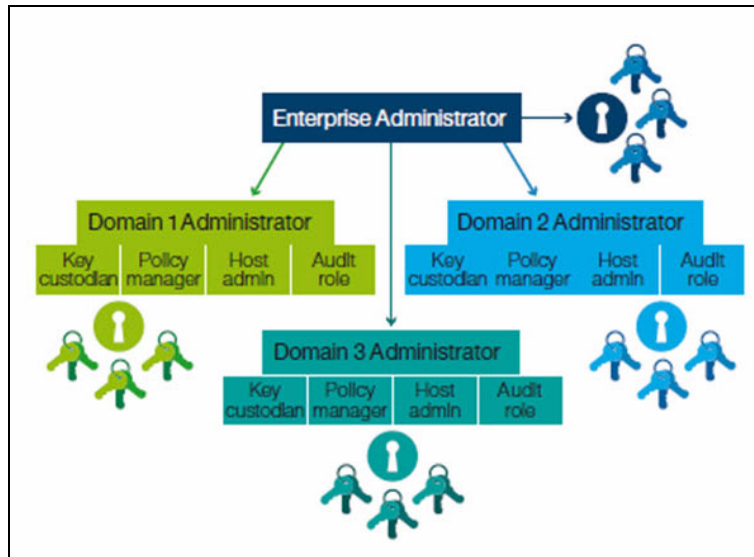


Figure 7-10 Role-based administration and domains

### ***Masking (obfuscating) and anonymizing sensitive data***

Masking, also known as obfuscation or anonymization, is the process of substituting sensitive information with fictitious yet realistic-looking data. Although masking originated as a method for testing test environments, it has rapidly evolved to the concept of masking on demand (MOD). MOD provides consistent, repeatable masking services across all platforms, data sources, and use cases, whenever and wherever they are needed. For example, masking can be performed at rest or in flight, using Hadoop or traditional relational data, such as flat files and data sets, IBM IMS™, and VSAM, during query, load, display, processing, and so on.

The rationale behind MOD is that even production environments can use masked data and reduce the risk of sensitive data disclosure. For example, analytics against a population could use anonymized individual information, preserving individual privacy while still maintaining the utility of the underlying data. A policy-driven, on-demand masking approach enables organizations to proactively protect data privacy and support compliance. This approach is especially important in the current era of computing, where data is everywhere and is continuously growing in volume, variety, and velocity. On the surface, it appears that data masking is a simple concept. However, it is technically challenging to run.

Many organizations operate complex, heterogeneous IT environments that are composed of multiple, interrelated applications, databases, and platforms. A data masking solution must consistently and effectively mask sensitive data across related data sources while also preserving application and database integrity.

This section describes the following four scenarios for masking:

1. Masking data inside of Hadoop within a MapReduce job
2. Masking data before placing it by using the InfoSphere Optim Hadoop loader
3. Masking inside of a big data database using user-defined functions (UDFs)
4. Masking data before its ingestion into the traditional Data Warehouse by using either an Eclipse-based GUI or an ETL tool

For each of the four use cases, you can use the same consistent masking functions, also known as Privacy Providers. InfoSphere Optim Data Privacy Provider Library (ODPP) is a set of privacy algorithms that are referred to as “providers” and that have the following characteristics:

- ▶ Consistent behavior across platforms
- ▶ Data source independent design
- ▶ Dynamic invocation of masking services
- ▶ Written in cross-platform ANSI and C/C++
- ▶ Interfaces with other languages that support the C calling conventions
- ▶ Provides platform abstraction services
- ▶ Supports SBCS, MBCS, and Unicode
- ▶ Deploys as a number of shared libraries

Here are the Privacy Provider functions:

- ▶ AFF: Mask anything with format affinity.
- ▶ AGE: Age dates upwards or downwards.
- ▶ CCN: Mask credit card numbers.
- ▶ EML: Mask email addresses.



- ▶ HASH: Mask anything to a hash value.
- ▶ NID: Mask national IDs (CA, ES, FR, IT, UK, and US).
- ▶ Semantic Masking, which is masking in groups or buckets to preserve statistical results for a specific category, such as the following ones:
  - SWAP: Random, Class-, or Distance-based swapping
  - NOISE: Noise Addition
  - PRAM: Post Randomization

As of InfoSphere Optim Version 9.1 FP4, you can also use the industry standard Secure Hash Algorithm (SHA-256), which is available with the Hash, Hash\_Lookup, and Email Privacy Providers. The hashing seed can be provided as a Hash Message Authentication Code (HMAC) key exit for greater security in cryptography. A keyed-hash message authentication code is used to calculate a message authentication code (MAC) involving a cryptographic hash function in combination with a secret cryptographic key.

Here are the four scenarios for masking:

#### 1. Masking data inside of Hadoop within a MapReduce job

In this scenario, you apply InfoSphere Optim data masking to an IBM BigInsights HDFS data source using native MapReduce jobs. The input file is a comma-separated file that contains sensitive customer data, including credit card numbers, emails, and addresses, as shown in Figure 7-11.

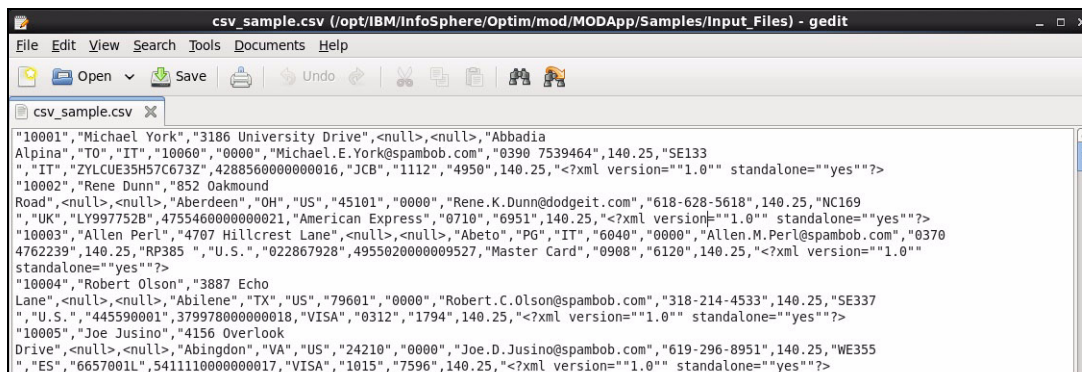


Figure 7-11 MapReduce example customer input file

To run masking functions, you must identify the information to be masked by using a configuration file. This file is the runtime configuration file that your MapReduce masking job uses to determine what data to mask and where to find the data source files for masking, the type of file, where to output the new masked files, what columns exist in the flat file, and what column data must be masked.

A snippet of the configuration file is shown in Figure 7-12, which shows masking of a credit card number by using the CCN masking provider.

```
# For fields that are neither masked nor participating in the key, a simple name will do.
field=ADDRESS1
field=ADDRESS2
field=LOCALITY
field=CITY
field=STATE
field=COUNTRY_CODE
field=POSTAL_CODE
field=POSTAL_CODE_PLUS4
field=EMAIL_ADDRESS
field=PHONE_NUMBER
field=YTD_SALES
field=SALESMAN_ID
field=NATIONALITY
field=NATIONAL_ID

# To mask a field, the mask keyword is used. The string after the equals sign should be enclosed in double-quotes (as is typical for an
ODPP/UDF/LUA masking string)
# The syntax used is identical to that used by ODPP, UDFs and Lua scripts, and if the datatypes and column-names match, the strings
should be copy-and-pasteable.
# The 9.1.0.4 release of the Masking On Demand Application supports 6 providers: CCN (credit card masking), AFF (affinity, or the "trans
col" masking algorithm),
# NID (national ID), AGE (the aging provider), EMAIL (E-mail masking), and HASH (generates an integer hash code from a string).
# All take the datatype WVARCHAR SZ, except for AGE, which also accepts the DATETIME SZ data type.
# The masking arguments are covered in The ODPP Syntax Guide, which is distributed in the Masking On Demand Application installation
package.
field=CREDITCARD_NUMBER,mask="PROVIDER=CCN,FLDDEF1=(NAME=CREDITCARD_NUMBER,DT=WVARCHAR_SZ,LEN=80),MTD=repeatabl,WHENINV=PRE"
field=CREDITCARD_TYPE
field=CREDITCARD_EXP
field=CREDITCARD_CVV
field=DRIVER_LICENSE
field=CREDITCARD_HISTORY
```

Figure 7-12 Sample MapReduce configuration file

After you complete the definition of the configuration file, it is time to run the masking MapReduce job. A shell script for running the job (**IOQMaskingTool.sh**) is in /opt/IBM/InfoSphere/Optim/mod/MODApp. This shell script requires little configuration (none in most cases). In this example, we configured the environment so that the default settings for this script suits our needs. To run the job, run the following command:

```
./IOQMaskingTool.sh -i Samples/Masking_OPTIM_CUSTOMERS.ioqmoda
```

We specified the .ioqmoda configuration file to direct the masking job to where to find the flat files for masking, where to output the newly generated masked flat files, and what fields to mask in the data. After the job completes, you can use a text editor, such as gedit, to compare the source and target files and observe that the masking was applied. For example, in the third record, you see that the email address, phone number, national ID, and credit card number was replaced with new values. The email address's domain was preserved, although it can easily be modified. The phone number was randomly generated with new numbers. The National ID was masked, but remains American. Lastly, although the credit card number was masked, the credit card number remains a Master Card, as indicated by the first four digits.

Figure 7-13 and Figure 7-14 show these changes.

```
"10003","Allen Perl","4707 Hillcrest Lane",<null>,<null>,"Abeto","PG","IT","6040","0000","Allen.M.Perl@spambob.com","0370 4762239",140.25,"RP385", "U.S.", "022867928", "4955020000009527" "Master Card", "0908", "6120", 140.25, "<?xml version='1.0' standalone='yes'?">
```

Figure 7-13 MapReduce example 3

```
"10003","Allen Perl","4707 Hillcrest Lane",<null>,<null>,"Abeto","PG","IT","6040","0000","email3@spambob.com", "3178 3242687", "140.25", "RP385", "U.S.", "022843740", "4955038576790023", "Master Card", "0908", "6120", "140.25", "<?xml version='1.0' standalone='yes'?">
```

Figure 7-14 MapReduce example 4

## 2. Masking data before placing it by using the InfoSphere Optim Hadoop loader

In the following scenario, we demonstrate how InfoSphere Optim can extract and convert data from heterogeneous data sources, including flat files and DBMSs, into Hadoop Distributed File System (HDFS) files. An important point to note is that there is never any unmasked data on the Hadoop system, which is often a requirement for highly secure data. The process occurs in the following order:

- a. Data is extracted from one or more source systems (such as Linux, UNIX, Windows, IBM System z®, or IBM System i®) by using the InfoSphere Optim Extract facility, either as a service that was created by using InfoSphere Optim Designer, which is the InfoSphere Optim Eclipse-based design-time tool, or run as a command-line script.
- b. A masking/conversion process that is known as a Convert Service, also created by using InfoSphere Optim Designer, applies the masking policies and formats the data into a CSV file. The Convert Service uses the same masking policies and windows.
- c. The InfoSphere Optim Hadoop Loader utility is then used to place the file on the Hadoop cluster. Optionally, you can then use Big SQL in IBM BigInsights and create Hive Tables from the flat files so that we can query them. Later in this chapter, we use Big SQL to query files that are archived to BigInsights.

The Hadoop Loader utility can be called within the InfoSphere Optim Service itself, allowing a seamless workflow of data extract, mask, and load. The loader can also be run from the command line in a script and is designed to convert InfoSphere Optim Extract and Archive files to CSV format and easily load them into a Hadoop Distributed File System (HDFS). The loader is implemented through a cross-platform loader called “pr0hdfs” that is primarily written in the Java programming language and packaged as a cross-platform Java Archive (JAR) that is named `pr0hdfs.jar`. It also includes a platform-specific executable wrapper (`pr0hdfs.exe` on Windows, and `pr0hdfs` on Linux and UNIX). The executable wrapper and JAR files must be installed collocated within the same directory and be named as above to function correctly.

The Loader also requires that the Hadoop “WebHDFS” REST web service interface is enabled on IBM InfoSphere BigInsights and is deployed on the Hadoop cluster's name node port 50070 or on the data nodes port 50075.

In this section, we assume that a subset of sales and customers data is extracted from Netezza and loaded into the InfoSphere BigInsights platform. Even though the section demonstrates conversion of the InfoSphere Optim extract file (.xf) into HDFS CSV files, the same process also works for loading InfoSphere Optim Archive data in Hadoop. Creation of a Convert Service requires the source extract file that was created in step a on page 169, masking mappings or policies, and identification of the output file format, which are Hadoop Distributed File System (HDFS) and Hadoop output files in our example. This file name is used only when data is converted into a single CSV file.

Because we convert multiple files in our example, we use file macros for auto-generated file names so that InfoSphere Optim can later create individual Hadoop CSV files for every table, as shown in Figure 7-15.

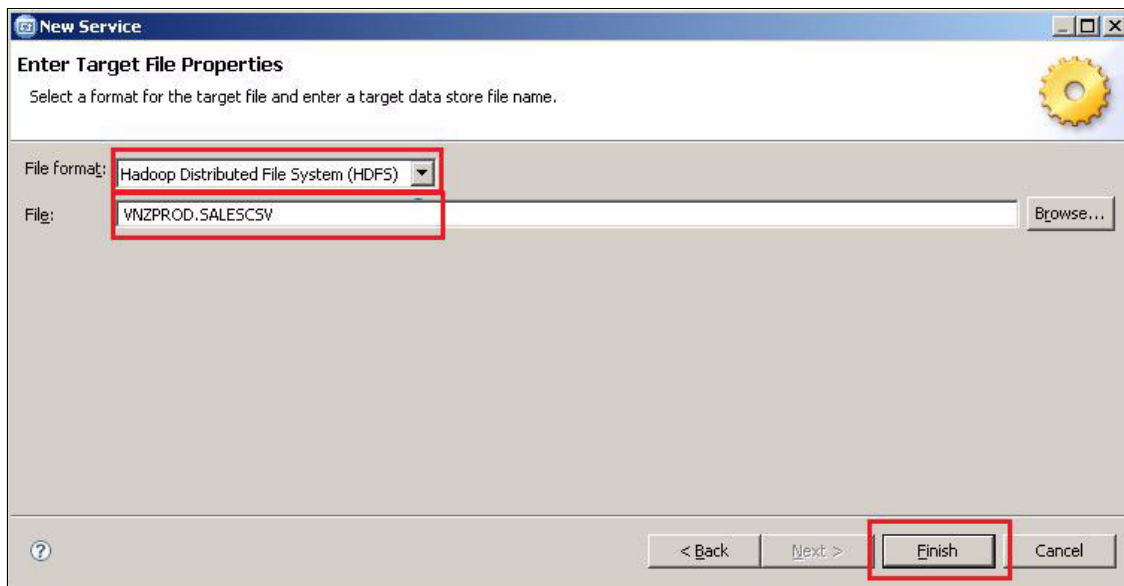


Figure 7-15 Auto-generated file names

As with all of the previous masking scenarios, the Convert Service applies the same masking policies, although in this case we use the InfoSphere Optim Designer GUI. For example, in Figure 7-16, we apply a credit card mask.

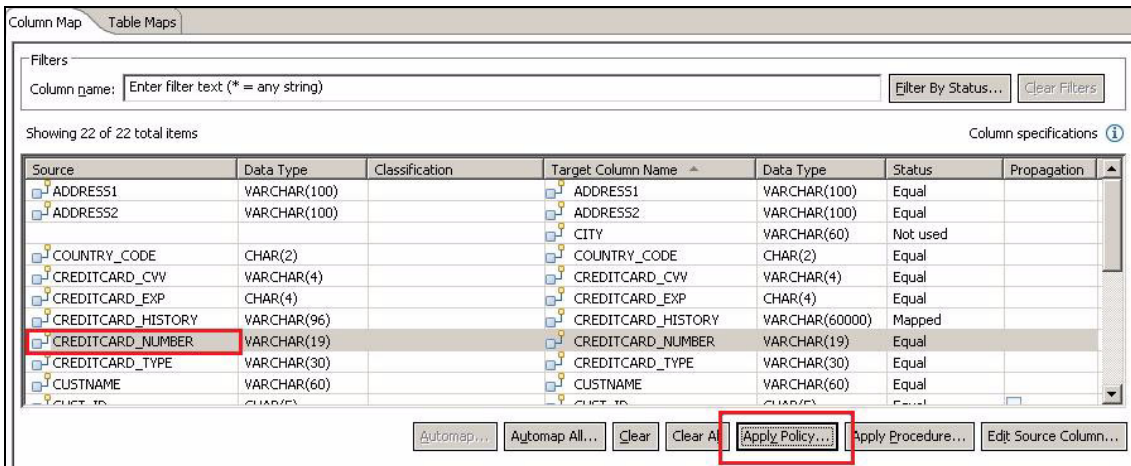


Figure 7-16 Apply credit card mask

After you define the masks to be applied, complete the service creation by adjusting various options in the Convert Service Editor. For example, to specify multiple output files, under the Target File Options tab, select the **Use file macro** check box and enter <\$objname> as the macro, as shown in Figure 7-17. InfoSphere Optim then converts data into individual files, with the object name (table) being the file name. (The default file extension is .csv.)

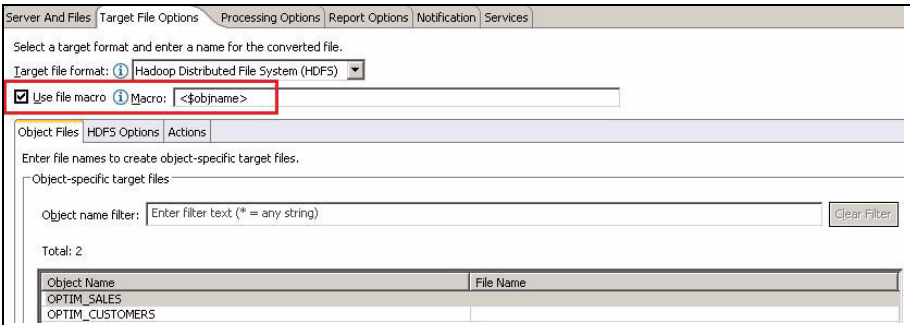


Figure 7-17 Convert Service Editor

Next, under the Target Files Options-Actions tab, call the Hadoop loader as part of the Convert process. In our example, we specify the loader executable location, `pr0hdfs`, which is in the BIN directory of the InfoSphere Optim installation. For the command-line options, we specify the following runtime parameters for the loader:

- **-u (url)**: The HTTP URL of the Hadoop cluster's name node. Notice that we are using port 50070, which is used by the WebHDFS service, which is needed for the InfoSphere Optim Hadoop Loader Utility.
- **-s (source)**: The source file to load into HDFS.
- **-d (destination)**: The destination directory in HDFS. The file name is derived from the source file name.
- **-o (overwrite)**: Determines whether the destination file may be overwritten.
- **-n (name)**: The name of the Hadoop user that is used to load in to HDFS.
- **-p (password)**: The password of the Hadoop user.

In Figure 7-18, we specify bigdata:50070 as the cluster name node and 50070 for the port. The file name is a macro that is passed in as “OPTIM\_FILE\_NAME, directory /user/biadmin, overwrite yes, Hadoop user biadmin, and password temp4now”.

**Convert Service Editor**

Use the editor to view and edit convert service properties.

Table Map: [Table Map \(Local\)](#) [Replace...](#) Server: (Local)  
Description: Source file: C:\Optim\Data\testextract.XF

Server And Files Target File Options Processing Options Report Options Notification Services

Object Files HDFS Options **Actions**

Use a utility to perform additional processing associated with a convert service action. For each action, enter a utility executable file loc  
[More...](#)

☒ Use actions

Action: [i](#) **End of Table** Optim variable delimiter: [i](#) **:**

Executable and command line options

Executable location: [i](#) C:\IBM\InfoSphere\Optim\rt\bin\pr0hdfs.exe

Command line options:

-u http://bigdata:50070 -s :OPTIM\_FILE\_NAME -d /user/biadmin -o yes -n biadmin -p temp4now

Figure 7-18 InfoSphere Optim Designer

After the service is saved and successfully run, the newly masked files are available on the Hadoop cluster.



You can also call the loader from the command line by running the utility once for each table. The password is suppressed from display, as shown in Figure 7-19.

```

C:\IBM\InfoSphere\Optim\rt\BIN>REM Run the Big Data Loader Utility to transfer t
he OPTIM_CUSTOMERS.csv file into HDFS

C:\IBM\InfoSphere\Optim\rt\BIN>pr0hdfs.exe -u http://bigdata.ibm.com:50070 -s C:
\Optim\Data\OPTIM_CUSTOMERS.csv -d /user/biadmin/mod/files/OPTIM_CUSTOMERS -o ye
s -n biadmin -p temp4now
[2013/10/26 01:26:49.156]I|: IBM Optim Hadoop loader version 9.1.0.3.
[2013/10/26 01:26:49.312]I|: WebHDFS URL is 'http://bigdata.ibm.com:50070'.
[2013/10/26 01:26:49.328]I|: Source path is 'C:\Optim\Data\OPTIM_CUSTOMERS.csv'
[2013/10/26 01:26:49.328]I|: Destination path is '/user/biadmin/mod/files/OPTIM
_CUSTOMERS'
[2013/10/26 01:26:49.328]I|: User name is 'biadmin'.
[2013/10/26 01:26:49.328]I|: Password has been specified (but not shown here).
[2013/10/26 01:26:49.328]I|: Overwrite is 'yes'.
[2013/10/26 01:26:49.359]I|: The file 'C:\Optim\Data\OPTIM_CUSTOMERS.csv' is 1,
751,676 bytes long.
[2013/10/26 01:26:49.578]I|: Redirect requested but followRedirects is disabled
[2013/10/26 01:26:49.734]I|: File 'C:\Optim\Data\OPTIM_CUSTOMERS.csv' (1,751,67
6 bytes) copied to '/user/biadmin/mod/files/OPTIM_CUSTOMERS/OPTIM_CUSTOMERS.csv'
(1,751,676 bytes) at 'http://bigdata.ibm.com:50070'.

C:\IBM\InfoSphere\Optim\rt\BIN>REM Run the Big Data Loader Utility to transfer t
he OPTIM_CUSTOMERS.csv file into HDFS

C:\IBM\InfoSphere\Optim\rt\BIN>pr0hdfs.exe -u http://bigdata.ibm.com:50070 -s C:
\Optim\Data\OPTIM_SALES.csv -d /user/biadmin/mod/files/OPTIM_SALES -o yes -n bia
dmin -p temp4now
[2013/10/26 01:26:49.906]I|: IBM Optim Hadoop loader version 9.1.0.3.
[2013/10/26 01:26:50.062]I|: WebHDFS URL is 'http://bigdata.ibm.com:50070'.
[2013/10/26 01:26:50.062]I|: Source path is 'C:\Optim\Data\OPTIM_SALES.csv'.
[2013/10/26 01:26:50.062]I|: Destination path is '/user/biadmin/mod/files/OPTIM
_SALES'
[2013/10/26 01:26:50.062]I|: User name is 'biadmin'.
[2013/10/26 01:26:50.062]I|: Password has been specified (but not shown here).
[2013/10/26 01:26:50.062]I|: Overwrite is 'yes'.
[2013/10/26 01:26:50.109]I|: The file 'C:\Optim\Data\OPTIM_SALES.csv' is 11,536
bytes long.
[2013/10/26 01:26:50.312]I|: Redirect requested but followRedirects is disabled
[2013/10/26 01:26:50.328]I|: File 'C:\Optim\Data\OPTIM_SALES.csv' (11,536 bytes
) copied to '/user/biadmin/mod/files/OPTIM_SALES/OPTIM_SALES.csv' (11,536 bytes)
at 'http://bigdata.ibm.com:50070'.

C:\IBM\InfoSphere\Optim\rt\BIN>_

```

Figure 7-19 Command line example

After the utility is run, you see that each CSV file is loaded into either the OPTIM\_CUSTOMERS or OPTIM\_SALES folder in the /user/biadmin/mod/files directory in HDFS on the BigInsights Hadoop cluster. The file length in bytes is displayed after a successful run, as is a confirmation of the file copy completion.

You can also log in to the BigInsights web browser interface to view the masked files, as shown in Figure 7-20.

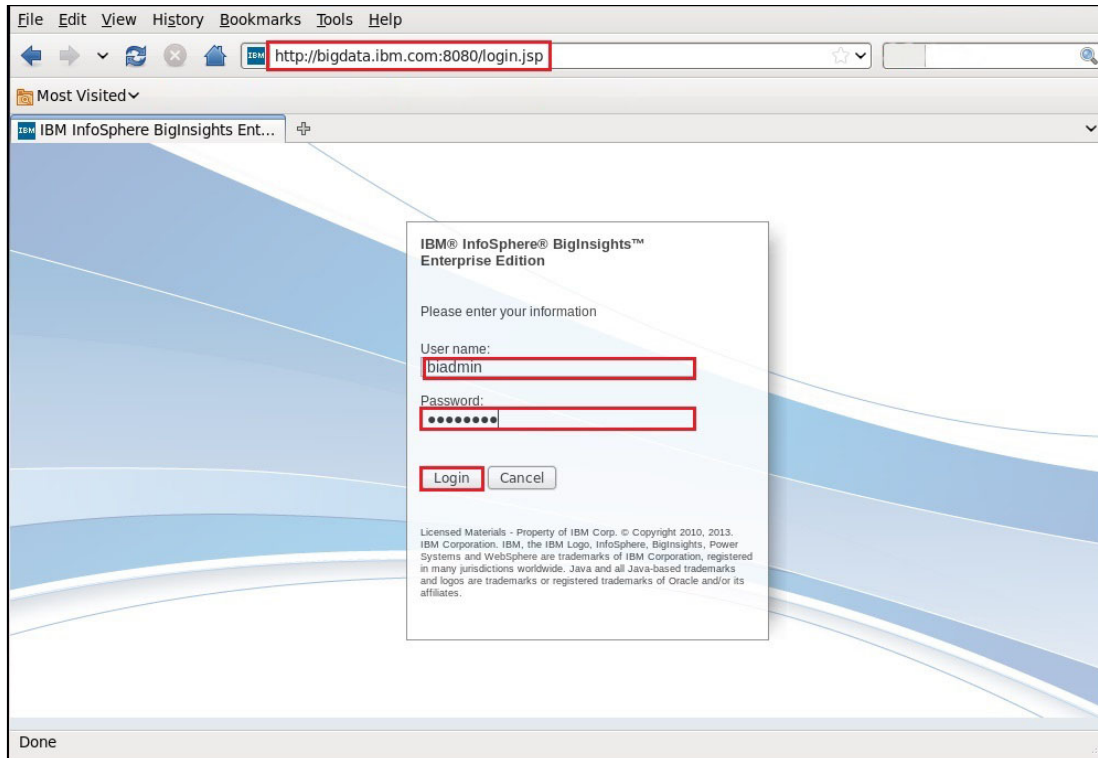


Figure 7-20 *BigInsights web browser*

To log in to the browser, complete the following steps:

- a. Log in to the browser.
- b. Click the **Files** tab under HDFS Explorer, find the masked file in the directory tree (the OPTIM\_SALES.csv file in our example, which is in the `hdfs://bigdata.ibm.com:9000/ user /biadmin / mod / files / OPTIM_SALES` directory). To view the file, highlight OPTIM\_SALES.csv and click the **Sheet** option in the portlet that contains the details about this flat file, as shown in Figure 7-21.

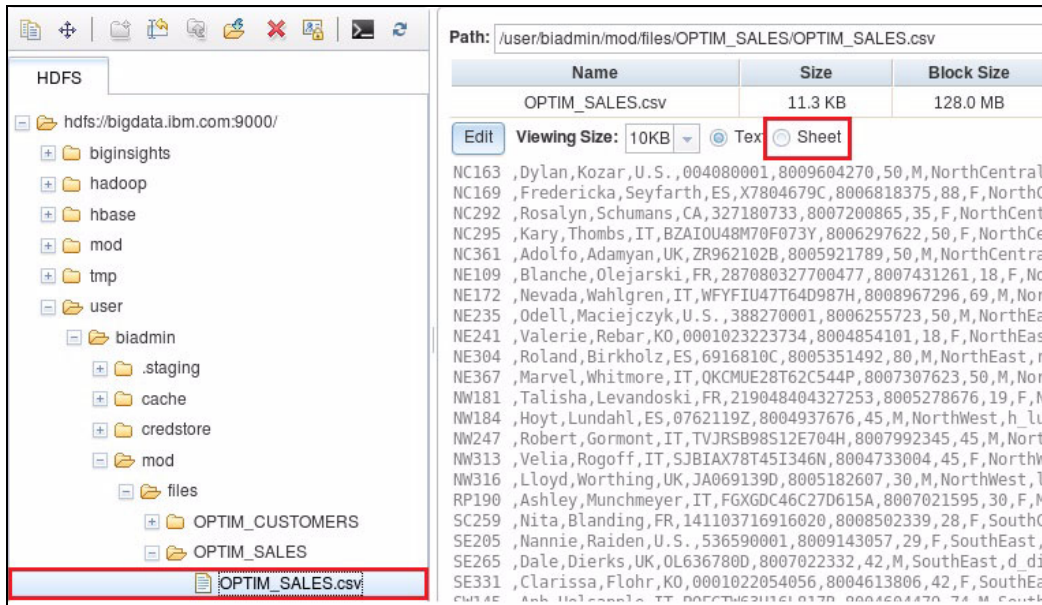


Figure 7-21 HDFS Explorer

### 3. Masking inside of a big data database by using user-defined functions (UDFs)

Data Warehouses are another important big data zone. It is not uncommon for warehouses to exceed 500 TB or more. In this high volume arena, the most efficient way to mask data is to do so within the database itself. InfoSphere Optim privacy functions can be directly called from database UDFs. With the InfoSphere Optim data privacy UDFs, you can include data masking functions in database SQL scripts or SQL statements to dynamically or statically mask data within the DBMS server framework. Sensitive information that is stored in the Data Warehouse can be masked in-place without data leaving the database. This action results in increased efficiencies while reducing masking complexity and storage requirements.

Masking through UDFs also provides dramatically faster performance and outperforms conventional ETL methods. In one lab test, we observed 1,120,000,000 elements that were masked in 60 seconds. Your results might vary. Finally, UDFs provide additional security because the function can be secured through database GRANTS.

This section demonstrates how InfoSphere Optim data masking UDFs can be tightly integrated in PureData for Analytics systems (which are by Netezza technology), which makes it simple to dynamically mask data in-place. Similar InfoSphere Optim data masking UDFs are available for other major DBMSs, including DB2, Oracle, and Teradata.

In this example, we work with the OPTIM\_CUSTOMERS table where the sensitive data is.

Masking with UDF is as simple as including the OPTIMMASK UDF function in SQL statements or scripts. Let us look at masking the PHONE\_NUMBER column in a simple **SELECT** statement:

```
VNZPROD(ADMIN)=> select CUST_ID, CUSTNAME, PHONE_NUMBER,  
OPTIMMASK(PHONE_NUMBER, 'PRO=AFFINITY, MTD=REPEATABLE, COPY=(1,4),  
FLDDEF1=(NAME=PHONE_NUMBER, DATATYPE=VARCHAR)') as MASKED_PHONE from  
OPTIM_CUSTOMERS limit 5;
```

In the **SELECT** statement, we include the CUST\_ID, CUSTNAME, the original PHONE\_NUMBER column, and also a “MASKED\_PHONE” column where we hold the values of the masked values that are generated by the OPTIMMASK privacy function.

Here is an explanation of the UDF syntax:

- PRO=AFF: Use the Affinity Masking Provider (a masking service to mask a value according to its original format).
- MTD=repeatable: Use a repeatable masking method (produces expected and consistent output values given a set of source values).
- COPY=(1,4): Indicates that the first four digits (area code) are kept in the masked data.
- NAME=PHONE\_NUMBER: Specifies the PHONE\_NUMBER column as the column to be masked.
- DT=VARCHAR: Specifies the column is in the VARCHAR data type.

A sample output is shown in Figure 7-22. In the query result, you can see that MASKED\_PHONE values are masked successfully, with the area codes unchanged.

CUST_ID	CUSTNAME	PHONE_NUMBER	MASKED_PHONE
10000	Emily Oswald	951-377-6099	951-896-3797
10001	Michael York	0390 7539464	0390 7341284
10004	Robert Olson	318-214-4533	318-920-0326
10005	Joe Jusino	619-296-8951	619-660-5962
10006	Rebecca Conner	0334 0902146	0334 6193302
(5 rows)			

Figure 7-22 Masked values query results

The following lines mask email addresses dynamically by using the following **SELECT** statement:

```
VNZPROD(ADMIN)=> select CUST_ID, CUSTNAME, EMAIL_ADDRESS,
OPTIMASK(EMAIL_ADDRESS, 'PRO=EML, FLDDEF1=(NAME=EMAIL_ADDRESS,
DATATYPE=VARCHAR)') as MASKED_EMAIL from OPTIM_CUSTOMERS limit 5;
```

In this **SELECT** statement, we include the CUST\_ID, CUSTNAME, the original EMAIL\_ADDRESS column, and a “MASKED\_EMAIL” column where we have the values of the masked values that are generated by the OPTIMASK privacy function. Here are the parameters that are used in this example:

- **PRO=EML**: Use the Affinity Masking Provider (a masking service to mask a value according to its original format).
- **NAME=EMAIL\_ADDRESS**: Specifies the EMAIL\_ADDRESS column as the column to be masked.
- **DT=VARCHAR**: Specifies the column is in VARCHAR data type.

The sample output is shown in Figure 7-23.

CUST_ID	CUSTNAME	EMAIL_ADDRESS	MASKED_EMAIL
10000	Emily Oswald	Emily.M.Oswald@mailinator.com	email1@mailinator.com
10001	Michael York	Michael.E.York@spambob.com	email2@spambob.com
10004	Robert Olson	Robert.C.Olson@spambob.com	email3@spambob.com
10005	Joe Jusino	Joe.D.Jusino@spambob.com	email4@spambob.com
10006	Rebecca Conner	Rebecca.R.Conner@spambob.com	email5@spambob.com
(5 rows)			

Figure 7-23 Sample SELECT output

In the query result, we can see that the MASKED\_EMAIL values are masked successfully. There are also different masking options to customize the format of the masked value.



#### 4. Masking data before its ingestion into the Data Warehouse

You can also use the InfoSphere Optim data masking functions by using the data transformation techniques in the “traditional” InfoSphere Optim Test Data Management Eclipse GUI workflow. In addition to masking capabilities, InfoSphere Optim also provides data subsetting functions, such as specifying filtering criteria for populating test data environments. InfoSphere Optim can maintain data in its own context and referential integrity when masking. The same comprehensive data masking capabilities are used to help companies ensure that privacy and compliance regulations are satisfied, even in non-production environments or for anonymized analytics applications.

For more information about implementing masking by using InfoSphere Optim Designer, go to the following website:

[http://pic.dhe.ibm.com/infocenter/optsol/v9r1/topic/com.ibm.nex.optimd.tdm.doc/topics/optdm-c-converting\\_data.html](http://pic.dhe.ibm.com/infocenter/optsol/v9r1/topic/com.ibm.nex.optimd.tdm.doc/topics/optdm-c-converting_data.html)

If you are licensed for IBM InfoSphere DataStage, you can also use the Masking Stage to implement your privacy functions directly from within the DataStage canvas.

#### ***Implementing data lifecycle management by archiving data***

This section describes archiving data by following the Data Minimization Principle. When the word *archiving* is used, you traditionally think of removing old data to save disk space, improve application performance, or to meet legal retention requirements. In the big data world, the first two items become much less important because storage costs are a fraction of traditional storage, more data is typically used for analysis, and MapReduce techniques provide scalability. Clearly, the legal retention requirements are still a factor. However, there is a fourth reason to archive, and that is to meet privacy requirements. Privacy requirements mean meeting the data minimization principle, which states that sensitive data must be kept only long enough to serve its purpose, after which time it must be securely destroyed. Consult with your organization to determine its privacy retention policies.

InfoSphere Optim has several capabilities for archiving data that is no longer needed. Data can also be moved from traditional transactional platforms to Hadoop to serve as a queryable archive. This capability provides a low-cost way to expand the capabilities of your analytics by providing a landing zone to capture a broader set of data from operational systems and streamline the delivery of data to reporting and analytical environments. Considering up to 80% of a Data Warehouse consists of stale and infrequently accessed data, moving this information to a lower-cost platform provides a useful method to augment your Data Warehouse environment and extend the system capacity. To meet privacy requirements, the data could also be anonymized to remove sensitive data, yet still retain its analytical utility.

### ***Using InfoSphere Optim to create the queryable archive for Hadoop***

In this use case, we demonstrate how InfoSphere Optim can be used to offload transactional and other data from traditional platforms to Hadoop by using policy-driven archiving so that the data can be used as a queryable data store.

To create a queryable archive, complete the following steps, which are further detailed later in this section:

1. Create an InfoSphere Optim Archive file by using the InfoSphere Optim Design/Runtime tool. The archive can consist of multiple data sources. For more information about creating an InfoSphere Optim Archive, see “Implementing an InfoSphere Optim Data Growth Solution”, which is available at the following website:

<http://www.redbooks.ibm.com/abstracts/SG247936.html?Open>

In this example, we assume that you have already accomplished step 1 and created the archive.

2. Create and run a Convert Process that directs the output to the Hadoop file system by using the InfoSphere Optim Hadoop loader.

The Convert process is similar to the Convert process that is covered in “Masking (obfuscating) and anonymizing sensitive data” on page 165, except that the input file has a file association of .AF (Archive File) versus .XF (Extract File). The Archive windows also look different.

In this Convert request that is shown in Figure 7-24, you see the source file of C:\Optim\Archive\DG.BigIDemo1.AF is converted to CSV format and moved to the destination file C:\Optim\Archive\DG.BigIDemo1out.csv. Also, the file format type is HDFS, indicating that you are using the loader.

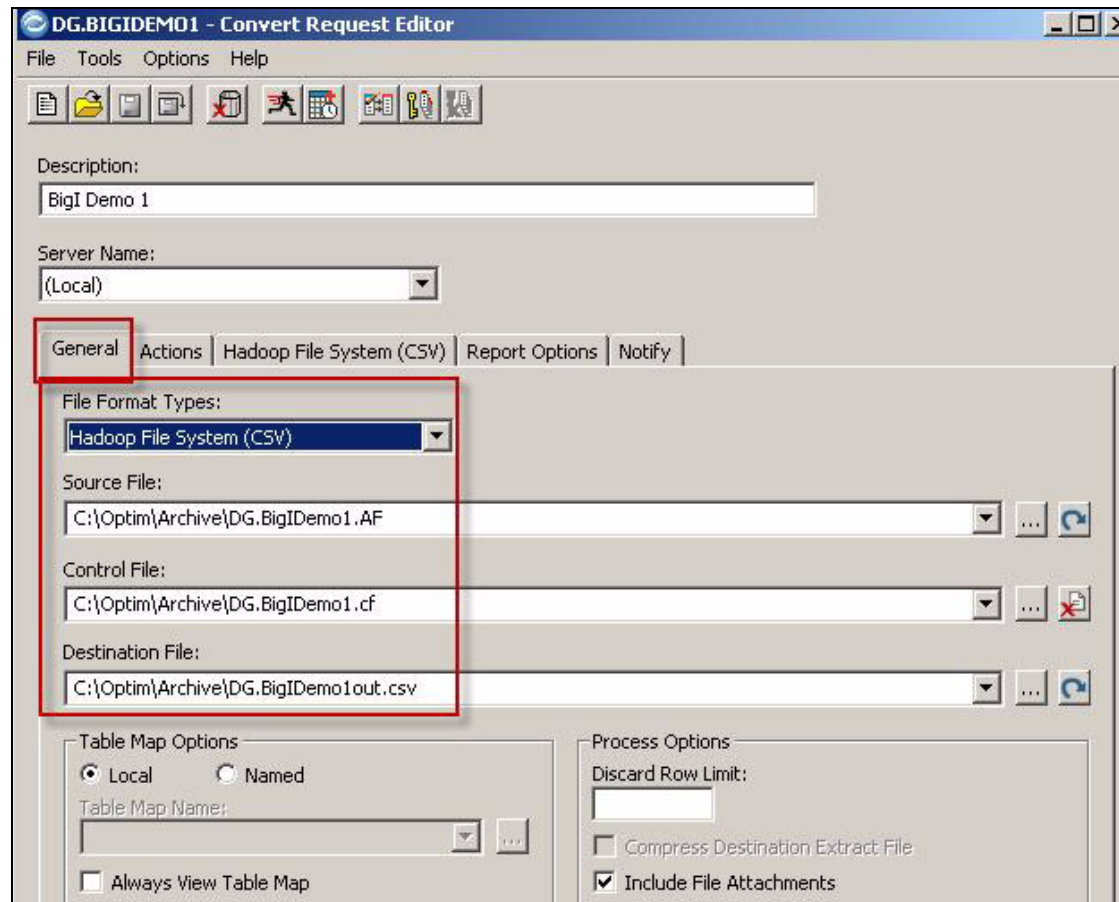


Figure 7-24 Convert request



In the Actions tab, check **Use Actions**, select the **End of Table** action, and provide the location for the InfoSphere Optim Hadoop loader, which in this example is `pr0hdfs.exe`. The command-line options are used to provide the calling parameters for the loader, such as the target cluster and port, Hadoop file and directory name, and Hadoop user connection information. You also provide a log file location, as shown in Figure 7-25. As with the Hadoop Loader masking example, you can also use the InfoSphere Optim file macro to have the utility create and process multiple files, one for each source table. The command syntax looks the following command:

```
-u http://edserver.ibm.com:50070 -s :OPTIM_FILE_NAME -d
/user/biadmin/optim/demo1/:OPTIM_OBJECT_NAME -o yes -n biadmin -p
biadmin
```

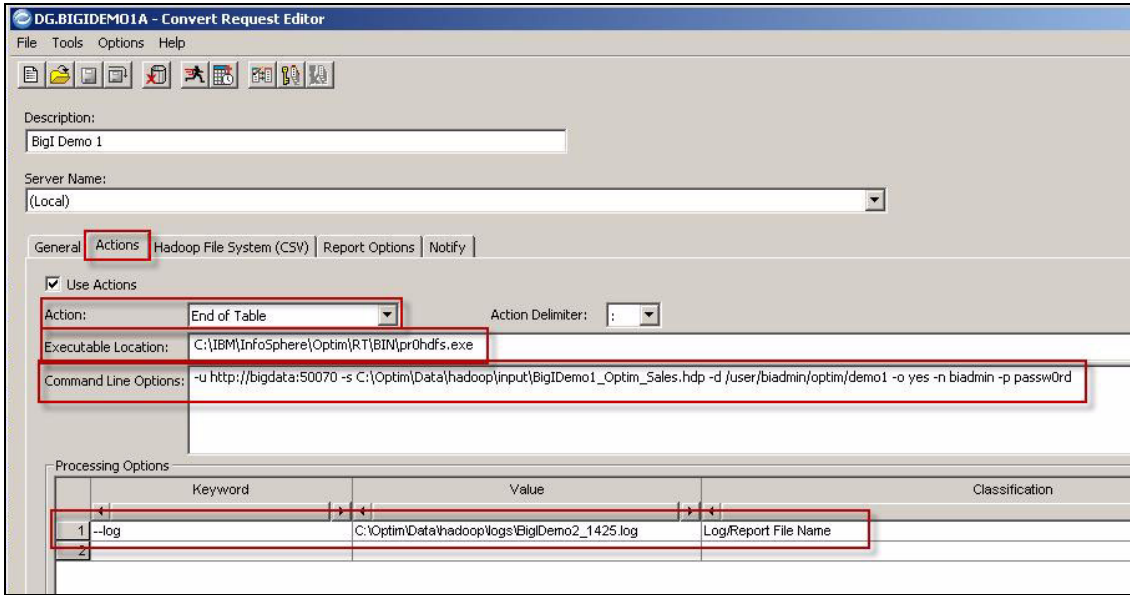


Figure 7-25 Archive options

In the Hadoop File System (CSV) tab, enter the path to the OPTIM\_SALES table that you specified as the target location, as shown in Figure 7-26.

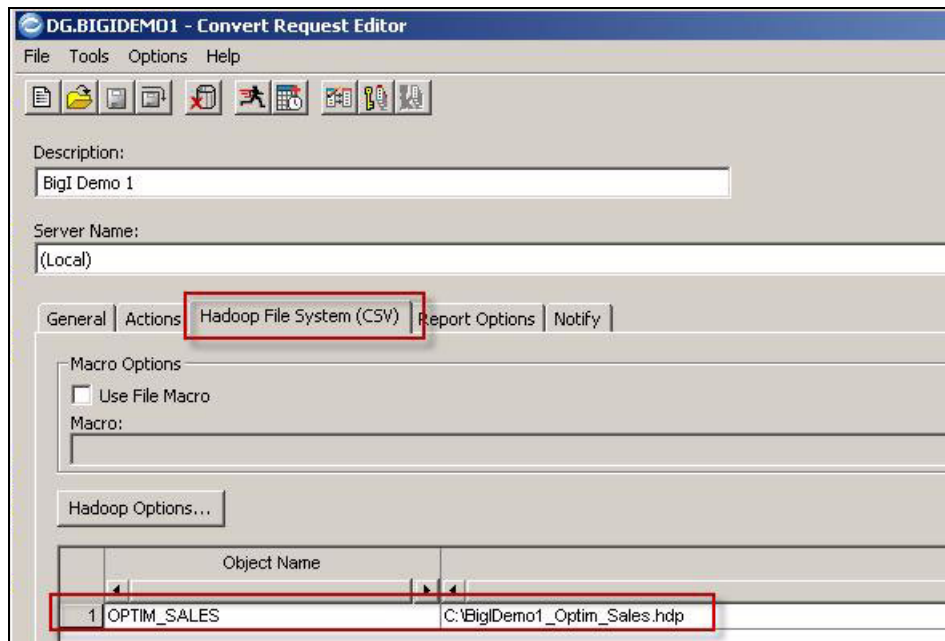


Figure 7-26 CSV tab

After the Convert request is run, the resultant file is placed in the Hadoop cluster. The Convert Process Report provides you with details about the runtime results. A sample report is shown in Example 7-1, including messages indicating that the copy was successful.

#### Example 7-1 Sample Convert Process Report

```
Request Name: DG.BIGIDEMO1A
Server Name: Local)
Source File: C:\Optim\Archive\DG.BigIDemo1.AF
Control File: C:\Optim\Archive\DG.BigIDemo1.cf
Destination File: C:\Optim\Archive\DG.BigIDemo1out.csv
Destination File Type: Hadoop Comma-Separated File
Table Map: Local
File Attachments: Processed
Client User ID: optim
Server User ID: optim
Time Started: 9/25/2013 11:06:54
Time Finished: 9/25/2013 11:06:55
Elapsed Time: 00:00:01
```

Control File: Retained  
Process Status: no errors, no warnings

Process Summary  
Tables Processed: 1  
Rows Extracted: 110  
Rows Converted: 110  
Rows with Errors: 0

Row Details:  
Extracted Converted Skipped Failed Table Name  
110 110 0 0 DB2.PRD.OPTIM\_SALES  
Hadoop File System (CSV) Settings  
File Macro: None  
Begin Label: None  
End Label: None  
Header Delimiter: None)  
LOB/XML directory location: None  
Null Column Value: None  
Field Delimiter:  
String Delimiter: None  
Eschar Char Delimiter: None  
Object Details:  
Object Name Destination File(s)  
OPTIM\_SALES C:\Optim\Data\hadoop\input\BigIDemo1\_Optim\_Sales.hdp

Action Settings  
End of Table  
Utility Name: C:\IBM\InfoSphere\Optim\RT\BIN\pr0hdfs.exe  
Utility Options: -u http://bigdata:50070 -s  
C:\Optim\Data\hadoop\input\BigIDemo1\_Optim\_Sales.hdp -d  
user/biadmin/optim/demo1 -o yes -n biadmin -p passw0rd --log  
C:\Optim\Data\hadoop\logs\BigIDemo2\_1425.log  
On Error Option: Continue Process

Actions Log Report  
End of Table  
2013/09/25 11:06:55.465][I]: IBM Optim Hadoop loader version 9.1.0.3.  
2013/09/25 11:06:55.637][I]: WebHDFS URL is 'http://bigdata:50070'.  
2013/09/25 11:06:55.637][I]: Source path is  
'C:\Optim\Data\hadoop\input\BigIDemo1\_Optim\_Sales.hdp'.  
2013/09/25 11:06:55.637][I]: Destination path is '/user/biadmin/optim/demo1'.  
2013/09/25 11:06:55.637][I]: User name is 'biadmin'.  
2013/09/25 11:06:55.637][I]: Password has been specified (but not shown here).  
2013/09/25 11:06:55.637][I]: Overwrite is 'yes'.

```
2013/09/25 11:06:55.684][I]: The file
'C:\Optim\Data\hadoop\input\BigIDemo1_Optim_Sales.hdp' is 11,536 bytes long.
2013/09/25 11:06:55.918][I]: Redirect requested but followRedirects is disabled
2013/09/25 11:06:55.934][I]: File
'C:\Optim\Data\hadoop\input\BigIDemo1_Optim_Sales.hdp' (11,536 bytes) copied to
'/user/biadmin/optim/demo1/BigIDemo1_Optim_Sales.hdp' (11,536 bytes) at
'http://bigdata:50070'.
```

---

You can also run the Hadoop loader from the command line:

- a. Open command line and go to InfoSphere Optim RT BIN directory by running the following command:

```
cd c:\IBM\Infoserver\Optim\RT\BIN
```

- b. Submit the following job from the command line:

```
pr0hdfs.exe -u http://bigdata:50070 -s
C:\Optim\Data\hadoop\input\BigIDemo1_Optim_Sales.hdp -d
/user/biadmin/optim/demo1 -o yes -n biadmin -p passwd --log
C:\Optim\Data\hadoop\logs\BigIDemo1.log
```

3. Define and configure your access method by using one of the following methods:
  - a. Use the BigInsights Console to view the files by using the built-in CSV converter to bring the data into BigSheets workbooks.
  - b. Use the Hive shell to use the InfoSphere Optim generated .xml metadata to design **HIVE** table create statements and run queries against the tables based on the InfoSphere Optim archived data content.
  - c. Use Big SQL from BigInsights to run queries and examine the output.
  - d. Use the BigSQL interface with JDBC/ODBC drivers to give tools SQL access to the files.

This step demonstrates several methods of accessing the archive file:

- You can examine the data with BigSheets. Using BigInsights Web Console, you can navigate through the file tree and find the Hadoop file. After you have the file open, click the pencil icon (circled in red) to view the data as a CSV file, as shown in Figure 7-27. Then, save the file as a BigSheets workbook for further analysis.

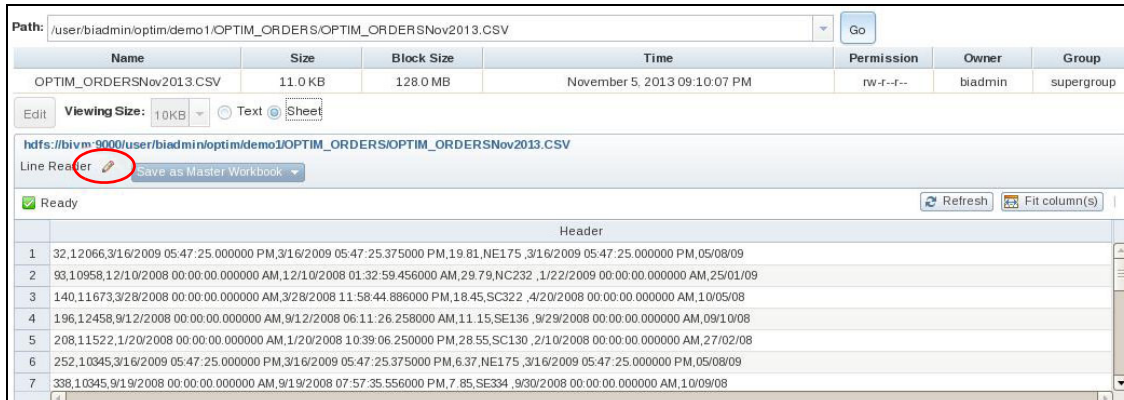
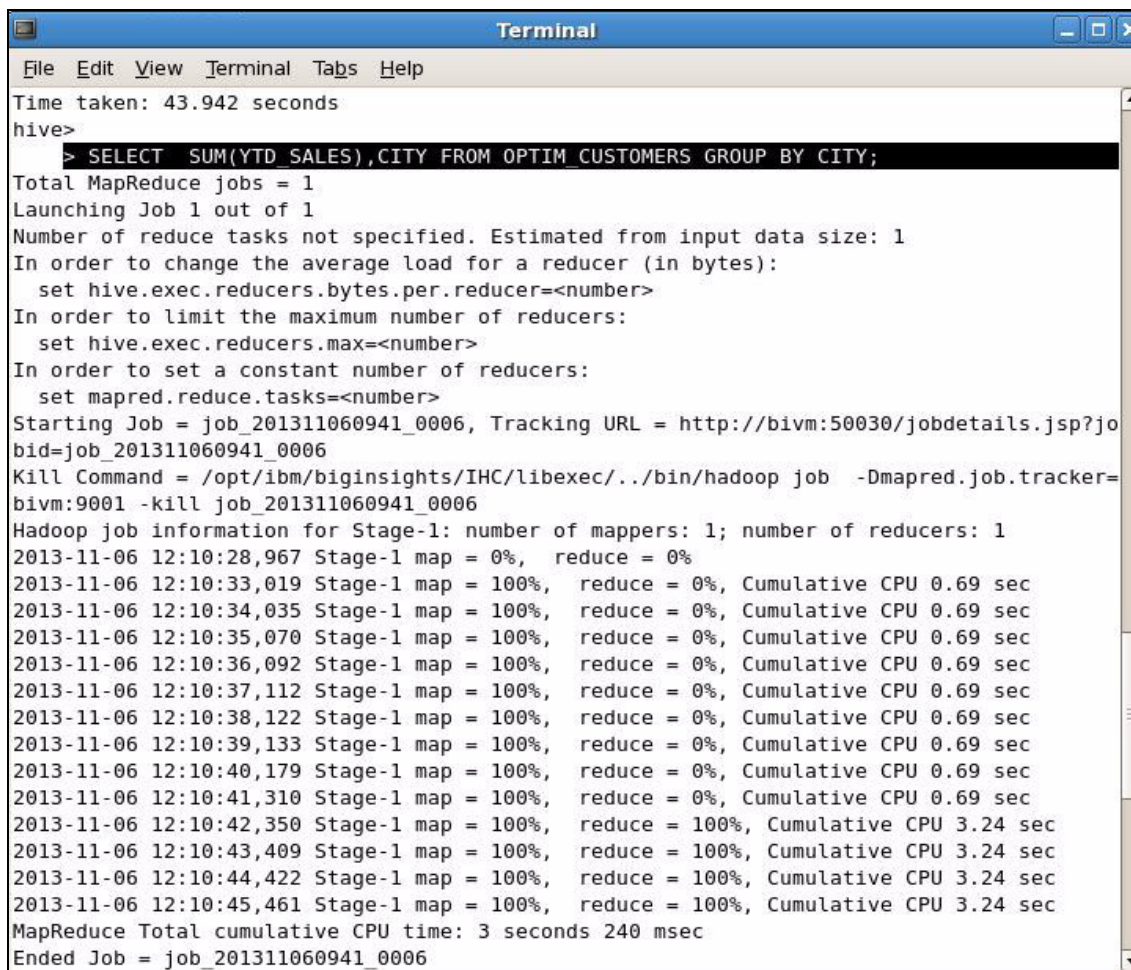


Figure 7-27 View file in BigInsights

- You can examine the data by using Hive. Hive is a widely used Data Warehouse-like infrastructure that provides data summarization and ad hoc querying in Hadoop. It provides a SQL-like interface for querying tables that are stored as flat files on HDFS, and has a metadata repository. After the files are populated to the Hadoop cluster, you need to create a Hive table to map the archives to Hive. For this step, you can use InfoSphere Optim generated .xml metadata to design your **HIVE** table create statements. With each InfoSphere Optim Archive or Extract, a special XML file is created that includes metadata that describes the contents in the archived/extracted data file. This file is used by InfoSphere Optim internal Manager and Dashboard applications, but can also be used by BigInsights / Hive.

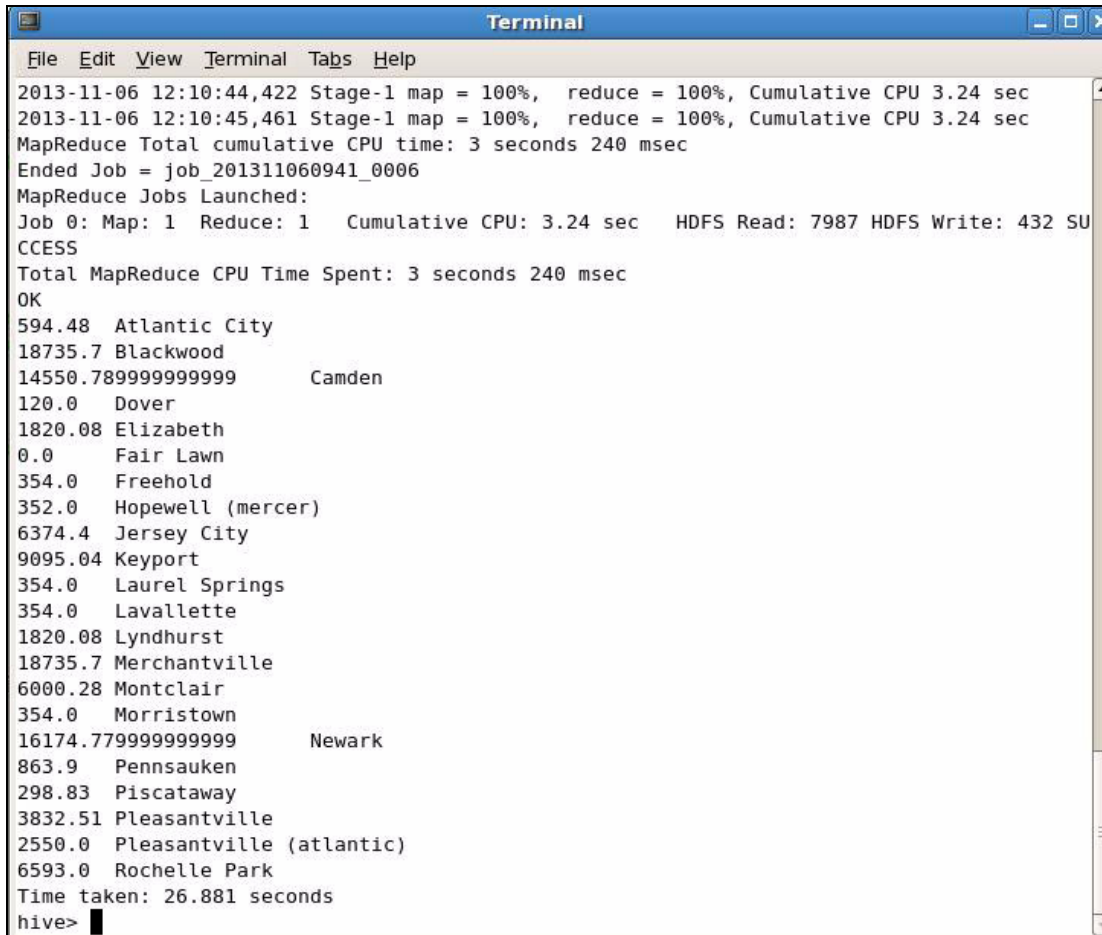
After you create your Hive table, you can use SQL statements to write queries.

Figure 7-28 and Figure 7-29 on page 189 show some sample Hive queries.

A terminal window titled "Terminal" with a menu bar (File, Edit, View, Terminal, Tabs, Help). The output shows the execution of a Hive query. It starts with "Time taken: 43.942 seconds" and "hive>". The query is "> SELECT SUM(YTD\_SALES),CITY FROM OPTIM\_CUSTOMERS GROUP BY CITY;". The output continues with "Total MapReduce jobs = 1", "Launching Job 1 out of 1", and "Number of reduce tasks not specified. Estimated from input data size: 1". It then provides instructions on how to change the average load for a reducer, limit the maximum number of reducers, or set a constant number of reducers. The job starts with "Starting Job = job\_201311060941\_0006, Tracking URL = http://bivm:50030/jobdetails.jsp?jobid=job\_201311060941\_0006". It shows the kill command and then the Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1. The output then shows a series of progress reports for Stage-1, indicating that the map task is 100% complete and the reduce task is 0% complete, with cumulative CPU time increasing from 0.69 sec to 3.24 sec. Finally, it shows "MapReduce Total cumulative CPU time: 3 seconds 240 msec" and "Ended Job = job\_201311060941\_0006".

```
Time taken: 43.942 seconds
hive>
> SELECT SUM(YTD_SALES),CITY FROM OPTIM_CUSTOMERS GROUP BY CITY;
Total MapReduce jobs = 1
Launching Job 1 out of 1
Number of reduce tasks not specified. Estimated from input data size: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapred.reduce.tasks=<number>
Starting Job = job_201311060941_0006, Tracking URL = http://bivm:50030/jobdetails.jsp?jobid=job_201311060941_0006
Kill Command = /opt/ibm/biginsights/IHC/libexec/./bin/hadoop job -Dmapred.job.tracker=bivm:9001 -kill job_201311060941_0006
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2013-11-06 12:10:28,967 Stage-1 map = 0%, reduce = 0%
2013-11-06 12:10:33,019 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 0.69 sec
2013-11-06 12:10:34,035 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 0.69 sec
2013-11-06 12:10:35,070 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 0.69 sec
2013-11-06 12:10:36,092 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 0.69 sec
2013-11-06 12:10:37,112 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 0.69 sec
2013-11-06 12:10:38,122 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 0.69 sec
2013-11-06 12:10:39,133 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 0.69 sec
2013-11-06 12:10:40,179 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 0.69 sec
2013-11-06 12:10:41,310 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 0.69 sec
2013-11-06 12:10:42,350 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 3.24 sec
2013-11-06 12:10:43,409 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 3.24 sec
2013-11-06 12:10:44,422 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 3.24 sec
2013-11-06 12:10:45,461 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 3.24 sec
MapReduce Total cumulative CPU time: 3 seconds 240 msec
Ended Job = job_201311060941_0006
```

Figure 7-28 Sample HIVE query



```
Terminal
File Edit View Terminal Tabs Help
2013-11-06 12:10:44,422 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 3.24 sec
2013-11-06 12:10:45,461 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 3.24 sec
MapReduce Total cumulative CPU time: 3 seconds 240 msec
Ended Job = job_201311060941_0006
MapReduce Jobs Launched:
Job 0: Map: 1 Reduce: 1 Cumulative CPU: 3.24 sec HDFS Read: 7987 HDFS Write: 432 SU
CESS
Total MapReduce CPU Time Spent: 3 seconds 240 msec
OK
594.48 Atlantic City
18735.7 Blackwood
14550.789999999999 Camden
120.0 Dover
1820.08 Elizabeth
0.0 Fair Lawn
354.0 Freehold
352.0 Hopewell (mercer)
6374.4 Jersey City
9095.04 Keyport
354.0 Laurel Springs
354.0 Lavallette
1820.08 Lyndhurst
18735.7 Merchantville
6000.28 Montclair
354.0 Morristown
16174.779999999999 Newark
863.9 Pennsauken
298.83 Piscataway
3832.51 Pleasantville
2550.0 Pleasantville (atlantic)
6593.0 Rochelle Park
Time taken: 26.881 seconds
hive>
```

Figure 7-29 Sample Hive query - part 2

- You can use Big SQL from BigInsights to directly run queries.

While connected to the BigInsights environment, open the BigInsights web console and navigate to the Welcome page. In the Quick Links section, click **Run Big SQL Queries**, as shown in Figure 7-30 on page 190.

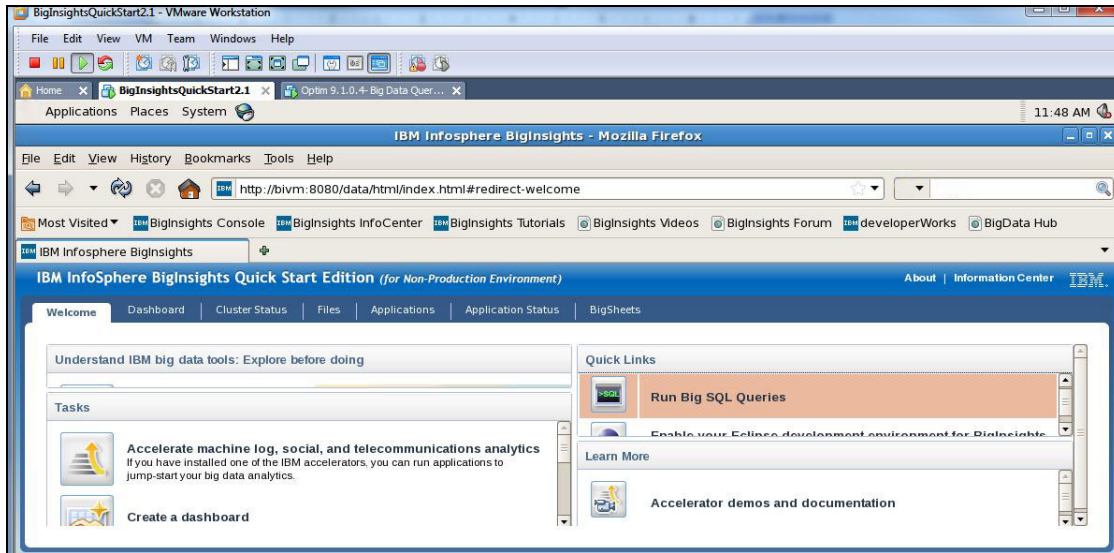


Figure 7-30 Using BigInsights to run queries

A new browser tab for Big SQL opens. Run some queries in this page by typing or pasting in SQL and clicking **RUN**, as shown in Figure 7-31 on page 191.



IBM InfoSphere BigInsights Enterprise Edition – Big SQL

```

SELECT C.CUSTNAME, O.ORDER_DATE, D.ITEM_QUANTITY, I.ITEM_DESCRIPTION, D.DETAIL_UNIT_PRICE ...
SELECT C.CUSTNAME, O.ORDER_DATE, D.ITEM_QUANTITY, I.ITEM_DESCRIPTION, D.DETAIL_UNIT_PRICE
FROM OPTIM_CUSTOMERS C, OPTIM_ORDERS O, OPTIM_DETAILS D, OPTIM_ITEMS I
WHERE C.CUST_ID = O.CUST_ID AND O.ORDER_ID = D.ORDER_ID AND D.ITEM_ID = I.ITEM_ID
ORDER BY C.CUSTNAME, O.ORDER_DATE;

```

Run

StatusResult

► Number of results returned: 200

custname	order_date	item_quantity	item_description	detail_unit_price
Amber Ingle	12/10/2008 00:00:00.000000 AM	4.0	Hog Lagoon Rally	14.0
Amber Ingle	6/26/2008 00:00:00.000000 AM	10.0	Keys to His Heart	22.0
Amber Ingle	6/26/2008 00:00:00.000000 AM	7.0	In the Depths of Midnight	19.0
Brenda Stadler	11/17/2008 00:00:00.000000 AM	20.0	Tiny Bubbles	32.0
Brenda Stadler	11/17/2008 00:00:00.000000 AM	3.0	Kansas City Killers	20.0
Brenda Stadler	5/12/2008 00:00:00.000000 AM	7.0	The Shower Sonata	22.0
Brenda Stadler	5/12/2008 00:00:00.000000 AM	3.0	Take Me For Granted	14.0

Figure 7-31 Big SQL

## 7.5.4 Consumption Phase (Sustain)

In this final phase, the big data application becomes integrated into the business with the expectation of providing business value. In this sense, you can think of this as “production mode”. All controls are in place, including monitoring to ensure security, compliance, and privacy requirements are met.

In this phase, you measure implementation progress against objectives. You want to consider which metrics are most useful for measuring success. For a security project, it might be the number of critical systems being monitored versus total systems, number of sensitive data elements, and so on.

## 7.6 Summary

This chapter has shown how the evolution of big data has resulted in changes to how you manage both security and privacy. You can expect to see further evolution as new applications and data sources arise, and the regulations to govern them. Using a framework to determine your security and privacy objectives, design appropriate countermeasures, and measure success assists you in reducing risk and exposure while gaining maximum value from your information.



## Information Quality and big data

Big data was initially described as driven by the three Vs: volume, variety, and velocity. Organizations were trying to address extreme volumes of information to incorporate a wider variety of information into their analysis, and to consume event-driven information arriving at high velocity. However, it was not long until a fourth V, veracity, was added to this list. In part, veracity represents a wanted state for all of the high-volume, high-velocity, and high-variety data. Organizations are using big data to drive analysis to achieve better and faster decision making. Therefore, they need confidence in the results of the analysis, for the results to be worthy of their trust.

There was a time when more data meant more trust. Business people wanted more information about their customers, their products, and their markets to support sound decisions. Today, though, in this new era of computing, you want more data up to a point; then, you move into this new and unsettling terrain where more data does not mean more trust. In fact, it is often the opposite. Addressing this issue of trust is the domain of Information Quality in the realm of big data.

This chapter explores key aspects of veracity in big data, specifically understanding big data content, and providing techniques to ensure and monitor that big data is fit for your purposes.

## 8.1 Information quality and information governance

Chapter 2, “Information Governance foundations for big data” on page 21 explored the Information Governance framework and its primary disciplines. Figure 2-1 on page 23, the Information Governance Maturity Model, showed that Information Quality is one of the core disciplines of the model. At the same time, the changing volume, variety, velocity, and veracity of data alter the requirements and risks for Information Quality. Data stewards, data scientists, and data analysts must be aware of these risks and considerations to ensure that information that is used in business decisions can be trusted and used with confidence.

Traditionally, Information Quality is defined as a set of identified dimensions, such as Completeness, Validity, Uniqueness (non-duplication), Timeliness, Accessibility, and Auditability (lineage).<sup>1</sup> When Information Quality is applied to big data, particularly to the largely unstructured set of social media data feeds, it is easy to dismiss Information Quality as irrelevant, that these measures do not apply, and there is no need to evaluate Information Quality. This argument assumes that these big data sources are either usable as is, present no risk, or both. Ultimately, big data is used to drive business decisions. If the information is incomplete or incorrect, it still impacts the business decision. You must step back and ask: “Given the context in which I want to use this set of data, what information about the set do I require to have trust or confidence in that data?”

This shift, or stepping back, allows you to ask relevant questions, identify potential risks, and define the relevant measures or dimensions for Information Quality for all types of data, whether within a big data or traditional data context. If you examine the big data use cases, you can identify the type of data that is used, some of the potential risks, and some of the relevant Information Quality measures for each type.

In this broader information governance context, several key roles must be considered:

- Data owners, often line-of-business executives, drive much of the usage and decision making based on the data. These individuals must communicate the broad requirements for the data, ask appropriate questions, and consider the risks with particular data in the decision-making process.

---

<sup>1</sup> *Improving Data Warehouse and Business Information Quality* by English, and *Journey to Data Quality* by Lee, et al.

- ▶ Data stewards put in place the practices and disciplines of information governance and information quality around the data sources. They evaluate which sources need regular monitoring for quality, particularly where those sources are used for ongoing business processes, and ensure that practices are in place to monitor the information quality measures.
- ▶ Data scientists and their teams focus on identifying which sources might be of use given the business requirements, which sources might provide additional insight or mitigate risk, review data content to identify deficiencies, and test out hypotheses to ensure that the data is of use and value in decision making.
- ▶ Data analysts act on behalf of the above individuals and teams to put appropriate information quality practices, rules, measures, and processes in place to ensure that the data continuously supports the needs of the organization and to review and remediate issues that are found with the data, rules, measures, or processes.

## 8.2 Exploring big data content

This section describes the activity of exploring your big data content.

### 8.2.1 Knowing your data

Since the early 2000s, the first capability (sometimes referred to as a “pillar”) for information integration has been “understanding”. Not only does this not change for big data, but in many cases you must dig deeper and cast aside typical assumptions from the world of structured data. Consider a typical traditional operational system, such as a payroll application. The data represents the salaries of your employees and what they have been paid each pay period. Your organization owns the data and controls how it is entered and stored (or your application system manual tells you those details), and you either have the metadata or can get it.

Now, consider an external source, perhaps statistics on typical employee salaries by various occupational classes over the last five years. Who created the source? What methodology did they follow in collecting the data? Were only certain occupations or certain classes of individuals included? Did the creators summarize the information? Can you identify how the information is organized and if there is any correlation at any level to information that you have? Has the information been edited or modified by anyone else? Is there any way for you to ascertain this information?

In addressing big data, an organization must address these questions, including aspects such as establishing the provenance (and possible lineage) of the data, the methods and biases that are used in data capture, and the methods (statistical or otherwise) that are used in data filtering and aggregation. Many of these aspects are assumed or treated as non-existent within traditional data sources, at least in relation to the contexts and purposes to which they are applied (such as reporting).

Although there is a broad range of possible data sources and types of data in big data, which is based on the use cases that are described in Chapter 4, “Big data use cases” on page 43, there are several common data types that can be used to illustrate the questions that are relevant to veracity. In our example, we look at Call Record Data, Sensor Data, Machine Data (for example, Logs), and Social Media Data.

## 8.2.2 Call detail records

Call detail records (CDRs) represent a large set of big data records. These records not only include data from phones and mobile phones, but also details that are related to Voice over Internet Protocol (VoIP) calls, such as conference calls and web-based seminars.

There are variations in formats depending on the provider (for example, telecommunications companies), but this data is also likely to be structured (for example, comma-separated value (CSV) or possibly hierarchical format). Information can be continuous, but most likely there are files or feeds batched from different sources at selected points in time (for example, daily). Such information can be used to integrate with other location-based data (for example, proximity of a shopper to a store), to assist with fraud or security analytics, or with solutions supporting cities or healthcare.

A call detail record can be considered a form of *transactional metadata*. It contains data fields that describe a specific instance of a telecommunication transaction, but does not include the content of that transaction. A call detail record contains at a minimum the following fields:

- ▶ The number making the call (party A)
- ▶ The number receiving the call (party B)
- ▶ The date and time when the call started
- ▶ The duration of the call
- ▶ The type of the call (such as voice and text)

In modern practice, though, there is much additional information that is provided. For example, with a mobile phone transaction, it is likely to include at least the global positioning system (GPS) of the cell tower the call went through. The content does differ based on the particular provider or service.

There are many different file formats or structures that are possible, such as comma-separated value (CSV) log files, extensible markup language (XML) files, or encrypted or encoded formats. These files, though, follow patterns that are familiar in traditional data.

## **Risks**

In most cases, call data is expected to occur daily and is used to drive business processes daily. Call data can be used in a number of use cases, possibly serving to identify demand or usage patterns, or to forecast trends. Missing data certainly impacts these situations. Incorrect data, such as code values, might skew information and subsequent decisions in unexpected and incorrect directions.

## **Relevant measures**

At a basic level, an organization must consider that when data is batched together as a file, there are quality checks for completeness of data that can occur at this level. Were all files received? Were any dates missed? Were any dates duplicated?

File type and size might provide a measure of structural integrity or consistency. It is feasible to have a broad range of file sizes, so it might be necessary to assess and understand what the typical variance is.

If the file is expected to have header or trailer records, then those records should be present, have the correct date range, and contain counts that match the detail record totals. These are measures of completeness and structural integrity of the batch file.

At the detail record level, typical information quality measures are used:

- ▶ **Completeness:** The required fields are populated.
- ▶ **Integrity:** The format should conform to a defined format.
- ▶ **Validity:** Many fields have standard acceptable values (for example, a field that is called Locale Country should contain a two character ISO Country Code).
- ▶ **Consistency:** There should be a detailed call end date within the start and end date of the header.

Additional measures might address the following items:

- ▶ Coverage and Continuity of Call Data: Receipt of data from expected sources at expected times.
- ▶ Receipt of batches/inputs per day over time.
- ▶ Gaps might indicate issues with call data source.
- ▶ Consistency of Call Data: Comparison versus data over time (for example, length of calls, number of calls, and overlaps of call data records for same user).
- ▶ Uniqueness of Call Data: Level of uniqueness for data within the batch or across batches.

Overall, Call Record Data, although of potentially high volume, falls into traditional data patterns that are related to information quality.

### 8.2.3 Sensor data

There is nothing like an Internet of Things (basically anything that has an embedded sensor) to help drive big data, and sensors are a significant part of this data explosion. It seems that practically any mobile device can become a sensor these days, not to mention the range of radio frequency identification (RFID) tags, machine sensors for weather, water, traffic, and so on. Apple's iPhones (Version 4 and later) include distinct sensors, such as an accelerometer, a GPS, a compass, and a gyroscope. These types of sensors are driving new initiatives such as IBM Smarter Cities®.<sup>2</sup> A good example of such use is SFpark,<sup>3</sup> a program helping drivers find parking spaces in San Francisco through 8200 parking sensors.

Although there are a wide variety of possible sensors for various purposes, there are several factors to consider:

- ▶ There are varied data formats depending on the sensor type, but, in general, there is likely a consistent format for each sensor type.
- ▶ The data format is likely to be structured or hierarchical data (such as XML, CST, and Really Simple Syndication (RSS)).
- ▶ There are multiple sensors of particular types.
- ▶ Most sensors produce continuous feeds in real time, although certain sensor data may be collected/delivered at selected or different points in time.

---

<sup>2</sup> [http://www.ibm.com/smarterplanet/us/en/smarter\\_cities/infrastructure/index.html](http://www.ibm.com/smarterplanet/us/en/smarter_cities/infrastructure/index.html)

<sup>3</sup> <http://sfpark.org/how-it-works/the-sensors/>



Here are some examples of sensors:

- ▶ Mobile device recordings
- ▶ Sensors for weather, traffic, water, motion, and so on
- ▶ Sensors for RFID tags

From an information quality or governance perspective, there is a large range of possible data that is generated, but an example, such as data from the National Weather Service,<sup>4</sup> is illustrative. The weather data comes from approximately 1800 tracking stations, and this data is generated at hourly intervals daily. Although it is feasible to look at some raw text data, there are two primary forms of data available: RSS and XML (the RSS is more truncated XML). An organization can obtain individual station data or compressed files of all the data for a time period.

For example, in a three-day period (25, 27, and 28, June 2013), the National Weather Service had 4165, 4169, and 4171 files respectively by date that were available in either the format *XXXX.xml* or *XXXX.rss* (where *XXXX* is the station identifier).

From an information processing perspective for an organization processing this basic sensor data, some immediate considerations about the data can be raised:

- ▶ Was the correct file type gathered?
- ▶ Did the contents match the stated file type?
- ▶ Given the lack of date in the file name, was this new data or contents already received?

As an example, look at the XML file that is shown in Example 8-1, which is modeled on the National Weather Service readings.

*Example 8-1 National Weather Service readings*

---

```
<?xml version="1.0" encoding="IA0-8859-1"?>
<?xml-stylesheet href="latest_ob.xsl" type="text/xsl"?
<current_observation version="1.0"version="1.0"
  xmlns:xsd="http://www.w3.org/2001/XMLSchema"
  xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
  <observation>2013-07-01 10 10:10</observation>
  <location>Centertown, KS</Location>
  <station_id>XMPL</statuib_id>
  <latitude>39.80</latitude>
  <longitude>-98.50</longitude>
  <weather>Cloudy</weather>
  <temperature_string>85.0 F (29.4 C)</temperature_string>
  <temp_f>85.0</temp_f>
```

---

<sup>4</sup> [http://w1.weather.gov/xml/current\\_obs/](http://w1.weather.gov/xml/current_obs/)

```
<temp_c>29.4</temp_c>
<relative_humidity>77</relative_humidity>
<wind_dir>Southwest</wind_dir>
<wind_mph>12.7</wind_mph>
</current_observation>
```

---

The contents include the following data:

- ▶ Location
- ▶ Station ID
- ▶ Latitude
- ▶ Longitude
- ▶ Observation time
- ▶ Temperature (both Fahrenheit and Celsius)
- ▶ Wind (direction and speed in miles per hour)

All of these features are structured content, which means they can be checked for typical information quality measures, such as the following ones:

- ▶ Completeness (Does the data exist?)
- ▶ Format (Does the data conform to an expected structure?)
- ▶ Validity (Is the data in the correct value set or range?)

There is some variation in the fields that are provided, which is common with XML documents, which indicates that additional checks might be made for consistency versus the XML schema or consistency over time intervals.

Although not occurring in these examples, but feasible for sensor data, is the possibility of sensor diagnostic or error codes. For example, a temperature value of -200.0 can be an indicator that the sensor had an error condition and uses that available field to pass on the diagnostic data. Depending on whether the sensor is external or internal, this may be an item to note as Incomplete or might be an item to trigger some notification/alert process.

## Risks

Mary Shacklett commented in a blog called Big Data Analytics: “But as this new machine-driven intelligence comes online, new risks are created that challenge the traditional thinking of IT risk management plans. Three major questions emerge:

- ▶ What happens if the mechanized automation or sensors fail?
- ▶ Is there a potential for greater security threats now that mechanized field operations flow over the Internet?
- ▶ Are there new information privacy risks?”<sup>5</sup>

Focusing on the first of these questions, what happens if a sensor fails? Perhaps nothing if the sensor is a weather station and there are other nearby stations. But if the sensor is monitoring a key component in an aircraft engine, a rail line, or a nuclear power plant, missing or incorrect information might create significant risk to public safety.

## Relevant measures

It is possible that an individual sensor reading that appears complete, correctly formatted, valid, and consistent might still have quality issues. Here are some additional factors to consider:

- ▶ Are there data points for all intervals or expected intervals? This is a measure of continuity for the data and can be applied to individual sensors and groups of sensors.
- ▶ Is there consistency of data across proximate data points? For example, regarding weather readings, if St. Paul, MN and Bloomington, MN both show temperatures at 84.0 F, but Minneapolis, MN shows a temperature of 34.0 F, the latter is probably an error, as you do not expect that sharp a temperature variant at locations in close proximity.
- ▶ Is there repetition/duplication of data across multiple recording intervals? The same data can come from a given sensor over multiple time periods, but is there a point at which this data becomes suspicious and suggests an issue with the sensor?
- ▶ Is there repetition/duplication of data across multiple sensors? There might be the same temperature, humidity, and wind for St. Paul, MN and Minneapolis, MN, but do you expect the exact same measurements between two sensors hour after hour? Some marginal variation in sensor readings is expected and consistent with different recording points.

Given the volume of data points and the velocity or frequency of delivery, these measures of continuity, approximate consistency, and repetition are as important as measures of information quality as completeness or validity are, if they are critical to analytic use. All of these measures can be evaluated and monitored and tracked over time as well, giving additional insight into trends of information quality.

---

<sup>5</sup> *Redefining risk for big data*, found at:  
<http://www.techrepublic.com/blog/big-data-analytics/redefining-risk-for-big-data/>

With an understanding of the data content and potential points of information quality failure, you come back to the question: What information about the set do you require to have trust or confidence in that data? If an organization is evaluating the impact of the weather on store-based sales versus online sales, they might want to correlate the hourly weather readings of stations that are close to their stores and close to particular customers' billing addresses. Hourly gaps might impact this analysis, but they might be able to smooth over such gaps with other nearby sensor readings.

If an organization is evaluating daily sales leading up to Christmas, they might care about only the aggregate weather for the day, such as minimum and maximum temperature and total precipitation. Two or three out of 24 possible data points might be sufficient for their needs, and the impact of specific data quality issues from a given sensor drops with an increase in available data points for the period or the general area. Conversely, if they have only one sensor with sporadic data near a given store or customer, the impact of data quality issues grows significantly.

The Internet of Things, the instrumentation of many devices, is having a profound impact on the variety, volume, and velocity of incoming data to evaluate. Although you looked at just one example of the type of information that is available from sensors, familiar data, such as weather observations, shows that not only do common information quality measures remain, but there are additional measures that can be put in place for ongoing monitoring. What becomes interesting is how the aggregation of such data might shift the quality requirements and their associated impact.

## 8.2.4 Machine data

Machine data fits the basic characteristics of big data: volume, variety, and velocity. Most applications and systems produce an ongoing series of data outputs (for example, logs) with variety or a different format for each one. The volume and variety of this data makes manual consumption difficult. Sifting and correlating data across multiple files is laborious and time consuming. The volume of this data continues to grow as more of the world is instrumented.<sup>6</sup>

System logs are one example of machine data. As an example, consider Figure 8-1 on page 203, which shows an IBM WebSphere Application Server log that is generated by an application of a fictitious company.<sup>7</sup>

---

<sup>6</sup> *IBM Accelerator for Machine Data Analytics, Part 1: Speeding up machine data analysis*, found at: <http://www.ibm.com/developerworks/data/library/techarticle/dm-1301machinedata1/>

<sup>7</sup> Ibid.

```
|-----10-----20-----30-----40-----50-----60-----70-----80-----|
[Sat July 14 03:58:12:906 PM] 00000048 InternalOracI I   DSR8205I: JDBC driver name :
Oracle JDBC driver
[Sat July 14 03:58:19:296 PM] 00000047 WSChannelFram A   CHF0019I: The Transport Channel
Service has started chain HttpsOutboundChain:service.Sample Outdoors.org
[Sat July 14 03:58:19:312 PM] 00000047 ExceptionUtil E   CNTR0020E: EJB threw an
unexpected (non-declared) exception
```

Figure 8-1 Sample WebSphere Application Server log

This log has a number of notable characteristics:

- ▶ A time stamp begins each record and has a specific format, such as Sat July 14 03:58:13 PM.
- ▶ The time stamp is preceded by a string represented by "[". A regular expression such as "((\\n)|(\\r))+\\[" can describe the repetitive occurrence of these characters at record boundaries.
- ▶ The time stamp is missing certain values, such as the year and the time zone. Such information may be available from the processing system.
- ▶ The log is missing other information, such as the name of the application or the server.

For further consumption and analysis, tools such as the IBM Accelerator for Machine Data Analytics might be needed to generate metadata and add in the relevant but missing information.

## Risks

The risks that are identified under sensor data might be applicable to machine data, such as logs, as well. At first glance, the privacy risk appears to be minimal or non-existent, but such logs often record user IDs. But, where logs record potential issues with operational systems, missing or incomplete data might mean that critical operational responses do not occur, resulting in network or system failure.

## Relevant measures

Generally, logs and similar machine data are expected to be produced at regular intervals. Completeness measures should address or look for the expected frequency of delivery and data frequencies. For example, a given log might on average produce 1000 lines of output per day. Variations above or below particular levels (such as beyond two standard deviations of the average) might signal an unusual error state that is generating many warning or error messages (when higher than expected) or the failure to record required data, such as when the log space is full (when there is a lower volume than expected).

Format and consistency checks can be applied to fields that are extracted from logs to ensure that the same type of information is produced. Extracted fields may contain defined values or codes that can be tested as valid. It is also possible to look at the raw log data for specific values or data patterns and compare them to the extracted data. Such an approach might help identify emerging situations or changes to the data that is produced by a specific application in its logs.

Evaluation of machine data, such as logs, is likely to focus on these areas of completeness and consistency to ensure that the correct information is extracted and provided to consuming applications, such as monitoring systems.

## 8.2.5 Social media data

Social media, such as tweets, blogs, and board messages, represent one of the rapidly expanding segments of big data, and one that organizations across many industries are taking significant advantage of, particularly for the 360° View of the Customer use case that is described in 4.3, “Enhanced 360° view of the customer” on page 49.

Many social media sites offer APIs that programmers can use to obtain public data, and there are software applications that can import various public data feeds. Often, these data feeds are returned or received in JavaScript Object Notation (JSON) format, an open standard that structures objects in a manner similar to XML. An example of data from a Twitter-based search on the topic “IBM Watson” is shown in Example 8-2. It shows the sample JSON record structure from a Twitter-based search.<sup>8</sup>

*Example 8-2 Sample JSON record for a Twitter-based search*

---

```
{
  "completed_in": 0.021,
  "max_id": 99999999111111,
  "max_id_str": "99999999111111",
  "next_page": "?page=2&max_id=99999999111111&q=IBM%20Watson",
  "page": 1,
  "query": "IBM+Watson",
  "refresh_url": "?since_id=99999999111111&q=IBM%20Watson",
  "results": [
    {
      "created_at": "Mon, 30 Apr 2012 18:42:37 +0000",
      "from_user": "SomeSampleUser",
      "from_user_id": 444455555,
```

---

<sup>8</sup> Query social media and structured data with InfoSphere BigInsights, found at:  
<http://www.ibm.com/developerworks/data/library/techarticle/dm-1207querysocialmedia/>

```

    "from_user_id_str": "444455555",
    "from_user_name": "Some Sample User",
    "geo": null,
    "id": 000000000000000001,
    "id_str": "000000000000000001",
    "iso_language_code": "en",
    "metadata": {
      "result_type": "recent"
    },
    "profile_image_url":

"http://a0.twimg.com/profile_images/222222/TwitterPic2_normal.jpg",
    "profile_image_url_https":

"https://si0.twimg.com/profile_images/222222/TwitterPic2_normal.jpg",
    "source": "<a href='http://news.myUniv.edu/'
rel='nofollow'>MyUnivNewsApp</a>",
    "text": "RT @MyUnivNews: IBM's Watson Inventor will present at
a conference April 12 http://confURL.co/xrr5rBeJG",
    "to_user": null,
    "to_user_id": null,
    "to_user_id_str": null,
    "to_user_name": null
  },
  {
    "created_at": "Mon, 30 Apr 2012 17:31:13 +0000",
    "from_user": "anotheruser",
    "from_user_id": 76666993,
    "from_user_id_str": "76666993",
    "from_user_name": "Chris",
    "geo": null,
    "id": 66666536505281,
    "id_str": "66666536505281",
    "iso_language_code": "en",
    "metadata": {
      "result_type": "recent"
    },
    "profile_image_url":

"http://a0.twimg.com/profile_images/3331788339/Mug_Shot.jpg",
    "profile_image_url_https":

"https://si0.twimg.com/profile_images/3331788339/Mug_Shot.jpg",

```

```

        "source": "<a href='http://www.somesuite.com'
rel='nofollow'>SomeSuite</a>",
        "text": "IBM's Watson training to help diagnose and treat
cancer
        http://someURL.co/fBJNaQE6",
        "to_user": null,
        "to_user_id": null,
        "to_user_id_str": null,
        "to_user_name": null
    },
    . . .
    "results_per_page": 15,
    "since_id": 0,
    "since_id_str": "0"
}

```

---

The JSON format includes a heading reference with the search criteria and trailing content. The core of the format is a series of tweets, including date and time, from and to user information, and the text of the tweet.

## Risks

In the aftermath of Hurricane Sandy in 2012, Kate Crawford noted in a Harvard Business Review blog, “The greatest number of Tweets about Sandy came from Manhattan. This makes sense given the city's high level of smartphone ownership and Twitter use, but it creates the illusion that Manhattan was the hub of the disaster. Few messages originated from more severely affected locations, such as Breezy Point, Coney Island, and Rockaway. As extended power blackouts drained batteries and limited cellular access, even fewer Tweets came from the worst hit areas. In fact, there was much more going on outside the privileged, urban experience of Sandy that Twitter data failed to convey, especially in aggregate. We can think of this as a signal problem: Data are assumed to accurately reflect the social world, but there are significant gaps, with little or no signal coming from particular communities.”<sup>9</sup>

She goes on to comment: “Data and data sets are not objective; they are creations of human design. We give numbers their voice, draw inferences from them, and define their meaning through our interpretations. Hidden biases in both the collection and analysis stages present considerable risks...”<sup>10</sup>

---

<sup>9</sup> *The Hidden Biases in Big Data*, found at:  
<http://blogs.hbr.org/2013/04/the-hidden-biases-in-big-data/>

<sup>10</sup> Ibid.



Consider the following questions from a risk perspective:

- ▶ Is there a bias in the collection method? People who have Twitter<sup>11</sup> accounts and like to express either where they were or an opinion about a particular subject (such as IBM Watson in the example above) make tweets. But there might be a large group of customers who do not express opinions through this channel. Increasing the number and diversity of these data sources helps to overcome bias.
- ▶ Was all relevant data collected? Suppose that you forgot to include critical hashtags? Maybe a common reference to the fictitious Sample Outdoors Company is #SampleOutdoor and a failure to include it significantly skews the results because that is the hashtag that is most commonly used by people complaining about the products. A comparison of search criteria that is used for potential or available variants might be needed to identify this gap.
- ▶ Was the geography of the social media information based on the user ID for the message, the identified location of the message, the place of business, or a reference in the body of the text? Is it possible to tell? In some cases, geographic references can be broad, such as a city, state, or country. Geocoding of these locations can end up with largely defaulted information that can skew results. A good example is a recent map of global protests over the last 40 years, which shows the center of protest activity in the United States in Kansas simply because it is the geographic center of the US.<sup>12</sup> Evaluation of skews in geographic data can be an important consideration for overall quality.

The bottom line is that biased or incomplete social media data sources can significantly impact business decisions, whether the business misses customer or population segments, misses product trends and issues, focuses attention on the wrong geography, or makes the wrong investments.

## Relevant measures

From an information quality perspective, there appears little to measure. Fields might or might not have values, there are some basic data formats, and there is little to check in terms of validity. The content that has value is the creation date, the user (assuming the information can be linked to some master data), the text of the tweet (for sentiment, as an example), and, if present, the geocode or language for the tweet. If that data is not present or is in an invalid format, it is not used.

---

<sup>11</sup> "Twitter is an online social networking and microblogging service that enables users to send and read "tweets", which are text messages limited to 140 characters. Registered users can read and post tweets, but unregistered users can only read them." <http://en.wikipedia.org/wiki/Twitter>

<sup>12</sup> *A Biased Map Of Every Global Protest In The Last 40+ Years*, found at: <http://www.fastcodesign.com/3016622/a-map-of-all-global-protests-of-the-last-40-years>

The crux of social media feeds is culling out data that you can pair with your own internal data, such as customers, products, and product sales.

Consider what you do know from this small example:

- ▶ The source
- ▶ The collection criteria
- ▶ The date and time the tweets were made
- ▶ Some identification of the user who sent the tweet
- ▶ Some content in the text that matched the collection criteria

By processing the collected file, you can also determine the following information:

- ▶ The number of tweets that are included
- ▶ The range of dates for the tweets
- ▶ The frequency of user identification with the tweets
- ▶ An analysis of the text content

This content becomes the core of its usage in data analysis. But over time, each feed can be assessed for varied measures of completeness, uniqueness, and consistency, similar to what you observed with call data and sensor data.

Measures of information quality might include or address:

- ▶ Coverage/continuity of social media data:
  - Comprehensiveness of the collection criteria. (Does the data include all relevant selections or leave out relevant content?)
  - Completeness of data gathering (for example, receipt of batches/inputs per day over time).
  - Gaps in times can indicate issues with the social media data source where data is expected to be fairly constant; otherwise, it might simply reflect movement on other, newer topics).
- ▶ Consistency or divergence of content in specific social media data feeds: Comparison versus data over time (for example, the average length of text, the number of records, and overlaps of social media records from same user)
- ▶ Uniqueness of social media data: Level of uniqueness for data within the batch or across batches (or are there many repetitions, such as re-tweets)

Overall, social media data is a case where the quality is as much about comprehensiveness as content. The content itself might or might not contain useful pieces of information, but if you do not have the content in the first place, then you might get biased or flawed analytics downstream. Beyond this point, there is a fine line between what might reflect a quality of data dimension and an analytical or business dimension.

## 8.3 Understanding big data

Chapter 4, “Big data use cases” on page 43 described big data use cases in detail. The first use case, Big Data Exploration, is focused on finding the data that you need to perform subsequent analysis. It incorporates the evaluation of all the types of data that are noted above and more to determine what data sources might be useful for inclusion in other key big data use cases.

### 8.3.1 Big Data Exploration

The questions that were raised when you looked at examples of big data are core aspects of the Big Data Exploration use case. As you and your organization look for the correct data to incorporate into your business decisions, you must consider how well you know the data. Where did it come from? What criteria were used to create it? What characteristics can you understand in the data contents?

Several tools in the IBM Big Data Platform are available to help you understand your big data.

#### IBM InfoSphere BigInsights

IBM InfoSphere BigInsights is a platform that can augment your existing analytic infrastructure, enabling you to filter high volumes of raw data and combine the results with structured data that is stored in your DBMS or warehouse. To help business analysts and non-programmers work with big data, BigInsights provides a spreadsheet-like data analysis tool. Started through a web browser, BigSheets enables business analysts to create collections of data to explore. To create a collection, an analyst specifies the wanted data sources, which might include the BigInsights distributed file system, a local file system, or the output of a web crawl. BigSheets provides built-in support for many data formats, such as JSON data, comma-separated values (CSV), tab-separated values (TSV), character-delimited data, and others.<sup>13</sup>

---

<sup>13</sup> *Understanding InfoSphere BigInsights*, found at:  
<https://www.ibm.com/developerworks/data/library/techarticle/dm-1110biginsightsintro/>

BigInsights is designed to help organizations explore a diverse range of data, including data that is loosely structured or largely unstructured. Various types of text data fall into this category. Indeed, financial documents, legal documents, marketing collateral, emails, blogs, news reports, press releases, and social media websites contain text-based data that firms might want to process and assess. To address this range of data, BigInsights includes a text processing engine and library of applications and annotators that enable developers to query and identify items of interest in documents and messages. Examples of business entities that BigInsights can extract from text-based data include persons, email addresses, street addresses, phone numbers, URLs, joint ventures, alliances, and others. Figure 8-2 highlights the BigInsights applications, such as board readers, web crawlers, and word counts, that help you explore a range of big data sources.

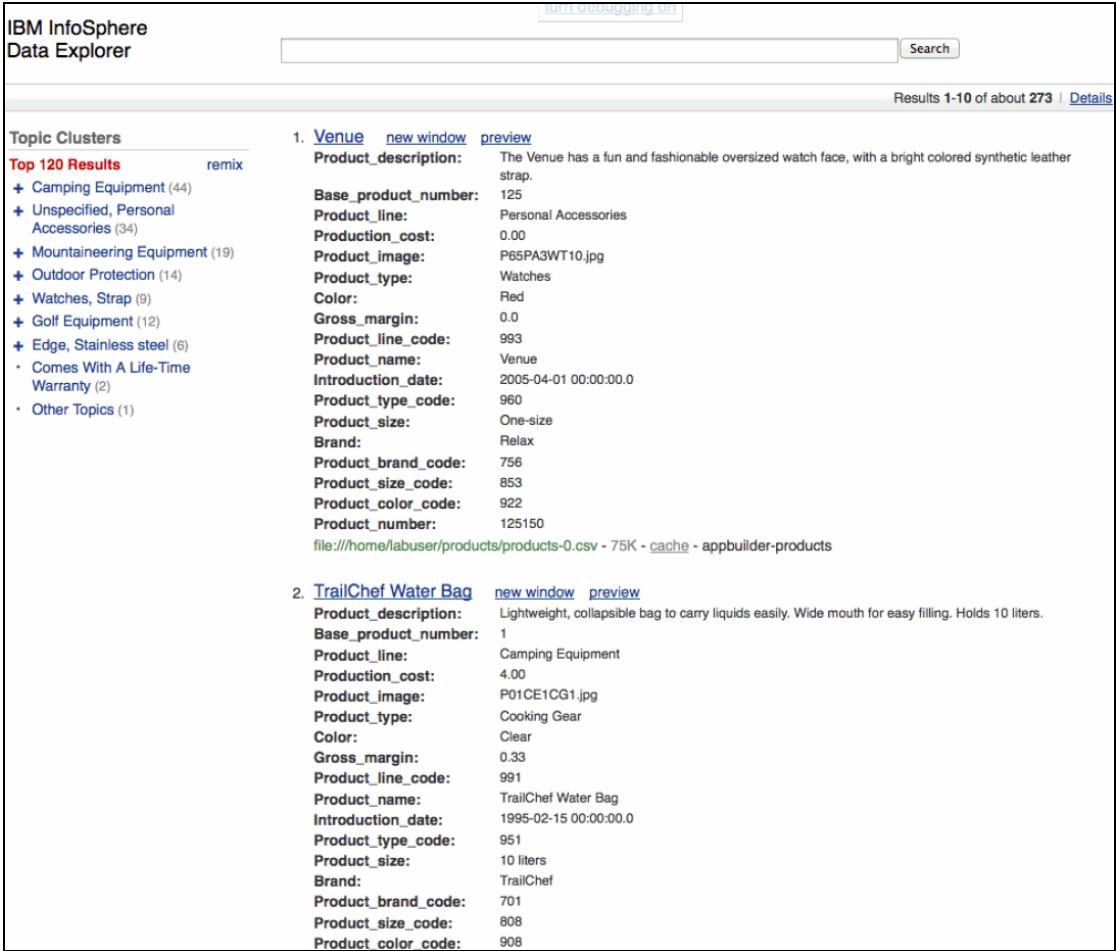


Figure 8-2 Applications for exploration in BigInsights

BigInsights helps build an environment that is suited to exploring and discovering data relationships and correlations that can lead to new insights and improved business results. Data scientists can analyze raw data from big data sources with sample data from the enterprise warehouse in a sandbox-like environment. Then, they can move any newly discovered high-value data into the enterprise Data Warehouse and combine it with other trusted data to help improve operational and strategic insights and decision making.

Capabilities such as sheets facilitate working with data in a common business paradigm. BigSheets can help business users perform the following tasks:

- ▶ Integrate large amounts of unstructured data from web-based repositories into relevant workbooks.
- ▶ Collect a wide range of unstructured data coming from user-defined seed URLs.
- ▶ Extract and enrich data by using text analytics.
- ▶ Explore and visualize data in specific, user-defined contexts.

Analysts usually want to tailor the format, content, and structure of their workbooks before investigating various aspects of the data itself. Analysts can combine data in different workbooks and generate charts and new “sheets” (workbooks) to visualize their data with a number of types of sheets available, as shown in Figure 8-3.

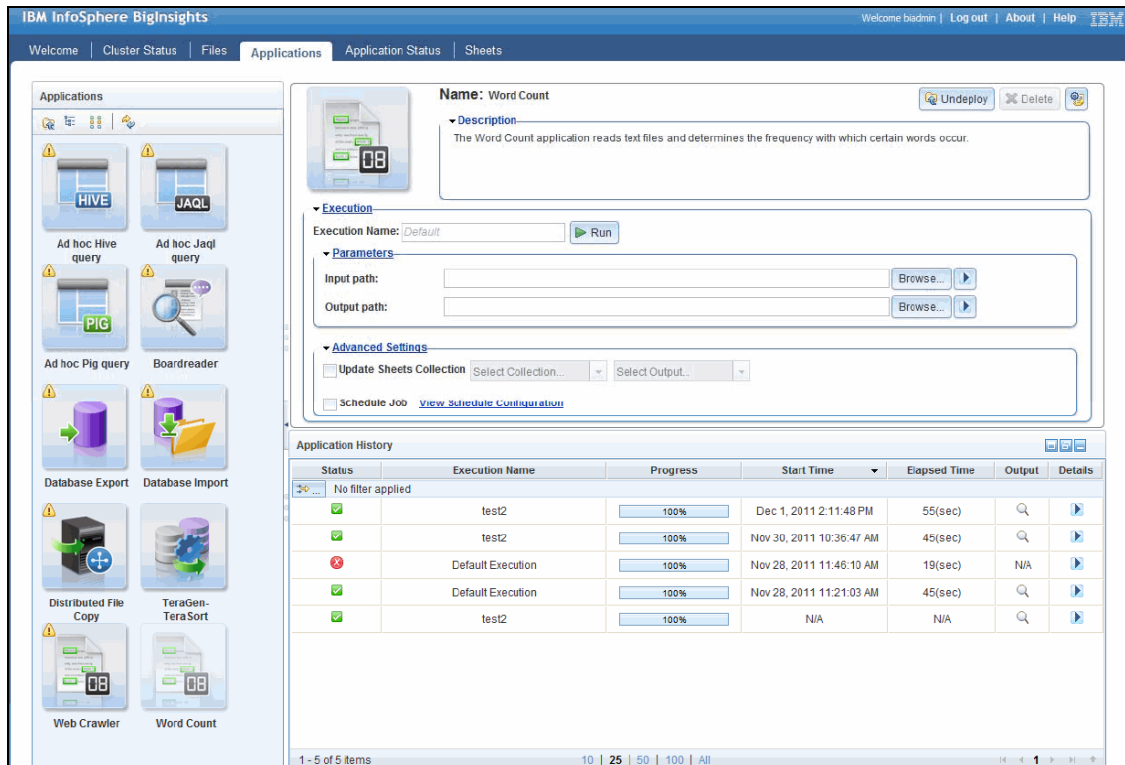


Figure 8-3 Types of exploratory spreadsheets in BigInsights

BigSheets provides a number of macros and functions to support data preparation activities, including the built-in operators to filter or pivot data, define formulas, apply macros, join or union data, and so on. BigSheets supports a basic set of chart types to help you analyze and visualize your results. This allows an iterative exploratory process as you discover new insights and decide to drill further into your data. Analysts can also export data into various common formats so that other tools and applications can work with it.

Data scientists and analysts can use the range of features in BigInsights to support Big Data Exploration, but can also address use cases such as Data Warehouse Augmentation and 360° View of the Customer.

BigInsights includes several pre-built analytic modules and prepackaged accelerators that organizations can use to understand the context of text in unstructured documents, perform sentiment analysis on social data, or derive insight out of data from a wide variety of sources.

### **IBM Accelerator for Machine Data Analytics**

IBM Accelerator for Machine Data Analytics is a set of BigInsights applications that speed the implementation of use cases for machine data. These applications use BigInsights runtime technologies to support their implementation.<sup>14</sup>

Machine data or log data typically contains a series of events or records. Some records are as small as one line, and others can span numerous lines. Typical examples of logs containing records that span multiple lines are application server logs, which tend to contain XML snippets or exception traces, database or XML logs, or logs from any application that logs messages spanning across multiple lines. Apache web access logs or syslogs are good examples of logs containing records fitting in one line.

---

<sup>14</sup> *IBM Accelerator for Machine Data Analytics, Part 1: Speeding up machine data analysis*, found at: <http://www.ibm.com/developerworks/data/library/techarticle/dm-1301machinedata1/>

Often times, machine data is configured to omit information for brevity. Information such as server name, data center names, or any other concept that is applicable to a business can be complementary when used during the analysis of the data. You can associate this information in the metadata to enrich further analysis. As shown in Figure 8-4, the log file information is extracted and standardized into a normal form with relevant metadata.

hdfs://bdvm238.svl.ibm.com:9000/demo/12\_4/output\_was/extract\_out/batch\_was.c

Comma Separated Value (CSV) Data Save as Master Workbook ▾

✓ Ready

	LogDatetimeNormalized	EventType	LogDatetime
1	Jul 14 2012 15:58:11.005 +0000'	A	Sat July 14 03:58:11:005 PM
2	Jul 14 2012 15:58:12.906 +0000'	I	Sat July 14 03:58:12:906 PM
3	Jul 14 2012 15:58:19.296 +0000'	A	Sat July 14 03:58:19:296 PM
4	Jul 14 2012 15:58:19.312 +0000'	E	Sat July 14 03:58:19:312 PM
5	Jul 14 2012 16:58:12.906 +0000'	I	Sat July 14 04:58:12:906 PM
6	Jul 14 2012 16:58:19.296 +0000'	A	Sat July 14 04:58:19:296 PM
7	Jul 14 2012 16:58:11.005 +0000'	A	Sat July 14 04:58:11:005 PM
8	Jul 14 2012 16:58:19.312 +0000'	E	Sat July 14 04:58:19:312 PM
9			

MethodIDWS	ShortnameWS	ThreadIDWS
WSVR0002I	WsServerImpl	0000000a
DSRA8205I	InternalOrad	00000048
CHFW0019I	WSChannelFram	00000047
CNTR0020E	ExceptionUtil	00000047
DSRA8200I	InternalOrad	00000048
CHFW0010I	WSChannelFram	00000047
WSVR0002I	WsServerImpl	0000000a
CNTR0020E	ExceptionUtil	00000047

Figure 8-4 Extracted log data with metadata

The Accelerator for Machine Data Analytics includes support for commonly known log data types and for other, “unknown” logs. To extract interesting fields for the known log types, several rules are provided as standard. Customization is supported to address additional fields. Whether the data is structured, semi-structured, or unstructured, if it is time series-based textual data, it can be used for analysis. Logs can be prepared as batches of similar data and then you can associate metadata by using the generic log type.



The accelerator provides an extraction application to take in previously prepared batches of logs and extract information from each of them. In addition to the metadata with each batch, the extraction application can take additional configurations that commonly apply to multiple batches. After log information is extracted, an indexing application is used to prepare data for searching. The accelerator provides additional applications for deeper analysis.

This rich set of capabilities allows for exploration and incorporation of a wide array of machine and log data, particularly for the Operations Analysis Big Data use case.

### IBM Accelerator for Social Media Data Analytics

IBM Accelerator for Social Data Analytics supports use cases for brand management and lead generation. It also offers generic applications that you can customize for your use case and industry.

At a general level, the accelerator supports the import of various social media data feeds, such as board readers and feeds from Twitter. This data can then be prepared, transformed into workbooks or BigSheets as with other data in IBM InfoSphere BigInsights, and then explored and analyzed, as shown in the core steps that are outlined in Figure 8-5.

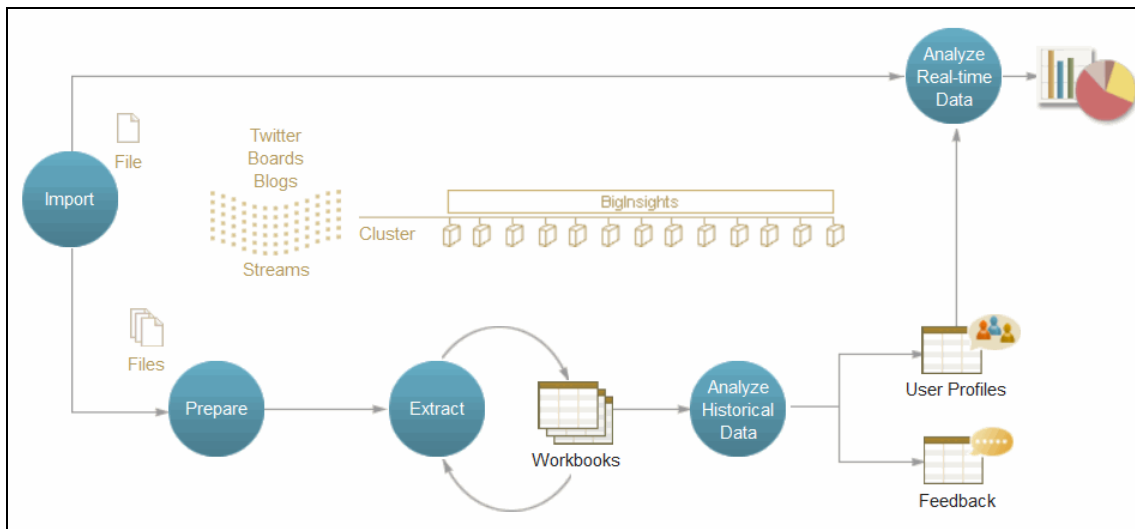


Figure 8-5 Processing with the Accelerator for Social Media Data Analytics

The range of exploratory options in BigSheets with the Accelerator for Social Media Data Analytics allows for working with the data in spreadsheets for sorting and manipulation, including the addition of functions, and the creation of chart views and visualizations of the data, as shown in Figure 8-6.

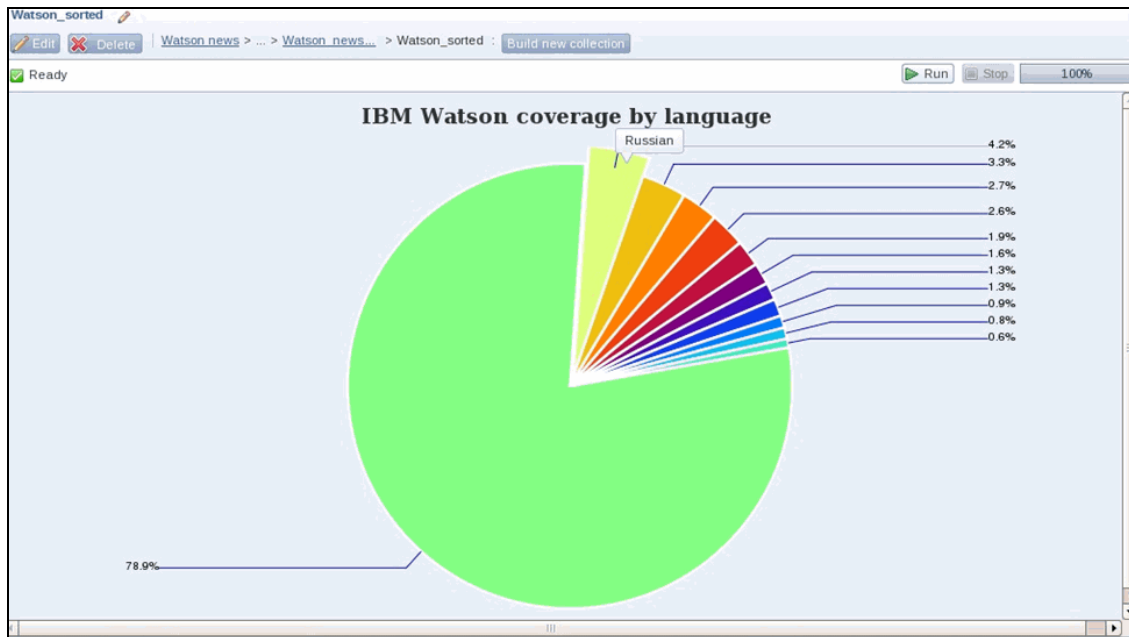


Figure 8-6 Visualization of social media data through BigSheets

Such visualization can be highly informative about specific dimensions, such as possible biases that can be exposed in the source content, and indicate where data must be supplemented by additional sources.

### IBM InfoSphere Information Analyzer

You use IBM InfoSphere Information Analyzer to understand the content, structure, and overall quality of your data at a certain point in time. Understanding the quality, content, and structure of your data is an important first step when you need to make critical business decisions. The quality of your data depends on many factors. Correct data types, consistent formatting, completeness of data, and validity are just a few of the criteria that define data of good quality.

Historically, InfoSphere Information Analyzer was used to provide understanding of traditional data sources, such as relational databases and flat files, but connectivity through Hive to big data sources allows you to bring the capabilities of data profiling and data rules to bear to further and extend your exploration of big data and begin the assessment of overall information quality. Coupled with the capabilities noted in BigInsights to extract and store data sources in HDFS, you can achieve insight into the completeness, consistency, and validity of the data sources, whether working with initial incoming data or sources that have been extracted into more usable formats.

InfoSphere Information Analyzer provides a core data profiling capability to look field-by-field at the characteristics of your data. Data profiling generates frequency distributions for each field that is analyzed. From those frequency distributions, as shown in Figure 8-7, information about the cardinality, format, and domain values is generated for additional assessment.

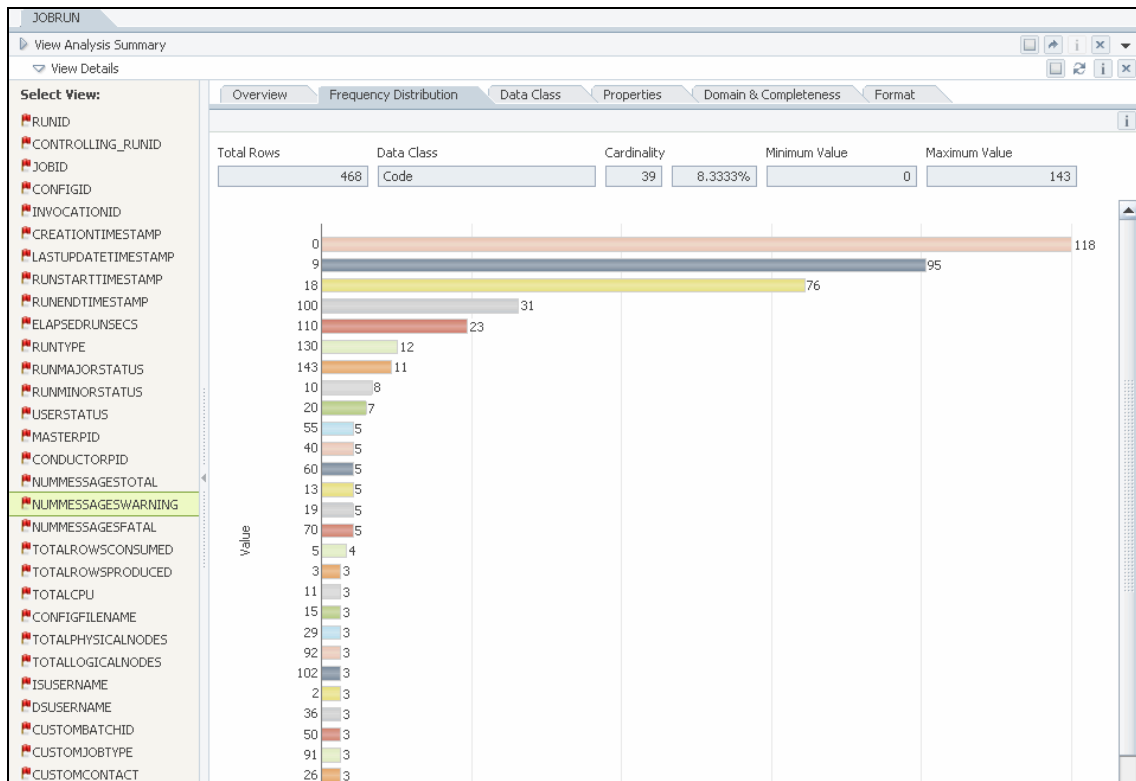


Figure 8-7 Frequency distribution from InfoSphere Information Analyzer column analysis

From this initial assessment, InfoSphere Information Analyzer supports the development of a broad range of information quality rules, including evaluation of completeness, uniqueness, and consistency of formats or across data sources, validity of domain values, and other more complex evaluations. As an example, information quality rules can test for the presence of common formats, such as a tax identification number that is embedded in text strings, such as log messages or other unstructured fields, as shown in Figure 8-8.

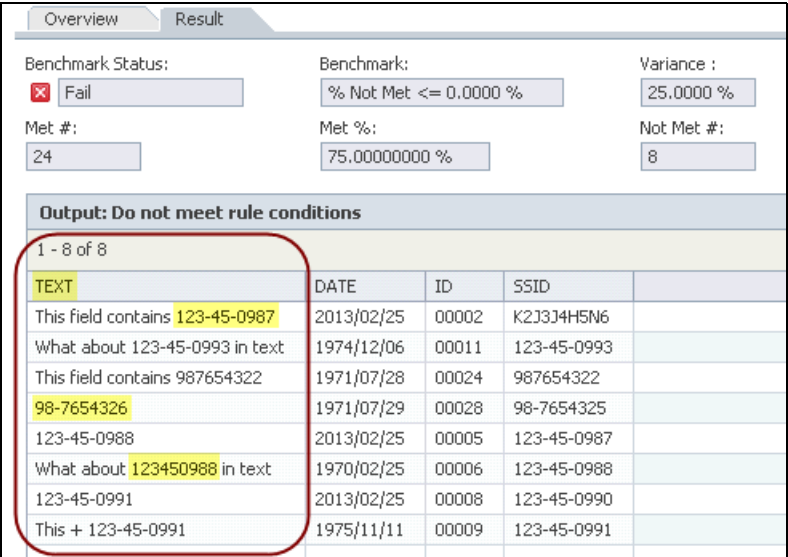


Figure 8-8 Example results from an InfoSphere Information Analyzer rule validation

InfoSphere Information Analyzer supports a disciplined approach to profiling and assessing data quality, and then identifying those conditions that must be monitored on an ongoing basis to provide confidence and trust in the data that is used for key business decision making.

## 8.4 Standardizing, measuring, and monitoring quality in big data

You have considered the aspects of exploring and understanding big data, including examples of the types of issues that might be encountered. Ultimately, this understanding drives you back to the original business requirements and goals to identify what data is relevant and what data is fit for its designated purpose.

## 8.4.1 Fit for purpose

What is the quality of a tweet or text message? Or a sensor stream? Or a log file? Or a string of bits that define an image? Does the presence or absence of specific data matter? You have considered these questions based on the type of data, some of the potential risks, and some possible ways to explore and measure the different data sources.

In the world of structured data, a payroll record is complete when the employee ID, payroll date, pay amount, the general ledger account, and certain other fields contain values. It has integrity when the values in those fields have the correct formats and correctly link to data in other tables. It has validity when the payroll date is the system date and the pay amount is in an established range. You set these rules when you established what was fit for purpose. For operational purposes, that means you can pay an individual and record the transaction in the general ledger. For financial reporting, that means you can summarize the transaction as an expense.

In the world of big data, though, with such a variety and volume of data coming in at high velocity, it is hard to ascertain what information quality means, and many of the traditional information quality measures seem to fall short. Is a message complete? Is it correctly formatted? Is it valid? In some cases, the questions appear nonsensical. So, you need to step back and ask “what is fit for your purpose?”, and that leads to another question: “What business objective am I trying to address and what value do I expect from that?” If you can answer this second question, you can start building the parameters that establish what is fit for your purpose, that is, your business requirements and your relevant measures.

In some instances, the business requirements are the same or similar to traditional information quality measures. In the Data Warehouse Modernization use case, which is described in Chapter 4, “Big data use cases” on page 43, organizations store structured data over many years into a Hadoop environment beyond the traditional database. The data is still structured, but the volume is high (for example, transactions over the last 12 years).

Similarly, with the recent development of the “Data Lake”<sup>15</sup> concept, organizations might require particular basic checks on incoming data before a file is allowed in (for example, to validate that key fields are present and, if available, to verify that the checksums match expectations). The most successful early adopters of these big data use cases have made information quality pre-eminent in their solution to ensure that the data lake does not become a “data swamp”.

---

<sup>15</sup> *Big Data Requires a Big, New Architecture*, found at:  
<http://www.forbes.com/sites/ciocentral/2011/07/21/big-data-requires-a-big-new-architecture/>

The intersection of understanding of the data with your business requirements brings you back to the point where you can establish the Information Quality, that is, the veracity that is needed for your big data initiative. These measurements might not be the traditional structured data measurements. Completeness might indicate that a message or tweet contains one or more hashtags that you care about; other tweets should be filtered out. You might need to look at continuity as a dimension with sensor readings; did you receive a continuous stream of information, and if not, is there a tolerable gap for the data? In other cases, the measurements may be the same as or close to traditional data measures, but scaled and calculated across a much higher volume.

## **8.4.2 Techniques for Information Quality Management**

As you evaluate these distinct types of data, it is important to keep in mind the techniques that are available to not only assess and measure Information Quality, but also to modify the Information Quality. You can consider this to be the domain of Information Quality Management.

### **Information Quality validation**

From a general perspective, you can assess Information Quality at four levels:

1. The field: A discrete piece of data
2. The record: A collection of related fields, usually of some specific type of data
3. The set: A collection of related records, stored together in a database table, a file, or just a sequence of messages with some designation at the start and end of the set
4. The group of sets: A collection of related sets, stored together in a database schema or a file directory and having some cross-relationship

Data at any of these levels can be assessed and measured at the point of input to a process, within a process, at the point of exit from a process, or within its storage location. Assessing and measuring within a process offers the opportunity to standardize, correct, or consolidate data. Assessing and measuring within a storage location can identify existing patterns of issues for reporting and monitoring, but standardizing or correcting, if needed, must be performed by some additional process.

Consider two types of data from a traditional perspective: master data and transactional data.

Master data includes records about domains, such as Customer, Patient, Product, or Location. Each record for one of these entities has some set of fields, such as a key or identifier, a description, perhaps some alternative references (for example, Name), and likely some set of codes that provide more description or some dates indicating when the record was created or modified.

- ▶ Requirements at a field level might indicate whether the field is required or optional, whether it must be complete, whether it must have a specific format, and what, if any, are the rules for the field.
- ▶ At a record level, further requirements might indicate whether the record has a valid key, whether the record is complete to a minimum extent, whether the values or relationships between the fields are consistent (for example, an active product must have a unit of measure), and whether there are any other conditions that must be met.
- ▶ A set of master data might also have requirements, such as whether each record is unique, whether the data values across records are consistent, and whether aggregated totals within the set are reasonable.
- ▶ Finally, there might be evaluations that are needed across sets containing similar master data, such as whether each set has complete and consistent record counts, whether each set's records have the same level of completeness, or whether there are issues of referential integrity across associated sets (for example, corresponding Address records for each Customer).

Transactional data includes records about domains, such as Customer Orders, Shipments, Guest Visits, Patient Stays, or Product Invoices. These records typically intersect at least two master domains, if not more, but represent specific instances often with identified quantities at particular dates and times. For example, an Order indicates the quantity of a given Product that is wanted by a Customer made on a specific date. Each record for one of these entities has some set of fields, such as a transaction identifier, keys to the associated master data, date and time of creation, and some quantity or amount of the transaction.

- ▶ Requirements at a field level might indicate whether the field is required or optional, whether it must be complete, whether it must have a specific format, and what, if any, are the rules for the field.
- ▶ At a record level, further requirements might indicate whether the record has valid keys for both the transaction and the relevant master data, whether the record is completed to a minimum extent, whether the values or relationships between the fields are consistent (for example, a positive quantity must have a positive price and amount), and whether there are any other conditions that must be met.

- ▶ A set of transactional data might also have requirements, such as whether each transactional record is unique (for example, you do not want duplicate orders), whether the data values across records are consistent, and whether aggregated totals within the set are reasonable.
- ▶ Finally, there might be evaluations that are needed across sets containing similar transactional data, such as whether each set has complete and consistent record counts (particularly to compare detail and summarized sets) or whether there are issues of referential integrity across the sets (for example, the set of Order transactions is not equal to the number of Shipped Item transactions).

Any or all of these conditions can be evaluated through specific data validation rules that are implemented through specific tools against the actual data sources.

Reviewing different types of big data reveals similar patterns of data. Both call data and sensor data exhibit patterns similar to transaction data at least to the record level.

- ▶ Requirements at a field level most likely indicate whether the field is required or expected or optional, whether it is expected to be complete, whether it is expected to have a specific format, and what, if any, are the rules for the field (for example, the range for Fahrenheit temperature readings should be roughly -60 to +130 degrees).
- ▶ At a record level, there may be no connection to any master data, but the requirements might include whether the record has an identifier (for example, RFID, other sensor tag, or call numbers), whether the record is completed to a minimum extent, whether the values or relationships between the fields are consistent (for example, call data record must have both a calling and called number and they cannot be equal), and whether there are any other conditions that must be met.

Volume and velocity impact whether any subsequent levels are used or available, although aggregated patterns might potentially substitute to evaluate for duplication and consistency.

- ▶ Call data records might be delivered as sets similar to other transactional data. There might be headers and trailers to such sets or batches indicating the set or batch number (an identifier) and the period that is included. Records can be counted to confirm that the set is complete. Records can be assessed for uniqueness or duplication. Periods can be compared between the records and the set to ensure that the correct contents are included. Such sets can also be compared to the prior set, if available, to ensure that the contents were not duplicated from one interval to the next.



- ▶ Sensor data is more likely to enter as a stream of content rather than a batch, particularly if real-time analysis of and responses to sensor data are required. However, it is feasible to collect specific information from sensor content and then generate aggregate statistics at intervals for use in Information Quality assessment.

Machine and social media data often have additional metadata added and data contents of large unstructured text that is parsed into smaller domains. This gives them the characteristic of fields within records for subsequent processing.

- ▶ Requirements at a metadata or field level most likely indicate whether the field is expected or not because these sources can be highly variable, how it may be represented as complete, whether a given field conforms to a targeted format (or contains extraneous and unexpected content), and what, if any, are the rules for the field (for example, an error code, if present, should be in a given range; a geocode, if present, should conform to a standard latitude/longitude coordinate pattern).
- ▶ At a record level, the content is likely to be fairly unique, although it might follow the pattern of the given type of source, whether it is a log or social media feed. These sources are unlikely to have an internal identifier, but the requirements might indicate that a generated identifier should be added (for example, log name and date), whether the record is complete to an extent that makes it usable (for example, tags exists in the social media text), whether the values or relationships between the fields are consistent, and whether there are any other conditions that must be met.

As with other big data, the volume and velocity likely preclude developing sets beyond initial exploratory work, although new data patterns may become important for training algorithms that use and process the data. Correlations may be applied across multiple records as through they were sets or groups of sets through more advanced analytical tools. This processing is beyond the scope of this book.

For our purposes, the field and record level content of these big data sources can be evaluated along known Information Quality dimensions, taking their requirements into context.

## **Information Quality cleansing and transformation**

Additional requirements might determine what should happen to these big data sources when the data does not satisfy the expected levels of Information Quality. A sensor reading of -200 °F is not valid or expected, but might indicate other issues with the sensor that must be handled, such as a repair or replacement of the sensor altogether. Another example is codes coming from call data sources that differ from those that are used within the organization. Beyond information validation, data might need to be variously filtered, mapped, transformed, standardized, deduplicated/linked/matched, consolidated, or aggregated, particularly to in-process data. These are all Information Quality techniques that serve to make the data usable for processes and downstream decision making. The application of these techniques can be validated as well, whether at the field, record, set, or grouped sets levels.

### ***Filtering***

With traditional data sources, all or most data is typically processed and subsequently stored. But with big data, there can be a much noise, which is extraneous information that is not required. Filtering is a basic technique to ignore or remove this extraneous content and process only that which is important. Filtering can occur on intake of data, or after particular steps are performed. Two primary filtering techniques are value-based selection and sampling.

With value-based selection, you assess a particular field or domain for specific content, such as a log date that is equal to today's date. Where the criteria are met, the data moves on to other steps. Where the criteria are not met, the data may be routed elsewhere or simply be ignored and dropped.

With sampling, you select an entire record using a sampling technique. This can be a specific interval, such as every 1000th record, or based on a randomized interval.

A third filtering technique is aggregation (see “**Aggregation**” on page 228), where detail records are summarized and only the summary records are processed.

### ***Mapping and transformation***

The most basic techniques of data cleansing are to change a format or map one value to another. A date field that is delivered as a string might need conversion to a specific date format. A code value of “A” might need to be changed to a code value of “1”. These mappings and transformations may also incorporate more complex processing logic across multiple fields. These techniques can be used to complete incomplete or missing data with default values.

A more complex transformation is a pivot of the data. In this case, data may be produced as an array of items or readings, but each item must be considered and processed individually.

Table 8-1 and Table 8-2 show some representative data before and after a horizontal pivot. Each Item is a field on a record in the first instance. The pivot converts each item, along with the other record data, into an individual and distinct record.

*Table 8-1 Data before a pivot*

OrderID	OrderDate	CustID	Name	Item1	Item2	Item3
123456	6-6-2013	ABCX	John Doe	notebook	monitor	keyboard

*Table 8-2 Data after a horizontal pivot*

PivotID	OrderID	OrderDate	CustID	Name	Item
1	123456	6-6-2013	ABCX	John Doe	notebook
2	123456	6-6-2013	ABCX	John Doe	monitor
3	123456	6-6-2013	ABCX	John Doe	keyboard

Data conditions typically requiring mapping and transformation include the following conditions:

► Lack of information standards

Identical information is entered differently across different information systems (for example, various phone, identifier, or date formats), particularly where the information source is outside the control of the organization, as with most big data. This makes the information look different and presents challenges when trying to analyze such information.

► Lack of consistent identifiers across different data

Disparate data sources often use their own proprietary identifiers. In addition, these sources may apply different data standards to their textual data fields and make it impossible to get a complete or consistent view across the data sources.

### ***Standardization***

Standardization is the process to normalize the data to defined standards. Standardization incorporates the ability to parse free-form data into single-domain data elements to create a consistent representation of the input data and to ensure that data values conform to an organizations standard representation.

The standardization process can be logically divided into a conditioning or preparation phase and then a standardization phase. Conditioning decomposes the input data to its lowest common denominators, based on specific data value occurrences. It then identifies and classifies the component data correctly in terms of its business meaning and value. Following the conditioning of the data, standardization then removes anomalies and standardizes spellings, abbreviations, punctuation, and logical structures (domains).

A traditional example is an address, such as “100 Main Street W, Suite 16C”. The variability in such data precludes easy comparison to other data or even validation of key components. A standardized version of this data can be what is shown in Table 8-3.

*Table 8-3 Data after standardization*

HouseNumber	Directional	StreetName	StreetType	UnitType	UnitNumber
100	W	Main	St	Ste	16C

Each component is parsed into specific metadata, which are unique fields that can be readily validated or used for additional processing. Further, data such as “Street” is conformed to a standard value of “St” (and “Suite” to “Ste”). Again, this facilitates validation and subsequent usage.

This approach is used to parse and work with machine data, such as log entries or social media text. Data conditions typically requiring standardization include the following ones:

► Lack of information standards.

Identical information is entered differently across different information systems (for example, various phone, identifier, or date formats), particularly where the information source is outside the control of the organization, as is the case with most big data. Where the domain is complex, standardization is a preferred and likely required technique versus mapping and transformation.

► Unexpected data in individual fields.

This situation describes a problem where data is placed in to the wrong data field or certain data fields are used for multiple purposes. For further data cleansing, the system must prepare the data to classify individual data entries into their specific data domains.

► Information is buried in free-form text fields.

Free-form text fields often carry valuable information or might be the only information source. To take advantage of such data for classification, enrichment, or analysis, it must be standardized first.

### ***Deduplication, linking, and matching***

Where mapping, transformation, and standardization focus on field level data cleansing, the processes of deduplication, linking, and matching focus on data cleansing at the record or set level. A simple example of deduplication is to identify records with the same keys or the same content across all fields. At the record level, this process can be a match to an existing data source to assess whether the record already is processed. At the set level, this process can be a comparison to all other records in the set or a match and comparison to a prior set of data.

With more free-form data, such as names and addresses, product descriptions, lengthy claim details, or call detail records with variable start and end times, such deduplication and matching becomes more complex and requires more than simple evaluation of character strings in one or more fields. Although mapping and standardization are used to reduce the complexity, aspects such as word order, spelling mistakes and variants, and overlapping date or value intervals require capabilities that use fuzzy logic and probabilistic record linkage technology.

Probabilistic record linkage is a statistical matching technique that evaluates each match field to take into account frequency distribution of the data, discriminating values, and data reliability, and to produce a score, or match weight, which precisely measures the content of the matching fields and then gauges the probability of a match or duplicate. The techniques can be used to both link data together or to exclude data from matching.

For big data, data conditions typically requiring deduplication, linkage, or matching include batched sets of data, including free-form text. Identical information might be received and processed repeatedly, particularly where the information source is outside the control of the organization. Where the information contains keys or time stamps, the comparisons might be straightforward by using deterministic matching. Where the information lacks keys, but contains multiple domains and fuzzy, variable, or unstructured content, then probabilistic matching is preferred. Social media tweets might be one example where filtering out duplicated information (for example, re-tweets) with unstructured text is wanted.

### ***Consolidation***

Consolidation (also called survivorship) creates a single representation of a record across multiple instances with the “best of breed” data. Consolidation works at the set level where more than one record is identified as similar or duplicated. The process of consolidation can be performed at:

- ▶ The field level
- ▶ The record level

- ▶ The logical domain level (that is, name, address, product, call data, and so forth)
- ▶ Any combination of these levels

For big data, data conditions typically requiring consolidation are the same as those requiring deduplication, linkage, or matching, and include batched sets of data, including free-form text. Identical information might be received and processed repeatedly, and consolidation of duplicate information is needed to filter out noise or redundancy that might skew subsequent analysis.

### ***Aggregation***

Aggregation generates summarized information of amounts/quantities, typically for a higher dimension than an individual record. Records are grouped by one or more characteristics. Aggregations may then be derived or calculated, including counts, sums, minimum/maximum/mean values, date intervals, and so on. Where aggregation is done on related, linked, or matched data, the summary might be done in tandem with a consolidation process. For example, you might group call data summaries both by day of the week and by month, and compute totals.

For big data, data conditions typically requiring aggregation include batched sets of data, whether entering as a batch or grouped through deduplication, linking, and matching. Sets of data must have aggregated summaries that are captured and stored for later use, whether for comparison, validation, or analysis.

## **IBM InfoSphere Information Server for Data Quality**

Chapter 6, “Introduction to the IBM Big Data Platform” on page 97 touched on IBM InfoSphere Information Server as a cornerstone for cleansing, standardizing, linking, consolidating, and ultimately validating data. Together with the tools that handle big data in motion and big data at rest, the InfoSphere Information Server capabilities allow you to take advantage of several core information quality capabilities.

### ***Filtering, mapping, standardization, and transformation***

InfoSphere Information Server for Data Quality provides many techniques to filter, standardize, and transform data for appropriate and subsequent use. Figure 8-9 on page 229 highlights some of the range of standardization and transformation capabilities that can be brought together in simple or complex information integration processes.

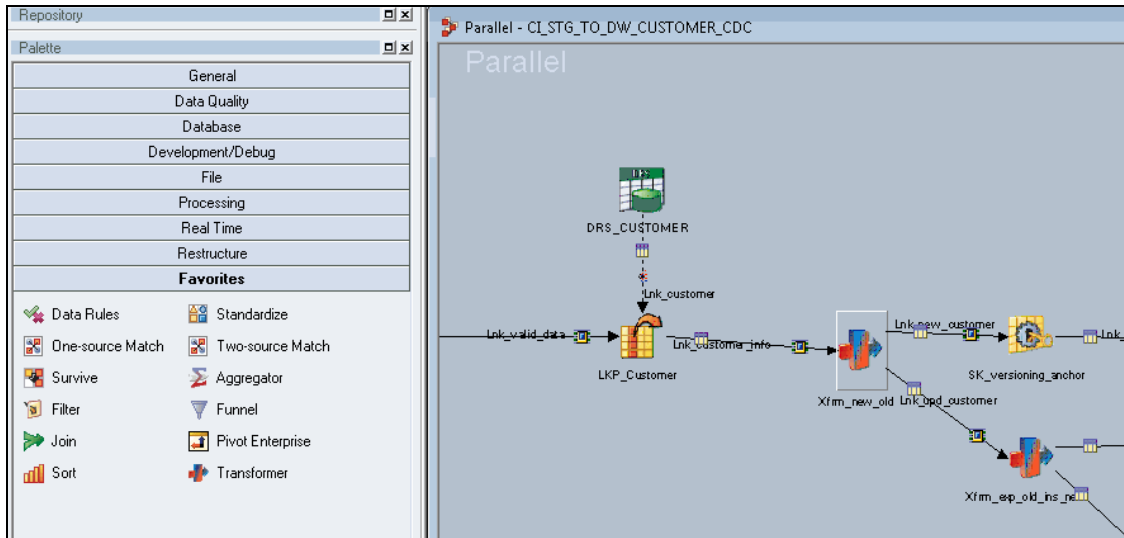


Figure 8-9 Example capabilities to standardize and transform data in InfoSphere Information Server

For example, you can apply a Filter stage to select and pass on only certain types of data. Other filtering can be applied through database and file connection stages. With big data sources, you can add an InfoSphere Streams connector to filter high-volume real-time streams before loading them into a Data Warehouse for subsequent analysis.

With a Transformation stage you can trim data by removing leading and trailing spaces, concatenate data, perform operations on dates and times, perform mathematical operations, and apply conditional logic. These capabilities may be applied field-by-field or based on more complex conditions.

In a Standardization stage, you can cleanse and normalize many types of data, including names, email, product descriptions, or addresses. For example, the addresses 100 W. Main St and 100 West Main Street both become standardized to 100 W Main St. This capability is critical to ensure that dissimilar presentations of the same information are appropriately and consistently aligned in support downstream decisions.

Through these varied capabilities, you can collect, standardize, and consolidate data from a wide array of heterogeneous data sources and data structures and bring big data and traditional sources together, or compare current files to prior files to identify differences in data volumes (or missing data segments or ranges of data).

### ***Deduplication and matching***

With the IBM QualityStage® component of InfoSphere Information Server for Data Quality, you can use one- or two-source matching capabilities to identify duplicate records for entities such as individuals, companies, suppliers, products, or events. Matching uses a probabilistic record linkage system that identifies records that are likely to represent the same entity through a broad set of matching algorithms, including fuzzy logic, strict character comparison, interval matching (such as the overlap of start and end dates for multiple records), and geocoded distances.

Before matching can take place, a data analyst configures the specific match conditions through the QualityStage Match Designer user interface.

### ***Survivorship and aggregation***

InfoSphere Information Server for Data Quality provides the Survive stage to address consolidation of fields and records that are based on specific grouping characteristics. The Aggregator stage is used to generate summarized values, such as counts and other statistics.

### ***Validation***

Through InfoSphere Information Analyzer, a component of InfoSphere Information Server for Data Quality, you can apply a broad range of data validation rules. These data quality evaluations include checks for the following items:

- ▶ Completeness (existence of data)
- ▶ Conformance (correct structure of data)
- ▶ Validity (valid values, valid ranges, valid combinations of data, and validation versus reference sources)
- ▶ Uniqueness/Occurrence (frequency of data occurrence in a set)
- ▶ Operational calculations (arithmetic and aggregated comparisons)
- ▶ Pattern or String identification (occurrence of specific data instances)

Although such rules can be run in a stand-alone mode against data sources, particularly during exploratory phases of work, they can also be embedded directly into processes to evaluate data in motion. Figure 8-10 on page 231 shows a simple example of a Data Rule stage reading a big data source, evaluating for completeness in the data, and then separating the data that met the conditions (including source identifiers and correct dates) from the data that failed.



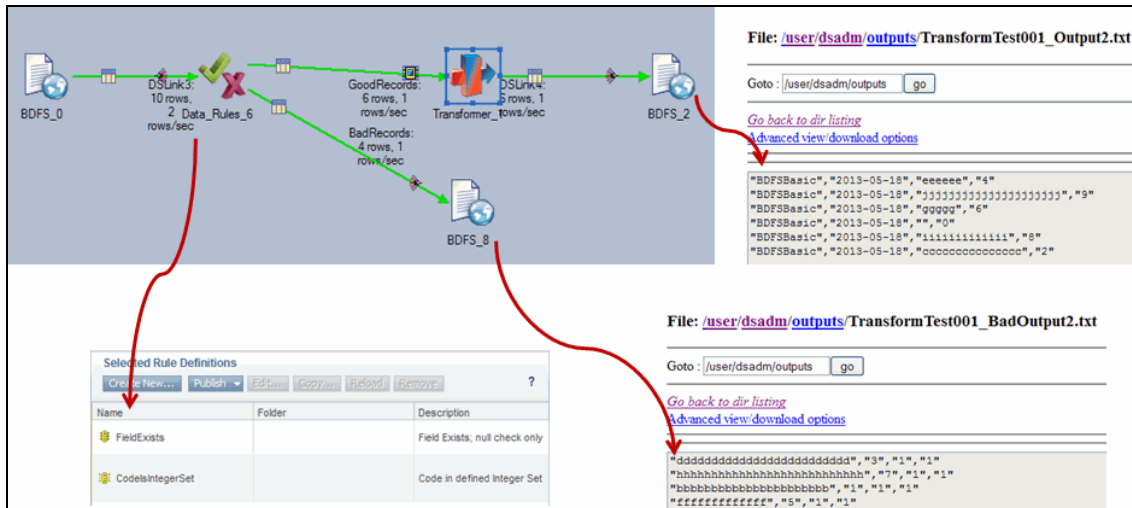


Figure 8-10 In-stream validation of data with InfoSphere Information Analyzer rules

By combining techniques for standardization and transformation, record matching, and data validation, many possible quality measures can be put together that meet the criteria for appropriately using big data, either by itself and in combination with traditional data sources, across the span of big data use cases.

### Monitoring exceptions

The IBM InfoSphere Data Quality Console component of InfoSphere Information Server for Data Quality allows data stewards to coordinate monitoring of exceptions to the standardization, transformation, matching, and validation of data. Any data that is considered to be an exception can be routed and stored for review and potential remediation.

When exceptions arrive in the console, reviewers, review managers, and business stewards can browse the exceptions and assign an owner, priority, and status, as shown in Figure 8-11. To record status details and aid collaboration, they can add notes to an exception descriptor. They can also export the exception descriptors or exceptions to view outside the console or share with others.

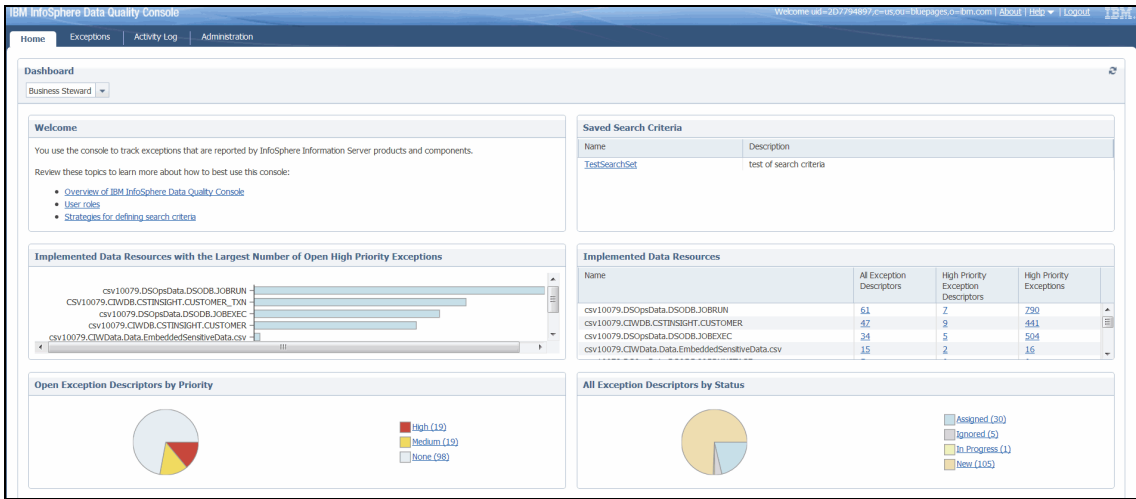


Figure 8-11 Monitoring quality measures in the InfoSphere Data Quality Console

The InfoSphere Data Quality Console provides a unified view of information quality at a detailed exception level across products and components. As you identify problems with information quality, you can collaborate with other users to resolve the problems, whether through correction of data source or data issues, modification of data validation rules, or updates to critical big data processes.

### Governing Information Quality

At the broader level, executives, line-of-business managers, and information governance officers must be aware of issues that might impact business decisions. The IBM InfoSphere Governance Dashboard provides a common browser-based view into the issues that are raised at lower levels and linked to relevant business policies and requirements. Figure 8-12 on page 233 shows one example of a selected set of information quality rules that are presented through the InfoSphere Governance Dashboard that is linked to their relevant policies.

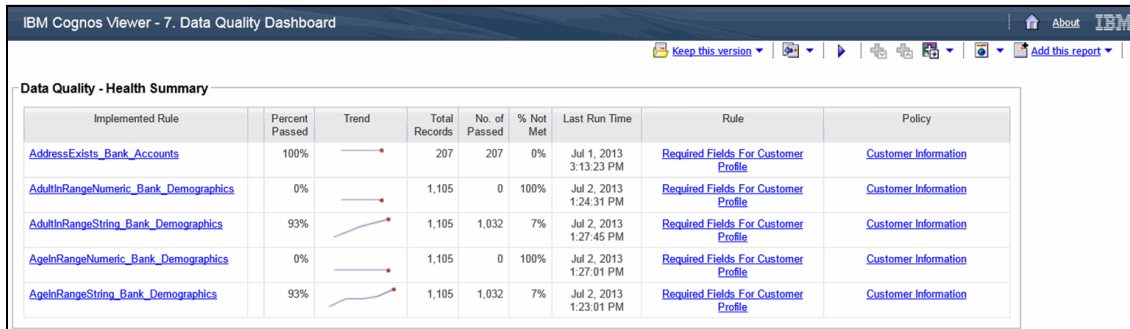


Figure 8-12 Information quality trends that are viewed through the InfoSphere Governance Dashboard

Coming back to the principles of Information Governance, this stage is an intersection of people and process with the tools used to report information. The presentation of results in an InfoSphere Governance Dashboard does not drive changes to Information Quality in your big data. It merely reflects a current state that is based on the policies and rules that you have put in place. Some of these measures can be used to track progress over time, but the people responsible for Information Governance must make decisions about what data drives business decisions, what issues and considerations for veracity must be taken into account, what data must be corrected, and what data according to what rules must be monitored and tracked.

### 8.4.3 Governance and trust in big data

Ultimately, the capabilities of standardizing, validating, measuring, and monitoring information quality can ensure trust in big data only if these capabilities are used in a governed environment. The creation of a data governance body is an essential piece to oversee, monitor, and mitigate information quality issues. Data owners are critical to establishing the correct policies and requirements that impact business decisions. Data stewards play a crucial role in ensuring that the processes that are involved incorporate the correct information quality components to meet those policies and requirements and ensure ongoing success of the big data initiatives. Data scientists are key to understanding the data, how it can be optimally used to support the business requirements, and in laying out the foundation of how the data must be standardized and validated to meet the requirements on an ongoing basis. Data analysts establish, test, and evaluate the rules and processes that are needed to establish the veracity of big data. All these individuals must work together within the context of the broader Information Governance framework to ensure that the big data that is used meets the expectations and minimizes the risks of the organization.





## Enhanced 360° view of the customer

Traditionally, enterprises rely on master data management (MDM) to create a single, accurate, and complete view across the enterprise of structured information about key business objects, such as customers, vendors, employees, and accounts. With the emergence of big data, the variety (email, chat, Twitter feeds, call center logs, and so on) and volume of information about customers must be incorporated into that 360° view of the customer.

This chapter provides an overview of MDM and the value it brings to a 360° view of structured data. Then, this chapter describes two approaches to using MDM in combination with big data technology to deliver an enhanced 360° view of the customer: using MDM and InfoSphere Data Explorer to create a more accurate view of customer information across structured and unstructured data, and using big data technology to deal with finding and merging duplicate customer information with large data sets.

## 9.1 Master data management: An overview

What is master data? In typical large organizations, information about key business objects, such as customers, vendors, products, accounts, and locations, are spread out across multiple IT systems. For example, in retail systems, information about a customer might be in a Customer Relationship Management (CRM) system, a help desk system, a billing system, and a marketing system (and likely more). Data that is listed above (customers, vendors, products, and so on) that is logically reused across multiple systems is known as *master data*.

Unfortunately, because master data is dispersed throughout the enterprise on disjointed systems, this unmanaged master data suffers from a number of data quality problems, such as the following ones:

- ▶ Data is out of date between systems (timeliness). For example, a female customer has gotten married and changed her name. This updated name change might be reflected in one system, such as Enterprise Resource Planning (ERP), but not in others, such as Legacy Billing.
- ▶ Data is inconsistent between the different systems. Values can be inconsistent between systems or data can have similar values, but a different representation or format. For example, you can have a name that represented by these elements on one system (courtesy title, first name, last name, middle name, and modifier), but only two elements on another system (first name and last name). Even if you have the same data elements, the values might be encoded or standardized differently, such as “35 W. 15th” versus “35 West 15th Street”. The data also might not be cleansed. Without consistent data rules, there might be inaccurate data in fields, such as Social Security numbers (999-99-9999 is entered by a clerk simply to complete a customer form).
- ▶ Data is incomplete across the systems. Data might be present in one application, but not in another. For example, the Customer Relationship Management (CRM) system might have information about a customer's preferred contact information, but that information might not be in the marketing system. Often, bringing together the complete set of information between the different systems is a challenge because those systems might not be easily extendable to add the full range of data values about a customer. In that case, there might not be an easy way to pull the data. In some cases, enterprises might integrate customer data through point-to-point application integration, but that process is hard to scale across the enterprise.

As a result, there is no consistent and accurate view of this master data in the enterprise. This can lead to wrong business decisions and alienation of customers. Recently, a major retailer used an incorrectly cleansed mailing list to send a solicitation to a customer with the following address:

John Doe,  
Daughter Killed in A Car Crash

The grieving customer revealed this to the press, creating a major reputation issue<sup>1</sup> for the retailer.

### 9.1.1 Getting a handle on enterprise master data

To address all these MDM problems and to deliver an accurate view of master data, enterprises must introduce an MDM solution. Enterprises employ MDM solutions as a central place to manage core business data that is used by multiple stakeholders. MDM solutions provide the tools and processes around these core data, allowing the enterprise to achieve these key operational goals:

- ▶ The core business data is consistent. For example, ensuring that each department uses the same customer name or product description.
- ▶ The data is governed so that the data is updated or changed in accordance to business policies.
- ▶ The data is widely available.

Essentially, MDM solutions exist as the authoritative source for these core data elements across different lines of business and across multiple business processes.

MDM solutions sit at the intersection of the information landscape. They consolidate core master data from a diversity of sources, and they distribute that consolidated view to a range of downstream consuming applications and systems. This central role is highlighted in the reference architecture that is described in Chapter 5, “Big data reference architecture” on page 69.

---

<sup>1</sup> <http://www.nydailynews.com/news/national/officemax-adds-daughter-killed-car-crash-letter-article-1.1583815>

To address the consolidation of core master data, the MDM solution must incorporate several core capabilities:

- ▶ **Standardization:** As with Information Quality solutions, standardization resolves many issues in linkage, integration, and consolidation of data.
- ▶ **Common rules and policies:** To keep the quality of the master data high, data rules establish the criteria that is acceptable or unacceptable to include. An example criteria is "Cannot create an account for a person under 18."
- ▶ **Linkage/matching:** At the heart of MDM solutions is the ability to connect pieces of data by using probabilistic matching techniques. This allows organizations to match both core master data and master data to related content (for example, people to organizations, people to households, and parts to products).
- ▶ **Aggregation and consolidation:** When the MDM solution is used to create a single consolidated view of the master data, this capability provides automated grouping (for example, two distinct records become one consolidated record), unlinking or ungrouping (for example, where new facts indicate that previously grouped records must be separated because they are distinct), and stewardship to address possible groups that cannot be automatically resolved.
- ▶ **Stewardship:** Besides the previously noted resolution of fuzzy possible matches, stewardship also entails resolution of other information quality issues that arise when attempting to consolidate data from diverse sources.

To support the outbound consumption of the consolidated master data, the MDM solution must also provide capabilities for other applications and systems to work with the master data, such as services and APIs. Standard service and APIs provide the ability for other consuming systems to request and retrieve the consolidated data. This might include the originating source systems that consume and resolve the consolidated or linked master data through updates, or the Data Warehouse, which can incorporate identifiers from transactions and associate them to the correct master data entity, or other analytic or reporting sources, which must work across groups of master data records.

There are also many capabilities that are needed to govern the MDM solution within the contexts that are outlined in Chapter 2, "Information Governance foundations for big data" on page 21 and Chapter 3, "Big Data Information Governance principles" on page 33, including policies for the following items:

- ▶ **Standard security and audit** (including restrictions on who can manage and merge data)
- ▶ **Information Privacy** (including protection and masking of sensitive information)



- Rules, process, and management for defining data elements, policies, security, standardization, matching/unlinking rules, and so on

With the shift towards big data, there are a broad set of data sources (particularly unstructured data, operational logs, and social media sources) that arrive directly in the mix of data, and the MDM solution itself becomes part of the central set of analytics sources. The value with this shift is two-fold:

1. New insight is possible. For example, internal operational data, such as customer interaction history or support logs, can be used to alert call-center operators to an unhappy customer or be mined for remediation actions. This data is also used in determining product placement either through customer micro-segmentations (for example, “customers like you also viewed ...”) or complex predictive models.
2. A real 360° view can be achieved. Enterprises know that their internal data presents a narrow view of the customer. Many enterprises seek to augment information about their customers by purchasing third-party information, which they can associate with individual customers. This data aids the enterprise’s segmentation, but are typically coarse (for example, average income by postal code). This data supports macro-level marketing, but do nothing for individual customer insight. Social media, and consumer willingness to create public data about themselves, changes that situation. Now, key life events (birthdays, anniversaries, and graduations), fine-grained demographics (age and location), interests, relationships (spouse and child), and even influence (followers) exist in public or opt-in sources for a growing segment of the population. This view also supports micro-segmentation, but also drives quick insight into emerging trends that organizations can tie in to or provide customized services at an individual level to enhance each possible revenue stream.

Organizations cannot realize this value without the combination of MDM solutions and big data; they must have both. Without master data, it is impossible to connect this incoming data with insight. As you have observed with big data, the volume, variety, and velocity of the arriving information, which is coupled with the need for consumers to respond immediately to new insights, requires new considerations for master data that is different from the traditional approach to MDM solutions in the enterprise.

There are three key focal areas for MDM solutions, which are shown in the following list: entity and event extraction from new data sources, particularly to support data mining and predictive models, rapid search and linkage of disparate data, particularly to support real-time interactions, and the usage of matching technologies deep in the big data architecture in response to rising data volumes. These three areas, in the context of the reference architecture, are shown in Figure 9-1.

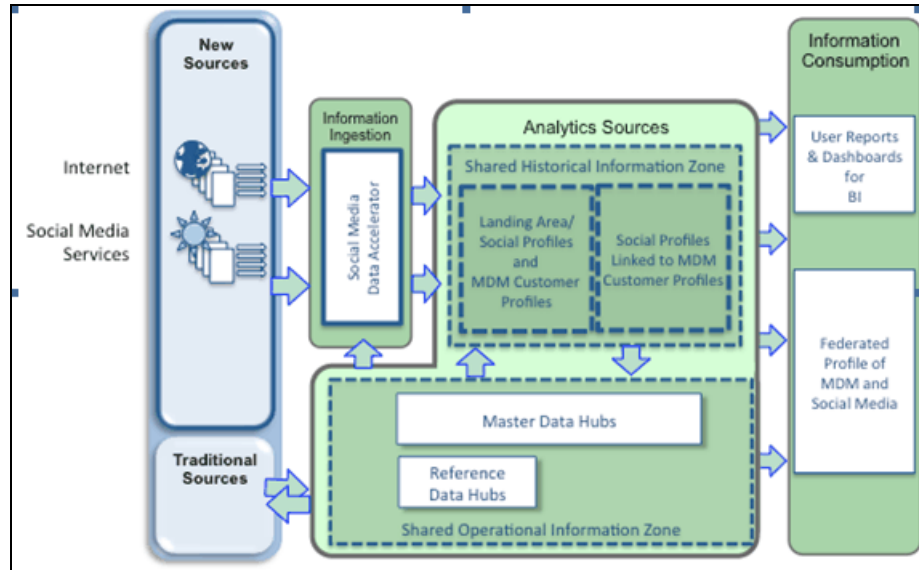


Figure 9-1 Intersection of big data with MDM solutions

1. Entity and event extraction<sup>2</sup>
  - a. There are many sources of new data, particularly unstructured data sources such as blogs, posts, and tweets.
    - i. Relationships (“Joe works for IBM, Jane is Joe’s sister;”)
    - ii. Life Events (“Joe was born on Feb 24, 1991”, “Joe just graduated from the University of Texas”)
    - iii. Real-Time Events (“Hey, I’m down near the music festival - is there a vegan restaurant nearby?”)
  - b. This data is high volume and only a certain percentage of the data is of use; the rest constitutes noise and extraneous content.

<sup>2</sup> For more information about social media data extraction, see Chapter 8, “Information Quality and big data” on page 193.

- c. Useful data about entities (for example, master data domains) and events must be extracted, structured, and then linked through standard MDM capabilities.
  - d. This structuring of data requires the application of text analytics tools with rule sets that are designed for social media data. As shown in Figure 9-1 on page 240, social media accelerators are used to take text segments and create structured information for further processing.
  - e. This data accumulates over time. Initially, the data can be integrated only around social media handles (such as social profiles), but as data accumulates in these profiles, it becomes rich enough to support matching to customer profiles or other social media profiles. Thus, continuous ingestion forms a key part of the social MDM solution.
  - f. Mining of the linked/discovered information can be used to create models about customers (Big Data Exploration).
2. Search and linkage of disparate data
- a. An increasingly broad range of information about an individual person (such as customer and patient) must be connected and displayed together for use by consumers (for example, a Customer Care center representative or a doctor).
  - b. The core of this information is in the MDM solution, but associated information exists in other structured sources (for example, account information and patient medical records), and in unstructured but accessible sources (for example, call logs).
  - c. Data in these cases must be quickly assembled and presented for the user, typically in real time and often with recent real-time events available for context.
  - d. Search across unstructured content and linkage to structured content is critical to this use case.
  - e. There is a federated approach that allows multiple sources to be brought together that include the MDM solution along with other relevant content.
3. Deep matching capabilities
- a. The volume of the big data, whether social profiles, call logs, or other content, precludes moving the data to the MDM system.
  - b. Where much of the information is ancillary to the needs of the MDM system, but important to the consumers, the data can be linked or matched where the data is rather than pushed into the MDM system.
  - c. The architecture supports both traditional MDM linkage/matching within the MDM solution, and a Big Match capability that can link and match the data where it is.

Let us look more specifically at the integration of the IBM Master Data Management Server with IBM InfoSphere Data Explorer to support the search and linkage of data, and the Big Match capability within IBM Big Insights to deeply link master information across various big data content.

## 9.1.2 InfoSphere Data Explorer and MDM

As described in Chapter 6, “Introduction to the IBM Big Data Platform” on page 97, InfoSphere Data Explorer is a powerful platform for searching, discovering, and linking data across a range of information repositories. Programmers can use the visual tools to create rich, searchable dashboards that deliver a 360° view of key business objects across structured and unstructured data, addressing the variety and volume of information that is related to, as examples, customers (or patient for healthcare, or citizen for public sector), products, locations (of stores or of customers), and partners.

InfoSphere Data Explorer can create these 360° views by harnessing a powerful four-tier architecture, as shown in Figure 9-2.

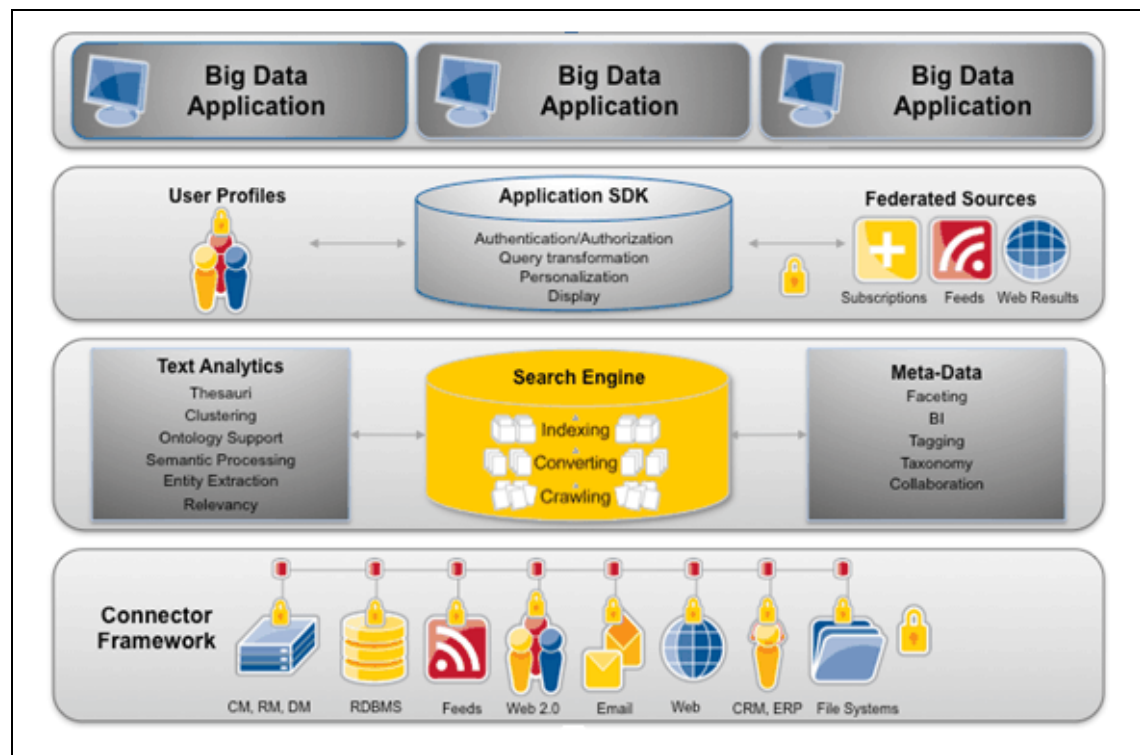


Figure 9-2 Architecture of InfoSphere Data Explorer

Starting from the bottom of Figure 9-2 on page 242, the architecture consists of the following layers:

1. A rich and extensible connector framework that can access a wide variety of source systems for both structured and unstructured data.
2. A processing layer that uses the connectors to crawl, index, and analyze large volumes of data. There are two functional aspects of the processing layer that are critical to creating the single view in InfoSphere Data Explorer:
  - a. An engine to discover and extract entities (such as customer and product) from the data sources.
  - b. A mechanism to keep the data fresh by querying the connectors for updates, and reindexing and analyzing as the data changes.
3. An application framework layer that is used by developers to create the overall entity model, manage the interaction and linking of discovered entities between data sources, and put all the related information together in a personalized portal.
4. The top layer is the set of Big Data Exploration and 360° view applications.

### **Governance concerns with InfoSphere Data Explorer**

Although InfoSphere Data Explorer can pull together an integrated and complete view of a business object, the accuracy and veracity (such as correctness) of that view depends on the quality of the data across the multiple sources. If the data suffers from the quality problems that are described in 9.1, “Master data management: An overview” on page 236, then InfoSphere Data Explorer cannot discern how all the entities are related across different data sources, and does not present a single view of the entity.

This issue is illustrated in some detail in Figure 9-3.

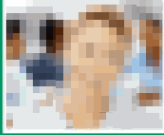


SOURCE SYSTEMS	
	<b>CRM</b>
	Name: J. Robertson
	Address 35 West 15 <sup>th</sup> Address Anywhere, US 12345
	<b>ERP</b>
	Name: James Robertson
	Address 35 West 15 <sup>th</sup> St. Address Anywhere, US 12345
	<b>Legacy</b>
	Name Jim Robert
	Address 36 West 15 <sup>th</sup> St. Address Anywhere, US 12345

Figure 9-3 Distinct data representations across three example source systems

The data from multiple systems that is consumed by InfoSphere Data Explorer has several inconsistencies and quality errors, which can produce incorrect views and interpretations in the portal that is built on InfoSphere Data Explorer. These inconsistencies include the following ones:

- Multiple representations of the same person across the different systems such that they cannot be correctly matched between all the systems (Name 1, Name 2, and Name3). So, the correct information is not retrieved and linked across all the systems. If users search on Name3, they get a different result than if they search on Name 2.
- Data is not cleansed or accurate (such as demographic information), so an enterprise user draws the wrong conclusions, or reaches out to the wrong person, or does not reach the correct person. In this example, we have the wrong name (Name 3) and the wrong home address for Name 2, and do not realize that Name 2 is married.

## **Using MDM to enhance an InfoSphere Data Explorer 360° view**

As described in 9.1.1, “Getting a handle on enterprise master data” on page 237, an MDM system explicitly addresses and remediates issues that are associated with the veracity and accuracy of master data. By using MDM as the primary data source of an InfoSphere Data Explorer 360° view of the customer/individual/product, an InfoSphere Data Explorer portal receives the needed veracity across the variety of big data in the portal, eliminating the business issues that are described in 9.1.1, “Getting a handle on enterprise master data” on page 237.

A typical integration of MDM with InfoSphere Data Explorer includes the following actions:

1. InfoSphere Data Explorer uses MDM to get the core resolved identity for a person.
2. InfoSphere Data Explorer queries MDM to get the core demographic information about the person (including privacy policies). This information also includes summary information of purchased products, accounts the customer holds, and so on.
3. InfoSphere Data Explorer takes the core identity information that is retrieved from MDM and uses it as a key to find the same person across the other source systems, ensuring that the correct information is returned from the other systems that are linked into the 360° view.

The results are shown in Figure 9-4.

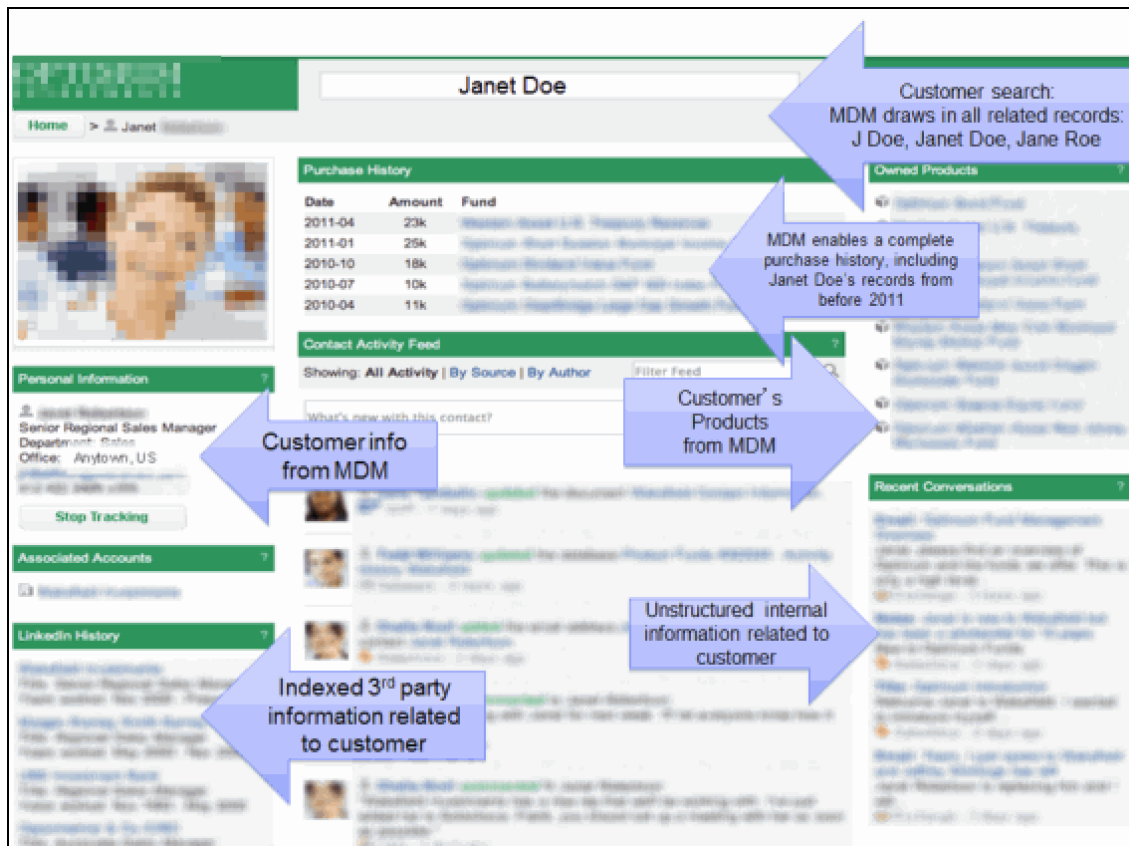


Figure 9-4 Integrated display of customer information in InfoSphere Data Explorer

## Architecture of InfoSphere Data Explorer and MDM integration

As shown in Figure 9-5 on page 247, the integration between InfoSphere Data Explorer and MDM relies on a tailored connector for MDM as the basis for a high-quality single view portal in InfoSphere Data Explorer. The portal developer uses MDM as the primary source for information about the customer (or other focus area) when searching other InfoSphere Data Explorer sources that are pulled in to the portal.

At the time of the writing of this book, the connector supports only the virtual style of MDM, which might limit what data is available.



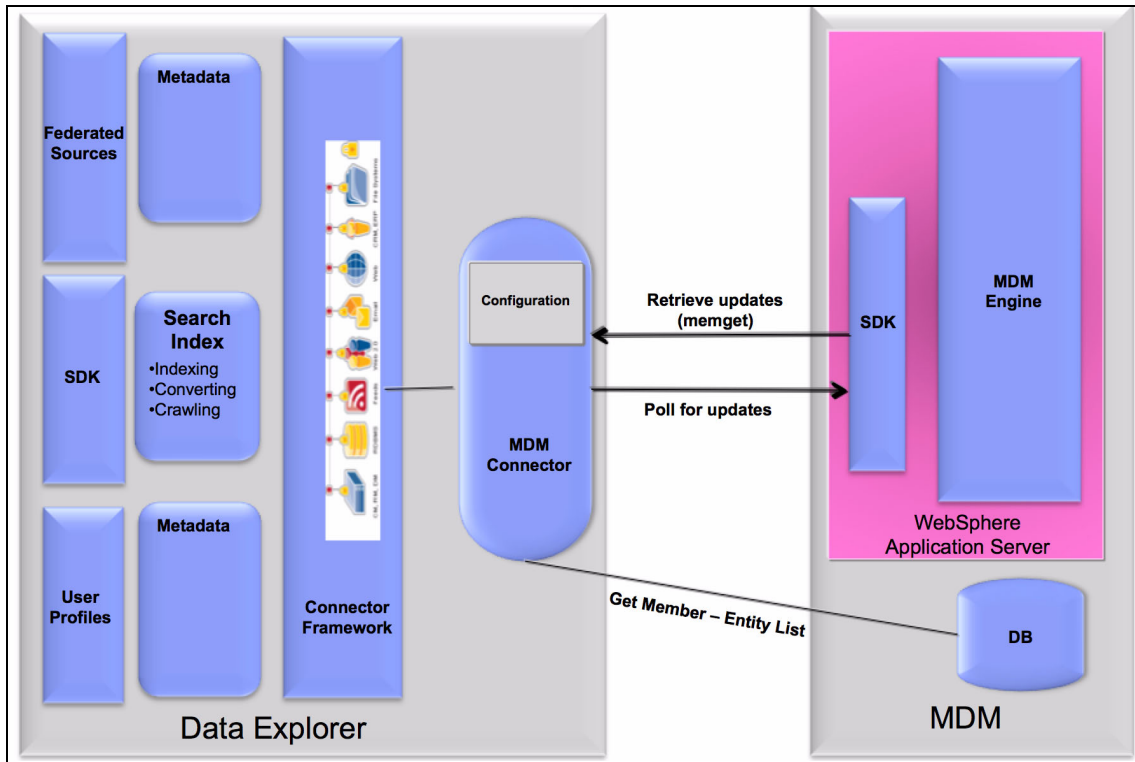


Figure 9-5 Architecture between InfoSphere Data Explorer and MDM

The integration consists of the following actions:

- ▶ When InfoSphere Data Explorer initially connects to MDM, InfoSphere Data Explorer crawls and indexes the MDM data.
- ▶ The index is periodically updated by pulling updates from MDM; the period can be specified in a configuration parameter to the connector.
- ▶ The MDM connector can retrieve the following information (the data elements (known as segments in MDM) are specified in the connector configuration):
  - Resolved entity and demographic information (in our previous example, Name 2 instead of Name 3, phone number, address, and so on).
  - Source records that contribute to the resolved entity. Based on the query, the connector can return all the records that are resolved for an entity along with the information of which system provided which record. In our example, this is Name 2, Name 1, and Name 3.
  - Householding information. Other person records related to the primary record (spouse, child, and so on).

- ▶ The developer creating the InfoSphere Data Explorer application performs the following actions:
  - Uses the MDM connector as the primary source of information about the person.
  - Uses the clean master data on the individual as the key to search other systems that are indexed by InfoSphere Data Explorer with confidence.

With InfoSphere Data Explorer and MDM, to get the correct information out of InfoSphere Data Explorer, you must consider several Information Governance preferred practices:

1. Follow your MDM governance processes:
  - a. Collapse duplicates through matching and linking, and deal with any records that are improperly linked by splitting them.
  - b. Standardize fields.
  - c. Establish relationships between entities:
    - i. Householding
    - ii. Person to organization
2. Apply the correct integration of InfoSphere Data Explorer and MDM:
  - a. Use the standard architecture for InfoSphere Data Explorer and MDM integration.
  - b. Ensure that you capture updates to MDM on a timely basis.
  - c. Establish the correct security to the MDM data (this is enforced by InfoSphere Data Explorer) to deliver only master data to authorized users.

## Big Match

As the volume and velocity of information about customers (and prospects, also known as potential customers) have grown dramatically in the age of big data, the ability of traditional matching technology to find and merge related and duplicate entity data has lagged behind the production of such data.

Furthermore, the variety of the sources of information (web registrations, third-party lists, social media profiles, and so on) cannot easily be integrated with the relational database model that is used in record matching by master data management systems. Fortunately, the underlying techniques that are used for entity management have been adapted to work in a big data environment.

The approach that is used by IBM InfoSphere Master Data Management Server is to take disparate sources of customer (or other party) data and match them through the Probabilistic Matching Engine (PME). PME matches records through three steps:

1. Standardization: Ensuring that data elements (names, addresses, phone numbers, social identification numbers, and so on) are consistently represented. This is the same capability (described in Chapter 8, “Information Quality and big data” on page 193) that is a key component in driving consistent information quality.
2. Candidate selection: In this step, the candidates are put into blocks based on the values in specific fields of the source records (grouping similar objects, such as all records with identical phonetic last names or with the same first three characters of a social identification number). The particular fields that are used in blocking are determined by record analysis, which looks at which fields are most likely to distinguish any records, and are blocked based on a hashing scheme to get a wide distribution of entries.
3. Candidate comparison: Finally, all the records in a block are compared on a field-by-field basis against each other, resulting in a matching score. The field scores have three components:
  - a. An overall weight of the field itself (matching names is more important than matching postal codes, matching phone numbers is more important than matching genders).
  - b. A quality of the match itself. Matches can be close (“Harald” versus “Harold”), exact, or even on nicknames (“Harry” versus “Harold”).
  - c. The probability of the presence of a value. This information is derived from the actual range of values in the data. For example, in North America, “Ivan” is a less prevalent name than “John”, so two names that match on “Ivan” have a higher weighting than two names that match on “John”.

Items that match over a threshold are automatically joined in a single record, with a predetermined mechanism for which fields are put into the common record. Items below that threshold, but above a floor (a lower or “clerical” threshold), can be sent to a data steward for resolution, and items below the floor level are kept as individual entities.

In a single-system probabilistic matching system, scalability is constrained by the following items:

- ▶ The amount of overall processing and parallel processing capability the system has.
- ▶ The requirement that all data elements be accessible on a common file system and in a common database.

- The distribution of items in the blocks (the less distributed the data elements are, the more comparisons the system must perform).

As the numbers of records start to move into the hundreds of millions, and with those records arriving daily, a traditional single-system probabilistic matching engine rapidly becomes overwhelmed. Either each block must hold more data, thus requiring an increasing number of comparisons, or the parameters for each block must be tightened and restricted to the point where many possible matches are missed because the blocks do not include relevant records. Furthermore, this does not include the processing cost for those records that must be transformed into a standardized format for processing.

### **Distributed probabilistic matching with Big Match**

The big data matching capability runs as a set of InfoSphere BigInsights applications within the InfoSphere BigInsights framework to derive, compare, and link large volumes of records, for example, 1,000,000,000 records or more. In general, the applications can either run automatically as you load data into your HBase tables or as batch processes after the data is loaded.

The overall approach lends itself to the distributed MapReduce paradigm that is used by big data:

- Mapping is used for standardization and blocking
- Shuffle is used for the distribution of blocks
- Reduce is used for the comparison of records in a block, and merging

The big data matching capability does not connect to or rely on the operational server that supports a typical InfoSphere MDM installation. Instead, the big data matching approach uses the existing matching algorithms (or allows you to build new ones) from InfoSphere MDM directly on the data that is stored in the big data platform.

Through the MDM workbench, you create a PME configuration that you then export for use within InfoSphere BigInsights. The chief component of the PME configuration is one or more MDM algorithms. In the realm of IBM InfoSphere MDM, an algorithm is a step-by-step procedure that compares and scores the similarities and differences of member attributes. As part of a process that is called derivation, the algorithm standardizes and buckets the data. The algorithm then defines a comparison process, which yields a numerical score. That score indicates the likelihood that two records refer to the same member. As a final step, the process specifies whether to create linkages between records that the algorithm considers to be the same member. A set of linked members is known as an entity, so this last step in the process is called entity linking. Potential linkages that do not meet the threshold are simply not linked together as entities.

The algorithms that you create differ depending on the data that you need to process. A broad range of algorithms exists, including functions to standard, group (bucket), and compare: names, addresses, dates, geographical coordinates, phonetics and acronyms, and the number of differences between two strings of data.

## Big Match architecture

From a governance perspective, it is useful to briefly consider the technical architecture for Big Match, particularly in relation to the big data reference architecture. As shown in Figure 9-6, when you install Big Match within your InfoSphere BigInsights cluster, the installer creates a component within the HBase master that is notified whenever a new table is enabled. When a table is enabled, the component creates a corresponding table in which to store the derivation, comparison, and linking data that is generated when you run the Big Match applications. The component also loads the algorithm configuration into the new matching processes on the HBase Region Server and into the Java virtual machines (JVMs) for MapReduce.

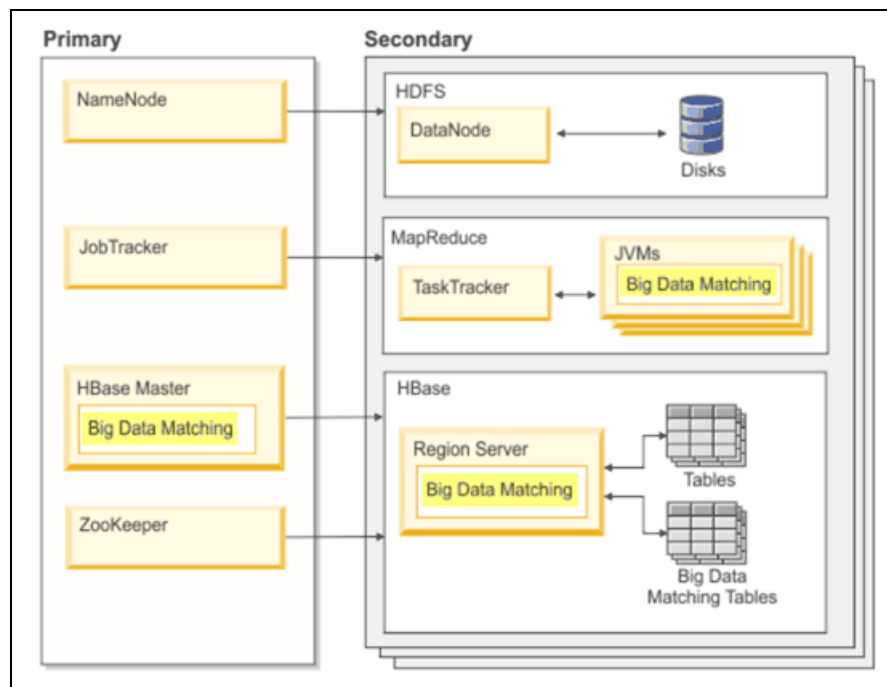


Figure 9-6 Architecture for Big Match within IBM BigInsights

After the components are installed and configured, you can run the applications in one of the following two ways:

- ▶ As automatic background processes that run as you load data into the HBase tables that you configured. As you write data into your InfoSphere BigInsights table, Big Match intercepts the data to run the algorithms.
- ▶ As manual batch processes that you run after you have loaded the data in to the HBase tables.

This architecture allows you to use existing investment and knowledge in match algorithms and provides the flexibility to support multiple additional algorithms on the data as your use cases expand.

The architecture also provides significant scalability. On a BigInsights cluster, aggregate comparisons per second are gated by available processor capacity, not by database I/O, and scale to much higher throughput rates. Improving performance becomes a task of simply adding data nodes to the InfoSphere BigInsights environment to handle higher volumes or reduce run time.

## 9.2 Governing master data in a big data environment

This section describes the governance risks for master data. Master data contains some of the most sensitive and regulated data within an organization regardless of where it is. From your earlier review of the big data reference architecture in Chapter 5, “Big data reference architecture” on page 69, you saw that the Shared Operational Zone, where master data is, is a core component of the Analytics Sources, and to support the emerging big data use cases, it must be available for the other zones. This creates the following risks:

1. Data privacy and security breaches. Basically, the master data is sensitive, so moving it around the enterprise in the following zones is an issue:
  - a. In the Exploration zone.
  - b. In the Deep Data Zone:
    - i. Entity matching.
    - ii. Sitting in Hadoop/HBase.
  - c. Moving data between zones.
  - d. In InfoSphere Data Explorer.
    - i. Getting access to unstructured data that was not previously traceable.
    - ii. The security policies are not appropriate for aggregated data.

2. Making unwarranted conclusions or linkages:
  - a. What if your data is not trustworthy or of high quality? Some examples, include false Yelp reviews, hiding of personal identifiable information (PII), bad master data, and not using enough information (no Big Match).
  - b. What if our analysis is wrong?
    - i. You link the wrong entities as a match (“Ivan Milx - Austin”, “Ivan Millx - Brazil”).
    - ii. You assert a relationship that is wrong. For example, Ivan Milx established a social media relationship with Deborah Milx, so you might assert that this is a family relationship (which is not true; Ivan just likes her work).
    - i. You miss a valid relationship. For example, Jane is the granddaughter of an important customer, but has an unrelated surname.
3. Surprising your customers with the following items:
  - a. Wrong insights.
  - b. Insights they did not want to share.
  - c. Creepiness factor (things they thought were private).

### **Information governance approaches**

To address these risks, organizations must take advantage of sound information governance practices:

- ▶ Security (access control) and encryption of data as it sits anywhere in the big data zones
- ▶ Masking of data as it moves into the Landing Area and Exploration zones
- ▶ Auditing of data access
- ▶ Redaction of sensitive data in real time (InfoSphere Guardium)
- ▶ Ensuring that InfoSphere Data Explorer security is correct
- ▶ Using stewardship practices within MDM to apply appropriate decisions for matching or delinking data
- ▶ Using data analysis tools to continuously monitor the quality of the data that is used in decision making, and the quality of a link or match of an entity.

These considerations tie back to the core principles that are outlined in Chapter 3, “Big Data Information Governance principles” on page 33. The usage of the big data reference architecture with an MDM solution allows an organization to achieve greater agility and use a greater volume and variety of information for more insight. The master data and ancillary information is core to an organization's ability to drive revenue and higher value, which means it is critical to protect that information and ensure that the people the organization interacts with understand and see that it is protected and valued by the organization. Furthermore, it is through appropriate stewardship, feedback, and compliance that this level of information governance can be maintained, even in the face of the volume, velocity, and variety of new data that is associated with the core master data.

In summary, you have looked at the integration of MDM solutions with the big data environment to support key big data use cases, particularly the enhanced 360° view. MDM technologies are coupled with capabilities for social media data extraction, for searching across structured and unstructured content for new enhanced views, and for matching big data in place. At the same time, the governance practices you have reviewed become critical for security and data privacy to conform to policy and avoid significant cost and risk to the organization, and for information quality to achieve the wanted level of veracity or confidence in the data and realize the potential value of the data and the insights that are derived from it.



# Related publications

The publications that are listed in this section are considered suitable for a more detailed discussion of the topics that are covered in this book.

## IBM Redbooks

The following IBM Redbooks publications provide additional information about the topic in this document. Some publications referenced in this list might be available in softcopy only.

- ▶ *Addressing Data Volume, Velocity, and Variety with IBM InfoSphere Streams V3.0*, SG24-8108
- ▶ *IBM Information Server: Integration and Governance for Emerging Data Warehouse Demands*, SG24-8126

You can search for, view, download, or order these documents and other Redbooks, Redpapers, Web Docs, draft and additional materials, at the following website:

[ibm.com/redbooks](http://ibm.com/redbooks)

## Other publications

These publications are also relevant as further information sources:

- ▶ *Analytics: The New Path to Value, a joint MIT Sloan Management Review and IBM Institute*, found at:  
<http://public.dhe.ibm.com/common/ssi/ecm/en/gbe03371usen/GBE03371USEN.PDF>
- ▶ *Analytics: The real-world use of big data - How innovative enterprises extract value from uncertain data*, found at:  
<http://www-935.ibm.com/services/us/gbs/thoughtleadership/ibv-big-data-at-work.html>
- ▶ *A Biased Map Of Every Global Protest In The Last 40+ Years*, found at:  
<http://www.fastcodesign.com/3016622/a-map-of-all-global-protests-of-the-last-40-years>

- ▶ *The Big Data Imperative: Why Information Governance Must Be Addressed Now*, found at:  
<http://public.dhe.ibm.com/common/ssi/ecm/en/iml14352usen/IML14352USEN.PDF>
- ▶ *Big Data: The management revolution*, found at:  
<http://blogs.hbr.org/2012/09/big-datas-management-revolution/>
- ▶ *Big Data Requires a Big, New Architecture*, found at:  
<http://www.forbes.com/sites/ciocentral/2011/07/21/big-data-requires-a-big-new-architecture/>
- ▶ *Data Profit vs. Data Waste*, IMW14664USEN
- ▶ *Data protection for big data environments*, IMF14127-USEN-01
- ▶ *Data virtualization: Delivering on-demand access to information throughout the enterprise*, found at:  
<http://www.ibm.com/Search/?q=IMW14694USEN.PDF&v=17&en=utf&lang=en&cc=us/>
- ▶ Densmore, *Privacy Program Management: Tools for Managing Privacy Within Your Organization*, International Association of Privacy Professionals, 2013, ISBN 9780988552517
- ▶ *The Digital Universe in 2020: Big Data, Bigger Digital Shadows, and Biggest Growth in the Far East*, found at:  
<http://idcdocserv.com/1414>
- ▶ *Digital Universe study*, found at:  
[https://www-950.ibm.com/events/wwc/grp/grp037.nsf/vLookupPDFs/ED\\_Accel\\_Analatics\\_Big\\_Data\\_02-20-2013%20v2/\\$file/ED\\_Accel\\_Analatics\\_Big\\_Data\\_02-20-2013%20v2.pdf](https://www-950.ibm.com/events/wwc/grp/grp037.nsf/vLookupPDFs/ED_Accel_Analatics_Big_Data_02-20-2013%20v2/$file/ED_Accel_Analatics_Big_Data_02-20-2013%20v2.pdf)
- ▶ English, Larry, *Improving Data Warehouse and Business Information Quality*, 1999, John Wiley and Sons, p.142-3.
- ▶ *The Evolution of Storage Systems*, found at:  
<http://www.research.ibm.com/journal/sj/422/morris.pdf>
- ▶ *The Hidden Biases in Big Data*, found at:  
<http://blogs.hbr.org/2013/04/the-hidden-biases-in-big-data/>
- ▶ *IBM Accelerator for Machine Data Analytics, Part 1: Speeding up machine data analysis*, found at:  
<http://www.ibm.com/developerworks/data/library/techarticle/dm-1301machinedata1/>

- ▶ “IBM Big Data Use Case Security/Intelligence Extension,” a PowerPoint presentation from IBM Corporation, February 19, 2013.
- ▶ *IBM InfoSphere BigInsights Enterprise Edition data sheet*, IMD14385-USEN-01
- ▶ *IBM InfoSphere Data Replication*, found at:  
<http://public.dhe.ibm.com/common/ssi/ecm/en/ims14394usen/IMS14394USEN.PDF>
- ▶ *IBM InfoSphere Master Data Management V11 Enterprise Edition*, IMF14126-USEN-01
- ▶ Lee, et al, *Journey to Data Quality*, MIT Press, 2009, ISBN 0262513358
- ▶ *The MDM advantage: Creating insight from big data*, found at:  
<http://www.ibm.com/common/ssi/cgi-bin/ssialias?infotype=PM&subtype=BK&htmlfid=IMM14124USEN/>
- ▶ *Query social media and structured data with InfoSphere BigInsights*, found at:  
<http://www.ibm.com/developerworks/data/library/techarticle/dm-1207querysocialmedia/>
- ▶ *Redefining risk for big data*, found at:  
<http://www.techrepublic.com/blog/big-data-analytics/redefining-risk-for-big-data/>
- ▶ *Simple Demographics Often Identify People Uniquely*, found at:  
<http://dataprivacylab.org/projects/identifiability/index.html>
- ▶ *Understanding Big Data - Analytics for Enterprise Class Hadoop and Streaming Data*, found at:  
<http://public.dhe.ibm.com/common/ssi/ecm/en/iml14296usen/IML14296USEN.PDF>
- ▶ *Understanding InfoSphere BigInsights*, found at:  
<https://www.ibm.com/developerworks/data/library/techarticle/dm-1110biginsightsintro/>

## Online resources

This website is also relevant as a further information source:

- ▶ IBM InfoSphere BigInsights - Bringing the power of Hadoop to the enterprise  
<http://www-01.ibm.com/software/data/infosphere/biginsights/>

## Help from IBM

IBM Support and downloads

[ibm.com/support](https://ibm.com/support)

IBM Global Services

[ibm.com/services](https://ibm.com/services)



## Information Governance Principles and Practices for a Big Data Landscape

(0.5" spine)  
0.475" <-> 0.875"  
250 <-> 459 pages







# Information Governance Principles and Practices for a Big Data Landscape

## Understanding the evolution of Information Governance

## Providing security and trust for big data

## Governing the big data landscape

This IBM Redbooks publication describes how the IBM Big Data Platform provides the integrated capabilities that are required for the adoption of Information Governance in the big data landscape.

As organizations embark on new use cases, such as Big Data Exploration, an enhanced 360 view of customers, or Data Warehouse modernization, and absorb ever growing volumes and variety of data with accelerating velocity, the principles and practices of Information Governance become ever more critical to ensure trust in data and help organizations overcome the inherent risks and achieve the wanted value.

The introduction of big data changes the information landscape. Data arrives faster than humans can react to it, and issues can quickly escalate into significant events. The variety of data now poses new privacy and security risks. The high volume of information in all places makes it harder to find where these issues, risks, and even useful information to drive new value and revenue are.

Information Governance provides an organization with a framework that can align their wanted outcomes with their strategic management principles, the people who can implement those principles, and the architecture and platform that are needed to support the big data use cases. The IBM Big Data Platform, coupled with a framework for Information Governance, provides an approach to build, manage, and gain significant value from the big data landscape.

## INTERNATIONAL TECHNICAL SUPPORT ORGANIZATION

## BUILDING TECHNICAL INFORMATION BASED ON PRACTICAL EXPERIENCE

IBM Redbooks are developed by the IBM International Technical Support Organization. Experts from IBM, Customers and Partners from around the world create timely technical information based on realistic scenarios. Specific recommendations are provided to help you implement IT solutions more effectively in your environment.

**For more information:**  
[ibm.com/redbooks](http://ibm.com/redbooks)