Compliments of informatica
Put potential to work.

Informatica Special Edition

Big Data Management FOR DUMMIES®

A Wiley Brand

Learn:

- Common big data project pitfalls and how to avoid them
- The difference between a big data lab and big data factory
- The three pillars of big data management
- Who you need on your big data team

Mike Wessler

About Informatica

Informatica is a leading independent software provider focused on delivering transformative innovation for the future of all things data. Organizations around the world rely on Informatica to realize their information potential and drive top business imperatives. Enterprises depend on Informatica to fully leverage their information assets. For more information, call +1 650-385-5000 (1-800-653-3871 in the U.S.), or visit **www.informatica.com**.

Big Data Management For Dummies

Informatica Special Edition

by Mike Wessler

For Dummies
A Wiley Brand

Big Data Management For Dummies®, Informatica Special Edition

Published by
John Wiley & Sons, Inc.
111 River St.
Hoboken, NJ 07030-5774
www.wiley.com

LIMIT OF LIABILITY/DISCLAIMER OF WARRANTY: THE PUBLISHER AND THE AUTHOR MAKE NO REPRESENTATIONS OR WARRANTIES WITH RESPECT TO THE ACCURACY OR COMPLETENESS OF THE CONTENTS OF THIS WORK AND SPECIFICALLY DISCLAIM ALL WARRANTIES, INCLUDING WITHOUT LIMITATION WARRANTIES OF FITNESS FOR A PARTICULAR PURPOSE. NO WARRANTY MAY BE CREATED OR EXTENDED BY SALES OR PROMOTIONAL MATERIALS. THE ADVICE AND STRATEGIES CONTAINED HEREIN MAY NOT BE SUITABLE FOR EVERY SITUATION. THIS WORK IS SOLD WITH THE UNDERSTANDING THAT THE PUBLISHER IS NOT ENGAGED IN RENDERING LEGAL, ACCOUNTING, OR OTHER PROFESSIONAL SERVICES. IF PROFESSIONAL ASSISTANCE IS REQUIRED, THE SERVICES OF A COMPETENT PROFESSIONAL PERSON SHOULD BE SOUGHT. NEITHER THE PUBLISHER NOR THE AUTHOR SHALL BE LIABLE FOR DAMAGES ARISING HEREFROM. THE FACT THAT AN ORGANIZATION OR WEBSITE IS REFERRED TO IN THIS WORK AS A CITATION AND/OR A POTENTIAL SOURCE OF FURTHER INFORMATION DOES NOT MEAN THAT THE AUTHOR OR THE PUBLISHER ENDORSES THE INFORMATION THE ORGANIZATION OR WEBSITE MAY PROVIDE OR RECOMMENDATIONS IT MAY MAKE. FURTHER, READERS SHOULD BE AWARE THAT INTERNET WEBSITES LISTED IN THIS WORK MAY HAVE CHANGED OR DISAPPEARED BETWEEN WHEN THIS WORK WAS WRITTEN AND WHEN IT IS READ.

For general information on our other products and services, or how to create a custom *For Dummies* book for your business or organization, please contact our Business Development Department in the U.S. at 877-409-4177, contact info@dummies.biz, or visit www.wiley.com/go/custompub. For information about licensing the *For Dummies* brand for products or services, contact BrandedRights&Licenses@Wiley.com.

ISBN: 978-1-119-25053-1 (pbk); ISBN: 978-1-119-25055-5 (ebk)

Manufactured in the United States of America

10 9 8 7 6 5 4 3 2 1

Publisher's Acknowledgments

Some of the people who helped bring this book to market include the following:

Project Editor: Carrie A. Johnson

Acquisitions Editor: Amy Fandrei

Editorial Manager: Rev Mengle

Business Development Representative: Karen Hattan

Special Help from Informatica: John Haddad, Murthy Mathiprakasam

Production Editor: Tamilmani Varadharaj

Table of Contents

Introduction 1

About This Book 1
Icons Used in This Book 2
Beyond the Book 2

Chapter 1: Identifying Big Data 3

Evolution of Data over the Years 3
Introducing Big Data 5
The Vs of Big Data 5
Identifying Different Sources of Data 6
How Big Data Is Used in Business 8

Chapter 2: Understanding the Challenges of Big Data 11

Identifying the Traditional Challenges of Big Data 11
Emerging Next-Generation Challenges 12
State of Big Data Projects 13
Understanding Why Businesses Are Struggling with Big Data 15
Introducing Big Data Management 16
Understanding the layers of big data 17
Defining big data management capabilities 17
Overcoming obstacles with big data management 18

Chapter 3: Building Blocks of Effective Data Management 19

Understanding a Big Data Laboratory versus Factory 19
Identifying the Three Pillars of Data Management 22
Integration 22
Governance 23
Security 23
Diving Deep into Big Data Management Processes 25
Empowering the Big Data Team 27

Chapter 4: Using Big Data Management in the Wild 29

Implementing Big Data Management in Business 30
Identifying Big Data Tools 32
Leveraging the Right Tools 34
Considering Commercial Tools Built atop Open Source Projects 35
Combining Management with Integration, Governance, and Security 36

Chapter 5: Ten Essential Tips for Succeeding with Big Data Management 39

Design Use Cases for Business Value 39
Automate and Centralize Your Data Management 40
Leverage Data Lakes 40
Create Collaborative Methods for Governance 41
Identify Data Quality Issues Early 42
See Your Data and Relationships with a 360-Degree View 42
Work with Expert Vendors to Accelerate Your Deployments 43
Look for Process Repeatability 43
Align Your Vocabulary 44
Automate Key Processes 44

Introduction

Big data is the subject of great energy and excitement, and for good reason. The prospect of channeling all the data in the universe (and that is a *lot* of data) into analytical engines to understand relationships between entities, identify illusive patterns, and predict future events is exciting! It is changing our lives and altering the way businesses see us as consumers. When used correctly, businesses find that big data unleashes a wealth of information and insights which translate to higher profits, reduced costs, and less risk; it is a win!

The downside is, despite all the hype, many big data projects struggle to deliver on those lofty promises. The fact is that while technology evolved and data grew at an exponential pace, the processes to manage big data were left behind. The result was frustration with many big data projects.

This book provides a solution through big data management. Based on three pillars of integration, governance, and security, big data management provides a set of processes and technologies that make big data projects successful. These pillars will deliver data that is clean, governed, and secure to discover insights and turn them into real business value.

About This Book

Big data management is the solution to struggling big data projects in business today. Too much emphasis is placed on individual technologies and not enough focus on foundational components of data integration, data governance, and data security. Applying these foundational components and their underlying processes will improve the effectiveness of big data projects that rely too much on new and unmanaged point-solution technology and "do-it-yourself" (DIY) manual processes.

The focus of this book is learning what big data management is and how to apply it to real-world big data projects. You will learn strategies, processes, and identify tools to implement big data management so you can deliver big data projects faster and with greater value.

Icons Used in This Book

Throughout this book, you will occasionally see special icons to bring your attention to a point that needs emphasized. I will keep them brief, and sometimes a little funny, but if you see one, take note because it's something you should know.

TIP

Tips indicate information that you may find useful. Often, they relate to an experience I had (or I wish I had at the time), or they add additional context to a topic.

REMEMBER

If you see this icon, it's probably something that will help you later. You won't find the meaning of life here, but you may find some advice that will make your life easier.

WARNING!

Warning means just that; be careful! I use warnings to alert you to common mistakes and serious issues for you to avoid.

TECHNICAL STUFF

I am a technical person at heart and I love to understand how and why things work (or don't). Yes, this is a *Dummies* book, but sometimes I delve deeper into a subject so you understand the "why" and "how" for a key topic.

Beyond the Book

This book can't teach you everything about big data, big data management, analytics, or exciting technologies like Hadoop or NoSQL. I encourage you to research these topics on your own, if not from a professional perspective, at least explore big data and the impact it will have on you as both a consumer and as a citizen in our ever-connected society.

One good place to visit is the Informatica Big Data Ready web page at `informatica.com/bigdataready`.

Chapter 1

Identifying Big Data

In This Chapter

- Understanding the evolution of data
- Explaining big data
- Leveraging big data in business

Data has evolved over the years and will continue to evolve. Originally a stable stream of well-structured data, the growth of technology has unleashed a flood of varied data from a myriad of sources. The flood of big data can overwhelm those who are unprepared, but for those ready for big data, many new business opportunities await. In this chapter, I explore how data has evolved into big data and how big data is used in business.

Evolution of Data over the Years

In the early days of data processing (early term for IT), data came from relatively few, well-defined sources; and once it came into the computers of the day, it was *structured.* That is, structured data is in a known format of data size and type; think alphanumeric data for customer names, identification numbers, and sales numbers. Over time, the size and number of data sources grew to be large, but they grew in a predictable manner and maintained their structured format.

Technologies to store and manage data such as Relational Database Management Systems (RDMS) evolved and excelled at managing this kind of structured data. Programming languages and Business Intelligence (BI) reporting tools were developed to glean value from the data. As structured data grew, iterative improvements in technology kept pace.

As new technologies emerged, they generated a new type of data: *unstructured* data. Unstructured data comes in a variety of data types, sizes, and formats. Examples of unstructured data include audio and video files, pictures and images, and unstructured text streams such as mobile texts or social media posts.

Some people use the term *semi-structured* data to further differentiate within unstructured data. Social media posts and mobile texts, which are loosely formatted text, are considered semi-structured data. Log files and sensor equipment are also semi-structured examples.

Technologies to manage structured data struggled to adapt to support unstructured data as well. Many established technologies offered support for unstructured data, but often the implementation wasn't as mature as what existed for structured data. In response, new technologies such as Hadoop and NoSQL emerged especially suited for unstructured data.

Hadoop is an open source software framework from the Apache Software Foundation used to store and process large non-relational data sets using a distributed architecture. NoSQL is a class of databases engineered to process large unstructured and semi-structured datasets. Commercialized and open source distributions of these technologies exist including Cloudera, Hortonworks, and MapR for Hadoop and Apache Cassandra, MongoDB, MarkLogic, and Couchbase for NoSQL.

Beyond formatting structure, data itself is categorized into different categories that are useful to understand:

- **Traditional data:** Data already existing in legacy systems, corporate databases, and local data stores such as Excel spreadsheets. This tends to be structured data and is well managed using existing technologies.
- **Enrichment data:** Data which is specific to a purpose and supplements traditional data. This is often external such as demographically available data about the customers already stored in traditional customer tables.

- **Emerging data:** Data which is new, external, and is often big data; usually in an unstructured, non-traditional format. Examples include social media, sensor, or log data. Data that's internal to the enterprise includes emails, documents and comments, and machine log files.

Data is continually evolving and will continue to do so as technology grows. Technology scientists, vendors, and businesses have the challenge of keeping up with this evolution, and the latest major evolutionary step is big data.

Introducing Big Data

In its simplest terms, big data isn't just a lot of data; it's a lot of data being generated very rapidly and in a lot of different formats. Big data is by its nature both structured and increasingly unstructured, and it's being generated at a very fast rate from a variety of data sources. These factors ensure that big data is "big" in that there's a lot of it, and it's increasing at an explosive rate.

Beyond raw size, big data exceeds the capacity of existing traditional systems to store and process it; new technologies and processes are required to make effective use of big data. Big data is very often unstructured and not stored within an organization's corporate databases; it's external and doesn't neatly fit into predefined formats.

Traditional technologies and methodologies are simply not suited to capture, store, or process big data. The new technologies and methodologies are critical for businesses and data professionals to understand when they enter into the world of big data.

The Vs of Big Data

Definitions of big data vary and will evolve because it's still a relatively young field. However, most reputable definitions include reference to the original "Three Vs of Big Data":

- **Volume:** The vast size of the data in terms of actual size and number of data items

- **Velocity:** How fast the data is being created and moves across networks
- **Variety:** Variation of data types including factors such as format, structure, and source

Recently, two additional Vs have been increasingly added as big data is better understood and used within business:

- **Veracity:** The trustworthiness of the data in terms of quality and accuracy
- **Value:** The benefit of the data to the organization and questions being asked; how it can be turned into business value.

It's a safe statement that big data is defined by its volume, velocity, variety, and veracity; these are all key factors to consider from the technical perspective. Including the "value" of big data in its definition recognizes that data varies in its business impact to an organization and that value is an important factor in determining the time horizon as to where to store and retain the data.

TIP

Advances in computing power and software are mitigating some of the impact of large data volume and velocity. What is increasingly important is the *management* of big data, which defines successful implementations. The Vs of big data are important, but the people, processes, and technology standardization and rationalization are where your projects will succeed or fail.

Identifying Different Sources of Data

To fully understand big data, why it's significant, and why it's challenging to manage, you must appreciate that

- Data is growing at an unprecedented, explosive rate.
- Technology is generating data in new and different ways.

How much is data growing? Consider these facts:

- The amount of data in the world doubles every two years based on multiple sources.
- By 2020, there will be 450 billion transactions on the Internet every day, according to International Data Corporation (IDC).
- As of 2013, there were between 2.5 to 3 zettabytes (ZB), but by 2020 there will be 44 ZB of data per IDC.

Where in the world is all this data coming from? New technologies, sensors on existing technologies, metadata (for example, data about data), and data about nearly everything a person or device does add to the data universe every second. Examples include the following:

- The over 7 billion mobile devices in use today allow for one device per person on Earth — wow! Furthermore, consider the many apps on each device that generate location, communication, purchase, picture, video, and social media data.
- Online activities for web users ranging from browsing, communications, and commerce are other common examples. Usage patterns and preferences contained in sessions yield a gold mine of useful data to be harvested.
- Sensors in the everyday devices you use in your lives. Increasingly, telemetry and location data within cars, home appliances, and entertainment devices are added as new features and capabilities are introduced.
- Sensors within medical, scientific, and manufacturing devices. As each device becomes more capable and networked, the data generated increases. Everything from hospital beds tracking a patient's detailed statistics to sensitive controllers on the factory floor are examples.

TIP

Networked chips embedded within appliances, manufacturing devices, mobile devices, and others are part of the *Internet of Things* (IoT). IoT is a growing contributor to big data and is itself a rapidly expanding technology.

While traditional data will always grow, unstructured enrichment and emerging data are growing even faster. Understanding the magnitude of data growth and having an appreciation of its sources better posture you to understand big data and its direction.

How Big Data Is Used in Business

A strong theme for businesses is to leverage technology to enhance operational effectiveness, reduce costs, and reach new and existing customers with better products and services faster than the competition. Much of the value proposition of big data is to be able to identify key information about your environment and customers so you can act more quickly to seize an opportunity.

Examples of how businesses can use big data are vast and industry specific, but common use cases include

- Business revenue generating and risk reduction operations to find new opportunities, reach out to new customers, increase customer loyalty, and increase operational efficiency.
- IT cost-reduction activities such as data warehouse optimization and offloading, and building centralized data lakes. Additionally, cost reductions can be realized by implementing projects to update existing IT infrastructure and operations to better leverage external data sources rather than reproducing in-house.
- Industry-specific activities to improve predictive maintenance in manufacturing environments, reduce healthcare costs while improving outcomes, service improvements in utilities and telecommunications, and identify and reduce fraud and default in insurance and financial industries.

The limit of how big data can be leveraged is increasingly less about the limits of technology and more about the imagination and management processes of its practitioners.

Serving up big data via the cloud

Cloud computing is about providing something useful *as a service*. Infrastructure as a Service (IaaS), Platform as a Service (PaaS), and Software as a Service (SaaS) are the most common forms of this computing architecture. Current examples of IaaS and PaaS from large vendors include Amazon Web Services (AWS) and Microsoft Azure.

Cloud computing in terms of big data includes accessing external data as a service. Rather than attempting to store data internally, you must access it via a service. It can be conceptually simple as customer demographic data or Twitter feeds or more complex industry-specific data in a private cloud. Another example is providing big data analytical processing capabilities as SaaS. Many companies can't establish a big data analytics capability in-house so they wisely use a big data SaaS offering to accelerate their implementation at lower cost and reduced complexity.

Chapter 2

Understanding the Challenges of Big Data

In This Chapter

- Highlighting the challenges of big data
- Learning why businesses struggle with big data
- Overcoming challenges with big data management

Big data promises incredible opportunities, but unfortunately, a lot of work and complexity are involved in unlocking those opportunities. Some challenges are obvious, while others are more subtle. Beyond technical obstacles, the operational and management challenges are often the most difficult to address. Fortunately, the intelligent use of a data management methodology can solve these challenges. In this chapter, I discuss the challenges of big data and introduce big data management as a solution.

Identifying the Traditional Challenges of Big Data

Some challenges of big data are obvious and tie back to the three Vs of big data: volume, variety, and velocity. These challenges manifest themselves as

- Large, ever expanding volume of data across a multitude of sources
- Different data types, especially with unstructured data
- Constant generation of new data to the point of system overload

New roles are emerging inside organizations like data scientists who need very quick access to data for discovering new insights. Automation systems and new types of data-driven applications also require constant access to trusted data assets. All these requirements can be challenging to deliver against as new data platforms, like Hadoop, emerge, which introduce new skills and new systems to integrate and manage.

These represent the traditional and obvious challenges of big data. On the positive side, while they are indeed daunting challenges, they are well understood by business and technical folks alike. Furthermore, evolution of technology and increases in processing power and storage help mitigate some of their impact. However, on the negative side, these obstacles don't reflect the full spectrum of challenges faced by big data practitioners; there's another set of more subtle yet equally important *next-generation* challenges.

Emerging Next-Generation Challenges

As companies delve into big data projects and technical and business specialists roll up their sleeves and "get dirty," they find a whole new class of challenges they may have never seriously considered before. These challenges are every bit as daunting as the more traditional challenges, and in fact they're often more difficult to overcome.

Next-generation challenges include the following:

- **Varied data sources:** Data comes from and resides in many different internal and external sources, including data warehouses, data marts, data lakes, generated reports, the cloud, and third-party resources. Data commonly originates from business transactions, web and machine log files, and social media.
- **Data silos:** Valuable data may not be accessible due to overly stringent polices and/or politics, or it may be so distributed, segregated, and walled-off that accessing it is too resource intensive and not repeatable. Or there may be useful data, but it's in a silo you aren't aware of; thus, you miss a full 360-degree view of the data.

- **Increasing security risks:** Data breaches are big news and legitimately can spell disaster for a company or organization; no one wants their CIO in the newspapers for a data breach. Sensitive data exists by itself and aggregation of seemingly non-sensitive data can become a security issue.
- **Lack of data governance:** In absence of a unified data governance policy, either too many controls or not enough as relate to data quality and data sharing are enforced, and there's no uniform process of accessing and managing (curating) data. At best, your efforts to access, prepare, and curate data are inconsistent and inefficient; at worst, you either can't access data, can't trust the data, or you create a potential security issue.
- **Too many emerging and changing technologies:** Big data is still evolving with new vendors, technology, and open source projects. Keeping track of this shifting landscape is difficult, and standardizing on a big data platform and methodology is both technically and often politically complex.
- **Value is difficult to unlock:** Data in itself has little value, but finding the important relationships within a data universe to identify actionable information is the real challenge. IT, business, and management stakeholders must be equipped with technology, policy, and the will to find and exploit opportunities from data before real value is achieved.

Next-generation big data challenges are often as process- and policy-driven as they are technical. In these cases, faster CPU chipsets won't help; it takes people breaking down institutional barriers, implementing new policies, and leveraging smarter toolsets that simplify work to overcome these obstacles.

State of Big Data Projects

Unfortunately, traditional- and next-generation challenges often encumber projects and frustrate businesses with a negative impact on the perception of big data. Industry experts claim big data will solve everyone's problems overnight, but many projects haven't experienced stellar results.

Most savvy technology, business, and executive folks understand the value of big data someday; regrettably, despite their efforts, that day is not *today*. Many big data projects have similar issues:

- **Increased complexity:** Pulling out the valuable insights from data is harder than originally imagined. Finding, accessing, integrating, and preparing diverse data can consume most of the project resources.
- **Extended delays:** Moving from inception to delivering a production-ready product takes too long. Subsequent projects also take nearly as long.
- **Moving beyond the Proof of Concept (POC):** The POC project either languishes indefinitely without showing results, or it's a one-off event that's difficult to reproduce on a larger scale.
- **Immature processes:** Business processes, data governance, and security compliance standards aren't yet mature enough to support the effort.
- **Unexpected cost:** Time and resources invested exceed what was initially expected.
- **Slow Time to Value:** The ROI being delivered is less than what was originally promised.

You can see in Figure 2-1 the complex ecosystem of big data projects with many data inputs, business outputs, cross-system processes, and multi-disciplinary stakeholders; no wonder these are difficult to manage.

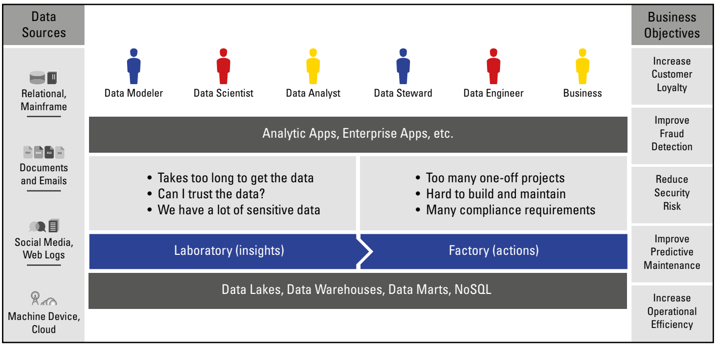

Data Sources
Relational, Mainframe
Documents and Emails
Social Media, Web Logs
Machine Device, Cloud
Data Modeler
Data Scientist
Data Analyst
Data Steward
Data Engineer
Business
Analytic Apps, Enterprise Apps, etc.
• Takes too long to get the data
• Can I trust the data?
• We have a lot of sensitive data
• Too many one-off projects
• Hard to build and maintain
• Many compliance requirements
Laboratory (insights)
Factory (actions)
Data Lakes, Data Warehouses, Data Marts, NoSQL
Business Objectives
Increase Customer Loyalty
Improve Fraud Detection
Reduce Security Risk
Improve Predictive Maintenance
Increase Operational Efficiency

Figure 2-1: Common problems with big data projects.

The ultimate effect of these problems is a degradation of faith in big data projects by business leaders. Executives conceptually understand big data can provide value, but for a multitude of reasons, they become hesitant to aggressively pursue future projects. This perception is unfortunate because once expectations are properly set and managed, coupled with the right data management methodology, great results are possible. However, to get past these initial hurdles, it's necessary to understand *why* big data projects experience problems.

TIP

Beware the magic "cure-all" solution regardless of the technology involved. Vendors love to sell companies tools that promise to erase complex problems, make everything faster, and reduce costs while bringing in huge profits, often with the single click of a button. The truth is, tools can bring you great outcomes, but it takes *work* to make it happen. Be realistic and know how to separate hype from reality.

Understanding Why Businesses Are Struggling with Big Data

Why the disconnect between the wonderful promises of big data versus the reality of many big data projects? The answer is usually a combination of factors, but often a set of common themes is found in struggling big data projects.

Companies struggling with big data projects typically have fallen prey to one or more of the following pitfalls:

- **Trying to do it all yourself:** Do It Yourself (DIY) is how many projects begin, but they struggle uphill as teams stumble through a steep learning curve for technology, processes, and governance. Inefficiencies, delays, and lost opportunities are common with this trial-and-error approach. Appealing at first as an attempt to keep costs low or as a POC effort, DIY seldom pays off in the long run.
- **Not enough trained and experienced people:** Closely tied to DIY, not having a sufficient pool of expertise on the project will cause delays and inefficiencies while driving up costs. Developing big data skillsets takes time and resources; it's not something staff can easily do in addition to their other duties.

- **Custom, hand-coded solutions:** A by-product of DIY, inexperienced teams are developing custom, hand-coded solutions, which aren't repeatable for future efforts. Common with data cleansing and integration processes, these home-grown solutions are hard to initially develop and aren't reusable when new datasets or projects are introduced. Thus, over time, a myriad of custom solutions are written at great expense with limited future benefit.
- **Not leveraging appropriate tools and best practices:** Reinventing the wheel time and time again while ignoring the expertise more experienced big data experts (and vendors) have developed is wasteful in terms of time and resources. Taking too much of a narrow view without leveraging the greater body of expertise and tools frequently leads to frustration, delay, and reduced positive results.

Did you notice these common errors are more about methodology than they are about raw technology? And their roots are in next-generation challenges more than traditional challenges? If you want to be successful with big data, you need to understand and embrace big data *management*.

REMEMBER

Is your big data project entirely IT driven or does it have a business leader as champion? Where is the funding (and hopefully there *is* funding) coming from and what and when is the projected ROI? And is the staff trained and allocated as a dedicated resource or is this yet another side project as time allows? Be sure to ask these questions and consider the consequences.

Introducing Big Data Management

Big data management is defined as the capture, integration, administration, and governance of big data for use in an organization's data-related applications, often with an emphasis on business intelligence and analytical applications. It spans both technical and non-technical aspects of access, integration, governance, and security of data being used within an organization to highlight relationships, identify trends, or otherwise provide a more complete view of the business environment. In particular, the focus is on the integration, governance, and security of big data. However, to really understand big data management, one must first comprehend the hierarchy of big data architectural layers.

Understanding the layers of big data

Operating on big data is best visualized as a three-layered architectural hierarchy:

- **Visualization and analysis:** This includes visualization tools like Tableau and Qlik and analytic tools with advanced statistics and machine-learning algorithms like R, SAS, and H2O.
- **Big data management:** This includes the technology needed to integrate, govern, and secure big data such as pre-built connectors and transformations, data quality, data lineage, data mastering, and data masking. Platform vendors for big data management include Informatica and a variety of startups with specific point solutions.
- **Storage persistence layer:** This includes persistence technologies and distributed processing frameworks like Hadoop, Cassandra, data warehouses, and MPP appliances.

The world of big data management exists in the middle, between the visualization and analytics and the storage and processing framework layers. Big data management interfaces with the data at the storage layer, processes that data, and provides datasets for visualization and analysis to the upper layer.

Defining big data management capabilities

Big data management performs several core functions within a defined hierarchy. Some functions are entirely technical, while others are non-technical. The core functions are

- **Integration:** Finding, accessing, integrating, cleansing, preparing, and delivering the data

- **Governance:** Managing and curating the data to ensure it is high quality, clean, trusted, and "fit-for-use"
- **Security:** Protecting the data from unauthorized access and manipulation

In Chapter 3, I dive deep into each function of big data management as you explore the details of integration, governance, and security.

Overcoming obstacles with big data management

Assume you have established a successful big data management solution; what problems can it solve? Big data management can help you

- Reduce the time to access, integrate, and prepare data
- Enhance the trustworthiness and quality of the data
- Protect sensitive data and reduce security breaches
- Enforce compliance and standards across the board
- Create repeatable processes to accelerate future efforts
- Bring projects from proof of concept to production faster, and with less risk and effort
- Adapt as technology and data evolves

Big data management is the solution needed to get you past the traditional and next generation stumbling blocks common to so many projects today. Using big data management won't magically solve your problems overnight, but it will move you much closer to meeting the expectations and achieving the ROI that has been promised for so long.

Chapter 3

Building Blocks of Effective Data Management

In This Chapter

- Exploring the differences between a big data laboratory and factory
- Learning the three pillars of big data management
- Detailing the processes within big data management
- Enhancing effectiveness through empowerment and intelligent resource management

If you've been reading this book straight through, at this point you know what big data is and why big data projects encounter problems; you are now ready to take a deep dive into big data management to understand why it is core to a big data strategy. I cover in detail the three pillars of big data management and what they do and why they're critical to your project's success. Next, you explore the processes associated with big data management in detail. Finally, I end the chapter with help on how to empower your team and make the most of the resources you already have in place. This chapter provides the foundational knowledge to truly understand big data management.

Understanding a Big Data Laboratory versus Factory

Before I delve into the details of big data management, I must apply context to the environments in which big data is used. Depending on the environment, the requirements for an effective implementation differ in several key ways; I explore those ways and why they matter.

In terms of environments, everyone often thinks in terms of development, test, and production, but with big data that isn't entirely accurate. Rather, think in terms of laboratory environment versus a factory environment.

Big data laboratory environments are typified by data scientists testing a series of datasets with a set of analytical algorithms in an attempt to identify key insights that bring potential value to the business. This is entirely scientific experimentation with the intent to find datasets and analytical models that can be passed to the operations team who will put the solution into production (aka "the factory"); that is where real business value will be derived. I use the term *factory,* but some organizations refer to it as *operations* or *production.*

Big data factories take the model provided by the laboratory and put those solutions into use as production. In big data analytics uses cases, these environments are managed by a team of IT specialists with business analysts reviewing and applying the resulting insights to the business. However, more common use cases (for example, next best offers, fraud detection, new data-driven products, predictive maintenance, and so on) strive to deliver actionable information directly to the end-user in real time. This eliminates the need for a business user acting as a middleman layer between the end-users and data. Big data factories are focused on data products that provide actionable information directly to end business users and consumers.

There is a relationship and dependency between laboratories and factories:

- Laboratories experiment until they have a solution that will provide business value. They pass that solution on to the factory for production.
- Factories implement the solution provided by the laboratories to generate real business value.
- Without factories, laboratories would have no reason to exist.
- Without laboratories, factories would have no solution to implement in production.

There are, however, critical differences between the needs of laboratories and factories:

- Self-service autonomy in the laboratory is critical because data scientists will be conducting many, many experiments until they find a winning solution. Data scientists need the freedom to set up and execute experiments themselves rapidly to pursue their research.
- Data in the laboratory is less subject to cleansing than in a factory environment. There's relatively little impact if data isn't fully curated in a laboratory environment as long as it's an authentic and accurate representation of the real world.
- Operational agility in factory environments is important because after a valuable insight is identified, the business must be nimble enough to exploit its temporary advantage. It does little good to have key information if you're unable to take advantage of that knowledge or if a competitor implements it first.
- Data integrity, timeliness, and trustworthiness are also important requirements in a factory environment; the only thing worse than no data is bad data. Taking action based on incorrect or outdated information can be costly in terms of time and resources; plus it degrades the trust the business has on big data for future operations.
- Laboratories and factories must conform to corporate security policies to protect sensitive data and adhere to regulatory compliance.

During the transition from laboratory to factory, make sure applicable governance and security policies are followed and put in place. On occasion, a less stringent laboratory environment introduces elements that aren't permitted in a controlled factory environment.

Why does this matter in terms of big data management? Assuming you accept that laboratories and factories are critical to the big data operations within a company, those two distinct environments must be managed appropriately. The needs and differences of the environments must be respected and carefully managed; recognizing these different approaches to integration, governance, and security is important when evaluating big data management platforms.

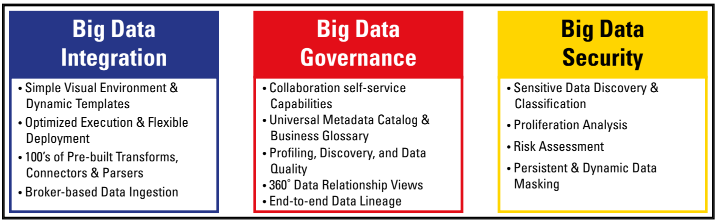
Fortunately, big data management can be architected to be flexible enough to meet the needs of laboratories and factories once you understand how it works.

Identifying the Three Pillars of Data Management

The power of flexibility of big data management comes from its architecture. Rather than attempting an overly complex architecture, big data management builds on three foundational pillars:

- Integration
- Governance
- Security

In Figure 3-1, you see the pillars of data management.



Big Data Integration

- Simple Visual Environment & Dynamic Templates
- Optimized Execution & Flexible Deployment
- 100's of Pre-built Transforms, Connectors & Parsers
- Broker-based Data Ingestion

Big Data Governance

- Collaboration self-service Capabilities
- Universal Metadata Catalog & Business Glossary
- Profiling, Discovery, and Data Quality
- 360° Data Relationship Views
- End-to-end Data Lineage

Big Data Security

- Sensitive Data Discovery & Classification
- Proliferation Analysis
- Risk Assessment
- Persistent & Dynamic Data Masking

Figure 3-1: Understanding the pillars of data management.

As you can see in Figure 3-1, there are only three pillars, but each pillar encompasses multiple processes.

Integration

Integration ingests and processes data to achieve a result; this processing must be scalable, repeatable, and agile. The longest delays in big data projects occur during integration; smarter integration will reduce these time frames, automate processes, and allow for rapid ingestion of new data. Key components of integration include

- Agile and high performance ingestion of next generation data
- Automated and scalable integration, cleansing, and mastering of next-generation data
- Optimized and readily usable tools for ingestion and processing coupled with repeatable processes

Governance

Governance defines the processes to access and administer data, ensures the quality of the data, how it is tagged and cataloged, and that it is fit-for-purpose. Essentially, the business and IT teams must have confidence their data is clean and valid. Key components of governance include

- Collaborative governance to allow everyone to participate in holistic data stewardship
- 360-degree view knowledge graph of data assets showing semantic, operational, and usage relationships
- Trust and confidence that the data is fit-for-purpose
- Data quality, provenance, end-to-end lineage and traceability, and audit readiness

Security

Security identifies and manages sensitive data with a 360-degree ring of risk assessment and analysis. Security must occur at the source, not just at the perimeter. Identifying which data is sensitive (credit card information, email addresses, addresses, Social Security numbers, and other personally identifiable information) and which data aggregated together becomes sensitive is a growing challenge. Key components of security are

- 360 degrees of sensitive data discovery, classification, and protection
- Data proliferation and risk analysis
- Masking and encryption for sensitive data
- Security policy creation and management

Security is huge, and many organizations rightfully protect their data like a grizzly bear protecting her cubs. This can become an obstacle for data access (ingestion as part of integration), especially if you can't prove you have sufficient security and governance controls in place.

It is far better to do the security, compliance, and governance work up front to alleviate data owners' concerns before requesting sensitive data. You must demonstrate you have appropriate security controls in place; otherwise data owners will block your efforts.

Exploring data governance strategy

Of the three pillars of data management, governance is often the most foreign to people with a technical background. Governance is about the policies, procedures, techniques, and technology you use to administer your data to ensure it is trustworthy, accurate, available, and actionable.

Governance is inherently a bureaucratic process; regulations, laws, and auditors require controls to exist. That frequently concerns people because they think governance must be a hindrance, and that perception isn't correct. Governance can either work for or against you, depending on how it is approached.

If you have weak governance, data will effectively become "locked up" because there is no established process to "free" it. Every time you want data, it's a battle to gain access.

If you establish strong governance processes allowing access to data across your enterprise, you will have created a standardized, repeatable process that will pay many dividends. Getting data becomes streamlined because you already have polices to access that data.

Governance also improves the quality and trustworthiness of your data, and it helps identify relationships within that data. In this context, you use governance techniques and technologies to enrich, curate, and tag metadata, thus making the data more useful and actionable.

Knowing the provenance (origin) of data, and tracing it from creation to its current state (end-to-end lineage), allows that data to be much more transparent and trustworthy to the business. Done correctly, you will continuously enrich a golden data record of your customers, and that brings real value to the table.

Diving Deep into Big Data Management Processes

Much of the heavy lifting of big data management occurs within integration. During integration, data ingestion, cleansing, preparation, and processing occur; however, security and governance also have processes as well. Understanding these processes will enhance your ability to manage big data more effectively.

Key big data management processes include

- **Access data:** Set up repeatable, well-managed processes to acquire data from both traditional and next generation data sources. Multiple data sources will be used, so having pre-configured access tools and connectors are a great timesaver.
- **Integrate data:** Establish processes to prepare and normalize data for a myriad of data sources. This process is often very challenging; resist the temptation to rely on manual methods, and leverage automation and repeatability as much as possible.
- **Cleanse data:** Review the data to ensure it's ready for use; that means checking for incomplete or inaccurate data and resolving any data errors that may bias analysis or negatively impact business operations and decision making. Beware this process can be tedious, and leverage automation options when available.
- **Master data:** Organize your data into logical domains that make sense to your business such as customers, products, and services. Furthermore, you can add enrichment data to further paint a clearer picture of your customers, products, and services and their relationships.
- **Secure data:** A mix of governance and security allows you to establish security rules and then implement those rules. First, you must determine how you will manage your sensitive data. Next, you must find and assess the risk of your sensitive data and implement rules via policy and technology. This process is very important but prone to be under-addressed by those inexperienced in big data management.

- **Explore and analyze data:** Implement a data laboratory to perform experiments with a clear business goal in mind. Based on your hypotheses, find what data exists and how it can be analyzed to create a model that delivers results. Then determine if the results are beneficial to the business; remember that providing actionable information and processes is the goal. Develop best practices to enhance agility and processes before pushing the solution into the factory.
- **Explore and analyze for business needs:** Test out data products to see if they provide a real value for the business; often you just need to try something to see if it works. It is common to use A/B testing to determine if a new data product adds value to the business. Make iterative improvements over time as you learn what works, what doesn't work, and what can be improved.
- **Operationalize the insights:** Automate and streamline your processes to create a steady pipeline of actionable insights to business users. It's not enough to have occasional production runs from the big data factory; the factory must be running regularly to be truly productive, meet business service-level agreements (SLAs), and achieve the expected ROI.

These processes aren't necessarily linear, although they have a general flow with reiteration and loopback as necessary. Really, these processes run as a cycle as data is brought into the system, processed, tested, and then implemented for the business; then the next data project or test is started.

The system will ingest data from data sources, clean, integrate, and manage that data, and then pass it to analytic applications for processing to develop insights and finally to business applications in the form of actionable information, all while applying big data management processes. Understanding the processes of big data management enables you to better manage environments.

Empowering the Big Data Team

It is not cliché to say a company's greatest asset is its people; it's the truth. The challenge is what can be done to increase their effectiveness and ability to produce results, and in this context I mean working with big data.

First, understand the role and needs of each team member or category of member. There will be a mix of data scientists, modelers, analysts, stewards, engineers, and business users, all with different perspectives, skill levels, and needs. Some will require greater self-service autonomy (in the laboratory environment), while others require operational agility (in the factory environment); your job is to identify their needs within the big data environment.

Next, incorporate the three pillars of big data management into the team members' operating principles and environment. Using a disciplined approach, ensuring that in particular governance and security processes are followed, is one of the biggest favors you can do for your team. Not having governance policies enabling hassle-free access to data will doom your team to needless headaches negotiating access to needed data. Failing to have necessary security controls in place also adds to data access issues, but worse yet it opens up the risk the team could be associated with a data breach. Make sure your team understands the value of governance and security and uses it to their advantage.

Next, get help for your team in terms of training, effective technology, outside experts, and vendor experience. Odds are your team is already overworked; why make them do things the "hard way" by denying those tools and expertise to increase their effectiveness? Forcing your team to work in isolation devoid of the great work already done with big data will send the team down a path of one-off, custom solutions, manual processes, and tedious work that is not reproducible. That DIY approach results in frustration for the team and costly lost opportunities for the business.

Finally, consider what you can do with what you already have by creating repeatable, automated processes and standardized technologies. Rather than re-inventing the wheel and expending resources for each new project or dataset, seize opportunities where you can

- Reuse existing infrastructure and tools
- Reuse skillsets, expertise, and processes
- Reuse previous projects' components

Working big data projects is initially complex work, but when quality big data management principles are followed, that work can be reused again and again to the benefit of the team and the business.

Taking steps to empower your big data staff isn't just right for them as employees, but it yields benefits for the company as well.

Your people are an investment, and those in the big data field know their value. There's an industry shortage of qualified data scientists, data engineers, and those who have knowledge and experience in the big data world, and that shortage is expected to increase in the near future. You must be willing to develop and retain your highly skilled big data workforce; otherwise they may go elsewhere under favorable market conditions.

Chapter 4

Using Big Data Management in the Wild

In This Chapter

- Identifying common use cases and drivers in big data management projects
- Learning more about big data management tools and techniques
- Highlighting several big data management toolsets in use today
- Reviewing real-world business experiences of big data management

Understanding the foundational concepts of big data management is essential, but you must also understand how the concepts exist in the practical world. Businesses initiate big data management projects for various purposes and from multiple perspectives; it's important to understand the drivers of those efforts. The ability to identify key attributes in big data management tools and how to effectively use those tools within the principles of big data management in production environments is critical information I provide. Finally, I identify some useful toolsets and highlight business experiences with those tools. In this chapter, you gain an understanding of how to merge the big data management principles with real-world business operations.

Implementing Big Data Management in Business

Companies initiate big data management projects for a variety of reasons. While the industries and circumstances may vary greatly, most projects originate from two emphases:

- Business centric
- IT centric

Business-centric big data management projects are just that: focused on generating a business benefit. Often, the intent is to generate new or additional revenue where it is relatively easy to calculate the ROI. In other cases, more subtle benefits occur such as better understanding the preferences or relationships of perspective customers or improving existing business processes. Even more subtle benefits are avoidance or detection of specific conditions such as fraud, claims, or preventing a component failure via the Internet of Things (IoT). These projects are frequently initiated by business analysts and executives.

Examples of common business-centric uses cases include

- Using credit card transaction and money transfer data for real-time fraud detection
- Analyzing website visitor behavior with clickstream data
- Processing customer data for a customer 360 initiative to gain a complete picture of customers' demographics, interests, patterns, and behavior
- Devising and implementing a program to increase customer loyalty
- Using predictive analytics to detect failing components and replace them before an assembly line is impacted
- Improving hospital patient outcomes and reducing the total cost of care

IT-centric big data management projects often seek to improve a process or provide an analytical or data capability that previously didn't exist. These projects provide infrastructure cost savings and an indirect business benefit and are more difficult to calculate ROI. For example, IT may create a new data lake or build a Hadoop cluster that provides enhanced capability for the organization, but in isolation these projects don't generate revenue unless they support a business process or initiative. These projects are often generated by IT as a consolidation or modernization initiative or in response to business requests to explore a new capability.

Examples of common IT-centric uses cases include

- Building a staging environment to offload data storage and Extract, Load, and Transform (ELT) workloads from a data warehouse
- Extending a data warehouse with a Hadoop-based Operational Data Store (ODS)
- Building a centralized data lake that stores all enterprise data required for big data analytics projects

These use cases are just the tip of the big data iceberg; below are examples of real companies' positive experiences:

- A large bank wanted to reduce the time required to detect fraudulent events. By better and faster preparation of data before processing with anti-fraud software, fraud events were detected quickly. The bank also integrated data from outside sources including credit card, money transfers, and mortgage payment information to further enhance their fraud analysis and detection.
- A well-known insurance firm sought to better understand their customers and to generate more carefully targeted and personalized marketing campaigns. By using big data management tools, the company ingested, cleaned, and matched data from customer profiles, partner data, previous history, web logs, and social media activity. This data created a 360-view of customers enabling more customized and effective marketing campaigns.
- A large healthcare insurer wanted to lower the cost of care while improving patient outcomes. The company leveraged big data management tools, improving its analytics infrastructure, which allowed more effective patient-provider collaboration and member engagement. These efforts improved patient outcomes and reduced costs, while retaining and increasing the number of those insured.

Putting big data management in the context of business and IT-centric projects is beneficial in understanding how big data management can help *your* business.

Identifying Big Data Tools

Within big data management architecture, there are processes and tools. Often the processes used are dependent on one or more tools for implementation. Rather than focusing on specific products (which can encompass several tools), you must first identify the categories of tools common in big data environments.

Big data tools are commonly separated into the following categories:

- **Data Ingestion:** Ingest (obtain, import, and process) data from different sources at various latencies (for example, including real-time), in an efficient usable manner leveraging pre-built connectors to simplify the data ingestion process.
- **Data Management:** The end-to-end tools and processes used to integrate, govern, secure, and administer the transformation of source data into data that is "fit-for-purpose" and in compliance with corporate and regulatory policies.
- **Data Integration:** Combine data from different sources using a variety of transformations such as filtering, joining, sorting, and aggregating while establishing relationships within the datasets to provide a unified view.
- **Data Quality:** Clean up data to ensure it's fit for its intended purposes to appropriately address incomplete or irrelevant entries, eliminate duplications, standardize and normalize data, and ensure data exceptions are handled properly, preferably in an automated and repeatable fashion.
- **Metadata Catalog:** A dedicated repository to collect, manage, and report on data assets, their relationships, and the processes used to integrate, govern, and secure those assets. A universal metadata catalog that spans the entire data infrastructure landscape is the foundation for big data management.

- **Master Data Management:** Enforce and ensure the accuracy and accountability of the critical data in an organization to provide a common point of reference and truth.
- **Data Masking:** De-identify, obfuscate, or otherwise obscure sensitive data, such as credit card numbers, so relational integrity is maintained, yet key sensitive values aren't accessible.
- **Data Security Analytics:** Analyze and assess the risk of a data security breach by identifying the location and proliferation, and tracking the usage of sensitive data.
- **Streaming Analytics:** Collect, process, and analyze multi-latency data (including real-time) to provide event-based insights and alerts within a time-interval of maximum business impact (often in real-time or near real-time).
- **Big Data Analytics:** Apply analytical formulas and algorithms to datasets to answer questions based on big data; these algorithms are employed by data experts to test hypotheses and validate analytic models used to improve business outcomes.
- **Data Lakes:** Collect and store all types of data as originally sourced for use as a live archive, data exploration, and an operational data store for pre-processing and preparing data for big data analytics.
- **Data Warehouses:** Collect and store structured data into a large repository for the purpose of applying analytics and generating reports.

When evaluating tools and software packages, ask "what does this actually do and where does it fit within my big data architecture?" Often, if you can't find a satisfactory answer of what a product does or how it complements or replaces an existing technology within your IT infrastructure, you should beware that it may have limited or no value. The same concept applies with big data management tools; if the perspective tool doesn't include functionality listed in the above categories, there may be more marketing hype than substance.

Leveraging the Right Tools

It isn't enough to simply have tools; to be effective you must have the *right* tools to complete the task at hand. The challenge is, how do you know what the right tools are? Specifically, what attributes make one tool more desirable over another tool? In an industry ripe with marketing hype and buzzwords, one must know how to identify real value.

One great way to start is applying the three pillars of big data management. If the tool relates to one or more processes within the pillars of integration, governance, or security, then odds are you are on the right track. Next, as discussed in the preceding section, determine what function or work the tool actually performs; it should be clearly defined with a demonstrated purpose or output. Finally, drill down into the specific features for each tool to determine which ones support forward looking, enterprise-grade features such as

- Automation and repeatability of key processes
- Reduced complexity with increased productivity via predefined and dynamic templates, connectors, transformations, rules and algorithms, and intuitive management tools
- Resiliency to underlying technology changes to preserve development and reduce maintenance
- Support for hybrid architectures such as cloud computing
- Agile and rapid deployment across multiple environments
- Leveraging existing skillsets and resources

Consistent with the focus on the three pillars of big data management, several key features under each pillar are

- Integration
    - High volume multi-latency ingestion
    - Optimized for powerful, scalable processing
    - Rapid deployment across varied environments

- Governance
  - Collaborative self-service approach
  - 360-view of data relationships
  - Fit-for-purpose data
- Security
  - Complete discovery and view of sensitive data
  - Analysis and assessment of security risks
  - Risk-centric and policy-based security

The ability to distinguish between okay versus great tools and needless fluff versus real features is important for any IT professional, not just those working in big data. By applying the methodologies above, you will more accurately identify quality tools warranting further investigation and discard tools providing lesser value.

Considering Commercial Tools Built atop Open Source Projects

Open source projects have given a wealth of powerful tools that drive the IT industry; examples include Apache web server, Tomcat application server, and Hadoop. By themselves, these open source products are used by many businesses, large and small, with great success.

However, no product is perfect, and vendors often use open source products as the basis for their offerings. Their reasoning is very compelling; take a proven open source package and add vendor-specific modifications to make a good open source package into a better commercial product. This is a common practice and has yielded many successful results.

Customers often prefer vendor solutions built atop open source for several reasons:

- Dedicated and accountable support for issues and bug fixes
- Regularly scheduled and accountable upgrades and security patches
- A single point of contact for issues, training, and expertise

- Features and enhancements made possible only by vendors with extensive expertise and large R&D engineering departments
- Comfort level and compliance assurance that a paid vendor is behind the product

In the world of big data management, Informatica has taken a similar approach of integrating their data management expertise with open source packages.

For example, Informatica leverages several open source tools as part of their big data management stack. Specifically, Informatica leverages open source MapReduce, Spark, Hive on Tez, YARN, Navigator, and Sqoop. As a user, you can expect the functionally of each open source package, but additional enhanced capabilities from Informatica.

Combining Management with Integration, Governance, and Security

Big data management pillars of integration, governance, and security form the overarching hierarchy of management processes. Those processes are implemented via technologies — often open source technologies. In Figure 4-1, you see how integration, governance, and security ride atop big data management processes powered by open source and vendor technologies.

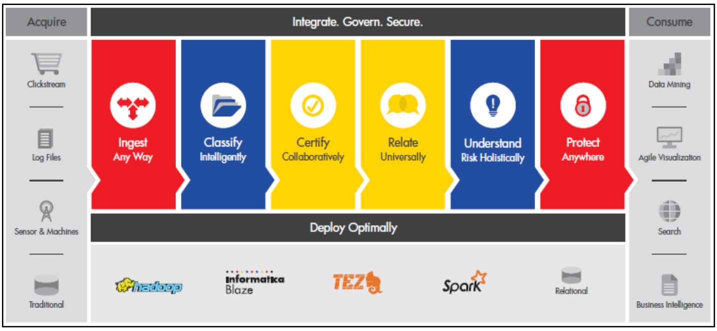


Figure 4-1: Integrating big data management processes and technology.

Introducing Informatica big data management products v10

Informatica, a leader in data management technology, has recently released its v10 family of new and upgraded products. Providing the three pillars of big data management, these tools merit investigation.

Tools of particular interest to the big data management practitioner include

- **Big Data Management v10:** Complete delivery of integration, governance, and security built on Hadoop
- **Secure at Source v10:** Enables data security intelligence allowing identification, analysis, and mitigation of security risks
- **Master Data Management (MDM) v10:** Provides MDM capabilities to provide a single view of data and 360-degree view of relationships and transactions

To learn more about these big data management products, visit the Informatica website at `www.informatica.com`.

The approach shown in Figure 4-1 is common in the big data world. Using a mix of open source tools (Hadoop, Spark) with a vendor big management engine (Informatica Blaze) and universal metadata catalog (Informatica Live Data Map), big data management processes are applied to technology to deliver integration, governance, and security.

Chapter 5

Ten Essential Tips for Succeeding with Big Data Management

In This Chapter

- Designing use cases with business value
- Leveraging data lakes
- Governing data collaboratively
- Working with expert vendors
- Automating key processes

Managing big data is the key to successful big data projects. Beyond technology, the management techniques deployed make the difference between success and failure. In this chapter, I identify tips and techniques to make you more effective at managing big data in the real world.

Design Use Cases for Business Value

Delivering value early and often is a function of your overall strategy. Sure, you can develop a very large and ambitious plan that promises a substantial payout at the end of the project, but that approach is fraught with risk and is often difficult to sell to senior leadership. Overly sized projects are difficult for new teams to tackle, and if issues occur (as they usually do), the future of the project is jeopardized.

A better way is to establish uses cases that deliver smaller victories earlier in the process. Using smaller, agile teams who focus on rapid, iterative development practices to show value early has many benefits. First, agile development does solve many of the challenges and mitigate risks found in larger projects; plus agile is the current, favored development methodology in many organizations. Next, a project that shows value early is easier to "sell" to management initially and to sustain as the project continues. Finally, smaller, more realistic goals are easier to achieve while building the confidence and capability of agile teams. When designing your uses cases and assembling your teams, focus on quicker wins that show a benefit rather than risking an overly ambitious project.

Automate and Centralize Your Data Management

Big data *management* practices are the defining factor in a project's success. Too often, organizations find themselves with fragmented, disparate teams attempting manual or one-off processes under the guise of big data management. These efforts often evolve from a lack of centralized direction from the top and never define an enterprise toolset to manage big data. In these situations, despite individuals' best efforts, success is often elusive.

To avoid this situation, at the outset define a core team to manage big data for the organization. Empower this team and break down organizational barriers to their success, often related to data ownership and access. Next, give them the enterprise-class tools required to do the heavy lifting of integration, governance, and security in an automated manner. The goal is to develop a team of data experts who focus their time and the organization's resources on solving data challenges rather than battling with other internal groups or struggling with inefficient tools and labor-intensive processes.

Leverage Data Lakes

Not all efforts require the same characteristics of data quality or access, or are even used for the same purpose. In some cases, data scientists use data experimentation, visualization, and advanced analytic tools to explore possibilities. In other cases, business analysts use reporting and BI tools to make key decisions. Many other examples exist, and they all have different data requirements.

Data lakes provide powerful capabilities to meet different requirements from the same repository of raw data. A data lake stores a large amount of raw data in its native format; it's tagged with unique identifiers and metadata, but it's still raw. When a business question is asked, the data is queried and the appropriate, smaller subset of data is returned to answer the question. This is in contrast to data marts, which offer a pre-built subset of data designed for a specific use case; in many situations the siloed nature of data marts is a liability.

The power of the data lake is that the same repository of raw data can be used for different use cases. Data scientists can use the data lake for their research while business analysts access more curated and governed datasets for their operational requirements. Sharing the same data lake for different purposes adds flexibility to the organization without the overhead cost of redundant, purpose-specific data marts.

Create Collaborative Methods for Governance

Establishing data governance is not a one-time event. As with most policies, after creation it is necessary to continuously monitor results and periodically revise policies to ensure they are relevant and make sense for the organization.

Establishing data governance methods is a collaborative approach between multiple stakeholders, but the two core groups are IT staff and business experts. The work itself is a continual cycle with key phases being discovery, definition, execution, and continuous monitoring. During discovery, automated tools discover data domains, and data stewards profile the data to gain insight into the quality of the data. During definition phase, the business glossary, metadata, data taxonomy, data quality, data matching, data access, and retention rules are established. In the execution phase, data stewardship, master data management, and provisioning policies are applied. The final phase uses measurement and continuous monitoring of results. The cycle is repeated as needed, and changes are identified and implemented. Throughout this cycle, careful collaboration between business and IT stakeholders occurs.

TIP

A universal metadata catalog is a key capability supporting data governance processes. It provides the ability to holistically understand and manage data assets and their relationships facilitating search, discovery, collaboration, and automation.

Identify Data Quality Issues Early

Few problems get better by themselves with age; the same concept holds true with data quality issues. Data quality issues degrade the integrity and trust in big data output. The sooner issues are identified, the sooner they can be addressed.

Tools and processes exist to identify problem data. After data has been initially ingested, cleansed, and processed, the method of applying data scorecards begins. You must define data profiles for the data quality scorecards and rules to be applied to the data. Once applied, you can address exceptions in the data both automatically and through alerts that require human intervention, and monitor scorecard results. Use of data scorecards will help ensure data quality is maintained.

See Your Data and Relationships with a 360-Degree View

Seeing the complete picture of your data provides the greatest opportunity to identify opportunities and mitigate risks. Multiple views of data exist, but two primary perspectives to consider are security and relationships.

Viewing data from a security perspective entails several different factors. First, you need to identify and protect your sensitive data across the whole of your operations, not just in one area. Siloed data protection is not effective if one area isn't protected; take a holistic approach. Set up effective security policies with audit triggers and notifications for key events. These efforts will give you a much more complete view and, therefore, control of your data.

Identifying relationships within datasets is where business value and opportunities exist; you must see these relationships to seize the greatest benefit possible. Place your efforts into identifying relationships between customers, intermediate parties, and products to know where best to leverage sales opportunities or anticipate specific events.

Work with Expert Vendors to Accelerate Your Deployments

A common error in companies engaged in big data projects is they try to do too much themselves without assistance. Big data management and analytics is an inherently complex discipline; it's not easy. Furthermore, most IT staff are almost entirely occupied just keeping current operations running; engaging in major projects outside their area of expertise often results in long-running projects with suboptimal results.

A wise solution is to intelligently engage vendors specializing in big data technologies, management, and analytics from the very start of your project. Include in-house IT and business experts because they understand business processes and IT peculiarities of the organization better than anyone else. However, leverage the deep expertise and technologies offered by reputable vendors to do the heavy lifting specific to big data management. Leveraging the right mix of in-house knowledge with outside expertise will yield positive results faster and with less risk for your organization.

Look for Process Repeatability

The most complex and time-consuming steps related to big data management are accessing, cleansing, and integrating the data. It is a given that the first time you perform these processes, it will be tricky. That said, successful organizations position themselves to endure this only once, rather than repeating the exact same struggles every time a new dataset is identified and accessed.

Implementing process repeatability is a key to success. First, avoid custom programming solutions unique to a situation and manual processes. Just as code reuse is a key to effective programming, standardizing and reusing processes and logic are highly beneficial with data management. Processes and logic related to data cleansing and integration are often good candidates to standardize and reuse. Find and document patterns in processes and logic that your teams can reuse time and time again to speed up delivery and reduce their workload so they can focus on more meaningful efforts. Leverage these reusable patterns and logic for tasks such as data ingestion, web log processing, ELT offloading, address validation, masking credit card numbers, and so on.

Align Your Vocabulary

An area that is often overlooked is aligning your processes and documentation along a common vocabulary. A popular exercise for many in school was to whisper a simple message around a circle in the classroom. Once the message was verbally received by the last student in the classroom, that message was compared to what was originally sent. Usually the message was distorted and often it was radically different. If that happens with common language, why would you expect different results with technical and business terms?

When starting a data management project, start off right by creating a glossary of accepted business terms and relevant technical terms. Be sure the terms are defined within the context of the data, processes, and project. Particularly when using business or technical terms, make sure they are fully defined and given the appropriate context. Distribute the glossary among relevant stakeholders and ensure their vocabulary is aligned with the contents of the glossary. Enforce rigor and adherence to the glossary throughout the project to reduce drift in language and prevent issues later.

Automate Key Processes

"I'd like to automate that long, manual process, but it is just too complex" is a statement commonly heard in IT operations. Sometimes a process truly cannot be automated, but too often there are other factors at play including Fear, Uncertainty, and Doubt (the FUD factor) or simply not having enough time or resources, which prevent automation.

Identifying your repeatable processes and automating those processes yield great benefits in terms of greater efficiency, faster output by IT staff, fewer human errors, and freeing up staff so they can apply their talents to more beneficial tasks. As mentioned throughout this book, data integration and governance can easily consume 80 percent of your analytics staff's time, so leverage tools wherever possible to simplify and automate key processes involved in managing your data. Consider your time spent automating key processes to be an investment, and it is a wise investment to make.

120,00
238,680
Indianapolis
95,890
Durham
Long Beach

To get accurate insights from big data you must first manage it

Big data holds some amazing promises for businesses to identify illusive patterns, predict future events, and gain a competitive advantage. But getting to these important discoveries is often a struggle without the right technology infrastructure, organization, and process for managing big data. This book shows you how to

- ***Overcome the challenges* — change big data challenges into small ones**
- ***Better manage big data* — make big data valuable for any use case**
- ***Integrate big data* — generate agile, automated, and optimized data**
- ***Govern big data* — ensure it's clean and trusted**
- ***Secure big data* — make it safe and compliant**

Open the book and find:

- **What makes big data big and how is it used**
- **Why businesses struggle with big data**
- **How big data management overcomes big data challenges**
- **Real-world business experiences of big data management**
- **Ten do's and don'ts of managing big data**

Go to Dummies.com®
for videos, step-by-step examples, how-to articles, or to shop!

FOR DUMMIES®
A Wiley Brand

Also available as an e-book

ISBN: 978-1-119-25053-1
Not for resale

WILEY END USER LICENSE AGREEMENT