Making Everything Easier!™

BMC Software Special Edition

Managing Big Data Workflows

FOR DUMMIES®

A Wiley Brand

Learn to:

- Select the best big data technologies for the task at hand
- Solve business problems using big data technologies
- Navigate in the world of big data workflows
- Hire the best big data candidates

Brought to you by

bmc

Joe Goldberg
Lillian Pierson, PE

About BMC Software

BMC is a global leader in software solutions that help IT transform traditional businesses into digital enterprises for the ultimate competitive advantage.

Managing Big Data Workflows

FOR DUMMIES®
A Wiley Brand

BMC Software Special Edition

by Joe Goldberg and
Lillian Pierson, PE

FOR DUMMIES®
A Wiley Brand

Managing Big Data Workflows For Dummies®, BMC Software Special Edition
Published by
John Wiley & Sons, Inc.
111 River St.
Hoboken, NJ 07030-5774
www.wiley.com

ISBN 978-1-119-25904-6 (pbk); ISBN 978-1-119-25905-3 (ebk)

Manufactured in the United States of America

10 9 8 7 6 5 4 3 2 1

Publisher's Acknowledgments

We're proud of this book and of the people who worked on it. For details on how to create a custom *For Dummies* book for your business or organization, contact info@dummies.biz or visit www.wiley.com/go/custompub. For details on licensing the *For Dummies* brand for products or services, contact BrandedRights&Licenses@Wiley.com.

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

Project Editor: Martin V. Minner
Acquisitions Editor: Steve Hayes
Editorial Manager: Rev Mengle
Business Development Representative: Kimberley Schumacher
Production Editor: Kumar Chellappan

Table of Contents

Introduction

Most technical people have at least a basic idea about what big data is, but strikingly, many are still unaware of the tremendous impact big data is delivering to the enterprise. Simply put, big data technologies are a disruptive force that is rapidly changing the competitive landscape of business in every industry. While most organizations are taking their time in making the transition, other organizations are getting a significant head start through early adoption.

Although transitioning to big data technologies can be a massive and costly undertaking, it does not have to be that way. The good news is that you do not have to throw your old, traditional technologies away when you make the switch. Instead, you can continue using them in ways that enrich the results you get from a big data, augmented system. By augmenting your existing data system with big data sources and technologies, you can continue using the traditional infrastructure that has always delivered value to your organization. Furthermore, you will increase that value by orders of magnitude by integrating new voluminous streams of multivariety data that is relevant to your business's bottom line.

Likewise, you will be happy to know that many of the principles and processes you have used in traditional data management are relevant to big data technologies. Take batch processing and workflow management, for example; the vast majority of work that is done in big data systems is done by batch processing. This processing can be automated in the same way that traditional data processing has always been done — through workflow management tools. That means the transition into big data technologies is not nearly as steep as most would assume.

While there are some technical adaptations that should be put into place for big data technologies to be made enterprise-ready, those technologies are developed and readily available. Although big data is not the most costly IT initiative, certain

costs — factors such as how much data you store, and how many issues you choose to analyze — will determine project costs. The ROI that big data promises to deliver (when done correctly) should make the project more than worthwhile.

BMC Software knows just how daunting and complex it may seem when transforming your approach to data management and analytics. BMC sponsored this book to help make this journey simpler and easier for those looking to embark on your big data journey. Moreover, if you decide you could use some help putting big data measures into place, BMC Software has the expertise you need to get you going as quickly and painlessly as possible.

Foolish Assumptions

In writing this book, we have assumed that readers have a good understanding of how businesses operate and common requirements when it comes to information technology. The point of this book is to show readers how implementing big data technology will deliver value to the enterprise. It is not meant to be a technical how-to manual. If you are looking to expand your understanding of big data and its technologies so that you can make key decisions on how your organization should proceed, this book is for you.

Icons Used in This Book

Throughout this book, you will see the following margin icons:

The Tip icon marks tips and shortcuts that you can use to make subject mastery easier.

Remember icons mark the information that is especially important to know. To siphon off the most important information in each chapter, just skim through these icons.

The Warning icon tells you to watch out! It marks important information that may save you headaches.

Chapter 1

Defining Big Data and the Technologies Comprising It

In This Chapter

- Looking at what makes big data different
- Surveying big data's impact across industries
- Understanding Hadoop and how it works
- Relating Hadoop to other big data technologies

Although the term *big data* was originally coined back in the late 1990s, it wasn't until recently that data professionals began questioning the validity of the term. In heated data debates, professionals argue over whether *big data* is just another media-backed buzzword, or if the term serves some significant purpose. Opponents of the term "big data" argue that data is data, period. So why not just call it *data* and be done with it?

Although you may easily relate to opponents' resistance to marketing buzzwords, they're putting the cart before the horse by arguing against the term *big data* on this basis. In this chapter, you learn exactly what big data is, and what makes it different from *data* as that term traditionally has been defined. In addition, expect to walk away from this chapter with a solid understanding of the technologies that power big data solutions and how those technologies operate.

Big data is different from regular *data*, as it's traditionally been defined.

Grasping Big Data Basics

Big data is data of such volume, velocity, and variety that it cannot be handled in traditional relational database management systems (RDBMSs). Big data requires a new technological approach for handling and processing, and these different technologies are required to support *big data.*

Data is data . . . yes. But, when you hear the word *data,* you often can assume safely that this term refers to traditional data that is handled and stored using old technologies like RMDBMs and data warehouses. Big data is different because it requires entirely new data platforms to meet storage, handling, and processing requirements. At its essence, the term *big data* refers to an entirely new data technology paradigm.

Before going into a deeper discussion of big data technologies, it's important to fully understand the nature of the data volume, velocity, and variety that makes working with big data such a challenge.

Not all three of these factors are required. The presence of any one of them can drive a need for big data technology.

- **Volume:** A big data deployment can be easily distinguished by its scale. The term *data volume* is relative in that no specific amount of data mandates big data versus conventional technologies. A good guideline is that if your organization owns at least 1 terabyte of data — and it almost certainly does — then it has adequate data volume to justify a big data deployment.
- **Velocity:** Big data is characterized by high-velocity data. That is, batch data, real-time data or streaming data that's entering an IT system at a rate of anywhere from 30 Kbps up to roughly 30 Gbps. Since new developments in innovation and instrumentation continually make data generation and capture tasks easier and more affordable, enterprises of all sizes are seeing staggering and continual increases in average velocities of data entering a system.
- **Variety:** Big data deployments are also characterized by *high-variety data* — data that's comprised of any combination of structured, semi-structured or unstructured data types:

- **Structured data** is data in traditional systems that contain ordered tables with distinct rows and columns.
- **Semi-structured data** is non-relational data that's marked with tags by which it can be organized into hierarchies.
- **Unstructured data** is characterized by its complete absence of structure. It's data that's being generated everywhere, at all times — a natural by-product of peoples' daily activities, from engagement with social media networks and RSS feeds, to the generation of office emails, Word documents, and PDF files.

Another defining feature of big data is its *value.* Big data is low-value in raw form. It must be processed, reduced, rolled up, boiled down, cleaned up, and otherwise aggregated before it is useful for generating high-value insights.

Defining Who Uses Big Data and Why

In a recent study performed by Accenture and General Electric, almost 85 percent of organizations that were polled declared that big data and analytics dramatically affect their industries, what's required to stay competitive in those industries, and what's needed to maintain investor confidence. Moreover, without the competitive advantage that big data deployments deliver, these organizations suspect they'll lose market share and competitive advantage on a price basis to more analytics-savvy competitors.

While big data technology is being deployed across all sectors in the economy, it's making the strongest impact on the following industries.

- **Financial services:** Financial services firms are using big data analytics and predictive trading algorithms for the competitive advantage these methods generate. As a result, these organizations help their clients make better-informed investment decisions and enjoy more consistent returns.

- **Telecommunications:** Telecommunications companies are using big data and analytics to combat fraud and to build a 360-degree customer view that allows them to increase customer satisfaction rates and decrease churn.
- **Energy:** Oil and gas companies have been using big data, analytics, and Internet of Things technology to reduce operational and financial risks while increasing revenues. For example, big data is used to predict equipment failure. This enables proactive maintenance and equipment replacements, thus increasing production rates, decreasing down-time incidents, and driving incremental revenue.
- **Government:** Across the public sector, big data technologies are allowing local and national governments to increase tax collection compliance by identifying and seeking rectification for tax fraud incidents. Governments are also using big data technology to optimize labor cost through employee and contractor productivity monitoring.
- **Retail:** Big data is making quite the splash in retail by optimizing supply chain management to help companies reduce operating expense and improve inventory velocity. Big data is also being used to increase revenues by extending customers increasingly personalized offers and promotions through electronic in-store couponing technologies.

Although organizations are finding it easier than ever to collect all types of data from many different sources, they are finding it more difficult to use and interpret their data to optimize business processes and outcomes. One main reason for this is the relative shortage of data scientists whose specialized skills are required to generate the desired insights necessary to transform the business.

Surveying Big Data Technologies

From the technology perspective, Hadoop almost single-handedly brought the power of big data to the enterprise by making large-scale computing sufficiently flexible and affordable so that any size organization can leverage this technology. Where once the terms *big data* and *Hadoop* were used synonymously, the growth of Hadoop has changed things.

The widespread adoption of Hadoop to solve the computational challenge of big data has launched a groundswell of new innovations and supporting technologies for manipulating, managing, and extracting value from big data. In this section, you learn what Hadoop is and how it works, and then you get an introduction to other meaningful technologies that are important in the big data space.

Getting a handle on Hadoop

The common thread that underlies many big data solutions on the market today is Apache Hadoop. As an open-source framework for distributed computing, Apache Hadoop deploys commodity servers across a connected Hadoop cluster and uses those servers for in-parallel big data storage and processing. Apache Hadoop is comprised of a Hadoop Distributed File System (HDFS) and MapReduce, shown in Figure 1-1.

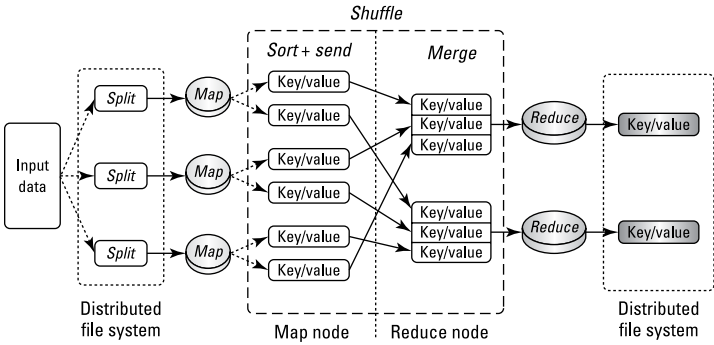

Figure 1-1: A schematic showing how HDFS and MapReduce work.

Looking into the HDFS

Hadoop's distributed file system (HDFS) facilitates scalable storage of massive amounts of big data on commodity servers distributed across a Hadoop cluster. Data files are broken into multiple blocks that are each replicated for redundancy and then stored across multiple servers in the cluster. Each server in the HDFS stores data from many different blocks, so data storage is distributed across several servers for built-in fault tolerance. The HDFS stores the input data for processing by MapReduce, and it stores output data after MapReduce has worked its magic.

Defining MapReduce

MapReduce is a batch-processing framework that performs distributed, parallel processing tasks on data that's stored in the HDFS. Its distributed processing tasks are carried out in-parallel, on specified servers that are distributed across the Hadoop cluster. MapReduce works by converting big data down into small sets of tuples that can then be organized and processed according to their key-value pairs. After converting the data down to tuples, the output is then stored back on the HDFS.

Big data technology can get really complicated. A very simple way to remember MapReduce is to think of it as a framework that processes and reduces raw big data into regular-size, tagged datasets that are much easier to work with.

Here is some terminology with which you should be familiar:

- **Map tasks:** The map task is where data is delegated to key-value pairs, transformed, filtered, and then assigned to nodes for reduce tasks. Map tasks are managed by Task Trackers and carried out on worker nodes.
- **Reduce tasks:** Reduce tasks are where data is reduced down to tuples of a key-value pair format. Reduce tasks are also managed by Task Trackers and carried out on worker nodes.

Relating Hadoop to other big data technologies

To make the best choice between technologies, you should have an eye on popular alternatives. The following is a brief description of some popular big data platforms and tools on the market:

- **Pig:** Pig is a procedural language for building big data applications. It is simpler for programmers and data scientists to use than writing MapReduce in Java. It generates MapReduce in the background.
- **Apache Hive:** Apache Hive is a data warehouse software that is useful for querying data from a Hadoop system using a language that is very similar to traditional SQL.

Hive can be used on huge volumes of data that sits in an HDFS, but it does not provide real-time data access and it also exhibits high latency.

- **Apache Sqoop:** Apache Sqoop is a tool that is used for transferring data between Hadoop and relational database management systems. This command-line interface application allows you to export data from Hadoop into structured data stores or to import data from structured systems directly into Hadoop.

- **In-memory databases:** In-memory computing offers low-latency, high-speed performance for use in generating real-time analytics from high-velocity streaming data. To be used in conjunction with, and not in place of, MapReduce batch processing, in-memory computing capabilities are just what you need if you are looking for quick insights from real-time streaming data sources.

  Apache Spark is a very popular platform for in-memory data processing and analysis, enabling real-time big data analytics. Apache Spark sits on top of the HDFS, but bypasses the MapReduce framework, thereby doing away with resource intensive disk accesses that MapReduce necessitates.

- **NoSQL databases:** Designed for storing and retrieving very large volumes of multi-variety data, NoSQL non-relational, distributed database systems offer a speed, scalability, and affordability that are unparalleled by traditional data management systems. NoSQL databases were designed to handle *big data* — in other words, large volumes of structured, semi-structured, and unstructured data. NoSQL databases come in four main flavors — column stores, document databases, key-value stores, and graph databases.

  Apache HBase is a popular type of NoSQL database. This column store runs on top of the HDFS and provides a good way for Hadoop to store sparse data. Although HBase is known for its consistency and scalability, it's not designed for use with traditional transactional systems, or for applications that require data analysis and aggregation across rows.

- **Massively Parallel Processing (MPP) platforms:** MPP platforms are scalable, high-performing, parallel processing alternatives to MapReduce. Unlike MapReduce, however, you can use MPP for the parallel processing of data sitting in a traditional data warehouse. You can even use SQL to query data from MPP systems.

  Because MPP databases deploy parallel processing on custom hardware, data ingestion and analysis speeds are incredibly fast. However, compared to MapReduce and its commodity servers, MPP systems are usually cost-prohibitive for large-scale projects.

Chapter 2

Managing Data Ingest and Data Pipeline Workflows

In This Chapter

- Getting oriented in the workflow world
- Seeing the commonalities and differences
- Taking stock of the tools
- Getting into data selection and ingestion
- Automating workflows for success

In the corporate world, big data is still a relatively new thing. Nonetheless, it's clearly revolutionizing business and its competitive landscape. Although big data is new to many, lots of big data features are common in the traditional data environment that most know so well. This chapter walks you through a detailed explanation of what big data workflows entail and the tools that are available to help manage them. You also learn some important best practices for data ingestion, data selection, and data workflow automation.

Walking Through Big Data Workflows

Workflows are nothing new. You may have heard them referred to as *batch, scheduling,* or *workload automation.* A *workflow* is a sequence of tasks that is performed by a computer. Workflows are managed by *workflow management tools* — tools that automate, execute, and manage the tasks that are defined by the workflow. Most workflows designate specific task sequences,

or specific conditions (either time and/or date conditions or a business event such as the completion of a transaction or the creation of a file) upon which the computer will be instructed to run (or halt) a series of processes.

Computers work in two major ways — transactional (or real-time) processing and batch processing:

- *Real-time processing,* as its name implies, involves performing computations or processing data that is arriving into the system in real-time. *Streaming* is another term used specifically to describe data constantly flowing into a computer system and being processed immediately. In real-time processing, computers interact with a human or another computer. In the case of human-computer interactions, the computer's work must be carried out while the person is waiting for a response.

  A good example of real-time processing is when you go to withdraw money using your ATM card. The banking system has to input your request, process it, and output to you a response almost instantaneously. A good example of streaming is in fraud detection where data representing a credit card transaction is "in flight" and the system must determine whether that transaction should be allowed or declined. This decision must be made while the transaction is pending and cannot be deferred to a later time.

- In *batch processing,* transactional data collects over a period of time. Later that data is processed as one unit, and then the output must be stored. Batch processing comprises the bulk of the work that's done by computers. Anything that isn't real-time processing is batch processing. Batch processing is not dependent on human-computer or computer-computer interactions to initiate a workflow. Batch processes work on their own, according to the instructions set forth in a workflow.

  In batch processing, separate programs define instructions for the data *input, processing,* and *output* phases. Payroll processing is an excellent example of batch processing at play in the real world. In this example, tasks in the input phase involve timecard completion and timecard data validation. In the processing phase, salaries are calculated, the payroll system is updated, and pay slips are produced. During the output phase, an electronic

interaction with the banking system executes the process for final payments. These processes are repeated every two weeks, in that particular order, for all employees working at an organization.

Understanding what comprises a big data workflow

Lots of new technologies and applications are present in the big data space, but the fundamentals of a big data deployment are not significantly different from those of a traditional data system. Because "big data," as its name implies, involves data, many of the same traditional processing and handling tasks are relevant to big data systems as well. The similarities are as follows.

- Both deploy extract, transform, and load processes.
- Both have input phases that deploy data processing on arrival.
- Both have processing phases that include data analysis tasks.
- Both have output phases that involve human consumption of data and the feeding of output data into new systems.

Despite their similarities, however, big data tools are quite different from those used in traditional data systems. These differences include:

- **Data lakes:** Big data systems often include *data lakes* — large flat data storage systems, each with its own processing engine. Data lakes are comparable to data warehouses in that these enterprise-wide data repositories both drive "conventional" analytics and decision-support systems. The frequently used distinction between a lake and warehouse is that the data in a warehouse is meticulously cleaned, extracted, and transformed, but because of technical and practical limitations, it represents only a small part of an organization's total data. A data lake, by comparison, has data in its raw (or nearly raw) state. A data lake includes big data technology that provides almost unlimited scale so that all of an organization's data can be stored in a single repository.

- **Apache Hive, Apache Sqoop, Apache Pig, and Apache Oozie:** These are data management tools that are included in the Hadoop ecosystem. Hive provides SQL-like access to the Hadoop Distributed File System (HDFS), Sqoop provides a way to import data into the HDFS environment, and Pig is a procedural language that gives programmers an alternative to developing MapReduce applications in Java (which is a complex task).

  Apache Oozie is a workflow tool that is intended to provide batch workflow management for the Hadoop ecosystem. However, Oozie is limited to supporting the technologies mentioned previously (Pig, Hive, Sqoop) and a few others. Any processing contained within big data workflows, other than the few supported natively by Oozie, have to be scripted, which is not a trivial process. All these technologies are relevant to Hadoop implementations, but have no place in the traditional data world.

- **New types of database structures:** One example of this is NoSQL — a non-relational, distributed database system that offers speed, scalability, and affordability for large volumes of multi-variety data.

- **Schema-on-read (not schema-on-write):** This attribute allows you to generate the schema you need when querying data for a specific task at hand, instead of having to work around a predetermined schema that was set in stone when the data was written into the database. The intended benefit is to enable new insights that are not "predetermined" or biased by data formats that are fixed in advance.

Traditional applications have a place among big data technologies. Traditional data is being input to big data systems from relational data structures, and also some of the data that is output from big data systems is being sent back into the traditional environment.

Managing big data workflows

When you are thinking about how you are going to manage a big data workflow, it is important to remember that these workflows are not comprised exclusively of big data technologies. Big data requires integration with new technologies and older, traditional technologies as well.

Since a big data deployment covers so many different technologies, one of the challenges with big data workflows is that you must design workflows that are capable of executing processes and tasks in a wide variety of tools and applications.

Batch processing is still critically important in big data systems because a vast majority of the work done by these systems is still being carried out through batch processing. Consequently, every new technology that arises in the big data space has been built to offer at least some functionality for batch processing. Think carefully about your big data workflow before designing it. Be prepared to encounter a number of different tools as well as a vast assortment of technology components. Plan the work carefully to avoid having too many tools in the system, and make sure the tools are synced and integrated.

Building too many tools into a process almost defeats the purpose of using tools at all. For each tool or application you design into your big data workflow, you must find a way to integrate, sync, and coordinate multiple tasks between it and the multitude of other tools you're using in the system. This process is complicated to build and manage. Most tools cannot give you the visibility and intelligence you need about what is happening in all other parts of the system.

Tool selection, integration, and management are long-standing challenges in the computing industry. Many people do lots of work figuring out what tools they want to use, only to discover that they have to do even more work figuring out how to use them, how to integrate them, and how to manage them. Once they have all of that figured out, they have to spend still more time and effort building the integration pieces they need to integrate the tool assortment. The explosion of newly available big data applications and tools has only accelerated the level of the challenges.

Now consider the notion of abstracting batch processing. Similar to using a high-level language for writing business applications, this language allows you to write the application without thinking about the type of computer on which it is going to be run. Apply the same concept to managing big data workflows. You know you are going to have a wide diversity of tools. You already have many tools in the traditional space — the complications have only increased with new big data tools.

Tools and workflow management are only going to continue getting more complex. Wouldn't it be great if you could find a level of abstraction where — no matter what applications you were running with the workflow management solution — you could seamlessly deal with them without doing lots of integration work on your own? Otherwise, you would have to learn a whole new set of methods just to manage the workflow instead of focusing on other important tasks.

When they talk about big data workflows, many people equate big data with Hadoop and they often talk about Oozie. Oozie is an Apache project and is generally considered a component to manage workflows in the Hadoop ecosystem. These days, the term Hadoop refers to more than the formal Apache Hadoop project, which consists only of YARN, MapReduce, and HDFS. The term usually refers to a collection or related applications that make up the Hadoop ecosystem. Each Hadoop vendor distributes a slightly different version of this ecosystem. Popular Hadoop vendors include:

- Cloudera, which adds Impala and Cloudera Manager
- Hortonworks, which adds Tez, Ambari, and most recently, Apache Ni-Fi relabeled as Hortonworks DataFlow
- MapR, which has a customized file system that is the heart of this differentiated solution

Because the relationships among tools within this expanding ecosystem is in a state of flux, the definition of *big data* can sometimes be a little complicated. To add to this complexity, many other tools and applications that are not directly related to Hadoop, such as NoSQL databases, other cluster managers, and many traditional data management tools, are also included in *big data.* You can see why the discussion about big data workflows can also get complex.

Taking a look at the tools and applications

For the moment, let's limit the discussion of batch processing and workflow management to the Hadoop ecosystem. Most practitioners are likely to be familiar with Oozie as a tool for managing Hadoop workflows. What immediately strikes a new user is that Oozie does not have a user interface for building workflows. Instead, XML documents are used to define a workflow. Additional configuration files provide parameters and arguments that instruct Oozie how to run the workflow. Because it is challenging for new users and inconvenient even

for seasoned technicians to manipulate XML, vendors have constructed layers on top of Oozie.

Cloudera has built a web application called HUE, which stands for Hadoop User Experience. Part of HUE is a job browser and editor. For Oozie users, it is probably the most common user interface for building Oozie workflows. Because HUE is from Cloudera, it was written for the Cloudera distribution. Other vendors, like Hortonworks, distribute HUE, but they have had to reengineer it to adapt it to their distribution.

Hortonworks has created a component called Ambari for managing a Hadoop environment. It has added a views capability into Ambari, and this feature allows users to view packages, files, definitions, instances, versions, and job status. Hortonworks has been working to deliver a jobs view, which is intended to replace HUE and to be a custom Hortonworks UI for Oozie.

MapR includes Oozie and HUE in its distribution, but it does not enable them. You have to do some additional post-installation configuration to enable them.

When you get Oozie, you just get a component to a workflow, but it does not have much of a UI. It relies on construction of XML files to define workflows. Even though XML is not the most difficult file format to deal with, it is not the easiest or most convenient either. For all these reasons, many people are opting for a different approach to Oozie.

Other components are available, like Apache Falcon, a feed management and data processing platform that uses Oozie as a scheduling engine and provides a UI to construct Oozie workflows. Some architects and developers use Falcon as a UI for Oozie, but people who do not use Oozie at all have a variety of different solutions. They can use third-party or commercial schedulers, cron, or some similar tool. Just about anything that can run a script works.

These tools simply run scripts that the user has built. This is how the DIY integration discussed earlier is achieved. These tools are not intended to manage workflows. Falcon's focus is on data lineage, Oozie is for workflow management but it is very difficult to use, and cron has no workflow ability whatsoever — it is a launch-and-forget mechanism, leaving you to do the real work for yourself.

Consider also some of the newer technologies. For example, Databricks Spark provides the JOBS component for managing Spark batch, but this facility is specific to Spark. As with other application-specific workflow management tools, JOBS does not know anything about what is happening in other components of the system.

Getting the lay of the land in Oozie

You can also do integration by way of scripts using Oozie because it has a script action. Oozie has a handful of things it can do natively. Oozie supports

- Apache Sqoop
- Apache Hive
- Apache Pig
- Apache HDFS
- File actions
- Shell actions

Basically, anything that is not supported natively is supported by a *shell action* — functionality that enables Oozie to run the scripts that you write for it.

Oozie is often referred to as the cron of the big data space. Just as cron offers only a very basic capability that is standard on every *nix box (any Linux or Unix), Oozie is very similar in that it is included in all Hadoop distributions but it offers only basic functionality. In fairness to Oozie, it provides a bit more functionality than cron provides.

Oozie is a basic tool that is not the easiest thing to use, even for big data technicians. It is definitely difficult to use for anyone who is not a hard-core technician. Oozie's complexity and lack of functionality force users to develop integration with other applications on their own. Integration is usually performed through shell scripting using bash or something similar. An unexpected and undesirable complication arising from integration development is the load it places on already strained data scientists and similarily limited resources.

One of the industry challenges is addressing a massive skills shortage of qualified big data practitioners. If you find someone

who is skilled in building the data science applications that you need, you don't want that person wasting time figuring out how to do scripting and integration with data engineering applications.

The tools discussed earlier in this chapter present the same problem. If you're looking for workflow management, these tools either offer nothing at all, or at best they offer something application-specific. In these cases, expect to have issues when trying to make a tool work with other tools.

Big data workflows, by definition, are comprised of a wide variety of tools. To build workflows, you must either do a significant amount of integration work on your own or use any number of diverse tools that have little in common. Neither of these is a satisfactory solution.

The current solution to this challenge is not much different from the solution in the traditional environments of decades past. Instead, you need a platform that gives you the best of both worlds — it should make building workflows easy, and it should offer you functionality for the easy integration of tools. It should enable you to integrate existing tools as well as the new big data tools that you will add to your system over time.

To easily accommodate new tools and the changes they bring to your workflows, your platform must be flexible and extensible. BMC Software has long-standing expertise in meeting these workflow management requirements, both in traditional and big data environments. That is why BMC is pleased to offer Control-M for use in managing workflows that include both traditional and big data technologies. This tool offers you the level of abstraction you need to make big data batch processing easy and hassle-free.

Ingesting and Processing Data in Big Data Systems

One of the major promises of big data is that it provides the ability to collect, in a single repository, all the data that an organization has, as opposed to traditional data management environments that either store data in siloes or store only a

relatively small subset of the entire body of data. Legitimate technology and financial constraints caused these conditions in the past, but the result is that most organizations are only using five to ten percent of their data to drive their analytics and decision-support systems. Making matters worse, data volumes are growing at exponential rates, making it harder for an organization to pull together all the data that it needs in order to generate deep data-driven insights.

Big data solutions offer you a way to pull all this data together. You can generate a 360-degree view of your customer (or patient, or product, or whatever other subject is of interest to your organization). The diversity and richness of big data sources in their original and near-raw state means that you can derive new insights that you may not have imagined were possible.

Different vendors refer to this data repository by different names. Whether you call it "enterprise data hub," "data lake," or "modern data architecture," Hadoop and big data solutions break the barriers of how much data you can store and process in a reasonably effective way. These solutions increase the limits on how you can scale your data environment and how you can continually increase the amount of data your organization can process. In this context, the notion of data ingestion becomes very important.

Even though Hadoop and other big data solutions provide an opportunity to increase your organization's data storage and processing capabilities, you still must be mindful of what kind of data you are bringing in and why. Now that you have so many options, you need to think carefully about all the factors that are relevant to data ingestion. Be sure to consider the following:

- What data do you want to bring into the system?
- At what speed will the data enter the system?
- How much total storage capacity do you expect the data to consume?
- What other datasets are also relevant to this data? What contextual datasets are required to make sense of the data?

- Is the data truly relevant to the business problem at hand? Be reasonable, but not overly restrictive. You may not know the answer to this question until you start your data exploration.
- What sort of data cleanup or transformation processes will be required after the data has been ingested?

Keep in mind that the data that is being ingested into big data systems is often not just new data. It may be traditional data that, until now, an organization has not been able to process.

Identifying relevant data sources and types

Big data systems are responsible for managing a rich, complex diversity of data. In general, within the big data space people think that they can store and process all the data in the world. Actually, there is still a price to pay. It may not make sense to store all your data in Hadoop databases.

Make sure to do sanity checks. Some organizations are paying $20,000 per node in a Hadoop cluster. Yes, that is much cheaper than if you increased to a Teradata environment or an HP Vertica environment, but it is not free! Take storage and processing costs into consideration.

One of the benefits of the HDFS is that it tolerates failure. It expects to be running on commodity hardware. It expects that some of that hardware is going to fail, so the data is replicated. By default, data is replicated three times in a cluster. Although you can change that replication factor up or down, when you think you are getting a cluster that has a total capacity of, for example, one petabyte, you really only have one-third of a petabyte of actual storage space. Even though you are using commodity hardware, storing data is still not free. It is cheaper than conventional enterprise-grade storage systems, but not free.

That is why you must identify what data you want to use to attack your business problem. This decision should be driven by your defined business use case. Use your business needs as your guide when selecting the data to bring into the system, not the other way around.

Here is an easy analogy to help you remember this guideline. You would not go out and get all the Twitter data in the world if you are not going to analyze Twitter sentiment, right?

Additionally, when selecting what data will bring the most value to your organization, do not forget traditional data as an important source. Big data is not limited to social media data, video data, image data, sensor data, and so on. Big data also includes traditional data that may have been dark data until recently. Big data can include archived data that was not previously accessible for analysis. This data is valuable, and now that it can be revived and analyzed, it makes a great data source in a big data environment.

Automating data workflows

In view of all this complexity, especially with respect to the data sources you are using, where they are coming from, and what tools you are running, obviously you need to automate your data workflows.

You should automate your big data workflows for two important reasons:

1. Data workflows involve lots of repetitive work which, when performed manually, is a waste of man-hours.
2. It is critical that the work being done by workflows is performed correctly. Performing that work manually greatly increases the chance of introducing human error.

As we mention earlier, make sure you are bringing in only the data that is relevant to the analysis at hand. Even more importantly, when you have identified that data, you must consider its processing requirements. For data to be stored in a state that is useful for analysis, a number of data cleanup processes are required. Data processing should be automated.

Before you begin your data analysis, check that all the data has arrived into the system and that cleanup processes are complete.

People think of analytics and data science as some sort of exercise where you are exploring data and trying to develop insightful algorithms. Although that is true, when you get an

algorithm or an analytics result that you like, you'll probably decide to repeat it on a regular basis. You need to be able to automate that analytic process, so that it does not waste the valuable time of data scientists. This is another classic batch processing workflow requirement.

After this big data processing and analysis is done, many organizations want to push the results back into traditional environments. This process can be automated with batch processing. Another process that requires automation includes setting limits on service levels or business requirements by date or time and then issuing notification if this does not occur. You need tools that enable you to automate all these processes.

After you have automated all the processes that you can automate in a big data environment, the rules that apply in traditional automation of batch processing begin to apply to big data automation. Some of these rules are:

- Processes must be set to happen on some sort of a regular basis. If the process does not run, then the right person or group must receive an automatic notification or email.
- If errors occur throughout a process, they must be captured for analysis. This information must be made available to the people who do the analysis, and the analysis should not require a huge knowledge of all the technology pieces in the big data system. Your workflow management tool should be able to manage these tasks correctly and execute this automatically.
- If people at your organization do not run analytics processes on a regular basis, you still want them to be able to run ad hoc analyses in a way that makes their data analysis easier. Your workflow management tool must be able to help these people run their analyses when they want, to check that each analysis ran successfully, and to manage analyses to whatever extent they want.

When this has become a part of the normal operational environment, you may also need to automate some of the technology requirements, such as performing audits, viewing responsible employees, making budgetary allocations, and authorizing chargebacks. These tasks require lots of automation management, and that applies to the big data environment as well.

To achieve the level of automation that big data projects require, you must realize that you are dealing with a set of processes that span many different types of technologies. The automation must be holistic and run as one whole piece. You do not want people working on these processes manually, and you definitely do not want people building their own integrations to make it work.

Seeing the possibilities of the Internet of Things

Just when you thought we couldn't get any more data than we already have, along comes the Internet of Things (IoT). We are beginning to reach a saturation point for people creating data as we move toward ubiquitous mobile phone and Internet access for the world's population. However, we are only starting to instrument devices like cars, refrigerators, stoves, utility meters, and thermostats, not to mention clothing and even humans themselves.

The volume expected from all these "things" is already being described as the next wave of data volume, velocity, and variety. It will make the current big data challenges trivial by comparison.

New tools are being introduced to help manage all this data. One that may be of particular interest is Apache Ni-Fi. This technology was built by the National Security Agency in 2006 and was shared with the Apache open source community in 2014. It has since been acquired by Hortonworks and released as Hortonworks DataFlow (HDF). It remains an open source technology.

HDF has great capabilities for capturing data from myriad devices, large and small, that may be deployed anywhere on the globe — even in hostile conditions.

However, like so many of the tools and applications we've discussed, HDF is limited to the specific functions related to collecting data. It cannot ensure downstream processing is initiated once the data is received, and it cannot have the data collection initiated on the basis of preprocessing that occurs. Although HDF can perform an important function, it ultimately adds to the complexity and the collection of tools that may be used within big data applications.

Chapter 3

Solving Business Problems with Big Data

In This Chapter

- Getting straight to the point of big data
- Seeing the benefits big data brings
- Identifying challenges along the way
- Exploring applications in the real world

In this chapter, you get a good look at exactly how big data is benefiting the bottom line of businesses in many different industries.

Identifying the Business Problem Solved

The problem solved by big data implementations is the same one that traditional data warehouse platforms address, but big data solutions are orders of magnitude more powerful and efficient. Both traditional and big data solutions aim to enhance decision making by supplying decision makers with the key information they need. That is called *data-informed decision making*.

Although the goal has always been to predict the future, big data systems are better at managing data volume and velocity. Consequently, they allow better, more accurate trend predictions from deeper sets of historical data. Real-time analytics, like those generated by Apache Spark, deliver fresh and current insights from streaming data sources, so organizational decision makers have the insights they need for timely, preemptive decision making. What is more, organizations now can

generate more reliable predictive models based on historical and real-time data. They can use those models to improve satisfaction rates, optimize expenditures, and increase overall revenues.

Before looking at some spectacular big data use cases, you should understand some of the pros and cons of going big with big data.

Weighing the Pros and Cons

Big data is a big investment. It is important to understand the benefits you can expect, as well as some of the challenges you will face along the way.

Browsing big data benefits

Big data analytics provide superior decision support — it is as simple as that. From a high-level perspective, big data analytics help key stakeholders make decisions based on real-time or predictive insights. This helps ensure that your organization remains competitive among its industry peers. But, from a lower-level perspective, take a look at how this result is achieved:

- **Gain a more comprehensive picture of the business and its customers:** Because big data systems are designed to integrate all varieties and volumes of data at a great velocity, a well-designed system provides the broadest view possible of your business.
- **Improve predictive models:** Big data systems store massive amounts of historical and current data — just what you need to generate accurate models to predict future performance. Although traditional systems are helpful in providing an idea of what has already happened and why, they struggle to keep pace with current conditions. Predictive models based on big data are designed to predict what will happen for your business based on what is happening now and what happened in the past.
- **Monitor for real-time events:** Using big data technology to connect and monitor systems has tremendous value, in and of itself. Early event detection is key in helping avert disasters. Because traditional data systems were not designed to support this level of real-time data ingestion and analysis, they are not feasible alternatives for real-time event monitoring.

Considering the challenges

- **Departmental barriers:** Although this challenge is not unique to big data projects, lack of consensus across departments and overall resistance to change often slow progress. To overcome this challenge, create a strategy that will help you elicit the support of management and support staff from across departments.
- **Security challenges:** Measures used to secure traditional data systems are inadequate in the face of incredibly dynamic and complex big data ecosystems. New security issues arise and must be addressed. Luckily for you, you learn all about those in Chapter 4.
- **Talent shortages:** Even the most user-friendly big data solutions are highly technical. Although some of your organization's existing staff are capable of working with big data, expect that most of them need extra training to get up to speed. More than likely, you need to take on some new hires, and those individuals probably will be hard to find and costly to retain.

Issues about data and staffing are extremely relevant if you want to create a well-thought plan for your big data project. Be sure to consider answers to the following questions:

- From where is your organization's data coming? Who owns that data? What data types will your system need to handle? How much data volume and velocity will your system need to accommodate?
- What skill sets are available through your organization's existing human capital? What skill sets do you need to effectively support a big data implementation? How will your organization acquire these skills? Will you make new hires? Will you train existing staff?

Bringing Big Data to Life in Business

The following five use cases demonstrate the powerful impact big data technologies are having on businesses' bottom line.

Reaching sales records with recommendation engines

For e-commerce retailers that have implemented a product recommendation engine, independent researchers have estimated that anywhere from 2 to 20 percent of online revenues are attributable to this system alone. These systems are among the more popular offshoots of today's big data technology. You can think of a recommender system as an engine that takes customer-specific information and sifts through mountains of big data to identify ideal products to recommend to that customer. When producing its recommendation, a recommender engine may consider any of the following:

- Previous purchases made by the customer
- Products that the customer has stored in his or her shopping cart, or alternatively, products that the customer previously viewed, or bookmarked as interesting
- General web browsing behavior, otherwise known as clickstream data
- Sentiment data from social media

Once the recommender has access to the data supporting these metrics, it sifts through a wide array of data and data systems to uncover similarities and trends and then makes a prediction about what other products the customer might also like to purchase. Although the exact effectiveness of any recommendation system is deployment-specific, the technology is considered to be of extremely high value in the retail industry.

Using predictive engines to improve supply predictions

To keep customers satisfied and stay profitable, fashion retailers have had to get exceptionally sophisticated with their demand forecasting methods. To avoid running out of stock on popular fashion products, and to avoid stock surplus, retailers need stable and reliable predictive engines.

Compared to traditional forecasting techniques, new big data-driven methods have been shown to produce a 50 percent increase in the quantity of items whose demand is accurately predicted. These new big data-backed predictive platforms incorporate many novel features, including machine learning algorithms, automated statistical modeling, traditional data sources (like historical data describing seasonal sales), and high volume and velocity near-real-time transactional data that reflects weekly sales.

Reducing fraud with big data analytics

New fraud detection methods combine the power of big data sources, machine learning, and analytics technologies to save banking institutions billions of dollars each year through real-time detection and alerts on suspicious banking activities. These systems integrate big data and traditional data sources. The big data mostly comes in the form of real-time streaming transactional data, including information on the transaction location, amount, payee/payor, social media data, and other relevant account information. The traditional data includes historical data from third-party sources to alert banks of the criminal and credit history issues of high-risk customers.

The systems integrate best-in-breed implementations of big data technologies. This includes Hadoop Distributed File System and MapReduce for batch processing, as well as Apache Spark or Apache Storm for real-time processing of streaming data. Lastly, (big) data science methods deployed by these systems include supervised and unsupervised machine learning methods, outlier detection, traditional rules-based analysis, text analysis, social network analysis, and many hybrids of these methods.

Increasing customer satisfaction with big data

Although telecommunications companies are no strangers to dealing with massive amounts of data, the data they are using and how they are using it has made all the recent difference. Using big data technologies to monitor and predict,

communications service providers have been able to produce a true 360-degree view of their customers and radically reduce customer churn rates by 50 percent. Newly redesigned, big data-driven churn management models incorporate data on all sorts of customer metrics, including:

- Usage patterns and network service quality
- A customer's account tenure and total spend
- Social media interconnectedness among customers
- A customer's overall social influence
- Clickstream data sources for customers

Results speak for themselves once again in the telecommunications industry, where service providers have deployed big data technologies to successfully increase customer satisfaction rates and predict future demand.

Saving billions with the IoT

IoT-created conservation practices are expected to save the utilities industry up to $200 billion by the year 2018 alone, according to a recent study by IDC Research Inc. The term *Internet of Things* (IoT) represents networks of interconnected data-producing devices that function as one system, according to instructions generated at IoT control centers. Although lean on predictive capabilities, IoT networks are designed for connectivity-based monitoring and operations optimization in the form of smart meters, grids, and intelligent assets.

Think of IoT as a big data product, because without connected big data sources, IoT would not exist. IoT data sources are high-velocity and high-variety, mostly comprised of vast quantities of machine-generated sensor data reporting machine pressures, temperatures, energy consumption, image data, video data, and traditional SCADA data, as well as a vast array of other streaming data sources. Real-time intelligence generated within IoT networks comes mostly via root-cause analysis, anomaly or outlier detection, and verification methods.

Chapter 4

Securing Big Data Systems

In This Chapter

- Understanding the evolving nature of Hadoop security
- Identifying weaknesses
- Shoring up a Hadoop system

The original purpose for Hadoop was large-scale public web data management — security concerns were not taken into account because they were not relevant for the project at hand. The first Hadoop distributions were not intended for enterprise-wide deployment. Although Hadoop technology continues to evolve, and security measures are becoming increasingly sophisticated, security concerns are still among the most widely reported obstacles to gaining the required authorizations for full-scale big data deployments across an organization. Simply put, many organizational leaders are not willing to assume the risk. In this chapter, you learn what the main security issues are, and what technologies are available to help you resolve those issues.

Spotting Security Concerns within the Hadoop Ecosystem

Although the original Hadoop distribution was fraught with security issues, the industry has made lots of progress. Before getting an overview of the solutions, take a look at the four main security obstacles.

The security issues implicit within Hadoop are real, but they should not be an impediment to success. Advances have been made to secure Hadoop's native weaknesses. You can remember the four main security challenges in Hadoop by referring to the four following concepts:

- Authentication
- Access
- Audits
- (Data) Protection

Authenticating users in Hadoop

When Hadoop was designed, developers Doug Cutting and Mike Caferalla assumed that Hadoop clusters would be used in a *trusted environment* — a secure area consisting of trusted machines and trusted users working in unison to achieve common goals. The original Hadoop release had no functionality for the authorization of users or services. Although a subsequent release added this functionality in the form of HDFS file permissions, this weak security measure was easily and commonly circumvented. In theory, in a truly trusted environment, issues of user impersonation would be no cause for concern.

Security must protect data both from intentional or malicious attacks as well as innocent errors. Both can be damaging. In practice, user impersonation can cause major detriment to trust between users, and consequently, to the project's overall well-being.

Controlling access levels within the system

Similar to its authentication deficiencies, Hadoop was not designed with functionality to limit user access to files in HDFS. With the original release, if a user had access to the HDFS, then he or she had access to the entire system. Hadoop offered no way to grant or restrict user access to only the portions of HDFS that were relevant to the user's task at hand.

Hadoop later released a distribution that allowed administrators to enforce file permissions to control user and user group access within HDFS, but this level of security is often inadequate for organizations with dynamic control policies. In efforts to overcome these security limitations, several solutions have emerged from a variety of sources. These are discussed later in this chapter.

Running audits on Hadoop user data

A third security problem with the original Hadoop distribution is that it had no native functionality for recording user access and system events. This makes it impossible to audit histories for data objects and access paths, thus preventing necessary tracking and record-keeping for who has been doing what, when, and where. Without this capability, organizations had no way to check historical data accesses and verify compliance with organizational control policies. With many users working inside a single system, this drawback could easily lead to a data management nightmare.

Protecting data that is moving and at rest

Because the original Hadoop distribution was designed for managing public web data, it had no need for encryption capabilities. Thus, data in the system was completely exposed, and that was perfectly okay. Anticipating that this deficiency would become an obvious problem, developers later released a version of HDFS that offered native data encryption capabilities for data at rest.

Still, the question of how to encrypt data that is moving within a system remained unresolved. Outside vendors have made progress by using network encryption methods to protect data that is in motion. Some of these include:

- MapReduce shuffles: SSL
- Java database connectivity clients (JDBC): SSL

- Network remote procedure calls (RPC): SASL
- HDFS protocol for data transfers

These network encryption methods are discussed later in this chapter, so keep reading!

Satisfying Security Requirements to Shore Up a Hadoop Solution

With respect to security, the Hadoop and big data community has made significant strides in a relatively short time — so much so that many organizations now feel sufficiently comfortable with Hadoop to implement it in their most sensitive environments. This section covers some of the more popular solutions for authenticating, controlling access, auditing, and protecting data in a big data system.

Authenticating systems and users

Kerberos, created several decades back by MIT, has been used for network security. Microsoft Windows security is based on Kerberos. Most agree that Kerberos is a robust, secure authentication mechanism. LDAP is a directory protocol that is widely used and has also been around for a long time. Many organizations rely on Microsoft Active Directory (AD) for user management — AD is based on LDAP. So these two protocols, Kerberos and LDAP, are broadly accepted in the industry and are close to being de facto standards. Early Hadoop implementations were not Kerberized, whereas today it's pretty much a standard feature of Hadoop products.

Although Kerberos is almost universally accepted, it is not simple to administer. Another complication is that each new application/technology that isn't directly based on Hadoop requires its own "interface" with Kerberos. So, for example, Cloudera Impala has had difficulties in the past working in a Kerberized cluster.

If your organization already has Active Directory Kerberos and LDAP authentication systems, you are in luck! You can use those to provide authentication capabilities for your big data system as well. If you do not have these authentication systems established, consider setting them up.

Controlling access to files and file parts

Establishing access control is the next line of enterprise-grade protection for big data. Designed specifically to secure the Hadoop cluster by controlling and limiting users' access rights, solutions like Apache Sentry or Apache Ranger allow administrators the ability to issue granular authorizations and to limit a user's access to specific servers, databases, tables, and views. These solutions also provide capabilities to assign user privileges based on a set of predefined roles.

Recording activity details for audits

Big data system administrators need user access and system event records and reports so that they can verify, track, and manage what is happening across the cluster. But because most users in the system should not be able to see all the activity details of the other users in the system, data masking becomes relevant when recording activity details for future audits. *Data masking* is an encryption method for hiding original data records so that they are not accessible to unauthorized users.

To better understand the concept, consider the following scenarios. One of the issues involved in building a rich data repository is that you may have complex "records" — any single record has data that should be visible to Person A, but not to Person B, and vice versa. You need to make sure that when Person A sees that record, the portion of the data that is restricted from that person is masked or removed. Similarly, if you want to generate test data, you may need to copy live data, but then mask some sensitive information so it is not visible to unauthorized users. These are two additional instances where data masking is important for protecting data in a system.

Encrypting data for better protection

The issue of how to encrypt data is another security concern that is key to shoring up Hadoop so that it is enterprise ready. As discussed earlier, there is data-at-rest and data-in-motion, each of which requires different security strategies.

Hadoop distribution vendors either already have encryption as a built-in function of HDFS or are aggressively moving toward providing it. By initially providing a "forked" or modified version of HDFS, MapR claimed built-in, high-speed encryption as one of its major differentiators. Hortonworks and Cloudera have responded with acquisitions of commercial solutions, which they are now incorporating into their offerings. Although enhanced and robust security is great, the different functions and capabilities mentioned in this chapter introduce administrative complexity to data security systems. The Hadoop community is moving to address this issue as well.

Chapter 5

Hiring and Getting Hired in the Big Data Space

In This Chapter

- Spotting who does what
- Selecting the best candidate
- Getting noticed in big data
- Preparing for your interviews

Now that you know what big data is, how it works, and the value it promises to deliver, it is time to see some insider tricks for hiring and getting hired in the big data space.

Peeking at Popular Job Titles

Before looking at the popular titles for positions in big data, you need to understand the difference between the data engineering and data science roles that comprise the field. Job titles that include the words "big data" are referring to *data engineering* positions — the design, building, and maintenance of big data systems that ingest, process, and store big data. Almost all positions with "big data" in the title are data engineering roles. *Data scientists* are also important in the big data space, but they have different responsibilities — they derive and communicate insights that create value from data. Without the work of data scientists, big data has very low value in its raw form. Popular job titles in the big data space include:

- Big data engineer
- Big data architect
- Big data software engineer

- Big data consultant
- Big data analyst
- Data scientist

Almost all jobs labeled as "big data" have minimum educational requirements that specify a BS or MS degree in computer science, math, or a related field. They almost all require substantial experience with the Apache big data stack (HDFS, Hive, HBase, Kafka, Pig, Oozie, and YARN). They also ask for shell scripting experience in either Unix shell or PowerShell. Most jobs require experience in design, implementation, and management of data ingestion into big data systems. They almost all require a solid knowledge of, and experience with, both SQL and NoSQL databases. Lots of these positions are requesting skills in Python or R scripting for use with Apache Spark.

To put this in perspective, jobs for "data scientists" generally request applicants to have a BS or MS degree in any quantitative discipline. These jobs require that applicants have experience using traditional relational databases, as well as coding skills in Python or R. Applicants should have solid experience in statistical analysis, and in many cases, a proficiency in machine learning. Job listings for data scientists call for good communicators who are personable and have a friendly personality.

Choosing the Best Candidate for the Job

When it comes to hiring in big data, there is good news and bad news. The bad news is that you will not easily find people who have the skills you need, and if you do, they will almost certainly be working for someone else. The good news is that you will have less work to do narrowing your options. After HR has screened your CVs and lined up interviews, your job is to personally evaluate the candidates' personality, drive, skills, and experience. Here is what to look for:

- **Drive:** The best big data engineers are the ones who are determined, with a high degree of initiative. Look for candidates who are self-taught; they have the personality to make things happen in spite of challenges.

- **Personality:** For a better workplace environment, look for someone who is serious and passionate about big data, but also excited to share that passion with others. These people are often involved in pro bono work, perhaps supporting an open source project, or volunteering on data for development projects. If you see this sort of experience on a CV, take note.
- **Skills and experience:** Remember that you are dealing with a major skills shortage when hiring big data talent. You should hire people who have at least one or two years working with tools in the Hadoop ecosystem. So, what do you do when you find a candidate with skills, but no experience in your industry? If you are hiring for big data engineering, do not let a lack of industry experience hold you back from hiring a candidate. Although industry-specific knowledge is important when hiring a data scientist, data engineers can switch between industries rather easily. If you see that a candidate has the relevant experience, get it while you can!

Getting Noticed as a Big Data Candidate

If you have read this far, then you have probably surmised some of the things you can do to increase your chances of getting hired as a big data professional. Just in case, this section is going to spell it out for you. You want to demonstrate that you have a driven and committed personality. If some of your big data skills are self-taught, do not hide that fact. It shows you have tenacity and are passionate about what you are doing. Similarly, if you have built some public projects or contributed to any big data user groups, do not let that go unnoticed. These show dedication, and also serve as a positive indicator of your personality.

Formal certifications and trainings are also valuable. Although self-taught skills are the earmark of a great candidate, do not underestimate the weight that proper certifications and credentials can have. Most Hadoop distributors offer online and in-person training and certification testing. For example, Cloudera has an excellent suite of courses and certifications that are focused on both the engineering and analytics aspects

of big data. (http://www.cloudera.com/content/www/en-us/training/certification.html):

- CCP: Data Engineer/CCP: Data Scientist (separate tracks in the Cloudera Certified Professional Program)
- Cloudera Certified Developer for Apache Hadoop (CCDH)/Cloudera Certified Administrator for Apache Hadoop (CCAH)/Cloudera Certified Specialist in Apache HBase (CCSHB)

Preparing for Interviews as a Big Data Professional

You have an interview all lined up and you are excited to go in there and knock their socks off, right? Here are some tips on how to do that. First, make sure you have covered the basics of Hadoop. They are certain to ask you about some elementary topics such as:

- Name the components of the Hadoop project and explain how each component functions.
- Describe the best hardware configuration for running Hadoop jobs, and explain any relevant nuances.
- What is the default replication factor for HDFS, and what is likely to happen if the replication factor is set to 1?

Next, get specific. Write out a list of the technologies required for the position and make sure you are prepared for any technology-specific question they throw at you. To see what questions to expect, take a look at the specific interview questions put together by DeZyre (http://www.dezyre.com/article/-top-100-hadoop-interview-questions-and-answers-2015/159). They will help you get prepared for technical questions on MapReduce, Hive, HBase, Pig, YARN, Flume, Sqoop, HDFS, ZooKeeper, and more.

If your interviewers throw you some questions from left field, do not let that get you off base. They may be testing your candor, or they may not know much about the technologies themselves. No one can reasonably be expected to be a master of all. Just show off what you know and be honest about your limits.

Chapter 6

Tips for Big Data Enthusiasts

In This Chapter

- Building your own big data library
- Convening on the best conferences
- Staying current in the big data blogosphere
- Following thought leaders

This chapter features tips about what books and blogs to read, what conferences to attend, and what thought leaders to follow if you want to stay up to speed with what is happening across the big data industry.

Knowing What Books to Read

Whether you are a business manager, a CIO, a data engineer, or a data scientist, one of the following four books is sure to pique your interest:

- ***Big Data: A Revolution That Will Transform How We Live, Work, and Think*** **(**`http://www.amazon.com/Big-Data-Revolution-Transform-Think/dp/0544227751`**):** This book provides a high-level discussion about big data and the big data phenomenon. It is well-suited for managers who want to understand how they can use big data sources to improve their business, but not for technical professionals who are already

knee-deep in data analysis. While this book only briefly discusses Hadoop and MapReduce, it is chock full of interesting big data business use cases.

- ***Hadoop: The Definitive Guide*** **(**`http://www.amazon.com/Hadoop-Definitive-Guide-Tom-White/dp/1491901632`**):** On the other end of the spectrum, *Hadoop: The Definitive Guide* is highly technical and not meant for the faint of heart. If you are a programmer and you want to learn all about the technical details involved in deploying Hadoop, this book is a great choice for you. If you are interested in learning the basics about other relevant technologies like YARN, Parquet, Flume, Crunch, and Spark, this book has what you need.
- ***The Signal and the Noise: Why So Many Predictions Fail — but Some Don't*** **(**`http://www.amazon.com/Signal-Noise-Many-Predictions-Fail-but/dp/0143125087/`**):** If you want to learn about how to use data for predictive forecasting, and all the nuances of forecasting for different types of real-life patterns, this book has you covered. This book is written for audiences of almost all skill levels, but if you are a highly skilled statistician or data scientist, the content in this book may be a little below your expectation.

Choosing Conferences to Attend

Whether you are looking to learn, network, or identify potential candidates in support of your own big data project, the following big data conferences have you covered:

- **Strata + Hadoop World (**`http://conferences.oreilly.com/strata/hadoop-big-data-ca`**):** *Strata + Hadoop World* truly is the world's premiere big data conference. Expect to see thousands of data people and organizations of all shapes and sizes, showing off and networking about all things data — big and small. Hiring for data talent? Take a look at the conference attendees. One word of warning, though — if you show up at this conference in a suit, you are going to be very out of place. Check out some past event coverage here: `http://www.youtube.com/user/OreillyMedia/playlists`.

- **Hadoop Summit (**`http://hadoopsummit.org/`**):** After Strata + Hadoop World, Hadoop Summit is the most popular big data conference in the world. Whether you are from the eastern or the western hemisphere, Hadoop Summit has something for everyone — with its April conference held in Dublin, Ireland, and its June conference held in San Jose, California. If you are looking for a venue where you can get up to date on the most cutting-edge and technical topics related to Hadoop, Hadoop Summit is the conference for you
- **Data Summit (**`http://www.dbta.com/DataSummit/2016/`**):** *Data Summit* is the business person's big data conference. It mostly features big data innovation as it relates to business and business management. If you are looking to network with other business-minded professionals who are serious about making big data work for them, this is the place for you.

Keeping Up with the Blogosphere

Blogs are another terrific way to stay on the pulse of what is happening within the big data industry. The following four websites are a great place to start:

- **Dataconomy (**`http://dataconomy.com/`**):** *Dataconomy* is focused on providing the latest news and expert opinions from the big data and data science sector. This content targets a more generalist audience, and topics featured include big data, data science, IoT, and financial technology.
- **O'Reilly Radar (**`http://radar.oreilly.com/data`**):** *O'Reilly Radar* data blog is the premiere place to stay up to date on the latest highly technical news from the big data industry.
- **Gigaom (Data Channel) (**`https://gigaom.com/channel/data/`**):** *Gigaom* delivers industry news about the most disruptive companies and people from the big data industry.

- **SiliconANGLE (Big Data Channel) (**`http://siliconangle.com/big-data/`**):** If you want all the latest on companies and people that are working in big data in the area "where computer science intersects social science," then the *SiliconANGLE* blog is the place for you.

Following Thought Leaders

Several names have been appearing over and over again in the arena of big data thought leaders. If you are not active on social media, now is as good a time as any to get started. And when you do, consider following these four recognized big data experts:

- **GregoryPiatetsky (@KDNuggets** `http://twitter.com/kdnuggets`**):** Piatesky's academic background is in computer science and he has been a consultant and entrepreneur in the data field for 18 years.
- **Vincent Granville (@analyticbridge** `http://twitter.com/analyticbridge`**):** Follow Granville, and consider joining his data community, *Data Science Central,* for the latest news on happenings, events, and networking opportunities in the big data space.
- **Kirk Borne (@KirkDBorne** `http://twitter.com/KirkDBorne`**):** Borne has an academic background in astrophysics and spent most of his career teaching in that field. He is a major promoter of data literacy and has more than 50,000 data-enthused followers on Twitter.
- **Bernard Marr (@BernardMarr** `http://twitter.com/BernardMarr`**):** Marr is a highly influential author, speaker, consultant, and thought leader in the big data space. With more than 85,000 Twitter followers, he must be doing something right.

bmc

BMC Control-M

Simplify & Automate Big Data Workflows

—

Bring IT to Life at bmc.com/hadoop

Leading organizations implement big data workflow automation to stay ahead

The big data era is fully underway. Leading organizations are quick to get ahead by finding ways to automate big data workflows. This book guides you through the details of what is involved in big data workflow automation. Don't miss out on its helpful tips for big data technology selection, security, and choosing the best candidates for your big data requirements.

- ***Learn about the technologies that comprise the big data landscape, and which ones are most appropriate for your organizational needs.***
- ***See how big data and data science are dramatically affecting businesses' bottom line.***
- ***Understand what parts of big data technologies can be automated to decide on the most efficient ways for your organization to proceed.***
- ***Identify the best conferences, books, blogs, and thought leaders with which to engage to keep informed on progress in the fast-moving big data space.***

Open the book and find:

- **Concise overviews on the most appropriate technologies for your big data requirements**
- **How to optimize organizational resources by automating big data workflows**
- **What is needed to ensure the security of your big data system**
- **Tips and tricks on places to go and things to look for when making big data hires**

Go to Dummies.com®
for more

Joe Goldberg is a solutions marketing consultant at BMC with more than 30 years of technical marketing and management experience. He specializes in large systems architecture, systems management software, and enterprise solutions. **Lillian Pierson, PE,** is a leading expert in big data and analytics. She's authored three technical books by John Wiley & Sons. Through Data-Mania, she offers online and face-to-face training courses in big data, analytics, and data science.

FOR DUMMIES®
A Wiley Brand

Also available as an e-book

ISBN: 978-1-119-25904-6
Not for resale

WILEY END USER LICENSE AGREEMENT

Go to www.wiley.com/go/eula to access Wiley’s ebook EULA.