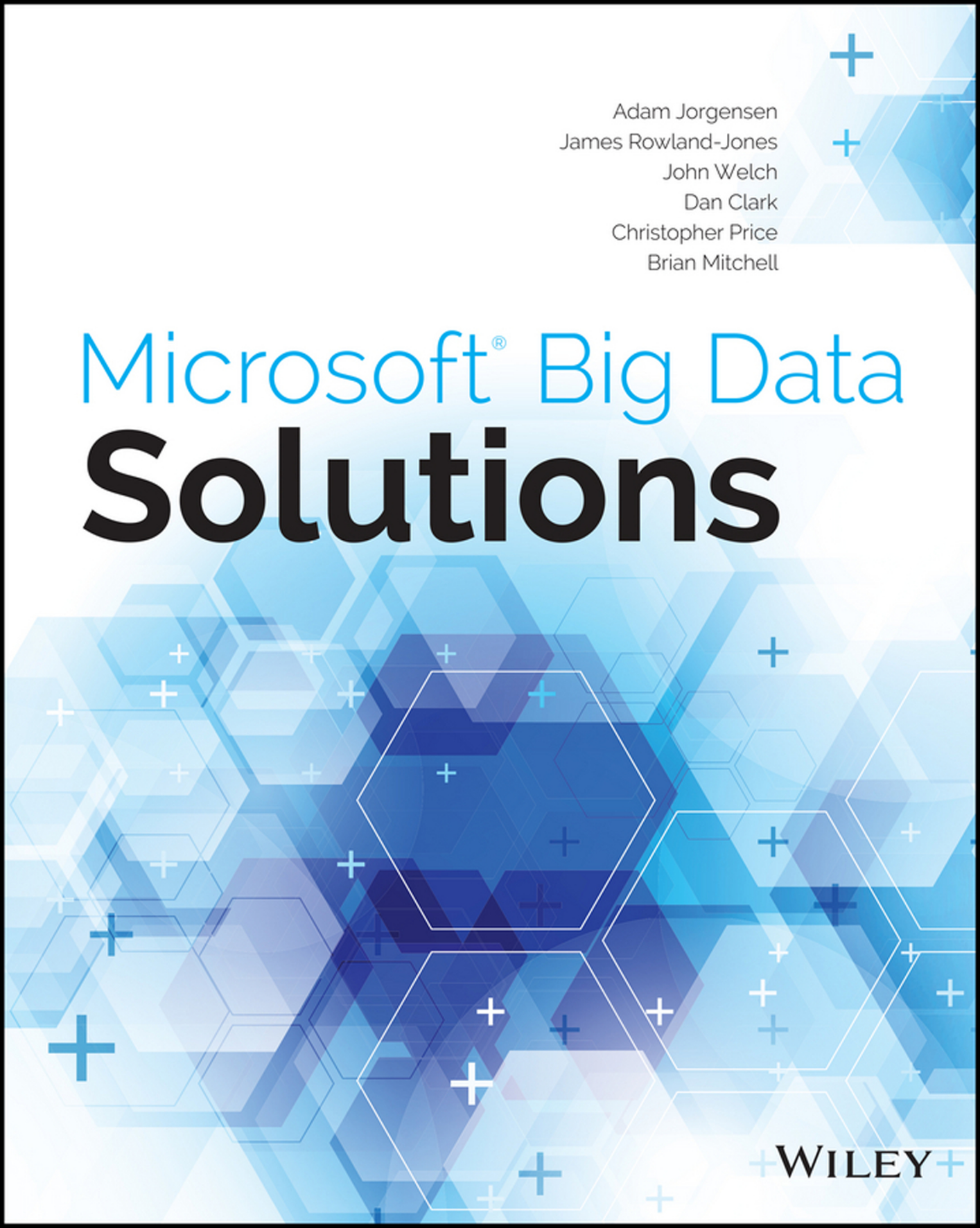
Adam Jorgensen
James Rowland-Jones
John Welch
Dan Clark
Christopher Price
Brian Mitchell

Microsoft® Big Data Solutions



WILEY

Microsoft® Big Data Solutions

Adam Jorgensen
James Rowland-Jones
John Welch
Dan Clark
Christopher Price
Brian Mitchell

WILEY

Microsoft® Big Data Solutions

Published by
John Wiley & Sons, Inc.
10475 Crosspoint Boulevard
Indianapolis, IN 46256
www.wiley.com

ISBN: 978-1-118-72908-3
ISBN: 978-1-118-74209-9 (ebk)
ISBN: 978-1-118-72955-7 (ebk)

Manufactured in the United States of America

10 9 8 7 6 5 4 3 2 1

Library of Congress Control Number: 2013958290

Executive Editor
Robert Elliot

Project Editor
Jennifer Lynn

Technical Editors
Rohit Bakhshi
John Hoang
Josh Luedeman

Production Editor
Christine Mugnolo

Copy Editor
Keith Cline

Editorial Manager
Mary Beth Wakefield

Freelancer Editorial Manager
Rosemarie Graham

Associate Director of Marketing
David Mayhew

Marketing Manager
Ashley Zurcher

Business Manager
Amy Knies

Vice President and Executive Group Publisher
Richard Swadley

Associate Publisher
Jim Minatel

Project Coordinator, Cover
Todd Klemme

Proofreader
Sarah Kaikini, Word One New York

Indexer
Robert Swanson

Cover Image
©traffic_analyzer/iStockphoto.com

Cover Designer
Ryan Sneed/Wiley

I am honored to dedicate this book to my author team who pulled together and created a wonderful project for the community they love as I do.

— Adam Jorgensen

For my beautiful and eternally patient wife, Jane, and our three children Lucy, Kate, and Oliver. I will love you all forever.

— James Rowland-Jones

To my lovely wife, Marlana, and my children, Kayla and Michael, thanks for the support and understanding during the late nights while I was writing.

— John Welch

To my family, thank you for your unconditional support throughout this process. I'd especially like to thank my wife Shannon for believing in me.

— Brian Mitchell

Acknowledgments

I would like to thank my Lord and Savior Jesus Christ for all He has blessed me with. Thank you to my wife and family for their support and love. This author team has been incredible and has responded to constantly changing platforms and market factors to deliver a great title on this fast changing subject area. Special thanks to the tech editors and vendor editors from Hortonworks and Microsoft. Last, but certainly not least, thank you to the readers and those data professionals whose passion makes a book like this worthwhile. It's for you we do what we do!

— Adam Jorgensen

I have come to know many people through the SQL community. Some I know professionally while others I know a bit better than that and consider friends. Adam definitely falls into the latter category. I'd like to start by thanking him not only for the opportunity to collaborate on this book, but also for his friendship, especially as we've journeyed on the PASS rollercoaster together.

I'd like to thank my copy editor Keith Cline and my official technical reviewers, Josh Luedemen (InfoTech), Michael Reed (PragmaticWorks), and John Hoang (Microsoft Azure CAT). John's help and support in particular has been invaluable to me for several years: I'd like to take this moment to give him a special mention and express my sincere thanks.

I'd also like to thank our oft-suffering editor Jennifer Lynn (www.pageoneediting.com). I know I didn't make your life very easy on this one, Jennifer, and would like to take this moment to both thank you and apologize for my seemingly never-ending list of excuses.

Last but by no means least, I'd like to thank my unofficial reviewer who gave up her time to offer me her feedback, insights, and sagely advice. Lara Rubbelke, you are nothing if not thought provoking and inspirational. Thank you so much for taking the time to help me shape my thoughts, bounce ideas, and for being my naïve friend.

—James Rowland-Jones

About the Authors

Adam Jorgensen is the president of Pragmatic Works and the executive vice president of the Professional Association for SQL Server (PASS). He has gained extensive experience with SQL Server, SharePoint, and analytics over the past 13 years. His primary focus is helping organizations and executives drive value through new technology solutions, management techniques, and financial optimization. He specializes in the areas of cloud and big data analytics and works on solutions to make those technologies real for enterprises. He lives in Jacksonville, Florida, with his wife, Cristina.

James Rowland-Jones is a principal consultant for The Big Bang Data Company. His focus and passion is to architect and deliver highly scalable, analytical platforms that are creative, simple, and elegant in their design. James specializes in big data warehouse solutions that leverage both SQL Server PDW and Hadoop ecosystems. James is a keen advocate for the SQL Server community, both internationally and in the United Kingdom. He currently serves on the board of directors for PASS and sits on the organizing committee for SQLBits (Europe's largest event for the Microsoft Data Platform). James has been awarded Microsoft's MVP accreditation since 2008 for his services to the community.

John Welch works at Pragmatic Works, where he manages the development of a suite of BI products that make developing, managing, and documenting BI solutions easier. John has been working with BI and data warehousing technologies since 2001, with a focus on Microsoft products in heterogeneous environments. He is a Microsoft Most Valued Professional (MVP), an award given due to his commitment to sharing his knowledge with the IT community, and an SSAS Maestro. John is an experienced speaker, having given presentations

at PASS conferences, the Microsoft Business Intelligence conference, Software Development West (SD West), Software Management Conference (ASM/SM), and others. He has also contributed to multiple books on SQL Server, including *Smart Business Intelligence Solutions with Microsoft SQL Server 2008* (Microsoft Press, 2009) and the *SQL Server MVP Deep Dives* (Manning Publications) series.

John writes a blog on BI and SQL Server Information Services (SSIS) topics at `http://agilebi.com/jwelch`. He is active in open source projects that help ease the development process for Microsoft BI developers, including ssisUnit (`http://ssisunit.codeplex.com`), a unit testing framework for SSIS.

Dan Clark is a senior BI consultant for Pragmatic Works. He enjoys learning new BI technologies and training others how to best implement the technology. Dan is particularly interested in how to use data to drive better decision making. Dan has published several books and numerous articles on .NET programming and BI development. He is a regular speaker at various developer/BI conferences and user group meetings, and enjoys interacting with the Microsoft developer and database communities.

Chris Price is a senior consultant with Microsoft based out of Tampa, Florida. He has a Bachelor of Science degree in management information systems and a Master of Business Administration degree, both from the University of South Florida. He began his career as a developer, programming with everything from Visual Basic and Java to both VB.Net and C# as he worked his way into a software architect role before being bitten by the BI bug. Although he is still passionate about software development, his current focus is on ETL (extract, transform, and load), Data integration, data quality, MDM (master data management), SSAS (SQL Server Analysis Server), SharePoint, and all things big data.

He regularly speaks at SQL Saturdays, PASS Summit, conferences, code camps, and other community events. He blogs frequently and has also authored multiple books and whitepapers and has served as technical editor for a range of BI and big data topics. You can follow Chris on his blog at `http://bluewatersql.wordpress.com/` or on Twitter at `@BluewaterSQL`.

Brian Mitchell is the lead architect of the Microsoft Big Data Center of Expertise. Brian focuses exclusively on data warehouse/business intelligence (DW/BI) solutions, with the majority of his time focusing on SQL Server Parallel Data Warehouse (PDW) and HDInsight. He has spent more than 15 years working with Microsoft SQL Server and Microsoft Business Intelligence. Brian is a Microsoft Certified Master–SQL Server 2008. You can find his blog on topics such as Big Data, SQL Server Parallel Data Warehouse, and Microsoft Business Intelligence at `http://brianwmitchell.com`. Brian earned his Master of Business Administration degree from the University of Florida. When he is not tinkering with SQL Server or Hadoop, Brian enjoys spending time exploring his adopted home state of Florida with his wife, Shannon, and their kids.

About the Technical Editors

Rohit Bakhshi is a product manager at Hortonworks, a leading provider of support and services for Apache Hadoop. Hortonworks builds and distributes the Hortonworks Data Platform (HDP), which is a 100% open source data management software powered by Hadoop available on Windows and Linux OS platforms.

Rohit is responsible for the HDP for the Windows product line, core Apache Hadoop components, and Platform Services for HDP. He has worked with Microsoft to bring the entire stack of Apache Hadoop components to Windows to enable Windows developers and system administrators to harness the full power of Apache Hadoop. Before Hortonworks, Rohit was a consultant in the Accenture Technology Labs R & D consulting group, where he focused on architecting and delivering big data solutions to Fortune 500 clients.

John Hoang is a senior program manager based out of Aliso Viejo, California, on the Azure Customer Advisory Team (AzureCAT). He has more than 20 years of experience working in various roles, including developer, business analyst, and project manager implementing software solutions to manufacturing, retail, and healthcare. He currently specializes in the SQL Server PDW. In his free time, John enjoys bike riding, tennis, and spending time with his two children.

Josh Luedeman has been working with SQL Server for more than eight years. He is currently a solutions architect with Data Structures, Inc., where he is working with customers to help them utilize business intelligence (BI) tools and big data. He has worked in IT for more than 10 years, holding positions in application support, database administration, and BI. In these industries, Josh has held integral roles in Fortune 500 companies, major institutions of higher education, small-medium businesses, and startups. Josh is a speaker at software development and data conferences including Code On The Beach and multiple SQL Saturdays. He is originally from Corning, New York, and currently resides in Orlando, Florida, with his wife and children. Josh can be found online at `www.joshluedeman.com`, `josh@joshluedeman.com`, `www.linkedin.com/in/joshluedeman`, and `@joshluedeman` on Twitter.

Michael Reed has a long history of designing innovative solutions to difficult business problems. During the last 14 years, he focused on database development and architecture, and more recently business intelligence and analytics. He is currently employed by Pragmatic Works as a Senor BI Consultant. Previously he was director of Insight and Analytics at a healthcare claim processor. Prior to that he held operations, data, and information delivery centric roles in Microsoft's Online Services Division; specifically the AdCenter Behavioral Targeting group, which is the primary research unit for mining social behaviors at Microsoft supporting the Bing decision search engine and BingAds advertising services.

In a prior life, he was co-owner of a multimillion dollar manufacturing business, grown from a startup, where he gained much of the business knowledge and insight he employs in his work today.

Contents

Introduction

This book was built for those of you who are searching. Those of you who are wondering. Searching and wondering what on earth big data will mean for your data world. IT takes a different approach, however, than the litany of titles designed to spend hundreds of pages beating you over the head telling you that you need big data, that everyone is doing it, and that you have to be "cool," too!

This author team wanted to create something that would be your go-to resource for moving from your existing relational world and provide you not only the roadmap forward but also practical experience for those of you who don't need the click here, move the mouse to the left, and click again level of instruction. We do explain some things in greater detail, but these are things that require this due to their newness or relative complexity.

We are focused on making sure you can ease your transition to using these tools and technologies because we have been where you are. Your boss came back from a conference and said, "We need a big data solution." When you inquire what he would like it to solve, he doesn't really know, but he knows how critical it is that the organization have one. You will become the responsible party for making these big data dreams come true.

Normally, this would entail training classes and long hours combing the Internet like you did when they told you they needed a data warehouse or a cube, those other words once foreign to you. You will learn through this text that big data is really big—no pun intended. It can do big things, solve big problems, and is a big ecosystem of tools and platforms. However, like most other ecosystems (RDBMSs, programming languages, mobile, and cloud), there are really only a few foundational things, and if you can come up to speed on those, you will be rocking and rolling when you need to apply more advanced tools, or automation, and so on.

Our Team

We have assembled a strong international team of authors to make sure that we can provide a sound perspective and knowledge transfer on the right topics (we'll discuss those shortly). Those topics include:

1. Accelerated overview of Big Data, Hadoop, NoSQL, and key industry knowledge
2. Key problems people are trying to solve and how to identify them
3. Delivering big data in a Microsoft world
4. Tool and platform choice
5. Installation, configuration, and exploration
6. Storing and managing big data
7. Working with, adding structure, and cleansing your data
8. Big data and SQL Server together
9. Analytics in the big data world
10. How this works in the cloud
11. Case studies and real world applications
12. Moving your organization forward in this new world

This team includes members of Pragmatic Works, a global leader in information services, software, and training; Microsoft Research; Microsoft Consulting Services; Azure Customer Advisory Team; and some other industry firms making a big impact in this expanding space.

All Kidding Aside

Big data is coming on strong. You will have these solutions in your environment within 24 months, and you should be prepared. This book is designed to help you make the transition with practical skills from a relational to a more "evolved" view of the data worlds. This includes solutions that will handle data that does not fit nicely into a tabular structure, but is nonetheless just as or more important in some cases as the data that you have curated so carefully for so many years.

You will learn some new terms as well. This will be almost as much a vocabulary lesson as a technical lesson.

Who Is This Book For?

This book is for those data developers, power users, and executives looking to understand how these big data technologies will impact their world and how to properly approach solutions in this new ecosystem. Readers will need a basic understanding of data systems and a passion for learning new technologies and techniques. Some experience with developing database or application solutions will be helpful in some advanced topic areas.

What You Need to Use This Book

We have designed this book to make extensive use of cloud resources so, as the reader, you will need to have a newer model computer PC or Mac that can access the Internet reliably. In addition, you will want to be able to install additional programs and tools as advised by the authors, so please ensure you have that access on the machine you're using. Different chapters will have different tools or data sets, so please follow the authors' instructions in those chapters to get the most out of your experience. Having access to a SQL Server database will be required in certain chapters, and if you wish to set up your environment on premise, then a virtualization technology such as Hyper-V, VMWare, or Virtual box is recommended.

Chapter Overview

Now we'll go through the chapters in this text and discuss what you'll be learning from each one.

- **Chapter 1: Industry Needs and Solutions**
  No book on big data would be complete without some coverage of the history, origins, and use cases in this ecosystem. We also need to discuss the industry players and platforms that are in scope for the book. Other books spend 5 to 6 chapters rehashing this information; we have done it efficiently for you so you can get to work on more fun topics!
- **Chapter 2: Microsoft's Approach to Big Data**
  Doing this in a Microsoft world is a little different that the traditional UNIX or Linux deployment. We chose this approach since we feel it makes this technology more accessible to millions of windows administrators, developers and power users. Many of the folks were surveyed before this writing, we heard overwhelmingly that we needed a Windows-focused solution to help the largest population of enterprise users access this new technology.

- **Chapter 3: Installing HDInsight**
  In this chapter, you'll get started configuring your big data environment.
- **Chapter 4: HDFS, Hive, HBase and HCatalog**
  These are some key data and metadata technologies. We'll make sure you understand when to use each one and how to get the most out of them.
- **Chapter 5: Storing and Managing data in HDFS**
  A distributed file system might be a new concept for most readers, so we are going to make sure we go through this core component of Hadoop and ensure you're prepared for designing with this incredible feature.
- **Chapter 6: Adding Structure with Hive**
  We need to go deeper into Hive because you'll use it a lot. Let's dive in with this chapter to make sure you understand commands and the logic behind using Hive efficiently.
- **Chapter 7: Expanding your Capability with HBase and HCatalog**
  Dealing with large tables and metadata requires some new tools and techniques. HBase and HCatalog will help you manage these types of challenges, and we're going to take you through using them. Get ready to put the BIG in big data.
- **Chapter 8: Effective Big Data ETL with SSIS, Pig, and Sqoop**
  We have to load this data, and there is no better way to do it than with our ETL expert authors. Come along while they take you through using favorite and familiar tools, along with some new ones, to load data quickly and effectively.
- **Chapter 9: Data Research and Advanced Data Cleansing with Pig and Hive**
  Now we've installed, configured, explored, and loaded some data. Let's get buys researching and cleansing this data with our new tools and platform.
- **Chapter 10: Data Warehouses and Hadoop Integration**
  How do SQL Server and business intelligence fit in with big data? Very closely. Most of the time they will work in tandem. We will show you when to use each solution and how they work together in scale-up and scale-out solutions.
- **Chapter 11: Visualizing Big Data with Microsoft BI**
  Now that we have the analysis, how do we visualize this for our users? Do we have new tools? Do we use our familiar tools? Yes! Let's do this together so we can understand how to combine these solutions for the best results for our users and customers.

- **Chapter 12: Big Data Analytics**
  You've heard about analytics. This chapter includes advanced statistical analysis, social sentiment analysis, forecasting, modeling, and much more! No PhD required.
- **Chapter 13: Big Data In the Cloud**
  Do you need a lot of servers in your data center to do the things in this book? No way! We can do it in the cloud in an elastic and scalable fashion.
- **Chapter 14: SQL Server Big Data Case Examples**
  How are other firms succeeding and failing in this ecosystem. We will take you through some of the best wins and losses and why these outcomes happened so you can model after them or avoid them.
- **Chapter 15: Building and Executing your Big Data Plan**
  How do we take what we've done and make it real? This chapter will help you write your big data plan.
- **Chapter 16: Operational Big Data Management**
  Administering these technologies and integrating them into your existing infrastructure will take planning and careful execution, just like your other critical systems. Let's plan this out together!

Features Used in This Book

The following features and icons are used in this book to help draw your attention to some of the most important or useful information in the book:

WARNING **Be sure to take heed when you see one of these asides. When particular steps could cause damage to your electronics if performed incorrectly, you'll see one of these asides.**

TIP **These asides contain quick hints about how to perform simple tasks that might prove useful for the task at hand.**

NOTE **These asides contain additional information that may be of importance to you, including links to videos and online material that will make it easier to following along with the development of a particular project.**

SAMPLE HEADING

These asides go into additional depth about the current topic or a related topic.

Part

I

What Is Big Data?

In This Part

CHAPTER

1

Industry Needs and Solutions

WHAT YOU WILL LEARN IN THIS CHAPTER:

- Finding Out What Constitutes "Big Data"
- Appreciating the History and Origins of Hadoop
- Defining Hadoop
- Understanding the Core Components of Hadoop
- Looking to the Future with Hadoop 2.0

This first chapter introduces you to the open source world of Apache and to Hadoop, one of the most exciting and innovative platforms ever created for the data professional. In this chapter we're going to go on a bit of a journey. You're going to find out what inspired Hadoop, where it came from, and its future direction. You'll see how from humble beginnings two gentlemen have inspired a generation of data professionals to think completely differently about data processing and data architecture.

Before we look into the world of Hadoop, though, we must first ask ourselves an important question. Why does big data exist? Is this name just a fad, or is there substance to all the hype? Is big data here to stay? If you want to know the answers to these questions and a little more, read on. You have quite a journey in front of you…

What's So *Big* About Big Data?

The world has witnessed explosive, exponential growth in recent times. So, did we suddenly have a need for big data? Not exactly. Businesses have been tackling the capacity challenge for many years (much to the delight of storage hardware vendors). Therefore the *big* in big data isn't purely a statement on size.

Likewise, on the processing front, scale-out solutions such as high-performance computing and distributed database technology have been in place since the last millennium. There is nothing intrinsically new there either.

People also often talk about unstructured data, but, really, this just refers to the format of the data. Could this be a reason we "suddenly" need big data? We know that web data, especially web log data, is born in an unstructured format and can be generated in significant quantities and volume. However, is this really enough to be considered big data?

In my mind, the answer is no. No one property on its own is sufficient for a project or a solution to be considered a big data solution. It's only when you have a cunning blend of these ingredients that you get to bake a big data cake.

This is in line with the Gartner definition of big data, which they updated in Doug Laney's publication, *The Importance of Big Data: A Definition* (Gartner, 2012): "High volume, high velocity, and/or high variety information assets that require new forms of processing to enable enhanced decision making, insight discovery and process optimization."

What we do know is that every CIO on the planet seems to want to start a big data project right now. In a world of shrinking budgets, there is this sudden desire to jump in with both feet into this world of big data and start prospecting for golden nuggets. It's the gold rush all over again, and clearly companies feel like they might miss out if they hesitate.

However, this is a picture that has been sharpening its focus for several years. In the buildup to this ubiquitous acceptance of big data, we've been blessed with plenty of industry terms and trends, web scale, new programming paradigms of "code first," and of course, to the total disgust of data modelers everywhere, NoSQL. Technologies such as Cassandra and MongoDB are certainly part of the broader ecosystem, but none have resonated as strongly with the market as Hadoop and big data. Why? In essence, unless you were Facebook, Google, Yahoo!, or Bing, issues like web scale really didn't apply.

It seems as though everyone is now building analytics platforms, and that, to be the king of geek chic, requires a degree in advanced statistics. The reason? Big data projects aren't defined by having big data sets. They are shaped by big ideas, by big questions, and by big opportunities. Big data is not about one technology or even one platform. It's so much more than that: It's a mindset and a movement.

Big data, therefore, is a term that underpins a raft of technologies (including the various Hadoop projects, NoSQL offerings, and even MPP Database Systems, for example) that have been created in the drive to better analyze and derive meaning from data at a dramatically lower cost and while delivering new insights and products for organizations all over the world. In times of recession, businesses look to derive greater value from the assets they have rather than invest in new assets. Big data, and in particular Hadoop, is the perfect vehicle for doing exactly that.

A Brief History of Hadoop

Necessity is the mother of invention, and Hadoop is no exception. Hadoop was created to meet the need of web companies to index and process the data tsunami courtesy of the newfangled Internetz. Hadoop's origins owe everything to both Google and the Apache Nutch project. Without one influencing the other, Hadoop might have ended up a very different animal (joke intended). In this next section, we are going to see how their work contributed to making Hadoop what it is today.

Google

As with many pioneering efforts, Google provided significant inspiration for the development that became known as Hadoop. Google published two landmark papers. The first paper, published in October 2003, was titled "The Google File System," and the second paper, "MapReduce: Simplified Data Processing on Large Clusters," published just over a year later in December 2004, provided the inspiration to Doug Cutting and his team of part-time developers for their project, Nutch.

MapReduce was first designed to enable Google developers to focus on the large-scale computations that they were trying to perform while abstracting away all the scaffolding code required to make the computation possible. Given the size of the data set they were working on and the duration of tasks, the developers knew that they had to have a model that was highly parallelized, was fault tolerant, and was able to balance the workload across a distributed set of machines. Of course, the Google implementation of MapReduce worked over Google File System (GFS); Hadoop Distributed File System (HDFS) was still waiting to be invented.

Google has since continued to release thought-provoking, illuminating, and inspirational publications. One publication worthy of note is "BigTable: A Distributed Storage System for Structured Data." Of course, they aren't the only ones. LinkedIn, Facebook, and of course Yahoo! have all contributed to the big data mind share.

There are similarities here to the SIGMOD papers published by various parties in the relational database world, but ultimately it isn't the same. Let's look at an example. Twitter has open-sourced Storm—their complex event processing engine—which has recently been accepted into the Apache incubator program. For relational database vendors, this level of open sharing is really quite unheard of. For more details about storm head over to Apache: `http://incubator.apache.org/projects/storm.html`.

Nutch

Nutch was an open source crawler-based search engine built by a handful of part-time developers, including Doug Cutting. As previously mentioned Cutting was inspired by the Google publications and changed Nutch to take advantage of the enhanced scalability of the architecture promoted by Google. However, it wasn't too long after this that Cutting joined Yahoo! and Hadoop was born.

Nutch joined the Apache foundation in January 2005, and its first release (0.7) was in August 2005. However, it was not until 0.8 was released in July 2006 that Nutch began the transition to Hadoop-based architecture.

Nutch is still very much alive and is an actively contributed-to project. However, Nutch has now been split into two codebases. Version 1 is the legacy and provides the origins of Hadoop. Version 2 represents something of a re-architecture of the original implementation while still holding true to the original goals of the project.

What Is Hadoop?

Apache Hadoop is a top-level open source project and is governed by the Apache Software Foundation (ASF). Hadoop is not any one entity or thing. It is best thought of as a platform or an ecosystem that describes a method of distributed data processing at scale using commodity hardware configured to run as a cluster of computing power. This architecture enables Hadoop to address and analyze vast quantities of data at significantly lower cost than traditional methods commonly found in data warehousing, for example, with relational database systems.

At its core, Hadoop has two primary functions:

- Processing data (MapReduce)
- Storing data (HDFS)

With the advent of Hadoop 2.0, the next major release of Hadoop, we will see the decoupling of resource management from data processing. This adds a third primary function to this list. However, at the time of this writing, Yarn, the Apache project responsible for the resource management, is in alpha technology preview modes.

That said, a number of additional subprojects have been developed and added to the ecosystem that have been built on top of these two primary functions. When bundled together, these subprojects plus the core projects of MapReduce and HDFS become known as a *distribution*.

Derivative Works and Distributions

To fully understand a distribution, you must first understand the role, naming, and branding of Apache Hadoop. The basic rule here is that only official releases by the Apache Hadoop project may be called *Apache Hadoop* or *Hadoop*. So, what about companies that build products/solutions on top of Hadoop? This is where the term *derivative works* comes in.

What Are Derivative Works?

Any product that uses Apache Hadoop code, known as *artifacts*, as part of its construction is said to be a *derivative work*. A derivative work is *not* an Apache Hadoop release. It may be true that a derivative work can be described as "powered by Apache Hadoop." However, there is strict guidance on product naming to avoid confusion in the marketplace. Consequently, companies that provide distributions of Hadoop should also be considered to be derivative works.

NOTE **I liken the relationship between Hadoop and derivative works to the world of Xbox games development. Many Xbox games use graphics engines provided by a third party. The Unreal Engine is just such an example.**

What Is a Distribution?

Now that you know what a derivative work is, we can look at distributions. A *distribution* is the packaging of Apache Hadoop projects and subprojects plus any other additional proprietary components into a single managed package. For example, Hortonworks provides a distribution of Hadoop called "Hortonworks Data Platform," or HDP for short. This is the distribution used by Microsoft for its product, HDInsight.

You may be asking yourself what is so special about that? You could certainly do this yourself. However, this would be a significant undertaking. First, you'd need to download the projects you want, resolve any dependencies, and then compile all the source code. However, when you decide to go down this route, all the testing and integration of the various components is on you to manage and maintain. Bear in mind that the creators of distributions also employ the committers of the actual source and therefore can also offer support.

As you might expect, distributions may lag slightly behind the Apache projects in terms of releases. This is one of the deciding factors you might want to

consider when picking a distribution. Frequency of updates is a key factor, given how quickly the Hadoop ecosystem evolves.

If you look at the Hortonworks distribution, known as Hortonworks Data Platform (HDP), you can see that there are a number of projects at different stages of development. The distribution brings these projects together and tests them for interoperability and stability. Once satisfied that the projects all hang together, the distributor (in this case, Hortonworks) creates the versioned release of the integrated software (the distribution as an installable package).

The 1.3 version made a number of choices as to which versions to support. Today, though, just a few months later, the top-line Hadoop project has a 1.2.0.5 release available, which is not part of HDP 1.3. This and other ecosystem changes will be consumed in the next release of the HDP distribution.

To see a nice graphic of the Hortonworks distribution history, I will refer you to `http://hortonworks.com/products/hdp-2/`. Hadoop is a rapidly changing and evolving ecosystem and doesn't rest on its laurels so including version history is largely futile.

Hadoop Distributions

Note that there are several Hadoop distributions on the market for you to choose from. Some include proprietary components; others do not. The following sections briefly cover some of the main Hadoop distributions.

Hortonworks HDP

Hortonworks provides a distribution of Apache Hadoop known as Hortonworks Data Platform (HDP). HDP is a 100% open source distribution. Therefore, it does not contain any proprietary code or licensing. The developers employed by Hortonworks contribute directly to the Apache projects. Hortonworks is also building a good track record for regular releases of their distribution, educational content, and community engagement. In addition, Hortonworks has established a number of strategic partnerships, which will stand them in good stead. HDP is available in three forms. The first is for Hadoop 1.x, and the second is for Hadoop 2.0, which is currently in development. Hortonworks also offers HDP for Windows, which is a third distribution. HDP for Windows is the only version that runs on the Windows platform.

MapR

MapR is an interesting distribution for Hadoop. They have taken some radical steps to alter the core architecture of Hadoop to mitigate some of its single points of failure, such as the removal of the single master name node for an alternative