Making Everything Easier!™

Big Data

FOR DUMMIES®

A Wiley Brand

Learn to:

- Leverage big data tools and architectures
- Explore how big data can transform your business
- Integrate structured and unstructured data into your big data environment
- Use predictive analytics to make better decisions

Judith Hurwitz
Alan Nugent
Dr. Fern Halper
Marcia Kaufman

Get More and Do More at Dummies.com®

Start with **FREE** Cheat Sheets

Cheat Sheets include

- Checklists
- Charts
- Common Instructions
- And Other Good Stuff!

To access the Cheat Sheet created specifically for this book, go to
www.dummies.com/cheatsheet/bigdata

Get Smart at Dummies.com

Dummies.com makes your life easier with 1,000s of answers on everything from removing wallpaper to using the latest version of Windows.

Check out our

- Videos
- Illustrated Articles
- Step-by-Step Instructions

Plus, each month you can win valuable prizes by entering our Dummies.com sweepstakes. *

Want a weekly dose of Dummies? Sign up for Newsletters on

- Digital Photography
- Microsoft Windows & Office
- Personal Finance & Investing
- Health & Wellness
- Computing, iPods & Cell Phones
- eBay
- Internet
- Food, Home & Garden

Find out "HOW" at Dummies.com

**Sweepstakes not currently available in all countries; visit Dummies.com for official rules.*


Big Data
FOR
DUMMIES
A Wiley Brand

by Judith Hurwitz, Alan Nugent, Dr. Fern Halper, and Marcia Kaufman

FOR DUMMIES
A Wiley Brand

Big Data For Dummies®
Published by
John Wiley & Sons, Inc.
111 River Street
Hoboken, NJ 07030-5774

www.wiley.com

Library of Congress Control Number: 2013933950

ISBN: 978-1-118-50422-2 (pbk); ISBN 978-1-118-64417-1 (ebk); ISBN 978-1-118-64396-9 (ebk);
ISBN 978-1-118-64401-0 (ebk)

Manufactured in the United States of America

10 9 8 7 6 5 4 3 2 1

About the Authors

Judith S. Hurwitz is President and CEO of Hurwitz & Associates, a research and consulting firm focused on emerging technology, including cloud computing, big data, analytics, software development, service management, and security and governance. She is a technology strategist, thought leader, and author. A pioneer in anticipating technology innovation and adoption, she has served as a trusted advisor to many industry leaders over the years. Judith has helped these companies make the transition to a new business model focused on the business value of emerging platforms. She was the founder of Hurwitz Group. She has worked in various corporations, including Apollo Computer and John Hancock. She has written extensively about all aspects of distributed software. In 2011 she authored *Smart or Lucky? How Technology Leaders Turn Chance into Success* (Jossey Bass, 2011). Judith is a co-author on five retail *For Dummies* titles including *Hybrid Cloud For Dummies* (John Wiley & Sons, Inc., 2012), *Cloud Computing For Dummies* (John Wiley & Sons, Inc., 2010), *Service Management For Dummies,* and *Service Oriented Architecture For Dummies,* 2nd Edition (both John Wiley & Sons, Inc., 2009). She is also a co-author on many custom published *For Dummies* titles including *Platform as a Service For Dummies,* CloudBees Special Edition (John Wiley & Sons, Inc., 2012), *Cloud For Dummies,* IBM Midsize Company Limited Edition (John Wiley & Sons, Inc., 2011), *Private Cloud For Dummies,* IBM Limited Edition (2011), and *Information on Demand For Dummies,* IBM Limited Edition (2008) (both John Wiley & Sons, Inc.).

Judith holds BS and MS degrees from Boston University, serves on several advisory boards of emerging companies, and was named a distinguished alumnus of Boston University's College of Arts & Sciences in 2005. She serves on Boston University's Alumni Council. She is also a recipient of the 2005 Massachusetts Technology Leadership Council award.

Alan F. Nugent is a Principal Consultant with Hurwitz & Associates. Al is an experienced technology leader and industry veteran of more than three decades. Most recently, he was the Chief Executive and Chief Technology Officer at Mzinga, Inc., a leader in the development and delivery of cloud-based solutions for big data, real-time analytics, social intelligence, and community management. Prior to Mzinga, he was executive vice president and Chief Technology Officer at CA, Inc. where he was responsible for setting the strategic technology direction for the company. He joined CA as senior vice president and general manager of CA's Enterprise Systems Management (ESM) business unit and managed the product portfolio for infrastructure and data management. Prior to joining CA in April of 2005, Al was senior vice president and CTO of Novell, where he was the innovator behind the company's moves into open source and identity-driven solutions. As consulting CTO for BellSouth he led the corporate initiative to consolidate and transform all of BellSouth's disparate customer and operational data into a single data instance.

Al is the independent member of the Board of Directors of Adaptive Computing in Provo, UT, chairman of the advisory board of SpaceCurve in Seattle, WA, and a member of the advisory board of N-of-one in Waltham, MA. He is a frequent writer on business and technology topics and has shared his thoughts and expertise at many industry events throughout the years.

He is an instrument rated private pilot and has played professional poker for the past three decades. In his sparse spare time he enjoys rebuilding older American muscle cars and motorcycles, collecting antiquarian books, epicurean cooking, and has passion for cellaring American and Italian wines.

Fern Halper, PhD, is a Fellow with Hurwitz & Associates and Director of TDWI Research for Advanced Analytics. She has more than 20 years of experience in data analysis, business analysis, and strategy development. Fern has published numerous articles on data analysis and advanced analytics. She has done extensive research, writing, and speaking on the topic of predictive analytics and text analytics. Fern publishes a regular technology blog. She has held key positions at AT&T Bell Laboratories and Lucent Technologies, where she was responsible for developing innovative data analysis systems as well as developing strategy and product-line plans for Internet businesses. Fern has taught courses in information technology at several universities. She received her BA from Colgate University and her PhD from Texas A&M University.

Fern is a co-author on four retail *For Dummies* titles including *Hybrid Cloud For Dummies* (John Wiley & Sons, Inc., 2012), *Cloud Computing For Dummies* (John Wiley & Sons, Inc., 2010), *Service Oriented Architecture For Dummies,* 2nd Edition, and *Service Management For Dummies* (both John Wiley & Sons, Inc., 2009). She is also a co-author on many custom published *For Dummies* titles including *Cloud For Dummies,* IBM Midsize Company Limited Edition (John Wiley & Sons, Inc., 2011), *Platform as a Service For Dummies,* CloudBees Special Edition (John Wiley & Sons, Inc., 2012), and *Information on Demand For Dummies,* IBM Limited Edition (John Wiley & Sons, Inc., 2008).

Marcia A. Kaufman is a founding Partner and COO of Hurwitz & Associates, a research and consulting firm focused on emerging technology, including cloud computing, big data, analytics, software development, service management, and security and governance. She has written extensively on the business value of virtualization and cloud computing, with an emphasis on evolving cloud infrastructure and business models, data-encryption and end-point security, and online transaction processing in cloud environments. Marcia has more than 20 years of experience in business strategy, industry research, distributed software, software quality, information management, and analytics. Marcia has worked within the financial services, manufacturing, and services industries. During her tenure at Data Resources, Inc. (DRI), she developed sophisticated industry models and forecasts. She holds an AB from Connecticut College in mathematics and economics and an MBA from Boston University.

Marcia is a co-author on five retail *For Dummies* titles including *Hybrid Cloud For Dummies* (John Wiley & Sons, Inc., 2012), *Cloud Computing For Dummies* (John Wiley & Sons, Inc., 2010), *Service Oriented Architecture For Dummies,* 2nd Edition, and *Service Management For Dummies* (both John Wiley & Sons, Inc., 2009). She is also a co-author on many custom published *For Dummies* titles including *Platform as a Service For Dummies,* CloudBees Special Edition (John Wiley & Sons, Inc., 2012), *Cloud For Dummies,* IBM Midsize Company Limited Edition (John Wiley & Sons, Inc., 2011), *Private Cloud For Dummies,* IBM Limited Edition (2011), and *Information on Demand For Dummies* (2008) (both John Wiley & Sons, Inc.).

Dedication

Judith dedicates this book to her husband, Warren, her children, Sara and David, and her mother, Elaine. She also dedicates this book in memory of her father, David.

Alan dedicates this book to his wife Jane for all her love and support; his three children Chris, Jeff, and Greg; and the memory of his parents who started him on this journey.

Fern dedicates this book to her husband, Clay, daughters, Katie and Lindsay, and her sister Adrienne.

Marcia dedicates this book to her husband, Matthew, her children, Sara and Emily, and her parents, Gloria and Larry.

Authors' Acknowledgments

We heartily thank our friends at Wiley, most especially our editor, Nicole Sholly. In addition, we would like to thank our technical editor, Brenda Michelson, for her insightful contributions.

The authors would like to acknowledge the contribution of the following technology industry thought leaders who graciously offered their time to share their technical and business knowledge on a wide range of issues related to hybrid cloud. Their assistance was provided in many ways, including technology briefings, sharing of research, case study examples, and reviewing content. We thank the following people and their organizations for their valuable assistance:

Context Relevant: Forrest Carman

Dell: Matt Walken

Epsilon: Bob Zurek

IBM: Rick Clements, David Corrigan, Phil Francisco, Stephen Gold, Glen Hintze, Jeff Jones, Nancy Kop, Dave Lindquist, Angel Luis Diaz, Bill Mathews, Kim Minor, Tracey Mustacchio, Bob Palmer, Craig Rhinehart, Jan Shauer, Brian Vile, Glen Zimmerman

Kognitio: Michael Hiskey, Steve Millard

Opera Solutions: Jacob Spoelstra

RainStor: Ramon Chen, Deidre Mahon

SAS Institute: Malcom Alexander, Michael Ames

VMware: Chris Keene

Xtremedata: Michael Lamble

Publisher's Acknowledgments

We're proud of this book; please send us your comments at http://dummies.custhelp.com. For other comments, please contact our Customer Care Department within the U.S. at 877-762-2974, outside the U.S. at 317-572-3993, or fax 317-572-4002.

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

Acquisitions, Editorial

Senior Project Editor: Nicole Sholly

Project Editor: Dean Miller

Acquisitions Editor: Constance Santisteban

Copy Editor: John Edwards

Technical Editor: Brenda Michelson

Editorial Manager: Kevin Kirschner

Editorial Assistant: Anne Sullivan

Sr. Editorial Assistant: Cherie Case

Cover Photo: © Baris Simsek / iStockphoto

Composition Services

Project Coordinator: Sheree Montgomery

Layout and Graphics: Jennifer Creasey, Joyce Haughey

Proofreaders: Debbye Butler, Lauren Mandelbaum

Indexer: Valerie Haynes Perry

Publishing and Editorial for Technology Dummies

Richard Swadley, Vice President and Executive Group Publisher

Andy Cummings, Vice President and Publisher

Mary Bednarek, Executive Acquisitions Director

Mary C. Corder, Editorial Director

Publishing for Consumer Dummies

Kathleen Nebenhaus, Vice President and Executive Publisher

Composition Services

Debbie Stailey, Director of Composition Services

Contents at a Glance

Table of Contents

Introduction

Welcome to *Big Data For Dummies.* Big data is becoming one of the most important technology trends that has the potential for dramatically changing the way organizations use information to enhance the customer experience and transform their business models. How does a company go about using data to the best advantage? What does it mean to transform massive amounts of data into knowledge? In this book, we provide you with insights into how technology transitions in software, hardware, and delivery models are changing the way that data can be used in new ways.

Big data is not a single market. Rather, it is a combination of data-management technologies that have evolved over time. Big data enables organizations to store, manage, and manipulate vast amounts of data at the right speed and at the right time to gain the right insights. The key to understanding big data is that data has to be managed so that it can meet the business requirement a given solution is designed to support. Most companies are at an early stage with their big data journey. Many companies are experimenting with techniques that allow them to collect massive amounts of data to determine whether hidden patterns exist within that data that might be an early indication of an important change. Some data may indicate that customer buying patterns are changing or that new elements are in the business that need to be addressed before it is too late.

As companies begin to evaluate new types of big data solutions, many new opportunities will unfold. For example, manufacturing companies may be able to monitor data coming from machine sensors to determine how processes need to be modified before a catastrophic event happens. It will be possible for retailers to monitor data in real time to upsell customers related products as they are executing a transaction. Big data solutions can be used in healthcare to determine the cause of an illness and provide a physician with guidance on treatment options.

Big data is not an isolated solution, however. Implementing a big data solution requires that the infrastructure be in place to support the scalability, distribution, and management of that data. Therefore, it is important to put both a business and technical strategy in place to make use of this important technology trend.

For many important reasons, we think that it is important for you to understand big data technologies and know the ways that companies are using emerging technologies such as Hadoop, MapReduce, and new database

engines to transform the value of their data. We wrote this book to provide a perspective on what big data is and how it's changing the way that organizations can leverage more data than was possible in the past. We think that this book will give you the context to make informed decisions.

About This Book

Big data is new to many people, so it requires some investigation and understanding of both the technical and business requirements. Many different people need knowledge about big data. Some of you want to delve into the technical details, while others want to understand the economic implications of making use of big data technologies. Other executives need to know enough to be able to understand how big data can affect business decisions. Implementing a big data environment requires both an architectural and a business approach — and lots of planning.

No matter what your goal is in reading this book, we address the following issues to help you understand big data and the impact it can have on your business:

- What is the architecture for big data? How can you manage huge volumes of data without causing major disruptions in your data center?
- When should you integrate the outcome of your big data analysis with your data warehouse?
- What are the implications of security and governance on the use of big data? How can you keep your company safe?
- What is the value of different data technologies, and when should you consider them as part of your big data strategy?
- What types of data sources can you take advantage of with big data analytics? How can you apply different types of analytics to business problems?

Foolish Assumptions

Try as we might to be all things to all people, when it came to writing this book, we had to pick who we thought would be most interested in *Big Data For Dummies.* Here's who we think you are:

- **You're smart.** You're no dummy, yet the topic of big data gives you an uneasy feeling. You can't quite get your head around it, and if you're pressed for a definition, you might try to change the subject.

- **You're a businessperson who wants little or nothing to do with technology.** But you live in the 21st century, so you can't escape it. People are saying, "It's all about big data," so you think that you better find out what they're talking about.
- **You're an IT person who knows a heck of a lot about technology.** The thing is, you're new to big data. Everybody says it's something different. Once and for all, you want the whole picture.

Whoever you are, welcome. We're here to help.

How This Book Is Organized

We divided our book into seven parts for easy reading. Feel free to skip about.

Part I: Getting Started with Big Data

In this part, we explain the basic concepts you need for a full understanding of big data, from both a technical and a business perspective. We also introduce you to the major concepts and components so that you can hold your own in any meaningful conversation about big data.

Part II: Technology Foundations for Big Data

Part II is for both technical and business professionals who need to understand the different types of big data components and the underlying technology concepts that support big data. In this section, we give you an understanding about the type of infrastructure that will make big data more practical.

Part III: Big Data Management

Part III is for both technical and business professionals, but it gets into a lot more of the details of different database options and emerging technologies such as MapReduce and Hadoop. Understanding these underlying technologies can help you understand what is behind this important trend.

Part IV: Analytics and Big Data

How do you analyze the massive amounts of data that become part of your big data infrastructure? In this part of the book, we go deeper into the different types of analytics that are helpful in getting real meaning from your data. This part helps you think about ways that you can turn big data into action for your business.

Part V: Big Data Implementation

This part gets to the details of what it means to actually manage data, including issues such as operationalizing your data and protecting the security and privacy of that data. This section gives you plenty to think about in this critical area.

Part VI: Big Data Solutions in the Real World

In this section, you get an understanding of how companies are beginning to use big data to transform their business operations. If you want to get a peek into the future at what you might be able to do with data, this section is for you.

Part VII: The Part of Tens

If you're new to the *For Dummies* treasure-trove, you're no doubt unfamiliar with The Part of Tens. In this section, Wiley editors torture *For Dummies* authors into creating useful bits of information that are easily accessible in lists containing ten (or so) elucidating elements. We started these chapters kicking and screaming but are ultimately very glad that they're here. After you read through the big data best practices, and the do's and don'ts we provide in The Part of Tens, we think you'll be glad, too.

Glossary

We include a glossary of terms frequently used when people discuss big data. Although we strive to define terms as we introduce them in this book, we think you'll find the glossary a useful resource.

Icons Used in This Book

Pay attention. The bother you save may be your own.

You may be sorry if this little tidbit slips your mind.

With this icon, we mark particularly useful points to pay attention to.

Here you find tidbits for the more technically inclined.

Where to Go from Here

We've created an overview of big data and introduced you to all its significant components. We recommend that you read the first four chapters to give you the context for what big data is about and what technologies are in place to make implementations a reality. The next two chapters introduce you to some of the underlying infrastructure issues that are important to understand. The following eight chapters get into a lot more detail about the different types of data structures that are foundational to big data.

You can read the book from cover to cover, but if you're not that kind of person, we've tried to adhere to the *For Dummies* style of keeping chapters self-contained so that you can go straight to the topics that interest you most. Wherever you start, we wish you well.

Many of these chapters could be expanded into full-length books of their own. Big data and the emerging technology landscape are a big focus for us at Hurwitz & Associates, and we invite you to visit our website and read our blogs and insights at `www.hurwitz.com`.

Occasionally, John Wiley & Sons, Inc., has updates to its technology books. If this book has technical updates, they will be posted at `www.dummies.com/go/bigdatafdupdates`.