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

Copyright © 2013 by John Wiley & Sons, Inc., Hoboken, New Jersey

Published simultaneously in Canada

No part of this publication may be reproduced, stored in a retrieval system or transmitted in any form or by any means, electronic, mechanical, photocopying, recording, scanning or otherwise, except as permitted under Sections 107 or 108 of the 1976 United States Copyright Act, without either the prior written permission of the Publisher, or authorization through payment of the appropriate per-copy fee to the Copyright Clearance Center, 222 Rosewood Drive, Danvers, MA 01923, (978) 750-8400, fax (978) 646-8600. Requests to the Publisher for permission should be addressed to the Permissions Department, John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030, (201) 748-6011, fax (201) 748-6008, or online at http://www.wiley.com/go/permissions.

Trademarks: Wiley, the Wiley logo, For Dummies, the Dummies Man logo, A Reference for the Rest of Us!, The Dummies Way, Dummies Daily, The Fun and Easy Way, Dummies.com, Making Everything Easier, and related trade dress are trademarks or registered trademarks of John Wiley & Sons, Inc. and/or its affiliates in the United States and other countries, and may not be used without written permission. All other trademarks are the property of their respective owners. John Wiley & Sons, Inc. is not associated with any product or vendor mentioned in this book.

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

For general information on our other products and services, 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.

For technical support, please visit www.wiley.com/techsupport.

Wiley publishes in a variety of print and electronic formats and by print-on-demand. Some material included with standard print versions of this book may not be included in e-books or in print-on-demand. If this book refers to media such as a CD or DVD that is not included in the version you purchased, you may download this material at http://booksupport.wiley.com. For more information about Wiley products, visit 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

Introduction 1

Part 1: Getting Started with Big Data 7

Chapter 1: Grasping the Fundamentals of Big Data 9
Chapter 2: Examining Big Data Types 25
Chapter 3: Old Meets New: Distributed Computing 37

Part 11: Technology Foundations for Big Data 45

Chapter 4: Digging into Big Data Technology Components 47
Chapter 5: Virtualization and How It Supports Distributed Computing 61
Chapter 6: Examining the Cloud and Big Data 71

Part 111: Big Data Management 83

Chapter 7: Operational Databases 85
Chapter 8: MapReduce Fundamentals 101
Chapter 9: Exploring the World of Hadoop 111
Chapter 10: The Hadoop Foundation and Ecosystem 121
Chapter 11: Appliances and Big Data Warehouses 129

Part 1V: Analytics and Big Data 139

Chapter 12: Defining Big Data Analytics 141
Chapter 13: Understanding Text Analytics and Big Data 153
Chapter 14: Customized Approaches for Analysis of Big Data 167

Part V: Big Data Implementation 179

Chapter 15: Integrating Data Sources 181
Chapter 16: Dealing with Real-Time Data Streams and Complex Event Processing 193
Chapter 17: Operationalizing Big Data 201
Chapter 18: Applying Big Data within Your Organization 211
Chapter 19: Security and Governance for Big Data Environments 225

Part VI: Big Data Solutions in the Real World 235
Chapter 20: The Importance of Big Data to Business 237
Chapter 21: Analyzing Data in Motion: A Real-World View 245
Chapter 22: Improving Business Processes with Big Data Analytics: A Real-World View 255
Part VII: The Part of Tens 263
Chapter 23: Ten Big Data Best Practices 265
Chapter 24: Ten Great Big Data Resources 271
Chapter 25: Ten Big Data Do's and Don'ts 275
Glossary 279
Index 295

Table of Contents

Introduction 1

About This Book 2
Foolish Assumptions 2
How This Book Is Organized 3
Part I: Getting Started with Big Data 3
Part II: Technology Foundations for Big Data 3
Part III: Big Data Management 3
Part IV: Analytics and Big Data 4
Part V: Big Data Implementation 4
Part VI: Big Data Solutions in the Real World 4
Part VII: The Part of Tens 4
Glossary 4
Icons Used in This Book 5
Where to Go from Here 5

Part 1: Getting Started with Big Data 7

Chapter 1: Grasping the Fundamentals of Big Data 9

The Evolution of Data Management 10
Understanding the Waves of Managing Data 11
Wave 1: Creating manageable data structures 11
Wave 2: Web and content management 13
Wave 3: Managing big data 14
Defining Big Data 15
Building a Successful Big Data Management Architecture 16
Beginning with capture, organize, integrate, analyze, and act 16
Setting the architectural foundation 17
Performance matters 20
Traditional and advanced analytics 22
The Big Data Journey 23

Chapter 2: Examining Big Data Types 25

Defining Structured Data 26
Exploring sources of big structured data 26
Understanding the role of relational databases in big data 27
Defining Unstructured Data 29
Exploring sources of unstructured data 29
Understanding the role of a CMS in big data management 31