

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





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



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 spare 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

Big Data For Dummies®

Visit www.dummies.com/cheatsheet/bigdata to view this book's cheat sheet.

Table of Contents

Introduction

[About This Book](#)

[Foolish Assumptions](#)

[How This Book Is Organized](#)

[Part I: Getting Started with Big Data](#)

[Part II: Technology Foundations for Big Data](#)

[Part III: Big Data Management](#)

[Part IV: Analytics and Big Data](#)

[Part V: Big Data Implementation](#)

[Part VI: Big Data Solutions in the Real World](#)

[Part VII: The Part of Tens](#)

[Glossary](#)

[Icons Used in This Book](#)

[Where to Go from Here](#)

Part I: Getting Started with Big Data

Chapter 1: Grasping the Fundamentals of Big Data

[The Evolution of Data Management](#)

[Understanding the Waves of Managing Data](#)

[Wave 1: Creating manageable data structures](#)

[Wave 2: Web and content management](#)

[Wave 3: Managing big data](#)

[Defining Big Data](#)

[Building a Successful Big Data Management Architecture](#)

[Beginning with capture, organize, integrate, analyze, and act](#)

[Setting the architectural foundation](#)

[Performance matters](#)

[Traditional and advanced analytics](#)

[The Big Data Journey](#)

[Chapter 2: Examining Big Data Types](#)

[Defining Structured Data](#)

[Exploring sources of big structured data](#)

[Understanding the role of relational databases in big data](#)

[Defining Unstructured Data](#)

[Exploring sources of unstructured data](#)

[Understanding the role of a CMS in big data management](#)

[Looking at Real-Time and Non-Real-Time Requirements](#)

[Putting Big Data Together](#)

[Managing different data types](#)

[Integrating data types into a big data environment](#)

[Chapter 3: Old Meets New: Distributed Computing](#)

[A Brief History of Distributed Computing](#)

[Giving thanks to DARPA](#)

[The value of a consistent model](#)

[Understanding the Basics of Distributed Computing](#)

[Why we need distributed computing for big data](#)

[The changing economics of computing](#)

[The problem with latency](#)

[Demand meets solutions](#)

[Getting Performance Right](#)

[Part II: Technology Foundations for Big Data](#)

[Chapter 4: Digging into Big Data Technology Components](#)

[Exploring the Big Data Stack](#)

[Layer 0: Redundant Physical Infrastructure](#)

[Physical redundant networks](#)

[Managing hardware: Storage and servers](#)

[Infrastructure operations](#)

[Layer 1: Security Infrastructure](#)

[Interfaces and Feeds to and from Applications and the Internet](#)

[Layer 2: Operational Databases](#)

[Layer 3: Organizing Data Services and Tools](#)

[Layer 4: Analytical Data Warehouses](#)

[Big Data Analytics](#)

[Big Data Applications](#)

Chapter 5: Virtualization and How It Supports Distributed Computing

Understanding the Basics of Virtualization

The importance of virtualization to big data

Server virtualization

Application virtualization

Network virtualization

Processor and memory virtualization

Data and storage virtualization

Managing Virtualization with the Hypervisor

Abstraction and Virtualization

Implementing Virtualization to Work with Big Data

Chapter 6: Examining the Cloud and Big Data

Defining the Cloud in the Context of Big Data

Understanding Cloud Deployment and Delivery Models

Cloud deployment models

Cloud delivery models

The Cloud as an Imperative for Big Data

Making Use of the Cloud for Big Data

Providers in the Big Data Cloud Market

Amazon's Public Elastic Compute Cloud

Google big data services

Microsoft Azure

OpenStack

Where to be careful when using cloud services

Part III: Big Data Management

Chapter 7: Operational Databases

RDBMSs Are Important in a Big Data Environment

PostgreSQL relational database

Nonrelational Databases

Key-Value Pair Databases

Riak key-value database

Document Databases

MongoDB

CouchDB

Columnar Databases

HBase columnar database

Graph Databases

Neo4J graph database

Spatial Databases

PostGIS/OpenGEO Suite

Polyglot Persistence

Chapter 8: MapReduce Fundamentals

Tracing the Origins of MapReduce

Understanding the map Function

Adding the reduce Function

Putting map and reduce Together

Optimizing MapReduce Tasks

Hardware/network topology

[Synchronization](#)

[File system](#)

[Chapter 9: Exploring the World of Hadoop](#)

[Explaining Hadoop](#)

[Understanding the Hadoop Distributed File System \(HDFS\)](#)

[NameNodes](#)

[Data nodes](#)

[Under the covers of HDFS](#)

[Hadoop MapReduce](#)

[Getting the data ready](#)

[Let the mapping begin](#)

[Reduce and combine](#)

[Chapter 10: The Hadoop Foundation and Ecosystem](#)

[Building a Big Data Foundation with the Hadoop Ecosystem](#)

[Managing Resources and Applications with Hadoop YARN](#)

[Storing Big Data with HBase](#)

[Mining Big Data with Hive](#)

[Interacting with the Hadoop Ecosystem](#)

[Pig and Pig Latin](#)

[Sqoop](#)

[Zookeeper](#)

[Chapter 11: Appliances and Big Data Warehouses](#)

[Integrating Big Data with the Traditional Data Warehouse](#)

[Optimizing the data warehouse](#)

[Differentiating big data structures from data warehouse data](#)

[Examining a hybrid process case study](#)

[Big Data Analysis and the Data Warehouse](#)

[The integration lynchpin](#)

[Rethinking extraction, transformation, and loading](#)

[Changing the Role of the Data Warehouse](#)

[Changing Deployment Models in the Big Data Era](#)

[The appliance model](#)

[The cloud model](#)

[Examining the Future of Data Warehouses](#)

[Part IV: Analytics and Big Data](#)

[Chapter 12: Defining Big Data Analytics](#)

[Using Big Data to Get Results](#)

[Basic analytics](#)

[Advanced analytics](#)

[Operationalized analytics](#)

[Monetizing analytics](#)

[Modifying Business Intelligence Products to Handle Big Data](#)

[Data](#)

[Analytical algorithms](#)

[Infrastructure support](#)

[Studying Big Data Analytics Examples](#)

[Orbitz](#)

[Nokia](#)

[NASA](#)

[Big Data Analytics Solutions](#)

[Chapter 13: Understanding Text Analytics and Big Data](#)

[Exploring Unstructured Data](#)

[Understanding Text Analytics](#)

[The difference between text analytics and search](#)

[Analysis and Extraction Techniques](#)

[Understanding the extracted information](#)

[Taxonomies](#)

[Putting Your Results Together with Structured Data](#)

[Putting Big Data to Use](#)

[Voice of the customer](#)

[Social media analytics](#)

[Text Analytics Tools for Big Data](#)

[Attensity](#)

[Clarabridge](#)

[IBM](#)

[OpenText](#)

[SAS](#)

[Chapter 14: Customized Approaches for Analysis of Big Data](#)

[Building New Models and Approaches to Support Big Data](#)

[Characteristics of big data analysis](#)

[Understanding Different Approaches to Big Data Analysis](#)

[Custom applications for big data analysis](#)

[Semi-custom applications for big data analysis](#)

[Characteristics of a Big Data Analysis Framework](#)

[Big to Small: A Big Data Paradox](#)

[Part V: Big Data Implementation](#)

[Chapter 15: Integrating Data Sources](#)

[Identifying the Data You Need](#)

[Exploratory stage](#)

[Codifying stage](#)

[Integration and incorporation stage](#)

[Understanding the Fundamentals of Big Data Integration](#)

[Defining Traditional ETL](#)

[Data transformation](#)

[Understanding ELT — Extract, Load, and Transform](#)

[Prioritizing Big Data Quality](#)

[Using Hadoop as ETL](#)

[Best Practices for Data Integration in a Big Data World](#)

[Chapter 16: Dealing with Real-Time Data Streams and Complex Event Processing](#)

[Explaining Streaming Data and Complex Event Processing](#)

[Using Streaming Data](#)

[Data streaming](#)

[The need for metadata in streams](#)

[Using Complex Event Processing](#)

[Differentiating CEP from Streams](#)

[Understanding the Impact of Streaming Data and CEP on Business](#)

[Chapter 17: Operationalizing Big Data](#)

[Making Big Data a Part of Your Operational Process](#)

[Integrating big data](#)

[Incorporating big data into the diagnosis of diseases](#)

[Understanding Big Data Workflows](#)

[Workload in context to the business problem](#)

[Ensuring the Validity, Veracity, and Volatility of Big Data](#)

[Data validity](#)

[Data volatility](#)

[Chapter 18: Applying Big Data within Your Organization](#)

[Figuring the Economics of Big Data](#)

[Identification of data types and sources](#)

[Business process modifications or new process creation](#)

[The technology impact of big data workflows](#)

[Finding the talent to support big data projects](#)

[Calculating the return on investment \(ROI\) from big data investments](#)

[Enterprise Data Management and Big Data](#)

[Defining Enterprise Data Management](#)

[Creating a Big Data Implementation Road Map](#)

[Understanding business urgency](#)

[Projecting the right amount of capacity](#)

[Selecting the right software development methodology](#)

[Balancing budgets and skill sets](#)

[Determining your appetite for risk](#)

[Starting Your Big Data Road Map](#)

[Chapter 19: Security and Governance for Big Data Environments](#)

[Security in Context with Big Data](#)

[Assessing the risk for the business](#)

[Risks lurking inside big data](#)

[Understanding Data Protection Options](#)

[The Data Governance Challenge](#)

[Auditing your big data process](#)

[Identifying the key stakeholders](#)

[Putting the Right Organizational Structure in Place](#)

[Preparing for stewardship and management of risk](#)

[Setting the right governance and quality policies](#)

[Developing a Well-Governed and Secure Big Data Environment](#)

[Part VI: Big Data Solutions in the Real World](#)

[Chapter 20: The Importance of Big Data to Business](#)

[Big Data as a Business Planning Tool](#)

[Stage 1: Planning with data](#)

[Stage 2: Doing the analysis](#)

[Stage 3: Checking the results](#)

[Stage 4: Acting on the plan](#)

[Adding New Dimensions to the Planning Cycle](#)

[Stage 5: Monitoring in real time](#)

[Stage 6: Adjusting the impact](#)

[Stage 7: Enabling experimentation](#)

[Keeping Data Analytics in Perspective](#)

[Getting Started with the Right Foundation](#)

[Getting your big data strategy started](#)

[Planning for Big Data](#)

[Transforming Business Processes with Big Data](#)

[Chapter 21: Analyzing Data in Motion: A Real-World View](#)

[Understanding Companies' Needs for Data in Motion](#)

[The value of streaming data](#)

[Streaming Data with an Environmental Impact](#)

[Using sensors to provide real-time information about rivers and oceans](#)

[The benefits of real-time data](#)

[Streaming Data with a Public Policy Impact](#)

[Streaming Data in the Healthcare Industry](#)

[Capturing the data stream](#)

[Streaming Data in the Energy Industry](#)

[Using streaming data to increase energy efficiency](#)

[Using streaming data to advance the production of alternative sources of energy](#)

[Connecting Streaming Data to Historical and Other Real-Time Data Sources](#)

Chapter 22: Improving Business Processes with Big Data Analytics: A Real-World View

[Understanding Companies' Needs for Big Data Analytics](#)

[Improving the Customer Experience with Text Analytics](#)

[The business value to the big data analytics implementation](#)

[Using Big Data Analytics to Determine Next Best Action](#)

[Preventing Fraud with Big Data Analytics](#)

[The Business Benefit of Integrating New Sources of Data](#)

Part VII: The Part of Tens

Chapter 23: Ten Big Data Best Practices

[Understand Your Goals](#)

[Establish a Road Map](#)

[Discover Your Data](#)

[Figure Out What Data You Don't Have](#)

[Understand the Technology Options](#)

[Plan for Security in Context with Big Data](#)

[Plan a Data Governance Strategy](#)

[Plan for Data Stewardship](#)

[Continually Test Your Assumptions](#)

[Study Best Practices and Leverage Patterns](#)

Chapter 24: Ten Great Big Data Resources

[Hurwitz & Associates](#)

[Standards Organizations](#)

[The Open Data Foundation](#)

[The Cloud Security Alliance](#)

[National Institute of Standards and Technology](#)

[Apache Software Foundation](#)

[OASIS](#)

[Vendor Sites](#)

[Online Collaborative Sites](#)

[Big Data Conferences](#)

[Chapter 25: Ten Big Data Do's and Don'ts](#)

[Do Involve All Business Units in Your Big Data Strategy](#)

[Do Evaluate All Delivery Models for Big Data](#)

[Do Think about Your Traditional Data Sources as Part of Your Big Data Strategy](#)

[Do Plan for Consistent Metadata](#)

[Do Distribute Your Data](#)

[Don't Rely on a Single Approach to Big Data Analytics](#)

[Don't Go Big Before You Are Ready](#)

[Don't Overlook the Need to Integrate Data](#)

[Don't Forget to Manage Data Securely](#)

[Don't Overlook the Need to Manage the Performance of Your Data](#)

[Glossary](#)

[Cheat Sheet](#)

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