



CENGAGE
Learning®

Professional • Technical • Reference

DATA DIVINATION

BIG DATA STRATEGIES



PAM BAKER

WITH SPECIAL CONTRIBUTION BY BOB GOURLEY



DATA DIVINATION: BIG DATA STRATEGIES

PAM BAKER

Cengage Learning PTR



Professional • Technical • Reference

Australia • Brazil • Japan • Korea • Mexico • Singapore • Spain • United Kingdom • United States

Data Divination: Big Data Strategies
Pam Baker

Publisher and General Manager,
Cengage Learning PTR: Stacy L. Hiquet

Associate Director of Marketing:
Sarah Panella

Manager of Editorial Services:
Heather Talbot

Product Manager: Heather Hurley

Project/Copy Editor: Kezia Endsley

Technical Editor: Rich Santalesa

Interior Layout: MPS Limited

Cover Designer: Luke Fletcher

Proofreader: Sam Garvey

Indexer: Larry Sweazy

© 2015 Cengage Learning PTR.

CENGAGE and CENGAGE LEARNING are registered trademarks of Cengage Learning, Inc., within the United States and certain other jurisdictions.

ALL RIGHTS RESERVED. No part of this work covered by the copyright herein may be reproduced, transmitted, stored, or used in any form or by any means graphic, electronic, or mechanical, including but not limited to photocopying, recording, scanning, digitizing, taping, Web distribution, information networks, or information storage and retrieval systems, except as permitted under Section 107 or 108 of the 1976 United States Copyright Act, without the prior written permission of the publisher.

For product information and technology assistance, contact us at
Cengage Learning Customer & Sales Support, 1-800-354-9706.

For permission to use material from this text or product, submit all
requests online at **cengage.com/permissions.**

Further permissions questions can be emailed to
permissionrequest@cengage.com.

All trademarks are the property of their respective owners.

All images © Cengage Learning unless otherwise noted.

Library of Congress Control Number: 2014937092

ISBN-13: 978-1-305-11508-8

ISBN-10: 1-305-11508-2

eISBN-10: 1-305-11509-0

Cengage Learning PTR

20 Channel Center Street

Boston, MA 02210

USA

Cengage Learning is a leading provider of customized learning solutions with office locations around the globe, including Singapore, the United Kingdom, Australia, Mexico, Brazil, and Japan. Locate your local office at:
international.cengage.com/region.

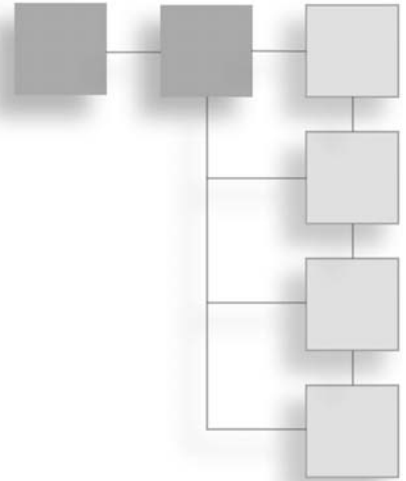
Cengage Learning products are represented in Canada by Nelson Education, Ltd.

For your lifelong learning solutions, visit **cengageptr.com.**

Visit our corporate website at **cengage.com.**

To my daughter Stephanie and my son Ben; you are my inspiration each and every day and the joy of my life. To my mother Nana Duffey; my profound gratitude for teaching me critical thinking skills from a very early age and providing me with a strong, lifelong education and living example of exemplary ethics.

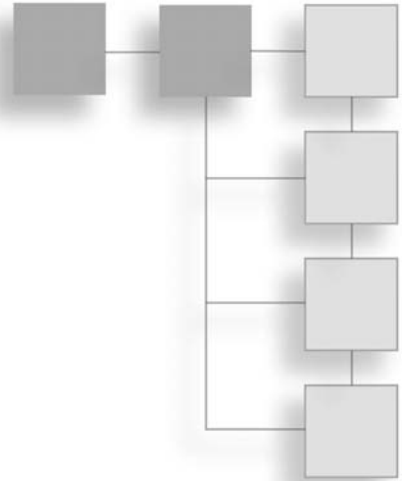
ACKNOWLEDGMENTS



First and foremost, I would like to thank the publishing team at Cengage Learning PTR for their hard work and patience during this time. Specifically, thank you Stacy Hiquet for publishing our text. Thank you Heather Hurley for facilitating this book and being the marvelous person that you are. Your professionalism is second to none but your demeanor is an absolute joy. Thank you Kezia Endsley for your calm management of the editing and deliverables, despite the challenges I inadvertently presented to you. Your many talents, eternal patience, and wise guidance were invaluable to this effort. You are by far the best editor I have had the pleasure of working with and I sincerely hope to have the honor of working with you again one day. Thank you Richard Santalesa for your thorough tech review and insightful suggestions. You are and always will be the greatest legal resource and tech editor a writer can ever have, not to mention the truest of friends. Thank you to the entire team at Cengage.

Many thanks also to my family for their patience and support during this time. Spending long and seemingly unending hours finishing the book is not only hard on the authors, but our families as well. A special thanks to two of my brothers—Steven Duffey and John Duffey—for pitching in to create screenshots and such to exacting specs to help speed completion of the book on such a tight deadline.

ABOUT THE AUTHORS



Pam Baker is a noted business analyst, tech freelance journalist and the editor of the online publication and e-newsletter, *FierceBigData*. Her work is seen in a wide variety of respected publications, including but not limited to *Institutional Investor* magazine, *ReadWriteWeb*, CIO (paper version), CIO.com, *Network World*, *Computerworld*, *IT World*, *LinuxWorld*, *iSixSigma*, and *TechNewsWorld*. Further she formerly served as a contracted analyst for London-based VisionGain Research and Evans Data Corp, headquartered in Santa Cruz, California. She has also served as a researcher, writer, and managing editor of *Wireless I.Q.* and *Telematics Journal* for ABI Research, headquartered in New York.

Interested readers can view a variety of published clips and read more about Pam Baker and her work on these websites: Mediabistro Freelance Marketplace at <http://www.mediabistro.com/PamBaker> and the Internet Press Guild at <http://www.netpress.org/ipg-membership-directory/pambaker>. There are also numerous professional references on her LinkedIn page at <http://www.linkedin.com/in/pambaker/>.

She has also authored numerous ebooks and several of the dead tree variety. Six of those books are listed on her Amazon Author Central page. Further, Baker co-authored two books on the biosciences for the Association of University Technology Managers (AUTM), a global nonprofit association of technology managers and business executives. Those two books were largely funded by the Bill and Melinda Gates Foundation.

Among other awards, Baker won international acclaim for her documentary on the paper-making industry and was awarded a Resolution from the City of Columbus, Georgia, for her news series on the city in *Georgia Trend Magazine*. The only other author to receive such recognition from the city was the legendary Carson McCullers. Baker is a member of the National Press Club (NPC) and the Internet Press Guild (IPG). You can follow or chat with her on Twitter where her handle is @bakercom1 or on Google + at [google.com/+PamBaker](https://plus.google.com/+PamBaker). You can also reach her through the contact form at FierceBigData where she is the editor (see <http://www.fiercebigdata.com/>).

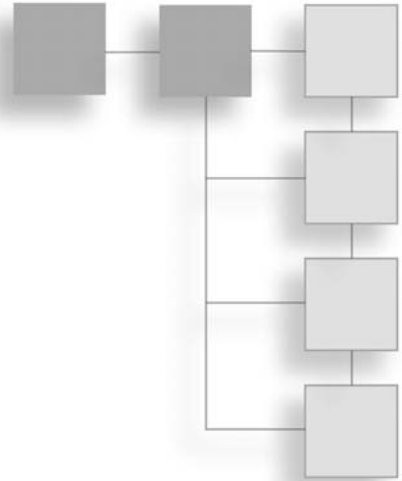
Bob Gourley is a contributing writer in *Data Divination*. He wrote a significant portion of the chapter on use cases in the Department of Defense and Intelligence Community given that is the area he focuses on in his work with big data. Gourley also wrote the chapter on Empowering the Workforce. He is the editor in chief of CTOvision.com and is the founder and Chief Technology Officer (CTO) of Crucial Point LLC, a technology research and advisory firm.

Bob was named one of the top 25 most influential CTOs in the globe by *InfoWorld*, and one of DC's "Tech Titans" by *Washingtonian*. Bob was named one of the "Top 25 Most Fascinating Communicators in Government IT" by the Gov2.0 community GovFresh. In 2012 Bob was noted as "Most Influential on Twitter for Big Data" by *Forbes*.

Bob holds three master's degrees including a master of science degree in scientific and technical intelligence from Naval Postgraduate School, a master of science degree in military science from USMC university, and a master of science degree in computer science from James Madison University. Bob has published over 40 articles on a wide range of topics and is a contributor to the January 2009 book titled *Threats in the Age of Obama*. His blog, CTOvision, is now ranked among the top federal technology blogs by *WashingtonTech*.

Bob is a founding member and member of the board of directors of the Cyber Conflict Studies Association, a non-profit group focused on enhancing the study of cyber conflict at leading academic institutions and furthering the ability of the nation to understand the complex dynamics of conflict in cyberspace. You can follow and chat with him on Twitter where his handle is @bobgourley. You can also find him on Twitter as @AnalystReport and @CTOvision and online at <http://ctovision.com/pro>.

CONTENTS



	Introduction	xv
Chapter 1	What Is Big Data, Really?	1
	Technically Speaking	1
	Why Data Size Doesn't Matter.	5
	What Big Data Typically Means to Executives.	5
	The "Data Is Omnipotent" Group	6
	The "Data Is Just Another Spreadsheet" Group	6
	Big Data Positioned in Executive Speak	7
	Summary	14
Chapter 2	How to Formulate a Winning Big Data Strategy	17
	The Head Eats the Tail	17
	How to End the "Who's on First" Conundrum	20
	Changing Perspectives of Big Data	20
	User Perception Versus the Data-Harvesting Reality	21
	The Reality of Facebook's Predictive Analytics	22
	Facebook's Data Harvesting Goes Even Further	23
	Using Facebook to Open Minds on the Possibilities and Potential of Big Data	24
	Professional Perceptions Versus Data Realities	24
	From Perception to Cognitive Bias.	25
	Finding the Big Data Diviners	26
	Next Step: Embracing Ignorance	28
	Where to Start	29

	Begin at the End.	31
	When Action Turns into Inaction.	33
	Identifying Targets and Aiming Your Sights.	34
	Covering All the Bases	35
	How to Get Best Practices and Old Mindsets Out of Your Way.	36
	Addressing People’s Fears of Big Data	37
	Ending the Fear of the Unknown	37
	Tempering Assurances for Change Is About to Come	38
	The Feared Machine’s Reign Is not Certain; Mankind Still Has a Role.	39
	Reaching the Stubborn Few.	40
	Answer the Questions No One Has Asked.	40
	Keep Asking What Is Possible	40
	Look for End Goals	41
	Cross-Pollinate the Interpretative Team	42
	Add Business Analysts and Key End Users to the Team	43
	Add a Chief Data Officer to Gather and Manage the Data.	43
	Start Small and Build Up and Out	45
	Prototypes and Iterations Strategies	46
	A Word About Adding Predictive Analytics to Your Data Strategy.	46
	Democratize Data but Expect Few to Use It (for Now).	47
	Your Strategy Is a Living Document; Nourish It Accordingly	48
	Summary.	48
Chapter 3	How to Ask the “Right” Questions of Big Data	49
	Collaborate on the Questions	51
	The Magic 8 Ball Effect.	52
	Translating Human Questions to Software Math	53
	Checklist for Forming the “Right” Questions	53
	Summary.	54
Chapter 4	How to Pick the “Right” Data Sources	55
	You Need More Data Sources (Variety), Not Just More Data (Volume)	55
	Why Your Own Data Isn’t Enough and Will Never Be Enough, No Matter How Large It Grows	56
	Data Hoarding versus Catch and Release	57
	One-Dimensional Traps	57
	The Mysterious Case of the Diaper-Buying Dog-Owner.	58
	The Value in Upsizing Transactional Data.	59
	The Limits to Social Media Analysis.	59
	The Monetary Value of Data Bought and Sold.	60
	Even Hackers Are Having Trouble Making Money on Data	61
	Evaluating the Source	62

	Outdated Models Invite Disruptors	63
	What to Look for When Buying Data	64
	Identifying What Outside Data You Need	64
	A Word About Structured vs. Unstructured Data	66
	Preventing Human Bias in Data Selection	68
	The Danger of Data Silos	69
	Steps to Take to Ensure You’re Using All the Data Sources You Need	70
	Summary	72
Chapter 5	Why the Answer to Your Big Data Question Resembles a Rubik’s Cube	73
	What Is Actionable Data Anyway?	74
	The Difference Among Descriptive, Predictive, and Prescriptive Analytics	77
	Descriptive Analytics	78
	Predictive Analytics	78
	Prescriptive Analytics	79
	Types of Questions That Get Straight Answers	81
	When Questions Lead to More Questions	82
	Types of Questions That Require Interpretation—The Rubik’s Cube	82
	Using Data Visualizations to Aid the Discovery Process	83
	Summary	85
Chapter 6	The Role of Real-Time Analytics in Rolling Your Strategy	87
	Examining Real-Time Delusions and Time Capsules	89
	Using Static versus Rolling Strategies	91
	A Word About Change Management in Moving to a Rolling Strategy	91
	Your Choices in Analytics	92
	Using Data from Human Experts’ Brains to Speed Analytics	95
	When Real-Time Is Too Late, Then What?	96
	Summary	97
Chapter 7	The Big Data Value Proposition and Monetization	99
	Determining ROI in Uncharted Territory	99
	The Lesson in the Skill of the Painter versus the Value of the Paintbrush	101
	Funny Money and Fuzzy ROI	102
	The Confusion in Cost	104
	Why Cost Isn’t an Issue	104
	Putting the Project Before the Business Case	105
	Calculating Actual Cost	106
	Where Value Actually Resides	107
	How to Make the Business Case from an IT Perspective	107
	How to Make the Business Case from a Non-IT Perspective	108

	Formulas for Calculating Project Returns	109
	Where the ROI Math Gets Simpler	111
	The Big Question: Should You Sell Your Data?	112
	Selling Insights	113
	Rarity Equals Cha-Ching!	113
	Summary	114
Chapter 8	Rise of the Collaborative Economy and Ways to Profit from It.	115
	Data Is Knowledge and an Asset	115
	Big Data’s Biggest Impact: Model Shattering	117
	The Sharing Economy	119
	The Maker Movement	121
	Co-Innovation	123
	Examples of New Models Emerging in the New Collaborative Economy	124
	Agile Is Out, Fluid Is In	126
	Using Big Data to Strategize New Models	129
	Summary	130
Chapter 9	The Privacy Conundrum.	131
	The Day the Whistle Blew and Killed the Myth of Individual Privacy	133
	Dangers in the Aggregate	134
	The Phone Call Heard Around the World	135
	How John Q. Public’s and Veterans’ Data Help Other Nations Plan Attacks	137
	Data Proliferation Escalates	140
	Drawing the Line on Individual Privacy	141
	The Business Side of the Privacy Conundrum	145
	The Four Big Shifts in Data Collection	145
	Data Invasiveness Changes	147
	Data Variety Changes	150
	Data Integration Changes	151
	Data Scope Changes	152
	The Business Question You Must Ask	158
	Who Really Owns the Data?	158
	The Role of Existing Laws and Actions in Setting Precedent	160
	The Snowden Effect on Privacy Policy	161
	The Fallacies of Consent	162
	Values in Personal versus Pooled Data	163
	The Fallacy in Anonymizing Data	165
	Balancing Individual Privacy with Individual Benefit	165
	When Data Collection Could Make You or Your Company Liable	166
	The Business Value of Transparency	168
	The One Truth That Data Practitioners Must Never Forget	170
	Summary	170

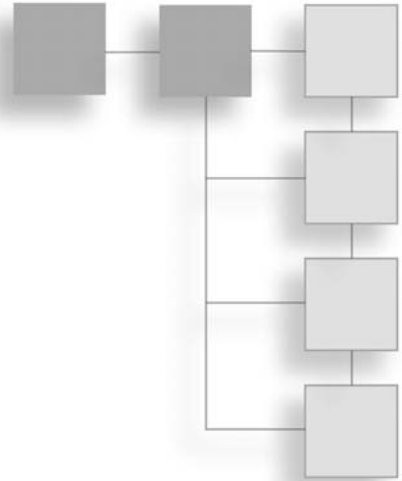
Chapter 10	Use Cases in the Department of Defense and Intelligence Community	171
	Situational Awareness and Visualization	172
	Information Correlation for Problem Solving (the “Connect the Dots” Problem)	174
	Information Search and Discovery in Overwhelming Amounts of Data (the “Needle in Haystack” Problem)	179
	Enterprise Cyber Security Data Management	182
	Logistical Information, Including Asset Catalogs Across Extensive/Dynamic Enterprises	183
	Enhanced Healthcare	184
	Open Source Information	186
	In-Memory Data Modernization	187
	The Enterprise Data Hub	187
	Big Data Use Cases in Weaponry and War	188
	Summary	189
Chapter 11	Use Cases in Governments	191
	Effects of Big Data Trends on Governmental Data	192
	United Nations Global Pulse Use Cases	194
	Federal Government (Non-DoD or IC) Use Cases	196
	State Government Use Cases	200
	Local Government Use Cases	204
	Law Enforcement Use Cases	206
	Summary	209
Chapter 12	Use Cases in Security	211
	Everything Is on the Internet	211
	Data as Friend and Foe	213
	Use Cases in Antivirus/Malware Efforts	214
	How Target Got Hit in the Bull’s Eye	217
	Big Data as Both Challenge and Benefit to Businesses	220
	Where Virtual and Real Worlds Collide	222
	Machine Data Mayhem	224
	The Farmer’s Security Dilemma	224
	The Internet of Things Repeats the Farmer’s Security Dilemma	
	Ad Infinitum	225
	Current and Future Use of Analytics in Security	226
	Summary	232
Chapter 13	Use Cases in Healthcare	233
	Solving the Antibiotics Crisis	234
	Using Big Data to Cure Diseases	235

	From Google to the CDC	236
	CDC’s Diabetes Interactive Atlas	239
	Project Data Sphere	244
	Sage Bionetworks	246
	The Biohacker Side of the Equation	247
	EHRs, EMRs, and Big Data	249
	Medicare Data Goes Public	251
	Summary	254
Chapter 14	Use Cases in Small Businesses and Farms	255
	Big Data Applies to Small Business	255
	The Line Between Hype and Real-World Limitations	256
	Picking the Right Tool for the Job	257
	Examples of External Data Sources You Might Want to Use	264
	A Word of Caution to Farmers on Pooling or Sharing Data	271
	The Claim that the Data Belongs to the Farmer	273
	The Claim that the Data Is Used Only to “Help” the Farmer Farm More Profitably	273
	The Claim that the Farmer’s Data Will Remain Private	274
	Money, Money, Money: How Big Data Is Broadening Your Borrowing Power	275
	PayPal Working Capital	277
	Amazon Capital Services	277
	Kabbage	278
	Summary	279
Chapter 15	Use Cases in Transportation	281
	Revving Up Data in a Race for Money	281
	The Disrupting Fly in the Data Ointment	282
	Data Wins Are Not Eternal	283
	Data Use in Trains, Planes, and Ships	284
	Connected Vehicles: They’re Probably Not What You Think They Are	286
	Data Leads to Innovation and Automation	290
	The Rise of Smart Cities	290
	Examples of Transportation Innovations Happening Now	291
	Data and the Driverless Car	293
	Connected Infrastructure	296
	Car Insurance Branded Data Collection Devices	299
	Unexpected Data Liabilities for the Sector	302
	Summary	304
Chapter 16	Use Cases in Energy	305
	The Data on Energy Myths and Assumptions	305
	EIA Energy Data Repository	308

	EIA Energy Data Table Browsers	309
	Smart Meter Data Is MIA	312
	The EIA's API and Data Sets	313
	International Implications and Cooperation	314
	Public-Private Collaborative Energy Data Efforts	315
	Utility Use Cases	316
	Energy Data Use Cases for Companies Outside the Energy Sector	317
	Summary	319
Chapter 17	Use Cases in Retail	321
	Old Tactics in a Big Data Re-Run	322
	Retail Didn't Blow It; the Customers Changed	323
	Brand Mutiny and Demon Customers	324
	Customer Experience Began to Matter Again	326
	Big Data and the Demon Customer Revival	326
	Why Retail Has Struggled with Big Data	328
	Ways Big Data Can Help Retail	329
	Product Selection and Pricing	330
	Current Market Analysis	332
	Use Big Data to Develop New Pricing Models	332
	Find Better Ways to Get More, Better, and Cleaner Customer Data	333
	Study and Predict Customer Acceptance and Reaction	333
	Predict and Plan Responses to Trends in the Broader Marketplace	338
	Predicting the Future of Retail	341
	Summary	342
Chapter 18	Use Cases in Banking and Financial Services	343
	Defining the Problem	343
	Use Cases in Banks and Lending Institutions	345
	How Big Data Fuels New Competitors in the Money-Lending Space	347
	The New Breed of Alternative Lenders	347
	PayPal Working Capital	347
	Prosper and Lending Club	348
	Retailers Take on Banks; Credit Card Brands Circumvent Banks	349
	The Credit Bureau Data Problem	350
	A Word About Insurance Companies	353
	Summary	355
Chapter 19	Use Cases in Manufacturing	357
	Economic Conditions and Opportunities Ahead	358
	Crossroads in Manufacturing	360
	At the Intersection of 3D Printing and Big Data	364

	How 3D Printing Is Changing Manufacturing and Disrupting Its Customers.	364
	WinSun Prints 10 Homes in a Single Day.	365
	The 3D Printed Landscape House	365
	The 3D Printed Canal House	367
	The Impact of 3D Home Printing on Manufacturing	367
	The Shift to Additive Manufacturing Will Be Massive and Across All Sectors.	368
	How Personalized Manufacturing Will Change Everything and Create Even More Big Data	370
	New Data Sources Springing from Inside Manufacturing	372
	Use Cases for this Sector.	372
	Summary.	373
Chapter 20	Empowering the Workforce	375
	Democratizing Data	376
	Four Steps Forward.	377
	Four More Steps Forward.	380
	Summary.	381
Chapter 21	Executive Summary	383
	What Is Big Data Really?	383
	How to Formulate a Winning Big Data Strategy	384
	How to Ask the “Right” Questions of Big Data	386
	How to Pick the “Right” Data Sources	386
	Why the Answer to Your Big Data Question Resembles a Rubik’s Cube	387
	The Role of Real-Time Analytics in Rolling Your Strategy	388
	The Big Data Value Proposition and Monetization	389
	Rise of the Collaborative Economy and Ways to Profit from It	390
	The Privacy Conundrum	391
	Use Cases in Governments	392
	Use Cases in the Department of Defense and Intelligence Community	393
	Use Cases in Security.	394
	Use Cases in Healthcare	395
	Use Cases in Small Businesses and Farms.	396
	Use Cases in Energy	397
	Use Cases in Transportation.	398
	Use Cases in Retail	400
	Use Cases in Banking and Financial Services	401
	Use Cases in Manufacturing.	402
	Empowering the Workforce	404
Index		407

INTRODUCTION



Amidst all the big data talk, articles, and conference speeches lies one consistently unanswered question: What can we actually do with big data? Sure, the answer is alluded to frequently but only in the vaguest and most general terms. Few spell out where to begin, let alone where to go with big data from there. Answers to related questions—from how to compute ROI for big data projects and monetize data to how to develop a winning strategy and ultimately how to wield analytics to transform entire organizations and industries—are even rarer. That’s why *Data Divination* was written—to answer all of those most pressing questions and more from a high-level view.

THIS BOOK IS FOR YOU IF

If you are interested in the business end of big data rather than the technical nuts and bolts, this book is for you. Whether your business is a one-man operation or a global empire, you’ll find practical advice here on how and when to use big data to the greatest effect for your organization. It doesn’t matter whether you are a data scientist, a department head, an attorney, a small business owner, a non-profit head, or a member of the C-Suite or company board, the information contained within these pages will enable you to apply big data techniques and decision-making to your tasks.

Further, many of the chapters are dedicated to use cases in specific industries to serve as practical guides to what is being and can be done in your sector and business. Ten industries are addressed in exquisite detail in their own chapters. There you’ll find use cases, strategies, underlying factors, and emerging trends detailed for the governments,

department of defense and intelligence community, security, healthcare, small businesses and farms, transportation, energy, retail, banking and insurance, and manufacturing sectors. However, it is a mistake to read only the chapter on your own industry, as changes wrought by big data in other industries will also affect you, if they haven't already.

If there is one thing that big data is shaping up to be, it is a catalyst of disruption across the board. Indeed, it is helping meld entire industries in arguably the biggest surge of cross-industry convergence ever seen. It therefore behooves you to note which industries are converging with yours and which of your customers are reducing or eliminating a need for your services entirely. It's highly likely that you'll find more than a few surprises here in that regard.

STRATEGY IS EVERYTHING

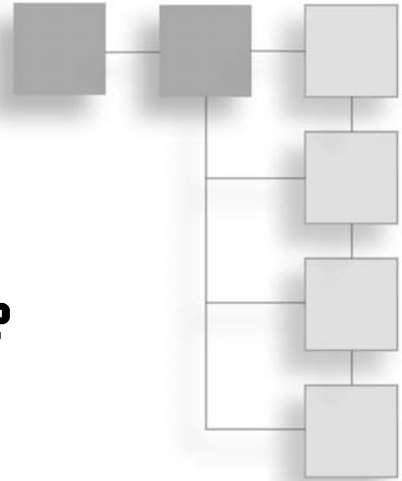
Data Divination is about how to develop a winning big data strategy and see it to fruition. You'll find chapters here dedicated to various topics aimed at that end. Included in these pages are the answers to how to calculate ROI; build a data team; devise data monetization; present a winning business proposition; formulate the right questions; derive actionable answers from analytics; predict the future for your business and industry; effectively deal with privacy issues; leverage visualizations for optimum data expressions; identify where, when, and how to innovate products and services; and how to transform your entire organization.

By the time you reach the end of this book, you should be able to readily identify what you need to do with big data, be that where to start or where to go next.

There are some references to tools here, but very few. Big data tools will age out over time, as all technologies do. However, your big data strategies will arch throughout time, morphing as needed, but holding true as the very foundation of your business. Strategy then is where you need to hold your focus and it is where you will find ours here.

From your strategy, you will know what tools to invest in and where and how you need to use them. But more than helping you pick the right tools and to increase your profits, your strategy will see you through sea changes that are approaching rapidly and cresting on the horizon now. The changes are many and they are unavoidable. Your only recourse is to prepare and to proactively select your path forward. We do our best to show you many of your options using big data in these pages to help you achieve all of that.

CHAPTER 1



WHAT IS BIG DATA, REALLY?

One would think that, given how the phrase “big data” is on the tip of nearly every tongue and top of mind for most, everyone knows what big data is. That’s not quite the case. Although there is a technical definition of sorts, most people are unsure of where the defining line is in terms of big versus regular data sizes. This creates some difficulty in communicating and thinking about big data in general and big data project parameters in particular.

This chapter considers the different interpretations of the meaning of the term “big data.”

TECHNICALLY SPEAKING

As discussed in more detail in the next chapter, big data does not mean more of the same data, simply boosting gigabytes to terabytes, although obviously it includes the expected growth of existing data sets. Rather, big data is a collection of data sets, some structured and some unstructured, some “onboarded” from physical sources to online sets, some transactional and some not, from a variety of sources, some in-house and some from third parties. Often it is stored in a variety of disparate and hard-to-reconcile forms. As a general rule, big data is clunky, messy, and hard, if not impossible, as well as significantly expensive, to shoe-horn into existing computing systems.

Furthermore, in the technical sense there is no widely accepted consensus as to the minimum size a data collective must measure to qualify as “big.” Instead the technical world favors a definition more attuned to data characteristics and size relative to current computing capabilities.

You’ll commonly hear big data defined as “containing volume, velocity, and variety” which is the three-legged definition coined by a 2001 Gartner (then Meta) report. These days, some people throw in a fourth “v,”—veracity—to cover data quality issues too.

But in essence big data is whatever size data set requires new tools in order to compute. Therefore, data considered big by today’s standards will likely be considered small or average by future computing standards.

That is precisely why attaching the word “big” to data is unfortunate and not very useful. In the near future most industry experts expect the word big to be dropped entirely as it fails to accurately describe anything essential to the concept. For what makes “big data” truly valuable are the “big connections” it makes possible—between people, places, and things—that were previously impossible to glean in any coherent fashion.

Even so, there are those who try to affix a specific size to big data, generally in terms of terabytes. However this is not a static measurement. The measure generally refers to the amount of data flowing in or growing in the datacenter in a set timeframe, such as weekly. Conversely, since data is growing so quickly everywhere, at an estimated rate of 2,621,440 terabytes daily according to the Rackspace infographic in Figure 1.1, a static measurement for a “big data” set is frequently meaningless after a very short time. (This infographic can also be found online at <http://www.rackspace.com/blog/exploring-the-universe-of-big-data-infographic/>.)

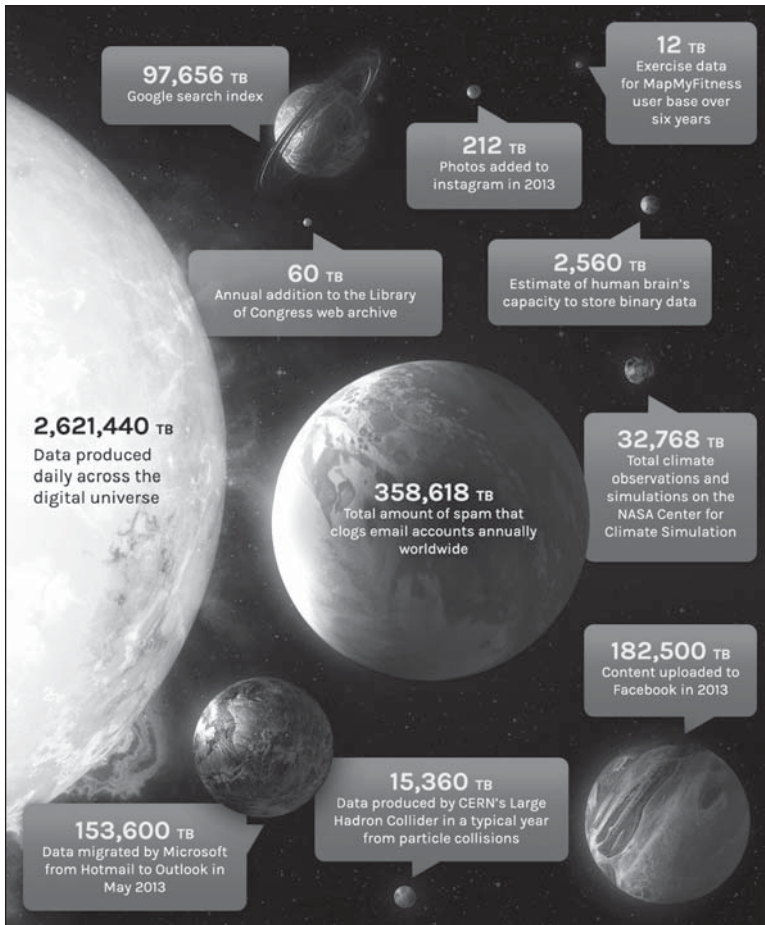


Figure 1.1

This interactive infographic has a counter at the top that shows how many terabytes of data were added to the digital universe since the user opened the infographic. The speed in which the counter counts gives you a good idea of just how fast data is growing overall. By rolling a mouse over the different planets, the user reveals the size of data in different categories relative to the size of all data generated (represented here as the sun) such as in email spam, in the Google search index, and in Facebook.

Source: Infographic courtesy of Rackspace. Concept and research by Dominic Smith; design and rendering by Legacy79.

Already we know that bigger data is coming. Data sets so big that we don't yet have a measuring term for it. But until then we'll use the measurements we do have: first up is zettabytes and then yottabytes. To give you an understanding of the magnitude of a yottabyte, consider that it equals one quadrillion gigabytes or one septillion bytes—that is a 1 followed by 24 zeroes. Consider Figure 1.2 for other ways to visualize the size of a yottabyte.

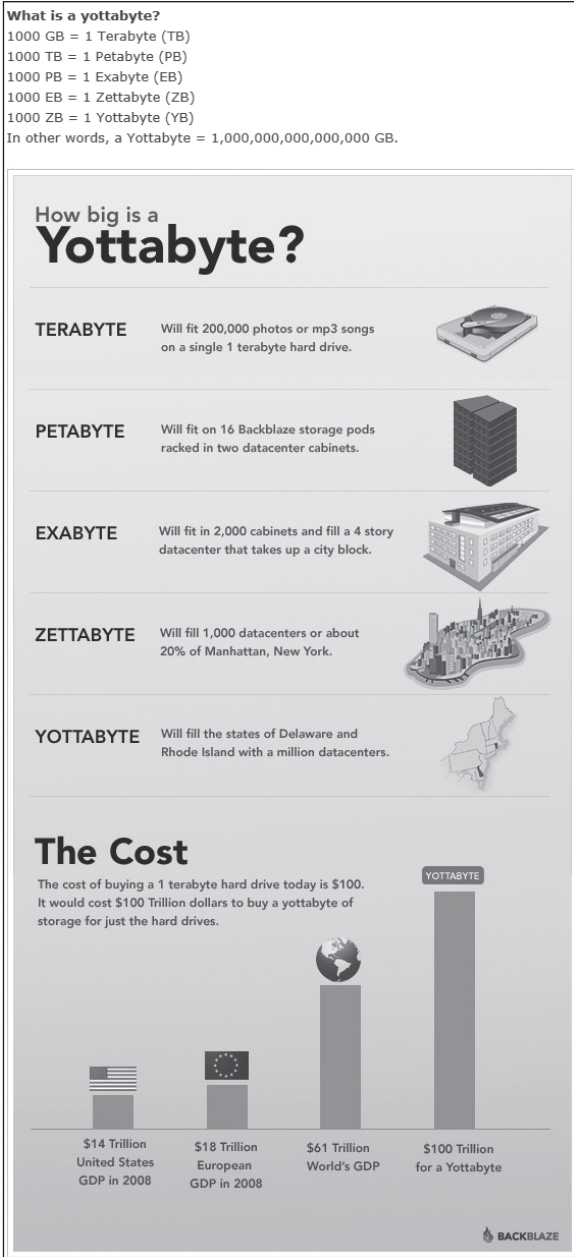


Figure 1.2
 This graphic and accompanying text visualize the actual size of a yottabyte.
 Source: Backblaze; see <http://blog.backblaze.com/2009/11/12/nsa-might-want-some-backblaze-pods/>.

As hard as that size is to imagine, think about what comes next. We have no word for the next size and therefore can barely comprehend what we can or should do with it all. It is, however, certain that extreme data will arrive soon.

WHY DATA SIZE DOESN'T MATTER

Therefore the focus today is primarily on how best to access and compute the data rather than how big it is. After all, the value is in the quality of the data analysis and not in its raw bulk.

Feel confused by all this? Rest assured, you are in good company. However, it is also a relief to learn that many new analytic tools can be used on data of nearly any size and on data collections of various levels of complexities and formats. That means data science teams can use big data tools to derive value from almost any data. That is good news indeed because the tools are both affordable and far more capable of fast (and valuable) analysis than their predecessors.

Your company will of course have to consider the size of its data sets in order to ultimately arrange and budget for storage, transfer, and other data management related realities. But as far as analytical results, data size doesn't much matter as long as you use a large enough data set to make the findings significant.

WHAT BIG DATA TYPICALLY MEANS TO EXECUTIVES

Executives, depending on their personal level of data literacy, tend to view big data as somewhat mysterious but useful to varying degrees. Two opposing perceptions anchor each endpoint of the executive viewpoint spectrum. One end point views big data as a reveal all and tell everything tool whereas the other end of the spectrum sees it is simply as a newfangled way to deliver analysis on more of the same data they are accustomed to seeing in the old familiar spreadsheet. Even when presented with visualizations, the second group tends to perceive it, at least initially, as another form of the spreadsheet.

There are lots of other executive perceptions between these two extremes, of course. But it is useful for your purposes here to consider the two extremes—omniscience and spreadsheet upgrade—in order to quickly assess the executive expectations. This will better prepare you to deliver data findings in the manner most palatable and useful to your individual executives.

The “Data Is Omnipotent” Group

For the first group, it may be necessary to explain that while big data can and does produce results heretofore not possible, it is not, nor will it ever be, omniscience as is often depicted in many movies. In other words, data, no matter how huge and comprehensive, will never be complete and rarely in proper context. Therefore, it cannot be omnipotent.

This group also tends to misunderstand the limitations of predictive analytics. These are good tools in predicting future behavior and events, but they are not magical crystal balls that reveal a certain future. Predictive analytics predict the future assuming that current conditions and trends continue on the same path. That means that if anything occurs to disrupt that path or significantly change its course, the previous analysis from predictive analytics no longer applies. This is an important distinction that must be made clear to executives and data enthusiasts. Not only so that they use the information correctly but they also understand that their role in strategizing is not diminished or replaced by analytics, but greatly aided by it.

Further, most big data science teams are still working on rather basic projects and experiments, learning as they go. Most are simply unable to deliver complex projects yet. If executives have overly high initial expectations, they may be disappointed in these early stages. Disappointment can lead to executive disengagement and that bodes ill for data science teams and business heads. This can actually lead to scrapping big data projects and efforts all together. Therefore, it’s important to properly and realistically manage executive expectations from the outset.

On the upside, executives in this group may be more open to suggestions on new ways to use data and be quicker to offer guidance on what information they most need to see. Such enthusiastic involvement and buy-in from executives is incredibly helpful to the initiative.

The “Data Is Just Another Spreadsheet” Group

At the other extreme end of the spectrum, the second group is likely to be unimpressed with big data beyond a mere nod to the idea that more data is good. This group views big data as a technical activity rather than as an essential business function.

Members of this executive group are likely to be more receptive to traditional visualizations, at least initially. To be of most assistance to this group of executives, ask outright

what information they wish they could know and why. Then, if they answer, you have a solid and welcomed way to demonstrate the value of the company's big data efforts by presenting exactly what was needed but heretofore missing.

If they can't or don't answer the question, work proactively to find ways to demonstrate the value of data analysis in ways that are meaningful to those executives.

Expect most executives to have little interest in how data is *cooked*—gathered, mixed, and analyzed. Typically they want to know its value over the traditional ways of doing things instead.

Whether executives belong to one of these two extreme groups or are somewhere in between, it is imperative to demonstrate the value of big data analysis as you would in any business case and/or present ongoing metrics as you would for any other technology.

However, your work with executives doesn't end there.

BIG DATA POSITIONED IN EXECUTIVE SPEAK

Although data visualizations have proven to be the fastest and most effective way to transfer data findings to the human brain, not everyone processes information in the same way. Common visualizations are the most readily understood by most people, but not always. Common visualizations include pie charts, bar graphs, line graphs, cumulative graphs, scatter plots, and other data representations used long before the advent of big data.

The most common of all is the traditional spreadsheet with little to no art elements. Figure 1.3 shows an example of a traditional spreadsheet.

Windy Crest Acres, Inc. (Compatibility Mode) - Excel

FILE HOME INSERT PAGE LAYOUT FORMULAS DATA REVIEW VIEW

Clipboard Font Alignment Number Styles

As of December 31, 2011

	Jan 2011	Feb 2011	Mar 2011	Apr 2011	May 2011	Jun 2011	Jul 2011	Aug 2011	Sep 2011	Oct 2011	Nov 2011
ASSETS											
Current Assets											
Bank Accounts	37,878.38	10,595.40	17,958.93	12,326.34	10,262.59	8,028.89	8,127.71	19,413.13	6,506.34	11,956.42	0.00
Bank Accounts	\$ 231.05	\$ 241.25	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Other Accounts	\$ 43,109.43	\$ 15,836.65	\$ 17,958.93	\$ 12,326.34	\$ 10,262.59	\$ 8,028.89	\$ 8,127.71	\$ 19,413.13	\$ 6,506.34	\$ 11,956.42	\$ 0.00
Total Bank Accounts											
Accounts Receivable	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Accounts Receivable	\$ 0.00	\$ 0.00	\$ 0.00	\$ 0.00	\$ 0.00	\$ 0.00	\$ 0.00	\$ 0.00	\$ 0.00	\$ 0.00	\$ 0.00
Other current assets	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Other current assets	\$ 0.00	\$ 0.00	\$ 0.00	\$ 0.00	\$ 0.00	\$ 0.00	\$ 0.00	\$ 0.00	\$ 0.00	\$ 0.00	\$ 0.00
Undeposited Funds	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Undeposited Funds	\$ 0.00	\$ 0.00	\$ 0.00	\$ 0.00	\$ 0.00	\$ 0.00	\$ 0.00	\$ 0.00	\$ 0.00	\$ 0.00	\$ 0.00
Total Other current assets											
Total Current Assets	\$ 43,109.43	\$ 15,836.65	\$ 17,958.93	\$ 12,326.34	\$ 10,262.59	\$ 8,028.89	\$ 8,127.71	\$ 19,413.13	\$ 6,506.34	\$ 11,956.42	\$ 0.00
TOTAL ASSETS	\$ 43,109.43	\$ 15,836.65	\$ 17,958.93	\$ 12,326.34	\$ 10,262.59	\$ 8,028.89	\$ 8,127.71	\$ 19,413.13	\$ 6,506.34	\$ 11,956.42	\$ 0.00
LIABILITIES AND EQUITY											
Liabilities											
Current Liabilities											
Accounts Payable	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	378.90
Accounts Payable	\$ 0.00	\$ 0.00	\$ 0.00	\$ 0.00	\$ 0.00	\$ 0.00	\$ 0.00	\$ 0.00	\$ 0.00	\$ 0.00	\$ 378.90

Figure 1.3

An example of a traditional spreadsheet with little to no art elements.

Source: Pam Baker.

Newer types of visualizations include interactive visualizations wherein more granular data is exposed as the user hovers a mouse or clicks on different areas in the visual; 3D visualizations that can be rotated on a computer screen for views from different angles and zoomed in to expose deeper information subsets; word clouds depicting the prominence of thoughts, ideas, or topics by word size; and other types of creative images.

Figure 1.4 is an example of an augmented reality image. Imagine using your phone, tablet, or wearable device and seeing your multi-dimensional data in an easy-to-understand form such as in this VisualCue tile. In this example, a waste management company is understanding the frequency, usage, and utility of their dump stations.

Figure 1.5 shows an example of a word cloud that quickly enables you to understand the prominence of ideas, thoughts, and occurrences as represented by word size. In this example, a word cloud was created on an iPad using the Infamous app to visualize news from several sites like FT, *Forbes*, *Fortune*, *The Economist*, *The Street*, and Yahoo! Finance. The size of the word denotes its degree of topic prominence in the news.



Figure 1.4

Augmented reality visualization. Imagine using your phone, tablet, or wearable device and seeing your multi-dimensional data in an easy to understand form such as in this VisualCue tile. In this example, a waste management company is understanding the frequency, usage, and utility of their dump stations.

Source: VisualCue™ Technologies LLC. Used with permission.