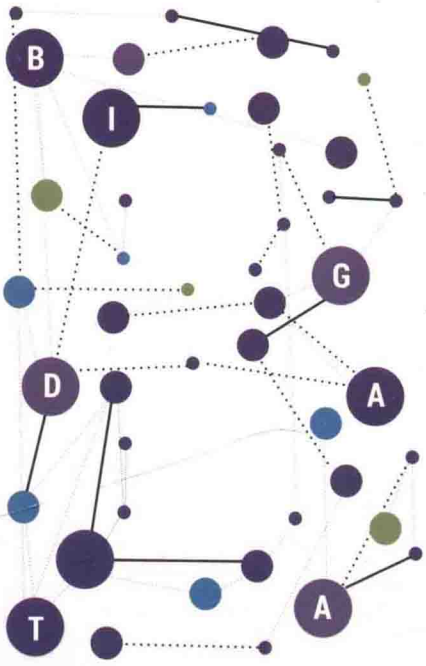
"For decades, all of the technologies that organizations used to measure and forecast their operations were a small niche in enterprise computing. That situation reversed itself a few years ago, and now the inevitable emergence of big data demands clear thinking and advice.

"Bill Schmarzo is the real deal. He shares his experience and know-how freely in a book that lays it out without hype."
—***Neil Raden,*** *CEO & Principal Analyst, Hired Brains Research*

Big Data

Bill Schmarzo

Understanding How Data Powers Big Business



WILEY

Big Data

Understanding How Data Powers Big Business

Bill Schmarzo

WILEY

Big Data: Understanding How Data Powers Big Business

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

Published simultaneously in Canada

ISBN: 978-1-118-73957-0
ISBN: 978-1-118-74003-3 (ebk)
ISBN: 978-1-118-74000-2 (ebk)

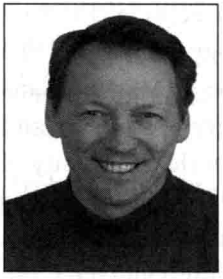
Manufactured in the United States of America

10 9 8 7 6 5 4 3 2 1

Library of Congress Control Number: 2013948011

Big Data

About the Author



Bill Schmarzo has nearly three decades of experience in data warehousing, Business Intelligence, and analytics. He was the Vice President of Analytics at Yahoo from 2007 to 2008. Prior to joining Yahoo, Bill oversaw the Analytic Applications business unit at Business Objects, Inc., including the development, marketing, and sales of their industry-defining analytic applications. Currently, Bill is the CTO of the Enterprise Information Management & Analytics Practice for EMC Global Services.

Bill is the creator of the Business Benefits Analysis methodology that links an organization's strategic business initiatives with their supporting data and analytic requirements. He has also co-authored with Ralph Kimball a series of articles on analytic applications. Bill has served on The Data Warehouse Institute's faculty as the head of the analytic applications curriculum. He has written several white papers and is a frequent speaker on the use of Big Data and advanced analytics to power an organization's key business initiatives.

His recent blogs can be found at `http://infocus.emc.com/author/william_schmarzo/`

You can also follow Bill on Twitter at @schmarzo.

About the Technical Editor

Denise Partlow has served in a wide variety of V.P. and Director of Product Marketing positions at both emerging and established technology companies. She has hands-on experience developing marketing strategies and "Go To Market" plans for complex product and service-based solutions across a variety of software and services companies. Denise has a B.S. in Computer Science from the University of Central Florida. She was a programmer of simulation and control systems as well as a program manager prior to transitioning into product management and marketing.

Denise is currently responsible for product marketing for EMC's big data and cloud consulting services. In that role, she collaborated with Bill Schmarzo on many of the concepts and viewpoints that have become part of *Big Data: Understanding How Data Powers Big Business.*

Credits

Executive Editor
Carol Long

Senior Project Editor
Adaobi Obi Tulton

Technical Editor
Denise Partlow

Production Editor
Daniel Scribner

Copy Editor
Christina Haviland

Editorial Manager
Mary Beth Wakefield

Freelancer Editorial Manager
Rosemarie Graham

Associate Director of Marketing
David Mayhew

Marketing Manager
Ashley Zurcher

Business Manager
Amy Knies

Production Manager
Tim Tate

Vice President and Executive Group Publisher
Richard Swadley

Vice President and Executive Publisher
Neil Edde

Associate Publisher
Jim Minatel

Project Coordinator, Cover
Katie Crocker

Proofreader
Sarah Kaikini, Word One

Indexer
Ron Strauss

Cover Image
Ryan Sneed

Cover Designer
Ryan Sneed

Acknowledgments

It's A Wonderful Life has always been one of my favorite movies. I always envisioned myself a sort of George Baily; someone who always looked for opportunities to give back. So whether it's been coaching youth sports, helping out with the school band, or even persuading my friend to build an ethanol plant in my hometown of Charles City, Iowa, I've always had this drive to give back.

When Carol Long from Wiley approached me about this book project, with the strong and supporting push from Denise Partlow of EMC, I thought of this as the perfect opportunity to give back—to take my nearly 30 years of experience in the data and analytics industry, and share my learnings from all of those years working with some of the best, most innovative people and organizations in the world.

I have been fortunate enough to have many *Forrest Gump* moments in my life—situations where I just happened to be at the right place at the right time for no other reason than luck. Some of these moments of serendipity include:

- One of the first data warehouse projects with Procter & Gamble when I was with Metaphor Computer Systems in the late 1980s.
- Head of Sales and Marketing at one of the original open source companies, Cygnus Support, and helping to craft a business model for making money with open source software.
- Creating and heading up Sequent Computer's data warehouse business in the late 1990s, creating one of the industry's first data warehouse appliances.
- VP of Analytic Applications at Business Objects in the 2000s, creating some of the industry's first analytic applications.
- Head of Advertising Analytics at Yahoo! where I had the great fortune to experience firsthand Yahoo!'s petabyte project, and use big data analytics to uncover the insights buried in all of that data to help Yahoo!'s advertisers optimize their spend across the Yahoo! ad network.
- A failed digital media startup, JovianDATA, where I experienced the power of cloud computing to bring unbelievable analytic power to bear on one of the digital media's most difficult problems—attribution analysis.
- And finally, my current stint as CTO of EMC Global Services' Enterprise Information & Analytics Management (EIM&A) service line, where my everyday job is to work with customers to identify where and how to start their big data journeys.

I hope that you see from my writing that I learned early in my career that technology is only interesting (and fun) when it is solving meaningful business problems and opportunities. The opportunity to leverage data and analytics to help clients make more money has always been the most interesting and fun part of my job.

I've always admired the teaching style of Ralph Kimball with whom I had the fortune to work with at Metaphor and again as a member of the Kimball Group. Ralph approaches his craft with very pragmatic, hands-on advice. Ralph (and his Kimball Group team of Margy Ross, Bob Becker, and Warren Thornthwaite) have willingly shared their learnings and observations with others through conferences, newsletters, webinars, and of course, their books. That's exactly what I wanted to do as well. So I've been actively blogging about my experiences the past few years, and the book seemed like a natural next step in packaging up my learnings, observations, techniques, and methodologies so that I could share with others.

There are many folks I would like to thank, but I was told that my acknowledgments section of the book couldn't be bigger than the book itself. So here we go with the short list.

- The Wiley folks—Carol Long, Christina Haviland, and especially Adaobi Obi Tulton—who reviewed my material probably more times than I did. They get the majority of the credit for delivering a readable book.
- Marc Demarest, Neil Raden and John Furrier for the great quotes. I hope the book lives up to them.
- Edd Dumbill and Alistair Croll from Strata who are always willing to give me time at their industry-leading data science conference to test my materials, and to the "Marc and Mark Show" (Marc Demarest and Mark Madsen) who also carve out time in their Strata track to allow me to blither on about the business benefits of big data.
- John Furrier and David Vellante from SiliconAngle and theCube who were the first folks to use the term "Dean of Big Data" to describe my work in the industry. They always find time for me to participate in their industry-leading, ESPN-like technology web broadcast show.
- Warren Thornthwaite who found time in his busy schedule to brainstorm and validate ideas and concepts from the book and provided countless words of encouragement about all things—book and beyond.

I'd like to thank my employer, EMC. EMC gave me the support and afforded me countless opportunities to spend time with our customers to learn about their big data challenges and opportunities. EMC was great in sharing materials including the data scientist certification course (which I discuss in Chapter 4) and the Big

Data Storymap (which I discuss in Chapter 12). EMC also gave me the time to write this book (mostly in airplanes as I flew from city to city).

I especially want to thank the customers over the past three decades with whom I have had the great fortune to work. They have taught me all that I share in this book and have been willing patients as we have tested and refined many of the techniques, tools, and methodologies outlined in this book.

I need to give special thanks to Denise Partlow, without whose support, encouragement, and demanding nature this book would never have gotten done. She devoted countless hours to reviewing every sentence in this book, sometimes multiple times, and arguing with me when my words and ideas made no sense. She truly was the voice of reason behind every idea and concept in this book. I couldn't ask for a better friend.

Of course, I want to thank my wife, Carolyn, and our kids, Alec, Max, and Amelia. You'll see several references to them throughout the book, such as Alec's (who is our professional baseball pitcher) help with baseball stats and insights. They have been very patient with me in my travels and time away from them. I know that a thank you in a book can't replace the missed nights tucking you into bed, long tossing on the baseball field or rebounding for you in the driveway, but thanks for understanding and being supportive.

Finally, I want to thank my Mom and Dad, who taught me the value of hard work and perseverance, and to never stop chasing my dreams. In particular, I want to thank my Mom, whose devotion to helping others motivated me to stick with this book even when I didn't feel like it. So in honor of my Mom, who passed away nearly 16 years ago, I will be dedicating proceeds from this book to breast cancer research, the disease that took her away from her family, friends, and her love of helping others. Mom, this book is for you.

Preface

Think Differently.

Your competitors are already taking advantage of big data, and furthermore, your traditional IT infrastructure is incapable of managing, analyzing and acting on big data.

Think Differently.

You should care about big data. The most significant impact big data can have on an organization is its ability to upgrade existing business processes and uncover new monetization opportunities. No organization can have too many insights about the key elements of their business, such as their customers, products, campaigns, and operations. Big data can uncover these insights at a lower level of granularity and in a more timely, actionable manner. Big data can power new business applications—such as personalized marketing, location-based services, predictive maintenance attribution analysis, and machine behavioral analytics. Big data holds the promise of rewiring an organization's value creation processes and creating entirely new, more compelling, and more profitable customer engagements. Big data is about business transformation, in moving your organization from retrospective, batch, business monitoring hindsights to predictive, real-time business optimization insights.

Think Differently.

Big data forces you to embrace a mentality of data abundance (versus data scarcity) and to grasp the power of analyzing all your data—both internally and externally of the organization—at the lowest levels of granularity in real-time. For example, the old business intelligence "slice and dice" analysis model, which worked well with gigabytes of data, is as outdated as the whip and buggy in an age of petabytes of data, thousands of scale-out processing nodes, and in-database analytics. Furthermore, standard relational database technology is unable to express the complex branching and iterative logic upon which big data analytics is based. You need an updated, modern infrastructure to take advantage of big data.

Think Differently.

Never has this message been more apropos than when dealing with big data. While much of the big data discussion focuses on Hadoop and other big data technology innovations, the real technology and business driver is the *big data economics*—the combination of open source data management and advanced analytics

software on top of commodity-based, scale-out architectures are 20 times cheaper than today's data warehouse architectures. This magnitude of economic change forces you to rethink many of the traditional data and analytic models. Data transformations and enrichments that were impossible three years ago are now readily and cheaply available, and the ability to mine petabytes of data across hundreds of dimensions and thousands of metrics on the cloud is available to all organizations, whether large or small.

Think Differently.

What's the biggest business pitfall with big data? Doing nothing. Sitting back. Waiting for your favorite technology vendor to solve these problems for you. Letting the technology-shifting sands settle out first. Oh, you've brought Hadoop into the organization, loaded up some data, and had some folks play with it. But this is no time for science experiments. This is serious technology whose value in creating new business models based on petabytes of real-time data coupled with advanced analytics has already been validated across industries as diverse as retail, financial services, telecommunications, manufacturing, energy, transportation, and hospitality.

Think Differently.

So what's one to do? Reach across the aisle as business and IT leaders and embrace each other. Hand in hand, identify your organization's most important business processes. Then contemplate how big data—in particular, detailed transactional (dark) data, unstructured data, real-time data access, and predictive analytics—could uncover actionable insights about your customers, products, campaigns, and operations. Use big data to make better decisions more quickly and more frequently, and uncover new monetization opportunities in the process. Leverage big data to "Make me more money!" Act. Get moving. Be bold. Don't be afraid to make mistakes, and if you fail, do it fast and move on. Learn.

Think Differently.

Introduction

Big data is today's technology hot topic. Such technology hot topics come around every four to five years and become the "must have" technologies that will lead organizations to the promised land—the "silver bullet" that solves all of our technology deficiencies and woes. Organizations fight through the confusion and hyperbole that radiate from vendors and analysts alike to grasp what the technology can and cannot do. In some cases, they successfully integrate the technology into the organization's technology landscape—technologies such as relational databases, Enterprise Resource Planning (ERP), client-server architectures, Customer Relationship Management (CRM), data warehousing, e-commerce, Business Intelligence (BI), and open source software.

However, big data feels different, maybe because at its heart big data is not about technology as much as it's about business transformation—transforming the organization from a retrospective, batch, data constrained, monitor the business environment into a predictive, real-time, data hungry, optimize the business environment. Big data isn't about business parity or deploying the same technologies in order to be like everyone else. Instead, big data is about leveraging the unique and actionable insights gleaned about your customers, products, and operations to rewire your value creation processes, optimize your key business initiatives, and uncover new monetization opportunities. Big data is about making money, and that's what this book addresses—how to leverage those unique and actionable insights about your customers, products, and operations to make money.

This book approaches the big data business opportunities from a pragmatic, hands-on perspective. There aren't a lot of theories here, but instead lots of practical advice, techniques, methodologies, downloadable worksheets, and many examples I've gained over the years from working with some of the world's leading organizations. As you work your way through this book, you will do and learn the following:

- Educate your organization on a common definition of big data and leverage the Big Data Business Model Maturity Index to communicate to your organization the specific business areas where big data can deliver meaningful business value (Chapter 1).
- Review a history lesson about a previous big data event and determine what parts of it you can apply to your current and future big data opportunities (Chapter 2).

- Learn a process for leveraging your existing business processes to identify the "right" metrics against which to focus your big data initiative in order to drive business success (Chapter 3).
- Examine some recommendations and learnings for creating a highly efficient and effective organizational structure to support your big data initiative, including the integration of new roles—like the data science and user experience teams, and new Chief Data Office and Chief Analytics Officer roles—into your existing data and analysis organizations (Chapter 4).
- Review some common human decision making traps and deficiencies, contemplate the ramifications of the "death of why," and understand how to deliver actionable insights that counter these human decision-making flaws (Chapter 5).
- Learn a methodology for breaking down, or functionally "decomposing," your organization's business strategy and key business initiatives into its key business value drivers, critical success factors, and the supporting data, analysis, and technology requirements (Chapter 6).
- Dive deeply into the big data Masters of Business Administration (MBA) by applying the big data business value drivers—underleveraged transactional data, new unstructured data sources, real-time data access, and predictive analytics—against value creation models such as Michael Porter's Five Forces Analysis and Value Chain Analysis to envision where and how big data can optimize your organization's key business processes and uncover new monetization opportunities (Chapter 7).
- Understand how the customer and product insights gleaned from new sources of customer behavioral and product usage data, coupled with advanced analytics, can power a more compelling, relevant, and profitable customer experience (Chapter 8).
- Learn an envisioning methodology—the Vision Workshop—that drives collaboration between business and IT stakeholders to envision what's possible with big data, uncover examples of how big data can impact key business processes, and ensure agreement on the big data desired end-state and critical success factors (Chapter 9).
- Learn a process for pulling together all of the techniques, methodologies, tools, and worksheets around a process for identifying, architecting, and delivering big data-enabled business solutions and applications (Chapter 10).
- Review key big data technologies (Hadoop, MapReduce, Hive, etc.) and analytic developments (R, Mahout, MADlib, etc.) that are enabling new data management and advanced analytics approaches, and explore the impact these technologies could have on your existing data warehouse and business intelligence environments (Chapter 11).

- Summarize the big data best practices, approaches, and value creation techniques into the Big Data Storymap—a single image that encapsulates the key points and approaches for delivering on the promise of big data to optimize your value creation processes and uncover new monetization opportunities (Chapter 12).
- Conclude by reviewing a series of "calls to action" that will guide you and your organization on your big data journey—from education and awareness, to the identification of where and how to start your big data journey, and through the development and deployment of big data-enabled business solutions and applications (Chapter 13).
- We will also provide materials for download on `www.wiley.com/go/bigdataforbusiness`, including the different envisioning worksheets, the Big Data Storymap, and a training presentation that corresponds with the materials discussed in this book.

The beauty of being in the data and analytics business is that we are only a new technology innovation away from our next big data experience. First, there was point-of-sale, call detail, and credit card data that provided an earlier big data opportunity for consumer packaged goods, retail, financial services, and telecommunications companies. Then web click data powered the online commerce and digital media industries. Now social media, mobile apps, and sensor-based data are fueling today's current big data craze in all industries—both business-to-consumer and business-to-business. And there's always more to come! Data from newer technologies, such as wearable computing, facial recognition, DNA mapping, and virtual reality, will unleash yet another round of big data-driven value creation opportunities.

The organizations that not only survive, but also thrive, during these data upheavals are those that embrace data and analytics as a core organizational capability. These organizations develop an insatiable appetite for data, treating it as an asset to be hoarded, not a business cost to be avoided. Such organizations manage analytics as intellectual property to be captured, nurtured, and sometimes even legally protected.

This book is for just such organizations. It provides a guide containing techniques, tools, and methodologies for feeding that insatiable appetite for data, to build comprehensive data management and analytics capabilities, and to make the necessary organizational adjustments and investments to leverage insights about your customers, products, and operations to optimize key business processes and uncover new monetization opportunities.

Contents