Big Data Analytics

Big Data Analytics

From Strategic Planning to Enterprise Integration with Tools, Techniques, NoSQL, and Graph

David Loshin

ELSEVIER

AMSTERDAM • BOSTON • HEIDELBERG • LONDON
NEW YORK • OXFORD • PARIS • SAN DIEGO
SAN FRANCISCO • SINGAPORE • SYDNEY • TOKYO
Morgan Kaufmann is an imprint of Elsevier

Morgan Kaufmann is an imprint of Elsevier
225 Wyman Street, Waltham, MA 02451, USA

First published 2013

Notices
Knowledge and best practice in this field are constantly changing. As new research and experience broaden our understanding, changes in research methods or professional practices, may become necessary.

Practitioners and researchers must always rely on their own experience and knowledge in evaluating and using any information or methods described herein. In using such information or methods they should be mindful of their own safety and the safety of others, including parties for whom they have a professional responsibility.

To the fullest extent of the law, neither the Publisher nor the authors, contributors, or editors, assume any liability for any injury and/or damage to persons or property as a matter of products liability, negligence or otherwise, or from any use or operation of any methods, products, instructions, or ideas contained in the material herein.

Library of Congress Cataloging-in-Publication Data
A catalog record for this book is available from the Library of Congress

British Library Cataloguing-in-Publication Data
A catalogue record for this book is available from the British Library

ISBN: 978-0-12-417319-4

Printed in the United States of America

13 14 15 16 17 10 9 8 7 6 5 4 3 2 1

FOREWORD

In the summer of 1995, I attended my first conference for information technology professionals. The event, called Interex, was an annual convention for users of the HP 3000, Hewlett-Packard's midrange business computer system known at the time for its reliability—and a devoted user base. More than 10,000 of these users gathered in Toronto that August to swap tips and pester HP executives for information about the future of their beloved system.

After spending several days talking with these IT managers, I came away with two observations:

1. The managers were grappling with rising technology demands from business executives and office workers who wanted more out of their IT investments.
2. While everyone in the Toronto Convention Center was talking about the HP 3000, there was a 100-foot-tall "WINDOWS '95" banner hanging from the nearby CN Tower, a prominent landmark in the city's skyline.

That banner was not part of the Interex event. But it would not be long before Microsoft's new desktop operating system would influence the work of just about everyone who used a PC.

It has always been this way. Innovations in computer hardware, communications technology, and software development regularly enter to challenge IT professionals to adapt to new opportunities and associated challenges. The latest edition of this recurrent dynamic is big data analytics, which takes advantage of advances in software programming, open source code, and commodity hardware to promise major gains in our ability to collect and analyze vast amounts of data—and new kinds of data—for fresh insights. The kinds of techniques that allow Google to index the Web, Facebook to build social graphs, and Netflix to recommend movies can be applied to functions like marketing (what's the next best offer for Marjorie?), risk management (the storm is tracking near our warehouse, better move the goods today), and equipment maintenance (the sensor says it's time to replace that engine part).

Those possibilities and many others have generated much interest. Venture capital is flowing to startups as database architects are cool again. Leaders in health care, finance, insurance, and other industries are racing to hire talented "data scientists" to develop algorithms to discover competitive advantages. Universities are launching master's programs in analytics in response to corporate demands and a projected skills gap. Statisticians are joining celebrity ranks, with one sparking cable news debates about presidential election predictions and another starring in TED Talk videos on data visualization design.

There is so much going on, in fact, that a busy IT professional looking for relevant help could use a personal guide to explain the issues in a style that acknowledges some important conditions about their world: They likely have a full list of ongoing projects. Their organization has well-defined IT management practices. They stipulate that adopting new technologies is not easy.

This is the kind of book you are reading now. David Loshin, an experienced IT consultant and author, is adept at explaining how technologies work and why they matter, without technical or marketing jargon. He has years of practice both posing and answering questions about data management, data warehousing, business intelligence, and analytics.

I know this because I have asked him. I first met David in 2012 at an online event he moderated to explain issues involved in making big data analytics work in business. I sought him out to discuss the issues in more detail as the editor of *Data Informed* (http://data-informed.com/) an online publication that chronicles these trends and shares best practices for IT and business professionals. Those early conversations led to David writing a series of articles for *Data Informed* that forms the basis for this book, on issues ranging from the market and business drivers for big data analytics, to use cases for these emerging technologies, to strategies for assessing their relevance to your organization.

Along the way, David and I have found ourselves agreeing about a key lesson from his years of working in IT (or, in my case, reporting on it): New big data analytics technologies are exciting, and represent a great opportunity. But making any new technology work effectively requires understanding the tools you need, having the right people

working together on common goals, and establishing the right business processes to create value from the work.

The teachings in this book go beyond this straightforward three-legged stool of technologies–skills–processes. At the end of each chapter, there are "thought exercises" that challenge you to consider the technology, business, and management concepts in the context of your organization. This is where David provides you the opportunity to answer the kinds of questions that will help you evaluate next steps for making the technologies covered here valuable to you.

These are like signposts to direct your work in adapting to the big data analytics field. It's much better than a 10-story banner blaring to a city that your world is about to change. Here, the signs come with full explanations and advice about how to make that change work for you.

Michael Goldberg

PREFACE

INTRODUCTION

In technology, it seems, what comes around goes around. At least in my experience, it certainly seems that way. Over recent times, the concepts of "big data" and "big data analytics" have become ubiquitous—it is heard to visit a web site, open a newspaper, or read a magazine that does not refer to one or both of those phrases. Yet the technologies that are incorporated into big data—massive parallelism, huge data volumes, data distribution, high-speed networks, high-performance computing, task and thread management, and data mining and analytics—are not new.

During the first phase of my career in the late 1980s and early 1990s I was a software developer for a company building programming language compilers for supercomputers. Most of these high-end systems were multiprocessor systems, employed massive parallelism, and were driven by (by the standards of the times, albeit) large data sets. My specific role was looking at code optimization, particularly focusing on increasing data bandwidth to the processors and taking advantage of the memory hierarchies upon which these systems were designed and implemented. And interestingly, much of the architectures and techniques used for designing hardware and developing software were not new either—much credit goes to early supercomputers such as the Illiac IV, the first massively parallel computing system that was developed in the early 1970s.

That is why the big data phenomenon is so fascinating to me: not the appearance of new technology, but rather how known technology finally comes into the mainstream. When the details of technology that was bleeding edge 20 years ago appear regularly in The New York Times, The Wall Street Journal, and The Economist, you know it has finally arrived.

THE CHALLENGE OF ADOPTING NEW TECHNOLOGY

Many people have a natural affinity to new technology—there is often the perception that the latest and shiniest silver bullet will not only eliminate all the existing problems in the organization will but also lead to the minting of a solid stream of gold coins enriching the entire organization. And in those organizations that are not leading the revolution to adoption, there is the lingering fear of abandonment—if they don't adopt the technology they will be left far behind, even if there is no clear value proposition for it in the first place.

Clearly, it would be unwise to commit to a new technology without assessing the components of its value—expected value driver "lift," as compared to the total cost of operations. Essentially, testing and piloting new technology is necessary to maintain competitiveness and ensure technical feasibility. But in many organizations, the processes to expeditiously mainstream new techniques and tools often bypass existing program governance and corporate best practices. The result is that pilot projects are prematurely moved into "production" are really just point solutions relying on islands of data that don't scale from the performance perspective nor fit into the enterprise from an architectural perspective.

WHAT THIS BOOK IS

The goal of this book is to provide a firm grounding in laying out a strategy for adopting big data techniques. It is meant to provide an overview of what big data is and why it can add value, what types of problems are suited to a big data approach, and how to properly plan to determine the need, align the right people in the organization, and develop a strategic plan for integration.

On the other hand, this book is not meant as a "how-to" for big data application development, MapReduce programming, or implementing Hadoop. Rather, my intent is to provide an overview within each chapter that addresses some pertinent aspect of the ecosystem or the process of adopting big data:

- Chapter 1: We consider the market conditions that have enabled broad acceptance of big data analytics, including commoditization

of hardware and software, increased data volumes, growing variation in types of data assets for analysis, different methods for data delivery, and increased expectations for real-time integration of analytical results into operational processes.

- Chapter 2: In this chapter, we look at the characteristics of business problems that traditionally have required resources that exceeded the enterprises' scopes, yet are suited to solutions that can take advantage of the big data platforms (either dedicated hardware or virtualized/cloud based).
- Chapter 3: Who in the organization needs to be involved in the process of acquiring, proving, and deploying big data solutions? And what are their roles and responsibilities? This chapter looks at the adoption of new technology and how the organization must align to integrate into the system development life cycle.
- Chapter 4: This chapter expands on the previous one by looking at some key issues that often plague new technology adoption and show that the key issues are not new ones and that there is likely to be organizational knowledge that can help in fleshing out a reasonable strategic plan.
- Chapter 5: In this chapter, we look at the need for oversight and governance for the data, especially when those developing big data applications often bypass traditional IT and data management channels.
- Chapter 6: In this chapter, we look at specialty-hardware designed for analytics and how they are engineered to accommodate large data sets.
- Chapter 7: This chapter discusses and provides a high-level overview of tool suites such as Hadoop.
- Chapter 8: This chapter examines the MapReduce programming model.
- Chapter 9: In this chapter, we look at a variety of alternative methods of data management methods that are being adopted for big data application development.
- Chapter 10: This chapter looks at business problems suited for graph analytics, what differentiates the problems from traditional approaches and considerations for discovery versus search analyses.
- Chapter 11: This short final chapter reviews best practices for incrementally adopting big data into the enterprise.

WHY YOU SHOULD BE READING THIS BOOK

You have probably picked up this book for one or more of these very good reasons:

- You are a senior manager seeking to take advantage of your organization's information to create or add to corporate value by increasing revenue, decreasing costs, improving productivity, mitigating risks, or improving the customer experience.
- You are the Chief Information Officer or Chief Data Officer of an organization who desires to make the best use of the enterprise information asset.
- You are a manager who has been asked to develop a big data program.
- You are a manager who has been asked to take over a floundering big data application.
- You are a manager who has been asked to take over a successful big data program.
- You are a senior business executive who wants to explore the value that big data can add to your organization.
- You are a business staff member who desires more insight into the way that your organization does business.
- You are a database or software engineer who has been appointed a technical manager for a big data program.
- You are a software engineer who aspires to be the manager of a big data program.
- You are an analyst of engineer working on a big data framework who aspires to replace your current manager.
- You are a business analyst who has been asked to join a big data application team.
- You are a senior manager and your directly reporting managers have started talking about big data using terminology you think they expect you to understand.
- You are a middle-level manager or engineer and your manager has started talking about big data using terminology you think they expect you to understand.
- You are just interested in nig.

How do I know so much about you? Because at many times in my life, I *was* you—either working on or managing a project for which I had some knowledge gaps, for an organization full of people not sure of why they were doing, what they were doing, with very few clear

success criteria or performance metrics. At times I would have loved to have had a straightforward book to consult for a quick lookup or a more in-depth read, without having to spend a huge amount of money on a technical book that only briefly addressed a topic of interest. And even more acutely, it is good to have an unbiased text to help differentiate marketing hype from reality.

OUR APPROACH TO KNOWLEDGE TRANSFER

As I have mentioned in the prefaces to my recent books ("Business Intelligence—The Savvy Manager's Guide, Second Edition," "Master Data Management," and "The Practitioner's Guide to Data Quality Improvement") I remain devoted to helping organizations strategically improve their capabilities in gaining the best advantage from what might be called "information utility." My prior experiences in failed data management activities drove me to quit my last "real job" (as I like to say) and start my own consulting practice to prove that there are better ways to organize and plan information-oriented program.

My company, Knowledge Integrity Inc. (www.knowledge-integrity.com), was developed to help organizations form successful high-performance computing, business intelligence, analytics, information quality, data governance, and master data management programs. As a way of distinguishing my effort from other consulting companies, I also instituted a few important corporate rules about the way we would do business:

1. Our mission was to develop and popularize methods for enterprise data management and utility. As opposed to the craze for patenting technology, methods, and processes, we would openly publish our ideas so as to benefit anyone willing to invest the time and energy to internalize the ideas we were promoting.
2. We would encourage clients to adopt our methods within their success patterns. It is a challenge (and perhaps in a way, insulting) to walk into an organization and tell people who have done their jobs successfully that they need to drop what they are doing and change every aspect of the way they work. We believe that every organization has its own methods for success, and our job is to craft a way to integrate performance-based information quality management into the existing organizational success structure.

3. We would not establish ourselves as permanent fixtures. We believe that information management is a core competency that should be managed within the organization, and our goal for each engagement is to establish the fundamental aspects of the program, transfer technology to internal resources, and then be on our way. I often say that if we do our job right, we work ourselves out of a contract.
4. We are not "selling a product," we are engaged to solve customer problems. We are less concerned about rigid compliance to a trademarked methodology than we are about making sure that the customer's core issues are resolved, and if that means adapting our methods to the organization's that is the most appropriate way to get things done. I also like to say that we are successful when the client comes up with our ideas.
5. Effective communication is the key to change management. Articulating how good information management techniques enhance organizational effectiveness and performance is the first step in engaging business clients and ensuring their support and sponsorship. We would invest part of every engagement in establishing a strong business case accompanied by collateral information that can be socialized within and across the enterprise.

With these rules in mind, our first effort was to consolidate our ideas for semantic, rule-oriented data quality management in a book, "Enterprise Knowledge Management—The Data Quality Approach," which was published in 2001 by Morgan Kaufmann. I have been told by a number of readers that the book is critical in their development of a data quality management program, and the new technical ideas proposed for rule-based data quality monitoring have, in the intervening years, been integrated into all of the major data quality vendor product suites.

Since 1999, we have developed a graduate-level course on data quality for New York University, multiple day-courses for The Data Warehousing Institute (www.tdwi.org), presented numerous sessions at conferences and chapter meetings for DAMA (the Data Management Association), course and online content for DATAVERSITY (www.dataversity.net), provided columns for Robert Seiner's Data Administration Newsletter (www.tdan.com), monthly columns for DM Review (www.dmreview.com), a downloadable course on data quality

from Better Management (www.bettermanagement.com), and hosting an expert channel and monthly newsletter at the Business Intelligence Network (www.b-eye-network.com) and TechTarget (www.TechTarget.com).

We are frequently asked by vendors across the spectrum to provide analysis and thought leadership in many areas of data management. We have consulted in the public sector for both federal, state, and other global government agencies. We have guided data management projects in a number of industries, including government, financial services, health care, manufacturing, energy services, insurance, and social services, among others.

Since we started the company, the awareness of the value of information management has been revealed to be one of the most important topics that senior management faces. In practices that have emerged involving the exploitation of enterprise data, such as Enterprise Resource Planning (ERP), Supply Chain Management (SCM), and Customer Relationship Management (CRM), there is a need for a consolidated view of high-quality data representing critical views of business information. Increased regulatory oversight, increased need for information exchange, business performance management, and the value of service-oriented architecture are driving a greater focus on performance-oriented management of enterprise data with respect to utility: accessibility, consistency, currency, freshness, and usability of a common information asset.

CONTACT ME

While my intention is that this book will provide a guide to a strategic plan for big data, there are situations where some expert advice helps get the ball rolling. The practices and approaches described in this book are abstracted from numerous real client engagements, and our broad experience may be able to jump-start your mission for deploying a big data application. In the spirit of openness, I am always happy to answer questions, provide some additional details, and hear feedback about the approaches that I have put in this book and that Knowledge Integrity has employed successfully with our clients since 1999.

We are always looking for opportunities to help organizations establish the value proposition, develop the blueprint, roadmaps, and

program plan, and help in implementing the business intelligence and information utilization strategy, and would welcome any opportunities to share ideas and seek out ways we can help your organization. I mean it, I really want to hear from you.

I can be reached via my e-mail address, loshin@knowledge-integrity.com, or through Knowledge Integrity's company web site, www.knowledge-integrity.com, via www.davidloshin.info, or through the web site I have set up for this book, www.bigdataliteracy.com.

David Loshin

ACKNOWLEDGMENTS

What is presented in this book is a culmination of years of experience in projects and programs associated with best practices in employing data management tools, techniques, processes, and working with people. A number of people were key contributors to the development of this book, and I take this opportunity to thank them for their support.

First of all, my wonderful wife Jill deserves the most credit for perseverance and for her encouragement in completing the book. I also must thank my children, Kira, Jonah, Brianna, Gabriella, and Emma, for their help as well.

Much of the material in this book has been adapted from a series of articles that I wrote for a great web site, www.data-informed.com. Editor Michael Goldberg and publisher Michael Nadeau were instrumental in motivating, editing, and enabling the publication of this book.

Abie Reifer has provided insight, guidance, and suggestions for improving the content.

Critical parts of this book were inspired by works that I was commissioned to assemble for vendors in the big data space such as HP/Vertica, SAP/Sybase, ParAccel/Pervasive, Teradata, and YarcData, as well as material presented through my expert channel at www.b-eye-network.com, adapted from presentation material at conferences hosted by Wilshire Conferences, DATAVERSITY, DebTech International, The Data Warehousing Institute (www.tdwi.org), and vendor-hosted webinars and live events.

CHAPTER 1

Market and Business Drivers for Big Data Analytics

1.1 SEPARATING THE BIG DATA REALITY FROM HYPE

There are few technology phenomena that have taken both the technical and the mainstream media by storm than "big data." From the analyst communities to the front pages of the most respected sources of journalism, the world seems to be awash in big data projects, activities, analyses, and so on. However, as with many technology fads, there is some murkiness in its definition, which lends to confusion, uncertainty, and doubt when attempting to understand how the methodologies can benefit the organization.

Therefore, it is best to begin with a definition of big data. The analyst firm Gartner can be credited with the most-frequently used (and perhaps, somewhat abused) definition:

> Big data *is high-volume, high-velocity and high-variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making.*[1]

For the most part, in popularizing the big data concept, the analyst community and the media have seemed to latch onto the alliteration that appears at the beginning of the definition, hyperfocusing on what is referred to as the "3 Vs—volume, velocity, and variety." Others have built upon that meme to inject additional Vs such as "value" or "variability," intended to capitalize on an apparent improvement to the definition.

The ubiquity of the Vs definition notwithstanding, it is worth noting that the origin of the concept is not new, but was provided by (at the time Meta Group, now Gartner) analyst Doug Laney in a research note from 2001 about "3-D Data Management," in which he noted:

[1]Gartner's IT Glossary. Accessed from <http://www.gartner.com/it-glossary/big-data/> (Last accessed 08-08-13).

Big Data Analytics. DOI: http://dx.doi.org/10.1016/B978-0-12-417319-4.00001-6

> *While enterprises struggle to consolidate systems and collapse redundant databases to enable greater operational, analytical, and collaborative consistencies, changing economic conditions have made this job more difficult. E-commerce, in particular, has exploded data management challenges along three dimensions: volumes, velocity and variety. In 2001/02, IT organizations must compile a variety of approaches to have at their disposal for dealing with each.*[2]

The challenge with Gartner's definition is twofold. First, the impact of truncating the definition to concentrate on the Vs effectively distils out two other critical components of the message:

1. "cost-effective innovative forms of information processing" (the means by which the benefit can be achieved);
2. "enhanced insight and decision-making" (the desired outcome).

The second is a bit subtler: the definition is not really a definition, but rather a description. People in an organization cannot use the definition to determine whether they are using big data solutions or even if they have problems that need a big data solution. The same issue impedes the ability to convey a value proposition because of the difficulty in scoping what is intended to be designed, developed, and delivered and what the result really means to the organization.

Basically, it is necessary to look beyond what is essentially a marketing definition to understand the concept's core intent as the first step in evaluating the value proposition. Big data is fundamentally about applying innovative and cost-effective techniques for solving existing and future business problems whose resource requirements (for data management space, computation resources, or immediate, in-memory representation needs) exceed the capabilities of traditional computing environments as currently configured within the enterprise. Another way of envisioning this is shown in Figure 1.1.

To best understand the value that big data can bring to your organization, it is worth considering the market conditions that have enabled its apparently growing acceptance as a viable option to supplement the intertwining of operational and analytical business application in light of exploding data volumes. Over the course of this book, we hope to

[2]Doug Laney. Deja VVVu: others claiming Gartner's construct for big data, January 2012. Accessed from <http://blogs.gartner.com/doug-laney/deja-vvvue-others-claiming-gartners-volume-velocity-variety-construct-for-big-data/>.

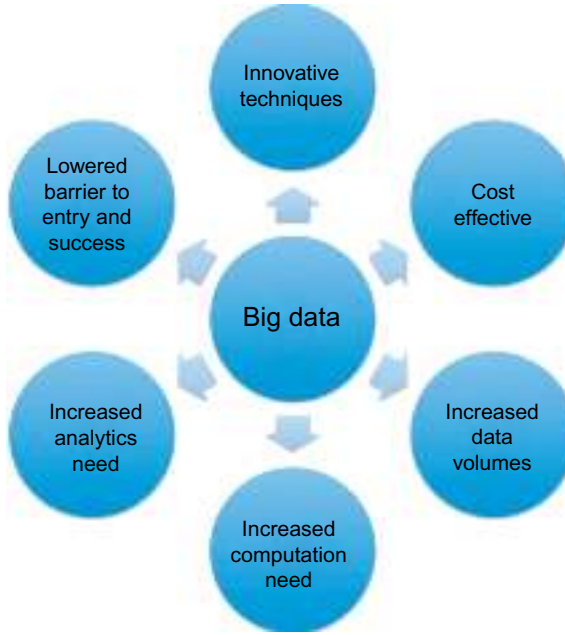

Figure 1.1 Cracking the big data nut.

quantify some of the variables that are relevant in evaluating and making decisions about integrating big data as part of an enterprise information management architecture, focusing on topics such as:

- characterizing what is meant by "massive" data volumes;
- reviewing the relationship between the speed of data creation and delivery and the integration of analytics into real-time business processes;
- exploring reasons that the traditional data management framework cannot deal with owing to growing data variability;
- qualifying the quantifiable measures of value to the business;
- developing a strategic plan for integration;
- evaluating the technologies;
- designing, developing, and moving new applications into production.

Qualifying the business value is particularly important, especially when the forward-looking stakeholders in an organization need to effectively communicate the business value of embracing big data platforms, and correspondingly, big data analytics. For example, a business

justification might show how incorporating a new analytics framework can be a competitive differentiator. Companies that develop customer upselling profiles based on limited data sampling face a disadvantage when compared to enterprises that create comprehensive customer models encompassing *all* the data about the customer intended to increase revenues while enhancing the customer experience.

Adopting a technology as a knee-jerk reaction to media buzz has a lowered chance of success than assessing how that technology can be leveraged along with the existing solution base as away of transforming the business. For that reason, before we begin to explore the details of big data technology, we must probe the depths of the business drivers and market conditions that make big data a viable alternative within the enterprise.

1.2 UNDERSTANDING THE BUSINESS DRIVERS

The story begins at the intersection of the need for agility and the demand for actionable insight as the proportion of signal to noise decreases. Decreasing "time to market" for decision-making enhancements to all types of business processes has become a critical competitive differentiator. However, the user demand for insight that is driven by ever-increasing data volumes must be understood in the context of organizational business drivers to help your organization appropriately adopt a coherent information strategy as a prelude to deploying big data technology.

Corporate business drivers may vary by industry as well as by company, but reviewing some existing trends for data creation, use, sharing, and the demand for analysis may reveal how evolving market conditions bring us to a point where adoption of big data can become a reality.

Business drivers are about agility in utilization and analysis of collections of datasets and streams to create value: increase revenues, decrease costs, improve the customer experience, reduce risks, and increase productivity. The data explosion bumps up against the requirement for capturing, managing, and analyzing information. Some key trends that drive the need for big data platforms include the following:

- **Increased data volumes being captured and stored**: According to the 2011 IDC Digital Universe Study, "In 2011, the amount of

information created and replicated will surpass 1.8 zettabytes, … growing by a factor of 9 in just five years."[3] The scale of this growth surpasses the reasonable capacity of traditional relational database management systems, or even typical hardware configurations supporting file-based data access.

- **Rapid acceleration of data growth**: Just 1 year later, the 2012 IDC Digital Universe study ("The Digital Universe in 2020") postulated, "From 2005 to 2020, the digital universe will grow by a factor of 300, from 130 exabytes to 40,000 exabytes, or 40 trillion gigabytes (more than 5,200 gigabytes for every man, woman, and child in 2020). From now until 2020, the digital universe will about double every two years."[4]
- **Increased data volumes pushed into the network**: According to Cisco's annual Visual Networking Index Forecast, by 2016, annual global IP traffic is forecast to be 1.3 zettabytes.[5] This increase in network traffic is attributed to the increasing number of smartphones, tablets and other Internet-ready devices, the growing community of Internet users, the increased Internet bandwidth and speed offered by telecommunications carriers, and the proliferation of Wi-Fi availability and connectivity. More data being funneled into wider communication channels create pressure for capturing and managing that data in a timely and coherent manner.
- **Growing variation in types of data assets for analysis**: As opposed to the more traditional methods for capturing and organizing *structured* datasets, data scientists seek to take advantage of unstructured data accessed or acquired from a wide variety of sources. Some of these sources may reflect minimal elements of structure (such as Web activity logs or call detail records), while others are completely unstructured or even limited to specific formats (such as social media data that merges text, images, audio, and video content). To extract usable signal out of this noise, enterprises must enhance their existing structured data management approaches to accommodate semantic text and content-stream analytics.
- **Alternate and unsynchronized methods for facilitating data delivery**: In a structured environment, there are clear delineations of the

[3]2011 IDC Digital Universe Study: extracting value from chaos, <http://www.emc.com/collateral/demos/microsites/emc-digital-universe-2011/index.htm>.

[4]The Digital Universe in 2020, <http://www.emc.com/collateral/analyst-reports/idc-the-digital-universe-in-2020.pdf>.

[5]See Cisco Press Release of May 30, 2012, <http://newsroom.cisco.com/press-release-content?type=webcontent&articleId=888280>.

discrete tasks for data acquisition or exchange, such as bulk file transfers via tape and disk storage systems, or via file transfer protocol over the Internet. Today, data publication and exchange is full of unpredictable peaks and valleys, with data coming from a broad spectrum of connected sources such as websites, transaction processing systems, and even "open data" feeds and streams from government sources and social media networks like Twitter. This creates new pressures for rapid acquisition, absorption, and analysis while retaining currency and consistency across the different datasets.

- **Rising demand for real-time integration of analytical results**: There are more people—with an expanding variety of roles—who are consumers of analytical results. The growth is especially noticeable in companies where end-to-end business processes are augmented to fully integrate analytical models to optimize performance. As an example, a retail company can monitor real-time sales of tens of thousands of Stock Keeping Units (SKUs) at hundreds of retail locations, and log minute-by-minute sales trends. Delivering these massive datasets to a community of different business users for simultaneous analyses gives new insight and capabilities that never existed in the past: it allows buyers to review purchasing patterns to make more precise decisions regarding product catalog, product specialists to consider alternate means of bundling items together, inventory professionals to allocate shelf space more efficiently at the warehouse, pricing experts to instantaneously adjust prices at different retail locations directly at the shelf, among other uses. The most effective uses of intelligence demand that analytical systems must process, analyze, and deliver results within a defined time window.

1.3 LOWERING THE BARRIER TO ENTRY

Enabling business process owners to take advantage of analytics in many new and innovative ways has always appeared to be out of reach for most companies. And the expanding universe of created information has seemed to tantalizingly dangle broad-scale analytics capabilities beyond the reach of those but the largest corporations.

Interestingly, for the most part, much of the technology classified as "big data" is not new. Rather, it is the ability to package these techniques in ways that are accessible to organizations in ways that up until recently had been limited by budget, resource, and skills constraints, which are typical of smaller businesses. What makes the big data

concept so engaging is that emerging technologies enable a broad-scale analytics capability with a relatively low barrier to entry.

As we will see, facets of technology for business intelligence and analytics have evolved to a point at which a wide spectrum of businesses can deploy capabilities that in the past were limited to the largest firms with equally large budgets. Consider the four aspects in Table 1.1.

The changes in the environment make big data analytics attractive to all types of organizations, while the market conditions make it practical. The combination of simplified models for development, commoditization, a wider palette of data management tools, and low-cost utility computing has effectively lowered the barrier to entry, enabling a much wider swath of organizations to develop and test out

Table 1.1 Contrasting Approaches in Adopting High-Performance Capabilities

Aspect	Typical Scenario	Big Data
Application development	Applications that take advantage of massive parallelism developed by specialized developers skilled in high-performance computing, performance optimization, and code tuning	A simplified application execution model encompassing a distributed file system, application programming model, distributed database, and program scheduling is packaged within Hadoop, an open source framework for reliable, scalable, distributed, and parallel computing
Platform	Uses high-cost massively parallel processing (MPP) computers, utilizing high-bandwidth networks, and massive I/O devices	Innovative methods of creating scalable and yet elastic virtualized platforms take advantage of clusters of commodity hardware components (either cycle harvesting from local resources or through cloud-based utility computing services) coupled with open source tools and technology
Data management	Limited to file-based or relational database management systems (RDBMS) using standard row-oriented data layouts	Alternate models for data management (often referred to as NoSQL or "Not Only SQL") provide a variety of methods for managing information to best suit specific business process needs, such as in-memory data management (for rapid access), columnar layouts to speed query response, and graph databases (for social network analytics)
Resources	Requires large capital investment in purchasing high-end hardware to be installed and managed in-house	The ability to deploy systems like Hadoop on virtualized platforms allows small and medium businesses to utilize cloud-based environments that, from both a cost accounting and a practical perspective, are much friendlier to the bottom line

high-performance applications that can accommodate massive data volumes and broad variety in structure and content.

1.4 CONSIDERATIONS

While the market conditions suggest that there is a lowered barrier to entry for implementing big data solutions, it does not mean that implementing these technologies and business processes is a completely straightforward task. There is a steep learning curve for developing big data applications, especially when going the open source route, which demands an investment in time and resources to ensure the big data analytics and computing platform are ready for production. And while it is easy to test-drive some of these technologies as part of an "evaluation," one might think carefully about some key questions before investing a significant amount of resources and effort in scaling that learning curve, such as:

- **Feasibility**: Is the enterprise aligned in a way that allows for new and emerging technologies to be brought into the organization, tested out, and vetted without overbearing bureaucracy? If not, what steps can be taken to create an environment that is suited to the introduction and assessment of innovative technologies?
- **Reasonability**: When evaluating the feasibility of adopting big data technologies, have you considered whether your organization faces business challenges whose resource requirements exceed the capability of the existing or planned environment? If not currently, do you anticipate that the environment will change in the near-, medium- or long-term to be more data-centric and require augmentation of the resources necessary for analysis and reporting?
- **Value**: Is there an expectation that the resulting quantifiable value that can be enabled as a result of big data warrants the resource and effort investment in development and productionalization of the technology? How would you define clear measures of value and methods for measurement?
- **Integrability**: Are there any constraints or impediments within the organization from a technical, social, or political (i.e., policy-oriented) perspective that would prevent the big data technologies from being fully integrated as part of the operational architecture? What steps need to be taken to evaluate the means by which big data can be integrated as part of the enterprise?

- **Sustainability**: While the barrier to entry may be low, the costs associated with maintenance, configuration, skills maintenance, and adjustments to the level of agility in development may not be sustainable within the organization. How would you plan to fund continued management and maintenance of a big data environment?

In Chapter 2, we will begin to scope out the criteria for answering these questions as we explore the types of business problems that are suited to a big data solution.

1.5 THOUGHT EXERCISES

Here are some questions and exercises to ponder before jumping headfirst into a big data project:

- What are the sizes of the largest collections of data to be subjected to capture, storage, and analysis within the organization?
- Detail the five most challenging analytical problems facing your organization. How would any of these challenges be addressed if the volume of data is increased by a factor of 10 and 100, respectively?
- Provide your own definition of what big data means to your organization.
- Develop a justification for big data within your organization in one sentence.
- Develop a single graphic image depicting what you believe to be the impact of increased data volumes and variety.
- Identify three "big data" sources, either within or external to your organization that would be relevant to your business.