Service Systems and Innovations in Business and Society Collection

Jim Spohrer and Haluk Demirkan, *Editors*

Obtaining Value from Big Data for Service Delivery

Stephen H. Kaisler
Frank Armour
J. Alberto Espinosa
William H. Money

BEP BUSINESS EXPERT PRESS

Obtaining Value from Big Data for Service Delivery

Obtaining Value from Big Data for Service Delivery

Stephen H. Kaisler, Frank Armour,
J. Alberto Espinosa, and William H. Money

BEP BUSINESS EXPERT PRESS

Obtaining Value from Big Data for Service Delivery

First published in 2016 by
Business Expert Press, LLC
222 East 46th Street, New York, NY 10017
www.businessexpertpress.com

ISBN-13: 978-1-63157-222-7 (paperback)
ISBN-13: 978-1-63157-223-4 (e-book)

Business Expert Press Service Systems and Innovations in Business and Society Collection

Collection ISSN: 2326-2664 (print)
Collection ISSN: 2326-2699 (electronic)

Cover and interior design by Exeter Premedia Services Private Ltd., Chennai, India

First edition: 2016

10 9 8 7 6 5 4 3 2 1

Printed in the United States of America.

I would like to dedicate this book to my wife, Chryl, who has encouraged me and supported me in its preparation.

—Stephen H. Kaisler

I would like to thank my wife, Delphine Clegg, for her support and encouragement in preparing this book.

—J. Alberto Espinosa

To my wife, Rose, my parents, my children, and my grandchild: you make everything beautiful.

—Frank Armour

To my wonderful wife, Libby, whose continuous support and strength allow me to devote more energy to research and writing, and to my daughter, Katy, whose enduring dedication and commitment have taught me how to successfully close a project.

—William H. Money

Abstract

Big data is an emerging phenomenon that has enormous implications and impacts upon business strategy, profitability, and process improvements. All service systems generate big data these days, especially human-centered service systems such as government (including cities), healthcare, education, retail, finance, and so on. It has been characterized as the collection, analysis and use of data characterized by the five Vs: volume, velocity, variety, veracity, and value (of data). As the plethora of data sources grows from sensors, social media, and electronic transactions, new methods for collecting or acquiring, integrating, processing, analyzing, understanding, and visualizing data to provide actionable information and support integrated and timely senior and executive decision-making are required. The discipline of applying analytic processes to find and combine new sources of data and extract hidden crucial decision-making information from the oceans of data is rapidly developing, but requires expertise to apply in ways that will yield useful, actionable results for service organizations. Many service-oriented organizations that are just beginning to invest in big data collection, storage, and analysis need to address the numerous issues and challenges that abound—technological, managerial, and legal. Other organizations that have begun to use new data tools and techniques must keep up with the rapidly changing and snowballing work in the field. This booklet will help middle, senior, and executive managers to understand what big data is; how to recognize, collect, process, and analyze it; how to store and manage it; how to obtain useful information from it; and how to assess its contribution to operational, tactical, and strategic decision-making in service-oriented organizations.

Keywords

analytic science, Big Data, business analytics, business intelligence, data science, descriptive analytics, enterprise architecture, NoSQL, predictive analytics, service delivery, service-oriented architecture

Contents

Purpose

This booklet is directed to senior executives and managers who need to understand the basic principles of Big Data as it is used to support service delivery, the tools and technology to develop and implement a Big Data support group within one's own organization, the roles and skills of personnel who will comprise such a group, and some of the challenges they will face in deploying and operating such an organization.

About the Contributors

J. Alberto Espinosa is currently the Chair of the Information Technology Department, a Full Professor and a Kogod Research Professor at the Kogod School of Business, American University. He holds PhD and Master of Science degrees in Information Systems from the Tepper School of Business at Carnegie Mellon University, a Master's degree in Business Administration from Texas Tech University, and a Mechanical Engineering degree from Pontificia Universidad Catolica, Peru. His research focusses on coordination and performance in global technical projects across global boundaries, particularly distance and time separation (e.g., time zones). His work has been published in leading scholarly journals, including Management Science, Organization Science, Information Systems Research, the *Journal of Management Information Systems*, IEEE Transactions on Software Engineering, IEEE Transactions on Engineering Management and Communications of the ACM.

J. Alberto Espinosa, PhD
Professor and Chair
Kogod School of Business
American University
Washington, DC
alberto@american.edu

William H. Money, Associate Professor, School of Business, The Citadel, Charleston, SC joined the Citadel in August 2014. Dr. Money served as an Associate Professor of Information Systems at the George Washington University, and as the director of the Executive Master of Science in Information Systems program. He joined the George Washington University, School of Business and Public Management faculty in September 1992 after acquiring over 12 years of management experience in the design, development, installation, and support of management information systems (1980–1992). His publications over the last 6 years and recent research interests focus on collaborative solutions to complex business problems; business process engineering and analytics; and information system development, collaboration, and workflow tools and methodologies. Previous teaching experience includes Purdue, Kent State, and American Universities. Dr. Money's academic training includes the PhD, Organizational Behavior/Systems Engineering, 1977, Northwestern University, Graduate School of Management; the MBA, Management, 1969, Indiana University; and a BA, Political Science, 1968, University of Richmond. Dr. Money has had numerous speaking engagements at professional meetings and publishes in information systems and management journals. Dr. Money has significant consulting experience in private, federal organizations, including the Department of State, D.C. Government, Department of Transportation, Coast Guard, and Department of the Navy.

William H. Money, PhD
Associate Professor
School of Business Administration
The Citadel
Charleston, SC
wmoney@citadel.edu

Frank Armour is an assistant professor of information technology at the Kogod School of Business, American University and is the faculty program director for the MS in Analytics degree program. He received his PhD from the Volgenau School of Engineering at George Mason University. He is also an independent senior IT consultant and has over 25 years of extensive experience in both the practical and academic aspects applying advanced information technology. He has led initiatives on, and performed research in: business analytics, big data, enterprise architectures, business and requirements analysis, agile system development cycle development (SDLC), and object-oriented development. He is the coauthor of the book, Advanced Use Case Modeling, Addison-Wesley, and is the author or coauthor of over 30 papers in the information technology discipline. He is primary co-chair for the enterprise architecture minitracks at both the HICSS and AMCIS conferences.

Frank J. Armour, PhD
Kogod School of Business
American University
Washington, DC
farmour@american.edu

Stephen H. Kaisler has been a senior scientist and senior software architect with several small technology and consulting firms over the past 10 years, where he has focused on machine learning, big data and advanced analytics, natural language processing, video processing, and enterprise architecture. Previously, he was Director of Systems Architecture and Technical Advisor to the Sergeant At Arms of the U.S. Senate. Prior to that position, he served as a Science Advisor consulting for the Internal Revenue Service, Chief Scientist for Analytics, and a Program Manager in Strategic Computing in the Defense Advanced Research Projects Agency. Dr. Kaisler has previously published four books and over 35 technical papers. He has taught part-time at George Washington University in the Depts. Of Computer Science and Information System Technology Management for over 35 years, where he is an Adjunct Professor of Engineering.

He is co-chair for the Enterprise Architecture minitrack and primary co-chair of the Big Data and Analytics minitrack at HICSS.

Stephen H. Kaisler, DSc
SHK and Associates
Principal
Laurel, M2 20723
Skaisler1@comcast.net

Acknowledgment

First, we would like to thank Haluk Demirkan for providing us with the opportunity to prepare this book and Jim Spohrer for his encouragement in seeing it to completion. We would also like to thank Haluk and Christer Carlsson for the opportunity to organize a Big Data minitrack and to provide encouragement and support for our tutorials on Big Data and Advanced Analytics at the 46th through 49th Hawaii International Conferences on Systems Science (HICSS). We also thank the many paper presenters and attendees at our minitrack and tutorials who have provided many insights and cogent comments on Big Data and Analytics. Finally, we wish to thank the staff, especially Kumar, of Business Expert Press for their efforts in producing this book.

List of Acronyms

ACO	Accountable Care Organization
AnBOK	Analytics Body of Knowledge
ANOVA	Analysis of Variance
API	Application Programming Interface
APT	Advanced Persistent Threats
B2B	Business-to-Business
BDA	Big Data Analytics
BI	Business Intelligence
CAO	Chief Analytics Officer
CAP	Certified Analytics Professional
CDO	Chief Data Officer
CHIP	Children's Health Insurance Program
CIFS	Common Internet File System
CMM	Capability Maturity Model
CMMS	Centers for Medicare and Medicaid Services
CSV	Comma Separated Values
DARPA	Defense Advanced Research Projects Agency
DB	Data Base
DBMS	Data Base Management System
EA	Enterprise Architecture
ETL	Extract, Transform, Load
FEAF	Federal Enterprise Architecture Framework
GPU	Graphics Processing Unit

HHS	Health and Human Services
HPC	High Performance Computing
HTML	HyperText Markup Language
IBM	International Business Machines Corporation
IDC	International Data Corporation
IDE	Interactive Development Environment
IDS	Intrusion Detection System
INFORMS	Institute for Operations Research and Management Sciences
IoT	Internet of Things
IPv6	Internet protocol, version 6
IT	Information Technology
LAMP	Linux-Apache-MySQL-Perl/PHP/Python
LEAP	Linux-Eucalyptus-AppScale-Python
MAMP	MacOS-Apache-MySQL-Perl/PHP/Python
MOOC	Massive Open Online Course
MPI	Message Passing Interface
NAP	National Academies Press
NFL	National Football League
NFS	Network File System
NIST	National Institute of Science and Technology
NLP	Natural Language Processing
NSF	National Science Foundation
OLAP	OnLine Analytical Processing
OMB	Office of Management and Budget
OODA	Observe, Orient, Decide, Act
OSGI	Open Source Gateway Initiative
OSS	Open Source Software
PhD	Doctor of Philosophy

RDBMS	Relational Data Base Management System
RETMA	Radio Electronics Television Manufacturers Association
RFID	Radio Frequency Identification
RTBDA	Real-Time Big Data Analytics
SDM	Service Delivery Model
SLA	Service Level Agreement
SOA	Service Oriented Architecture
SOC	Security Operations Center
SSN	Social Security Number
STEM	Science, Technology, Engineering and Mathematics
TOGAF	The Open Group Architectural Framework
UPC	Universal Product Code
VPN	Virtual Private Network
WAMP	WindowsOS-Apache-MySQL-Perl/PHP/Python
WEKA	Waikato Environment for Knowledge Analysis
WSJ	Wall Street Journal
XML	eXtended Markup Language
ZB	Zettabytes

CHAPTER 1
Introduction

The Internet, the World Wide Web, and the concept of service delivery have revolutionized the way commercial, academic, governmental, and nongovernmental organizations deal with their supplies and their clients and customers. Individuals and organizations are overwhelmed with data produced by IT systems that are so pervasive throughout society, government, and business. The wide variety and huge numbers of data sources including sensors, cell phones, tablets, and other devices is increasing at a seemingly exponential rate. Estimates (2010) were that all sources of data, including replicated data such as retweets and resends of e-mail, amount to tens of exabytes per month—that is 10^{18} or 1,000,000,000,000,000,000 bytes. The numbers are staggering, and, obviously, no one knows for sure. In 2012, the International Data Corporation (IDC) stated there were 2.8 zettabytes (ZB) and forecasted that we will generate 40 ZB by 2020 (http://www.webopedia.com/quick_ref/just-how-much-data-is-out-there.html). Our data generation is growing exponentially.

Individuals and organizations do not actively collect, own, process, or analyze this much data themselves. However, many individuals and organizations acquire and deal with gigabytes of data, and many organizations utilize terabytes and petabytes of data per year. Senior executives and managers in government, academia, and business operations are grappling with the deluge of data available to them and trying to make sense—decisions and conclusions based upon it. A critical area is how to collect, organize, process, store, and analyze this flood of data in order to deliver superior service to their client and customer base—both internal and external.

Much of this data is generated from the services sector of the economy: health, manufacturing, marketing, telecommunications, and so on. To address this wealth of data and the underlying technology and practices, IBM pioneered the term *service science* to encompass the broad spectrum of business, teaching, and research expertise to develop the capabilities to sustain and advance the services environment that we live in today. Advances in technology have made large volumes of data available to users and providers within the services environment. This volume of data has come to be called Big Data and has its own business, teaching, and research expertise associated with it.

This book will describe how coordinating and integrating the expertise between the services environment and the Big Data environment has and is leading to enhanced service delivery to customers and clients and increasing revenue and profit in many industries. However, despite the attention given in the popular press and the blog-o-sphere, many more opportunities exist and even more will be invented so more organizations can derive benefits and value from the analysis of Big Data.

Defining Big Data

There are many definitions of Big Data. Our preferred definition, cited in Kaisler et al. (2013) is: *Big Data is the volume of data that cannot be efficiently organized and processed with the storage and tools that we currently possess.* Under certain circumstances, we can organize, process, and analyze Big Data. However, we cannot do it very efficiently or effectively. For example, because we cannot process a real-time data stream fast enough, we cannot generate results that will enable decision-making within a specified observe, orient, decide, and act (OODA) cycle.

While Big Data often implies many very large volumes of data, as Leslie Johnston (2013) noted it can also imply that "Big Data can most definitely mean small data files but a lot of them." These extremes present challenges to business and IT managers and owners on a continuing daily basis.

How do you know when you are facing Big Data? Well, the transition from organizational data and databases to Big Data is not exact, but there are a number of characteristics that can be used to help one understand when the transition occurs.

Big Data has been characterized by several attributes. We have defined the five Vs as described in Table 1.1. The initial three Vs were first stated by Doug Laney (2001). Based on our research and experience, we added the last two Vs.

Table 1.1 Five Vs of Big Data

V	Description
Data volume	*Data volume* measures the amount of data collected by and available to an organization, which does not necessarily have to own all of it as long as it can access it. As data volume increases, the value of different data records will decrease in proportion to age, type, richness, and quantity among other factors. It is estimated that over 2.5 exabytes (10^{18}) of data are created every day as of 2012 (Wikipedia 2013).
Data velocity	*Data velocity* measures the speed of data streamng, its aggregation, and its accumulation. Data velocity also has connotations of how quickly it gets purged, how frequently it changes, and how fast it becomes outdated. e-commerce has rapidly increased the speed and richness of data used for different business transactions (e.g., website clicks). Data velocity management is much more than a bandwidth issue; it is also an ingest issue (the extract-transform-load (ETL) problem).
Data variety	*Data variety* is a measure of the richness of the data representation—either *structured*, such as resource description framework (RDF) files, databases, and Excel tables or *unstructured*, such as text, audio files, and video. From an analytic perspective, it is probably the biggest obstacle to effectively using large volumes of data. Incompatible data formats, nonaligned data structures, and inconsistent data semantics represent significant challenges that can lead to analytic sprawl.
Data value	*Data value* measures the usefulness of data in making decisions. It has been noted that "the purpose of computing is insight, not numbers." Data science is exploratory and useful in getting to know the data, but "analytic science" encompasses the predictive power of Big Data. A large amount of data may be valueless if it is perishable, late, imprecise, or has other weaknesses or flaws.
Data veracity	*Data veracity* is the accuracy, precision, and reliability of the data. A data set may have very accurate data with low precision and low reliability based on the collection methods and tools or the data generation methods. The information and results generated by processing this data may then be seriously flawed or compromised.

Big Data has often been used to represent a large volume of data of one type, such as text or numbers or pixels. Recently, many organizations are creating blended data from data sources with varied types through analysis. These data come from instruments, sensors, Internet transactions, e-mail, social media such as Twitter, YouTube, Reddit, Pinterest, Tumblr, RFID devices, and from clickstreams. New data types may be derived through analysis or joining different types of data.

Getting Started with Big Data

Research and practical analysis have shown that there are many areas where you can focus your attention. We will review four process areas that are enormously fruitful for immediate analysis of processes using the tools that can best be applied to Big Data.

First, *data categorization* can aid in the analysis of Big Data because the tools available now permit machine-learning algorithms to explore large and varied data collections through machines that are trained or seeded with previously known classifications, such as process output or groupings. The value of these groupings is that they will provide classifications or labeled analysis variables that may then be used to discover a relationship or a predictor that shows or predicts the value of or output of the process. It may be the result of a hidden value that is a part—input, subprocess, step, or activity—within the process or a characteristic of the process input. This analysis is truly a directed discovery process and organizational learning experience enabled by Big Data (Deng, Runger, and Tuv 2012; Hwang, Runger, and Tuv 2007). The data does not have to be fully classified or categorized because specific techniques can be applied to assign grouping or to cluster data or process outputs that do not appear within the previous groupings (Zhang et al. 2010). The value is that this categorization develops insights required to understand the root-causes for underlying problems by discovering relations that were simply not previously visible. This process was used in cancer prediction, diagnosis, and in understanding its development (Cruz and Wishart 2006).

Secondly, *functional data analysis* can be implemented because it permits discrete calculations due to the continuous nature of production data (Ferraty and Romain 2011). The section will not attempt to describe this analysis in great detail, but the reader should be aware of the benefits of this form of Big Data analysis. This application analysis is closely related to profile monitoring and control charting that is employed when the quality of a process or product can be characterized by a functional relationship between a production measure of some output or output value and an explanatory variable(s). We can often "see" these relationships (but may not fully understand the value) in the graphs, curves, and visual depictions prepared when data are logged and drawn as graphs or curves. The business value is obvious in being able to predict strength, changes, and locate when relationships may begin and end.

The value of this can be recognized in the potential for assessing quality characteristics, such as size and quantity data, product shapes, geometric relationships, appearances of faults and imperfections, patterns, and surface finish, while these are happening and relate them directly to end results of processes. Data may be produced by visual sensors, image observation in many medical, military, and scientific applications, or other mechanisms (Megahed, Woodall, and Camelio 2011).

Thirdly, managers must now recognize that situations that can be "drawn" as graphs showing associations between various objects in the data can be analyzed as Big Data problems. Graphic representations (Cook and Holder 2006) are seen as two connected nodes with an edge, if the nodes possess a relationship. Researchers have used these data to assess social networks (e.g., Facebook, LinkedIn), intrusions for networks, disease, or product adoption, and rating or rankings for consumer e-commerce actions (Chakrabarti and Faloutsos 2012). For example, one can identify potential changes in social network datasets or communications network. McCulloh et al. (2008) found variances in the Al-Qaeda network prior to September 11. For the business, it could be important to know that sources of data and information acquisition are changing for clients or business customers.

Finally, in the era of Big Data, data may now be available simultaneously from numerous and unrelated sources. Historically, a control chart was employed to observe multiple sources of periodic data (Boyd 1950). Automated tests can now be used to detect the likelihood of changes in only one stream of data and concurrently monitoring multiple streams based on the stream features, assessing correlations among the streams, and the scope and size of any change sought (Jirasettapong and Rojanarowan 2011).

Adding Value to Organizations

Big Data has an impact in every field of human endeavor, if the data are available and can be processed. Impact is different from value. Impact helps to advance a field with new knowledge whereas value affects how useful the resulting actionable information is—whether predicting events, making a profit, discovering a new particle, or improving the lot of our fellow humans. Big Data can add value in several ways: (1) It can make information transparent and usable at a higher frequency. (2) As more accurate data is collected, it allows organizations to conduct more controlled experiments to assess efficiency and refine business operations. (3) It can focus attention on narrower segments of the customer community for precisely specifying products and services (market segmentation). (4) Given usage data, it can be used for the specification of new products and services.

In a research study by IBM and the Said Business School, Oxford University (Turner, Schroeck, and Shockley 2012), four areas for employing Big Data were identified as described in Table 1.2.

Outline of This Book

This chapter has provided a brief introduction to some of the issues and challenges that senior executives and managers must consider in using Big Data to assess and enhance their service delivery operations. The remaining chapters provide additional information on each of the major topics presented earlier and present a framework for gaining value from this growing phenomenon.

The remainder of this book is divided into six chapters as follows:

- *Chapter 2: Applications of Big Data to Service Delivery* will discuss how Big Data can be used within the service delivery paradigm to deliver better quality of service and to identify and target a larger universe of potential customers.
- *Chapter 3: Analyzing Big Data for Successful Results* explains the types of analytics that may be applied to Big Data to yield actionable information and identify different analytical packages.
- *Chapter 4: Big Data Infrastructure—A Technical Architectural Overview* identifies and presents some of the elements of an architecture necessary to support Big Data acquisition and creation, storage, management, analysis, and visualization.

Table 1.2 Areas for use of Big Data

Using Big Data	Brief description
Customer analytics	IBM and Said noted that 55 percent of the companies they surveyed focus their Big Data efforts on customer-centered objectives in order to improve service delivery to their diverse customer base. These companies want to improve their ability to anticipate varying market conditions and customer preferences in order to take advantage of market opportunities to improve customer service and increase customer loyalty in an agile manner.
Build upon scalable and extensible information foundation	IBM and Said noted that companies believe results can be obtained from Big Data only if the IT and information infrastructure can respond to evolving aspects of Big Data focused on the three Vs: variety, velocity, and volume. This means they must be able to evolve their IT and information infrastructure in an agile manner transparently to customer interactions. (Note: they only examined the original three Vs, but we believe that the information foundation must be focused on the five Vs.)
Initial focus is on gaining insights from existing and new sources of Big Data	IBM and Said found that most initial Big Data efforts are focused on analyzing existing data sets and stores in order to have a near-term effect on business operations. They suggest that this is a pragmatic approach to beginning to develop a Big Data usage capability. Most companies do not know what insights they will gain or how much useful and usable information they can extract from the information on hand. In many cases, the data has been collected, perhaps organized, and stored away for many years without ever being analyzed.
Requires strong analytics	Using Big Data requires a variety of analytics tools and the skills to use them. Typically, companies use such tools as data mining, online analytical processing (OLAP), statistical packages, and the like on structured data based on existing data stores, marts, and warehouses. However, as they accumulate unstructured data, the diversity of data types and structures requires new techniques for analysis and visualization. Existing tools can have trouble scaling to the volumes characteristic of Big Data and, often, cannot adequately analyze geospatial data, voice, video, or streaming data.

- *Chapter 5: Building an Effective Big Data Organization* discusses the organizational structure of a Big Data operation, how to staff it, and the key characteristics of the staff.

- *Chapter 6: Issues and Challenges in Big Data and Analytics* presents some of the issues and challenges facing an organization that is using or considering using Big Data in its business operations.
- *Chapter 7: Conclusion: Capturing the Value of Big Data Projects* discusses how to measure Big Data value and understand how Big Data contributes to an organization's business operations.

CHAPTER 2

Applications of Big Data to Service Delivery

Organizations can, and are, utilizing Big Data to improve and enhance the delivery of services to their customers. From retail marketing to financial services to health care and more, Big Data and analytics are changing the landscape of service delivery. We first outline a simple, generic service delivery model (SDM) and then discuss how Big Data can be applied within the model. We then outline several examples of how Big Data is being applied to service delivery.

Defining Services

A *service* is often described as a tangible or intangible benefit provided by a person, a group, an organization, or some other entity to a person, a group, an organization, or some other entity. There are many types of services: providing advice such as WebMD, or providing education such as Massive Open Online Courses (MOOCs), or selling airline tickets. In the latter case, the service is not the airline ticket, but the transaction(s) leading to the exchange of money for the airline ticket whether physical or electronic. Thus, a service is differentiated from a "good" in that it is nonmaterial and does not incorporate a notion of ownership. In the previous case, the airline ticket, whether paper or electronic, is the "good" that is owned by the person purchasing it.

Whereas services used to describe and document the exchange of information verbally or in written form between humans, today services are most often exchanged through e-mail, social media, and online systems such as Google, Wikipedia, YouTube, Twitter, product ordering systems, and so on. A service is frequently manifest as the co-creation of value between the entities. For example, the purchase of airline tickets through Expedia or Orbitz involves the exchange of items of value: electronic tickets, electronic payments via credit card or PayPal, a sense of accomplishment in completing a task, and a sense of anticipation at taking a trip.

Since definitions are often constraining, we will utilize a broad conceptualization of a service. In general, *if something provides value, and is almost or immediately usable—without modification and without understanding exactly how it functions—it is likely a service in the eyes of the purchaser*. Services can be modified by the delivering organization and are frequently enhanced and upgraded at customer's request or to match the benefits provided by competitors. Such changes may not be particularly important to the acquirer unless they cause an organization to change its operation, pay more, or remove something the client deems to be of significance or value.

Using our loose, but applied, definition as a huge net, one may characterize many businesses and organizations as service businesses. For example—generally accepted services organizations include those in the financial, medical, education, government services, utilities, and many other industries.

As an example, consider the recent development of the Healthcare Market Place. Many will agree that it meets the conditions established as being those of a service. It offers itself as a resource where you can learn about your health coverage options, compare health insurance plans based on costs, benefits, and other important features, choose a plan, and enroll in coverage. The marketplace establishes conditions for its service in that it delivers information on programs to low and moderate income individuals who may not have the "resources to pay for coverage."

The service provided through this market place includes information that describes how users may save on the monthly premiums, save out-of-pocket costs of insurance coverage (available through the Marketplace), Medicaid information, and Children's Health Insurance Program (CHIP) data. The service is conditional and rules-based in that it may be managed by a state or by the Centers for Medicare and Medicaid Services (CMMS), which is a part of the Department of Health and Human Services (HHS). Potential users may access HealthCare.gov, or if one lives in a state with its own exchange, one can be guided to their state's marketplace site. Conditions for the use of data are described, and conditions applying to the state sites are located on state exchanges. Finally, this service describes the use of information, sharing of data, questions, and user data that must be provided to obtain the service, requirements for identifying social security numbers (SSNs), rules about citizenship, outcome questions about end result determinations, and privacy rules (https://www.healthcare.gov/how-we-use-your-data/).

A second example in the financial services arena will illustrate the breadth and depth of the processes and business rules that govern the services provided in service industries and organizations. Financial services are products that deliver process management solutions that are both end results and stepping stones to process reengineering and technological change. Often the services may result in reduced cost and enhanced operational performance for the customers of the service organizations. The services will implement portions of the business models of many organizations to control costs and support growth. The huge number and character of the available financial services includes processes specific to: leasing and lending, maintenance for client accounts and transactions, asset management, asset tracking and assessments, delinquency tracking, deal conversions, backroom operations, outsourcing, customer remediation, revenue accounting and profitability assessment, security originations and settlements, transfers, reconciliation, mortgage origination, sales, underwriting, closing and funding, post close and quality control, customer collections, card services, payments, lending, deposits, account maintenance, setup, marketing as well as social media, risk and fraud, and accounting analytics. The concept of Big Data in the finance arena can become very complex when the transactions are enumerated with such a laundry list.

Service Systems

A *service system* is a complex system in which specific arrangements of people and technologies take actions that provide value for others. *Service delivery* is the set of mechanisms, including manual, semi-automated, and automated systems of hardware, software, and communications that convey a service from one entity to another. A *service-oriented architecture* (SOA) is a foundation for a service delivery system.

A service system is a human-made system that provides value in provider–customer interactions. Providers and customers may be humans or computers, or both as in semi-automatic systems. The advent of online systems meant that humans do not necessarily need to be "in the loop"; hence, the concept of service delivery as the exchange of electronic information between two computer systems, for example, in a *business-to-business* (B2B) service system.

Service Delivery

Understanding the growing role of Big Data in service delivery requires that one appreciate the scope, value, and importance of service processes, and how Big Data has impacted many services. We will illustrate the growing importance of the role of processes and their growing impact in the service industry in this section.

Managers and executive reading this text may come from many different service industries each with a growing impact of processes that are important for the provided services. Each industry, business, and organization will have its own vision and firm understanding of what comprises the business model, processes, operations and services, and the outcomes or customer benefits, either objective and perceived, of what is provided to clients.

Key processing issues may be predetermined by the business operation plan of the organization. The organization generally has as its business objective to provide a clear set of services for each customer. Note that services may provide intangible results, such as information or electronic tickets, or tangible results, such as a package showing up on your doorstep. The proposed benefit may be more or less clear—and may need to be specified in greater detail as a component of any service transaction with the client. A customer benefiting from the service will have an expectation of what is to be received or gained from a service and the conditions that may be necessary for the benefit to be received.

Thus, the entire service can be viewed as a process, with one who purchases a service having a clear understanding of what is being gained, what a service accomplishes, and conversely what is not intended to be a result. The conditions may be explicit—in that limits, costs, dates, durations, guidance for the purchase, support and even limits on the acquiring client may be explicitly stated. All may be determined by contracts, either explicitly stated or implied. The implications from all of this task specificity are clear. The service organization will be following processes and procedures governed by business rules that ensure the delivery of the service if the processes and conditions are correctly followed, for example, on the business side, the actions required to provide the service to the user are well understood and correctly, consistently, and completely implemented.

A key idea in service delivery is that services are produced and consumed at the same time. Thus, the customer receives the ticket or some notification of it at the same time that money is provided to pay for it. What is produced and consumed here is the transaction that affects the exchange. However, some services may be delayed in consumption as, for example, in parcel transportation and delivery where the customer provides the package and the funding, but it is some time later when the package is delivered to its destination thus completing the service.

The goal of service delivery is to provide value, which means the exchange of information and, sometimes, objects must be mutually beneficial to both or all parties to the set of interactions. In a service-oriented environment, the critical elements are: (1) service delivery to the focal point of customer interaction, (2) understanding the needs and requirements of the (possibly diverse) customer base, and (3) ensuring the privacy and security of customer interactions—whether human-to-business or business-to-business interactions.

Service science is the application of scientific understanding to advance our ability to design, improve, and scale *service systems* for business and societal purposes (Maglio and Spohrer 2008). Service systems are inherently dynamic, complex, and continually changing in modern society. Thus, there is no single type of service system, but many types, and more to emerge in the next several years as innovation in interactive, always on technology advances and drives the evolution of service.

A Service Delivery Model

While there are a multitude of service delivery approaches, we broadly define SDM as a set of principles, policies, and processes used to guide the design, development, operation, and retirement of services delivered by a service provider with a view to offering a consistent service experience to a specific user community. Services are both inward- and outward-facing. Inward-facing services include support services and operations management services. Outward-facing services to the customer base include a variety of product, management, analytics, and professional services.

Figure 2.1 outlines a simplified SDM that we use to represent basic service delivery components. Many implemented SDMs can become complex and their description is beyond the scope of this book.

Business customers are individuals and the communities and markets they reside in provide the services for them to interact. Depending on the number and makeup of customer groups, a service-based organization may segment them. Customers interact with a service interface, exchanging information, perhaps in multiple exchanges (such as in transaction processing where one moves through multiple web pages). Each interaction may activate one or more process steps to process the information and perform actions to satisfy the service. Ultimately, the service produces an outcome, such as an electronic ticket or bill of lading or shipping receipt, which is returned to the customer.

Service management ensures that the organization is offering the right services at the right quality and price levels within its market. Typically, the services are described and managed in a service portfolio or catalog. The service management component is responsible for the life

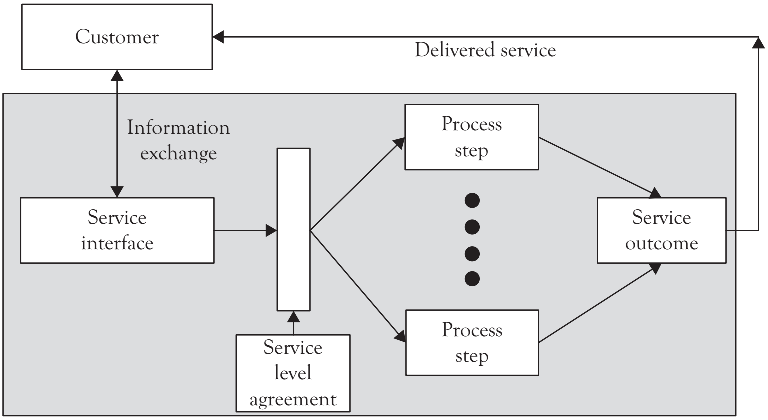


Figure 2.1 Simplified SDM

cycle management of services and their financial viability. Service level agreements (SLAs) can be used to guarantee performance-based service levels to the business customer.

Coupling Big Data to Service Delivery

Big Data has the potential to improve service delivery in a wide variety of ways. For example, Big Data customer and market analytics can be used to identify customers for new or existing services based on current buying patterns, geography, income status, and so on.

More and more organizations are concerned about how their services are being characterized in social media, such as the level of customer satisfaction. Social media review sites can be harvested and analyzed to determine how customers view the current service levels (e.g., excellent, good, fair, poor). Organizations utilize Big Data social media analytics to capture unstructured comments on blogs, review sites, media articles, and so on. They can perform reputational analysis on the data so the service provider better understands how the quality of their services is perceived by their customers and utilize that information to improve service delivery. Some examples include:

- For organizations, such as telecom companies, the levels of network availability and bandwidth can be tracked real time via Big Data.
- For package delivery, taxi, and other transportation-intense services, providers can utilize dynamic transportation and geolocation data to perform real-time tracking and issue alerts on current vehicle location, time to next appointment, current traffic patterns, and so on, helping to ensure performance-based SLAs.

Supporting Service Delivery with Big Data

As increasing amounts of data are generated and “almost unlimited” storage and computing capabilities become available (e.g., Clouds, HPC Clusters), Big Data has become the next major frontier in IT innovation. It offers unprecedented opportunities for discovering hidden intelligence and game-changing insights in business and other domains. Large search companies, such as Yahoo and Google, using data collected to support their users, but also collecting data from their users, have been able to provide deeper and better business insights than were previously available. The potential for generating new business value has led industry, academia, government, and commercial organizations to invest in Big Data analytics (BDA) in order to seek new business opportunities and to deliver new levels of service to their customers and clients.

An *analytic* is a process or procedure for finding some meaningful patterns in data. *Business analytics* is a set of methods or procedures for analyzing an organization's data to elicit information that can help the organization improve its business operations. Today, business analytics are largely statistically based processes applied to data to drive decision making. *Business intelligence* is a suite of tools and techniques for transforming raw data into meaningful and useful information to support business operations. Among the tools and techniques in this suite are reporting, data mining, text mining, and predictive analytics.

There are several categories of analytics. The current application focus is on descriptive and predictive analytics, with visual analytics receiving significant attention for their support to decision making through graphic presentations of data and results. *Descriptive analytics* looks at data and analyzes past events for insight as to how to approach the future. *Diagnostic analytics*, a subset of descriptive analytics, uses data to determine why something happened, for example, equipment failure, market downturns, and so on. *Predictive analytics* uses data to determine the probable future outcome of an event or a likelihood of a specific situation occurring. *Prescriptive analytics* goes beyond predicting future outcomes by also suggesting actions to benefit from the predictions and how to avoid bad situations and showing the decision maker the implications of each decision option. *Decisive analytics* goes beyond prescriptive analytics to operationally specify how to accomplish a goal or objective in some detail (akin to tactical and operational planning). *Visual analytics* has emerged from scientific visualization to focus on analytical reasoning facilitated by interactive visual interfaces.

Organizations have been using analytics to assess performance and facilitate decision making since the first organizations were formed. Up until the beginning of the electronic calculator age, circa the 1900s, most analytic systems were manual, for example, human computers. Analytics received a strong boost from Frederick Taylor and Edward Deming, among others, in the early years of the 20th century. Another strong proponent was Robert McNamara, first at Ford Motor Company, then at the Department of Defense. The advent of electronic computers, programming systems, and other tools moved analytics into the software arena.

Big Data in a vacuum cannot ensure that any services will change or that there will be service delivery improvements. Big Data must be analyzed within the context of a business, scientific, or other type of system. The *role of analytics* is to transform, process, and display Big Data and the results from its analysis to users to enable them to solve specific and particular problems.

Data-Driven Companies

International Data Corporation (IDC) suggests that companies that embrace Big Data are on track to be successful in this evolving era of Big Data. Some examples of analytics are provided below.

Under Armour, a company based in Baltimore, MD, is a successful sportswear manufacturer that continually seeks innovative changes to sportswear. For the NFL, they embedded sensors in the shorts worn by NFL candidates at their tryouts to help evaluate the parameters of the play. John Deere and Co., famed for its farm equipment among other machines, changed to an agile model of software development. It can now deploy new software on a monthly basis directly to tractors every month or so that helps them drive straight across plowed fields.

The *Harvard Business Review* (McAfee and Brynjolfsson 2012) noted that companies using data-driven decision-making are 5 percent more productive and 6 percent more profitable than their competitors. IDC noted that data-driven companies that use diverse data sources and diverse analytic methods are five times more likely to succeed in their projects than competitors that do not.

Retail Analytics

Walmart typically handles more than 1 million customer transactions every hour (2.5 petabytes of data) and utilizes analytics in most aspects of the sales and inventory processes.

However, analytics are not new for retailers, as they have been doing analysis of point of sale transactions and other business data for years. The first barcode data appeared in the 1970s. It was placed on a pack of Wrigley's Chewing Gum scanned via Universal Product Code (UPC) in the Marsh Supermarket, in Troy, OH, in 1974. It became the basis for tracking merchandise sales in supermarkets, which operated at a slim profit margin, and then expanded to many other stores. Today, UPC data is widely used and allows major retailers to track purchases by customers and predict their purchasing behavior.

Radio-frequency identification devices (RFIDs) are becoming ubiquitous and offer the ability to dynamically track items as they move. Similarly, with cellphones and Wi-Fi many companies can now detect cellphones within close proximity to a store, office, or other facility and, knowing something about the owner, transmit targeted advertising to the cellphone.

Examples of retail BDA include:

- Customer analytics
- Merchandising analytics
- Store operations
- Marketing analytics
- Return, fraud, and loss prevention analytics

Health Care

Health systems, hospitals, Accountable Care Organizations (ACOs), and physicians are becoming increasingly concerned about how the quality of service they provide to their patients is perceived in the marketplace. An industry shift toward transparency and consumerism is giving patients-as-consumers access to an unprecedented and growing number of health care data sources to search for and evaluate health care providers, while the proliferation of social media is changing industry dynamics. As patients-as-consumers flock to online media outlets such as rating/review sites and social media channels to share their patient experience, health care providers are now looking for better ways to monitor and analyze this feedback as value-driven government and commercial payers tie reimbursements to quality assessments in which patient satisfaction data is a factor. Consequently, pay-for-performance is driving demand from health care organizations for more data, better collection methods, standardized metrics, and more robust analytics applications to better understand patient satisfaction to improve their quality of service.

Binary Fountain™, a McLean, VA, company, is a provider of patient feedback management solutions uniquely designed for health care in a single cloud-based platform. Their proprietary healthcare-centric natural language processing engine mines patient feedback and commentary harvested from a growing number of public sources, then analyzes and benchmarks patient sentiment tied to 37 operational performance categories. This holistic approach to patient feedback management empowers operations, patient experience teams, marketing and advertising executives with insights that drive improvements to operational decision-making—resulting in reduced costs, strengthened brand positioning, higher revenues, better patient engagement, and quality of services.

Fraud Detection

FICO Falcon, a credit card fraud detection system, protects 2.1 billion active accounts worldwide. It is a BDA tool for the early detection of fraudulent activity on credit and debit cards. This system predicts the probability of fraud on an account by comparing current transactions (e.g., normal cardholder activity) to current fraud trends.

It utilizes real-time, transaction-based scoring with neural network models, and provides real-time decision making on suspected fraudulent transactions. It has adaptive models that can adjust to changing patterns in attempted fraud. The result is more secure and less costly credit card services to consumers, credit card providers, and the merchants that accept the cards.

Mass Transit

Bridj, a transit startup, has introduced a pop-up bus transportation system in Boston that adapts in real-time to ridership needs. It uses a network of express shuttles that offer efficient and flexible trips that are dynamic scheduled based on predictions of changing user transport. The system analyzes "between two and three billion data points to understand how Boston moves, from over 19 different data streams," including municipal data, census data, and social media data. It then targets neighborhoods it considers commuter pain points.

The system has cut some commute times in half by strategically offering bus service and routing in the city. For example, a ride from Coolidge Corner to Kendall Square, which would likely take 42 to 55 minutes on the Massachusetts Bay Transportation Authority, has taken 15 to 18 minutes on his company's buses. Bridj plans to set up pilot programs around the United States in the near future.

CHAPTER 3

Analyzing Big Data for Successful Results

> *Some companies have built their very businesses on their ability to collect, analyze and act on data.*
>
> —*Competing on Analytics* by Tom Davenport

Brad Peters (2014) has noted that the bloom of Big Data may be fading. Where it used to be all about the data, users and decision makers have realized that Big Data cannot deliver much value on its own. Now, the focus is on how to unlock value in the Big Data—it is all about the *analytics*.

Big Data without analysis yields no actionable information and intelligence that can be used to make decisions. Analytics seeks to discover patterns, tools, and techniques for uncovering useful insights into and extract meaning from high volume data streams and large datasets, producing as results valuable *data products*. Analytical software and service products for analyzing data abound. The results are typically presented in the form of derivative datasets and interactive graphical user interface displays, such as dashboards.

Numerous analytics have been applied to problems in business and the physical and social sciences. Up until 30 years ago, simple business models often sufficed for international business operational decision-making. The advent of globalization, brought on partly due to advances in digital technology, massive amounts of information available at our fingertips, coupled with a rapidly changing, even chaotic, international political environment, have up-ended these models. Globalization has increased the diversity and uncertainty in outcomes when complex systems such as financial flows and markets, regional economies and political systems, and transnational threats involving multiple actors are in constant flux.

These latter problems are often not quantitative, but qualitative, as the data to be processed is symbolic, textual, audio, imagery, and video.

Organizations have high expectations for Big Data. A Harvard Business Review study found that 85 percent of organizations reported that they have Big Data initiatives planned or in progress, with 70 percent of these initiatives being enterprise-driven. Eighty-five percent of the initiatives are sponsored by a C-level executive or the head of a line of business, the majority believe that initiatives will cross multiple lines of business or functions and the organizations expected an impact across multiple lines of business (Barth and Bean 2012).

However, the same survey found a significant capabilities gap, as the organizations perceived that they currently have limited abilities to address these expectations. Specifically, only 15 percent of respondents ranked their access to data today as adequate or world-class and only 21 percent of respondents ranked their analytic capabilities as adequate or world-class. Finally, only 17 percent of respondents ranked their ability to use data and analytics to transform their business as more than adequate or world-class (Barth and Bean 2012).

It is important to note that the word "analytics" is often misinterpreted. Analytics expertise is often confused with the ability to use tools like Tableau, Excel, or other business intelligence (BI) tools. Others think of analytics as simply doing quantitative analysis. But these are just parts of the whole. Analytics has a full life cycle, which begins by formulating an analytics question or problem that can be answered or resolved with data.

These analytics questions are generally not easy to answer and require some knowledge of the domain in which the analysis is taking place. For example, to answer a marketing question the analyst needs to have some knowledge and experience with marketing. That is, analytics does not happen in the vacuum. The analyst then needs to embark in a data quest to identify and gather all the necessary data to answer the question and then select the most appropriate modeling approach or combination of models most suitable to the task. Analysis and reporting of results comes at the very end of the cycle.

Figuring out an analytic approach to answer a question is not easy and this is where the analytics professional shines. Once an approach to answering the question has been identified, the analysis can then be automated with tools and this is the realm of "business intelligence."

As discussed earlier, in this book, Big Data has the potential to provide significant benefits to an organization, including, but not limited to:

- Creating visibility into operations and processes
- Enabling experimentation to discover needs
- Improving organizational performance
- Deeper insight into organizational processes due to greater granularity
- Population segmentation for targeted marketing actions
- Create new business models, innovative processes and products, and services

Big Data can improve decision-making because (1) there is more data to analyze and (2) better data management and analysis environments are becoming available all the time. The difference between the information that managers view as important and necessary for decision making and the information that their organizations can provide to them can have a significant impact on decision making. The gap on information about customer needs and preferences is estimated to average 80 percent (Pricewaterhouse Coopers 2009).

Big Data Analytics

Organizations will find there is little value in just storing large amounts of data. True business value is created only when the data is captured, analyzed, and understood: relationships, trends, and patterns discovered that result in insights that produce better decision-making and problem-solving. Big Data analytic applications need to be proactive, predictive, and have forecasting capabilities.

Currently the use of analytics as a competitive differentiator in selected industries is exploding. Disciplines such as marketing, sales, human resources, IT management, and finance are continuing to be transformed by the use of Big Data and analytics. Brynjolfsson (2010) studied 179 large companies and found that those adopting "data-driven decision making" achieved productivity gains that were 5 to 6 percent higher than companies that did not have such decision-making processes.

Before we start we need to address some of the confusions one hears about, especially with respect to terminology about different Big Data related disciplines associated with analytics. Moreover, analytics can have a different scope or meaning depending on an individual's background. For example, is traditional data mining or BI a form of analytics? What is data science? And how is it different from analytics?

Let us start with the a standard definition of *BI*: "Business Intelligence refers to the technologies, applications, and processes for gathering, storing, accessing, and analyzing data to help its users make better decisions" (Wixom and Watson 2010).

Next is *data mining*, which has traditionally been thought of as the computational process of discovering trends in data. It incorporates machine learning, statistics, and database systems. It is about extracting patterns and knowledge and identifying previously unknown relationships in the data. (e.g., using cluster analysis for anomaly detections). It is good for hypotheses development. Some view data mining as the data exploration work done to identify interesting hypotheses and analytics as the evaluation of those hypotheses.

You will also hear the term "Data Science" when Big Data is discussed. *Data science* "is a set of fundamental principles that guide the extraction of knowledge from data" (Provost and Fawcett 2013). A "*data scientist*" generally has deep knowledge of data mining analytics, a comprehensive view of the data, the discipline in which analysis is conducted (marketing, health care, social media, etc.), and the underlying core disciplines (e.g., statistics, mathematics, database, etc.).

Industry leaders in online education have argued that there is no consensus about what these terms mean, at least in the educational market. As a McKinsey Global Institute study noted (Manyika et al. 2011), the United States faces a talent shortage of 140,000 to 190,000 professionals with deep analytical skills, and about 1.5 million managers and decision makers with analytics skills. The online education market views these deep analytical professionals as "data scientists" who have a well-rounded education in data-related disciplines and often hold PhD degrees in mathematics, statistics, management sciences, or other quantitative fields.

"Analytics" professionals are viewed as those who also have deep quantitative backgrounds, but perhaps one step below the data scientist in terms of mathematical and data sophistication. "Business analytics professionals" appears to be the term of choice for the professionals required to make up the 1.5 million shortage described previously. These are either managers with deep expertise in the business domain of analysis (e.g., marketing, cyber security, public policy, etc.) who are trained in core analytics methods or savvy consumers of analytics reports produced by data scientists, but who can complement the reports with further analysis of their own.

Rather than trying to parse these different definitions, we adopt a definition of Big Data analytics as focusing on helping managers gain improved insight about their business operations and make better, fact-based decisions.

The *Institute for Operations Research and Management Sciences* (INFORMS), the leading professional and academic organization in the analytics community, defines Big Data analytics as "the scientific process of transforming data into insight for making better decisions"—see https://www.informs.org/About-INFORMS/What-is-Analytics. We will stick with that definition in this booklet.

Big Data analytics can be further be broken down as the use of:

- Large amounts of both structured and unstructured data, either at rest or in a streaming state
- The use of advanced information technology to support the analysis and modeling of data
- Statistical methods to provide rigor to the analysis
- Visual analysis to help discover patterns and trends in the data and present the results to key decision makers
- Other quantitative and qualitative methods, and mathematical or computer-based models

Following the infamous 3 Vs of Big Data, Big Data analytics (BDA) can also be viewed as doing analytics with large "volumes" of data, available in a "variety" of types and formats (e.g., structured, unstructured, multi-media, visual, etc.), possessing various degrees of "velocity" (e.g., speed at which the data becomes available, how fast it loses relevance, how quickly it changes values, etc.). We have added two Vs for Big Data: "veracity"—the accuracy and precision of data (e.g., how reliable is the data and the decisions or conclusions derived from it)—and "value"—the benefit it has to the organization's business operations.

Another way to distinguish BDA from plain analytics is the environment in which the analysis is conducted. One can do standard analytics work with Big Data by downloading the necessary data from large data warehouses and applying conventional analytics methods and tools. But this is not necessarily viewed as BDA. BDA is the practice of conducting analytics directly in the Big Data environments. This often requires programming skills, such as Java, Python, or R, and technologies like "in-memory" analytics, which are rapidly becoming mainstream. However, in some cases, when working with Big Data, analytics is not conducted directly on the Big Data per se. Instead, the necessary data is extracted from traditional and nontraditional sources into a structured data warehouse or database for further manipulation and analysis.

Examples of Big data analytics that can be applied in business organizations include:

- Management of customer relationships (free Wi-Fi)
- Financial and marketing activities (credit card drop)
- Supply chain management (find bottlenecks)
- Human resource planning (Hewlett Packard's flight risk score of its more than 300,000 employees)
- Pricing decisions (best price for a new product—Starbucks based on analysis of Tweets)
- Sport team game strategies (Moneyball)

However, Big Data is about more than just business analytics. For example, Big Data can transform:

- Medicine, including but not limited to processing 3-D hyperspectral high resolution images for diagnostics, genomic research, proteomics, and so on
- Demographic analysis
- Geointelligence: spatial analysis

Table 3.1 Examples of advanced analytics

Big Data production	Big Data exploration	Traditional data production	Traditional data exploration
Predictive maintenance	Text analytics	Credit risk	Customer buying patterns
Real-time fraud detection	Image recognition	Telecomm customer churn	Determine drivers of part failure
Predictive policing	Internet of Things	Product recommendations	Exploring anomalies, for example, new types of fraud
		Marketing response propensity	Customer segmentation

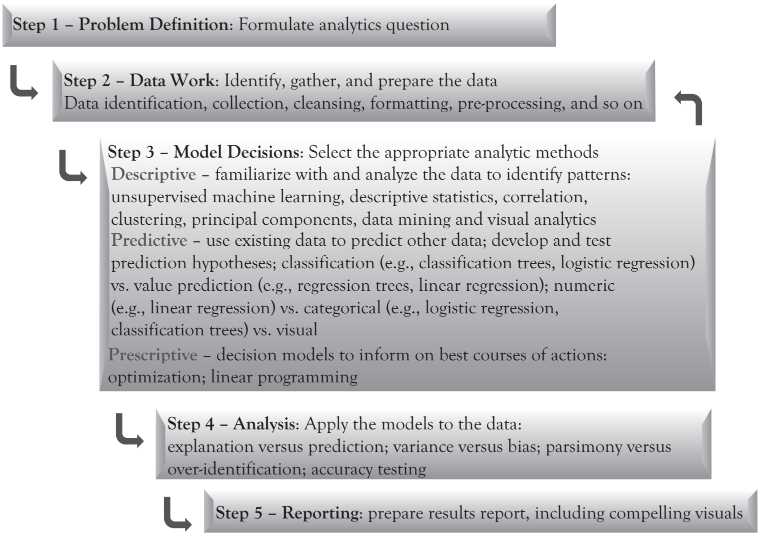
The term *advanced analytics* is utilized frequently in analytic circles. Advanced analytics uses both quantitative and qualitative methods, such as data mining, predictive analytics, and simulation and optimization techniques, to produce insights and information that traditional approaches cannot discover. Our view is that advanced analytics goes beyond merely statistical mechanisms to a variety of other analytical methods which, combined in various ways, can offer considerable analytical power and insight into large sets of data. Appendix A lists our taxonomy of analytics classes (Kaisler et al. 2014). A few of the examples of the application of advanced analytics to Big Data are included in Table 3.1.

Big Data Analytics Initiatives Need a Process

When approaching a BDA initiative, successful outcomes are more likely if a process or analytics life cycle is followed. While activities will vary based on the type of problem you are trying to solve, the nature of the solve and the nature of the data, they can be conducted in both agile or traditional environments, key activities include those that follow and are also presented in Figure 3.1.

The Analytics Cycle

The *analytics life cycle* has many steps, which may vary depending on the problem at hand, but it typically involves five distinct steps or phases, depicted in Figure 3.1: (1) problem definition, (2) data work, (3) modeling



Step 1 – Problem Definition: Formulate analytics question

Step 2 – Data Work: Identify, gather, and prepare the data
Data identification, collection, cleansing, formatting, pre-processing, and so on

Step 3 – Model Decisions: Select the appropriate analytic methods
Descriptive – familiarize with and analyze the data to identify patterns: unsupervised machine learning, descriptive statistics, correlation, clustering, principal components, data mining and visual analytics
Predictive – use existing data to predict other data; develop and test prediction hypotheses; classification (e.g., classification trees, logistic regression) vs. value prediction (e.g., regression trees, linear regression); numeric (e.g., linear regression) vs. categorical (e.g., logistic regression, classification trees) vs. visual
Prescriptive – decision models to inform on best courses of actions: optimization; linear programming

Step 4 – Analysis: Apply the models to the data:
explanation versus prediction; variance versus bias; parsimony versus over-identification; accuracy testing

Step 5 – Reporting: prepare results report, including compelling visuals

Figure 3.1 Key analytic lifecycle activities

decisions, (4) analysis, and (5) reporting. What is very telling about these five steps is that the actual analysis is only a small part of the full analytics cycle, and a lot of time needs to be devoted to analyzing the problem, preparing the data, and strategizing about the analytical approach to be used. We now discuss each of these steps in more detail.

Step 1—Problem Definition

What problem are you trying to solve? First, you need to formulate a business or domain question or problem that you wish to address. This will provide needed direction to efforts to select data and analytics. What level of problem are you focusing on: operational, tactical, or strategic? Managers have different perspectives on analytics that depend on their business operations responsibilities and level of decision making within an organization. Table 3.2 presents a brief description of these "Levels" of analytic initiatives, which focus on the type of decision making within an organization.

One important aspect of the problem definition has to do with functional domain knowledge. Before we can frame and articulate an analytics question, the analyst needs to understand the functional domain of analysis. For example, it would be very difficult to build a predictive model

Table 3.2 Business-focused analytics levels

Class	Description
Operational	Analytics that support day-to-day business operations, including monitoring and event data analytics leading to "here and now" metrics and near-term decision making. For example, dynamic product pricing based on customer purchasing patterns (Clifford 2012).
Tactical	Analytics focused on mid-term decision making and dealing with tactical problems, including simple predictive models based on historical data. For example, target's use of purchasing patterns to predict pregnancy in a teenage girl (Hill 2012).
Strategic	Analytics for long-term decisions, focused on organizational stability and directional decision making, including predictive, prescriptive, and comparative analytics.

for stock prices without having some fundamental knowledge of finance and the stock market. Similarly, in order to use analytics to detect and prevent cyber security breaches, the analyst needs to understand the cyber security domain.

The analytics question to be answered needs to have a well-defined objective but should also be somewhat general at this stage. For example, how can the company increase its market share by 1 percent? Which price will maximize profits for a particular type of new coffee beverage? As explained later in this chapter, the analyst will have to perform some descriptive analytics and be in a better position to translate the general analytics question into a number of more specific and testable analytics hypotheses.

Step 2—Data Work

This step involves the identification, gathering, cleansing, and preparing necessary data available to answer the question or problem. This can be the most time-consuming step in the entire analytics process. Analysts often spend 70 to 80 percent of their effort in this activity. As we have mentioned in earlier chapters, data can take many forms:

- Structured data "legacy"
- Unstructured data
- The "deep" web

- Data that is hidden in the enterprise
- Sensory data
- Analog devices turned digital

All data and analytics projects should begin with: What do you want to accomplish with your data? To organize and manage data, we must understand our data. When considering Big Data, we often only consider scale, for example, how much data do we have, the ability to acquire, store, and preserve data for future analysis. However, the complexity of data is often a large challenge as well: factors such as the complexity of semantics and the richness of data, the complex relationships among data elements and structures. Context and data reliability is normally required to understand the data and to establish accuracy, future usability, and so on.

A key question to answer is: What will yield more accurate results, sophisticated and complex models with poor quality data or simple models with high quality data? Naturally, this is a balancing act. But, the point most experts agree on is that nothing beats high-quality data. Bad data will only lead to poor predictions and no model sophistication will be able to correct for bad data. Complex models are very difficult to understand, tune, and update. Simpler, parsimonious models have the added advantage that they can provide more intuitive explanations of the relationships and patterns in the data, which are easier to explain and justify to a management audience. Regardless, analysts need to spend a high proportion of the analytic life cycle time working with the data.

Whether data is maintained internally or gathered externally, there are two important processes that will follow: data cleansing and data preprocessing. Data cleansing has to do with making corrections for things like incorrect data, inconsistent data, redundancies, outliers, and missing data among many other things.

Pre-processing has to do with the fact that raw data is rarely in the form needed for analytics. For example, often the necessary data is in multiple tables, in which case the analyst may have to join multiple data tables or link data elements as needed. Another type of pre-processing has to do with unstructured, text, and categorical data. Unstructured and text data are often pre-processed to construct structural variables for analysis. For example, if you are building a model to predict which e-mail messages are spam, the use of excessive capitalization is considered to be a good predictor. If you agree, then you could pre-process e-mail text to count the number of capitalized letters in the message.

With categorical data, the problem is that this type of data is not quantitative and therefore cannot be used as is with most statistical methods. But categorical data can often be converted into numerical data (e.g., male = 0, female = 1; urban = 0, suburban = 1, rural = 2). Pre-processing is often needed too when the evaluation of the model shows problems with the data. For example, if the predictor variable data is skewed, you may not be able to use most traditional models like ordinary least squares regression, which requires that the predictor variable be normally distributed. However, skewed data can often be corrected by taking the log of the data. Other pre-processing transformations often used include: squared terms (x^2), inversed variable ($1/x$), interaction terms ($x_1 * x_2$), and rank transformations (i.e., use the rank of the data rather than the data itself, which provides a more uniform distribution), among others.

Step 3—Modeling Decisions

In this step, modeling is performed to understand and explore the data as well as to then predict and prescribe outcomes based on the data. There are multiple approaches on how to model decisions, these include the following:

- *Analysis types*: descriptive, predictive, or prescriptive
- *Modeling type*: numerical associations or categorical
- *Analysis approach*: structured, unstructured, visual, or mixed
- *Statistical learning type*: unsupervised or supervised learning

Analysis Types

When modeling a decision via analytics, one typically has to answer three basic questions: What has happened? What is likely to happen? What can we do when it happens? To answer these questions, one uses several categories of analytical types of approaches, and these include descriptive, predictive, and prescriptive analytics to quantitatively and qualitatively model the data. Descriptive or visual models are used to cognitively improve the results of the models. Use these modeling approaches as appropriate for the specific business problem that is being solved. These types are discussed in Table 3.3.

Descriptive Analytics. *Descriptive analytics* allows you to consolidate Big Data into smaller, more useful subsets of information. Big Data is not suitable for human understanding, but the information we derive from the data is.

Descriptive analytics' real power is that it allows you to familiarize yourself with the data (utilize descriptive statistics, correlations, factor analysis, cluster analysis, etc. to better understand the date), as well as

Table 3.3 Analytic modeling categories

Type	Description
Descriptive	Descriptive: getting meaning from the data—for example, BlueFin Technologies: did viewers liked a particular TV show last night, based on Tweets? A set of techniques for reviewing and examining the data set(s) to understand the data, determine what has happened, and find the patterns: How many, when, where? Often uses data mining and clustering techniques.
Predictive	Predictive: using some variables in the data to predict the outcome of others—for example, Target—which product purchases are the best predictors that the customer is expecting a baby? A set of techniques that analyze current and historical data to determine what is most likely to happen or not. What could happen next if …? What actions are needed? Often uses regression methods, time series models, and statistical machine learning techniques.
Prescriptive	Prescriptive: using the data to recommend what to do to achieve outcomes—for example, Starbucks—what is the optimal price of a new coffee brand to maximize sales? A set of techniques for developing and analyzing alternatives computationally, which can become courses of action—either tactical or strategic. What are the possible outcomes? How can we achieve the best outcome? What if we do …? Often uses operations research methods, decision modeling, symbolic machine learning, simulation, and system dynamics.

starting to generate possible hypotheses (via data mining, patterns, trends, etc.). Any analytic method aimed at understanding the data or evaluating possible associations, classifications, or relationships falls in the category of descriptive analytics.

While data mining is not used exclusively for descriptive analytics, it is often used to uncover previously unidentified relationships in the data. There are many descriptive analytics methods and describing all of them is way beyond the scope of this booklet, but some examples include: descriptive statistics—which provides various statistical parameters for the data (e.g., mean, minimum, maximum, standard deviation)—frequency distributions, correlation analysis, factor analysis (through principal components), cluster analysis, and classification trees.

Predictive Analytics. *Predictive analytics* is by far the most popular type because the very reason why we embark on an analytics project is usually to be able to *anticipate* outcomes. Basically, you take data that you have to predict data you do not have. However, descriptive analytics should often precede predictive analytics because it is important that the analysts get immersed in the data to develop familiarity and identify trends and associations in the data that were previously unknown. Descriptive analytics should help the analyst translate the general analytics question(s) for the problem at hand into more specific testable hypotheses that can be evaluated with predictive models.

One analytics question may result in several testable hypotheses. For example, an analytics question like “what are the best predictors of stock price?” may lead into various analytic hypotheses, such as: a stock’s price-earnings ratio today will have a positive effect on the stock’s price one year from now; the company size is a predictor of stock price stability; the number of years in business is a predictor of stock price stability; technology stocks are generally overpriced; and so on. Each one of these hypotheses can be tested with a specific predictive model.

Most literature on predictive analytics emphasizes predictive accuracy. However, predictive accuracy needs to be well understood. For example, certain methods like classification trees can achieve 100 percent predictive accuracy through over-identification. The reason for this is that accuracy is often tested with the existing data. Data sets are often partitioned into a training set (i.e., the portion of the data used to build the model) and a testing set (i.e., the remaining data used to test the accuracy of the model). There are many different methods to partition the data into training and testing sets, and evaluate the model accuracy. For example, you could partition the data randomly using an arbitrary threshold (e.g., 80 percent training set, 20 percent testing set). Another method could be to construct multiple training/testing sets to evaluate the predictive accuracy with multiple partitions. Yet another one could just "leave one observation out" and train the model with the rest of the data, and then test the model with the one observation left out, and then repeat this process multiple times. And yet another approach is to conduct Monte Carlo experiments to analyze the dispersion of data values and their effect upon results.

Complex and sophisticated models can be developed such that the predictive model touches every data point in the sample and, therefore, any accuracy testing will yield 100 percent accuracy. These models are said to be *over-identified*. But this makes the model exceedingly complicated, making it very difficult to use it to explain relationships in the data. But more importantly, internal accuracy is no guarantee that the model will provide accurate predictions with new data. Testing models with new data is not about predictive accuracy, but about testing the external validity of the model.

Consequently, a sound predictive model needs to balance three things: (1) parsimony, that is, model simplicity that can provide useful intuitive explanations (e.g., a 10 percent decrease in a stock's price-earnings ratio leads to a 3 percent stock price increase a year later), (2) predictive accuracy, and (3) external validity.

At this point the analyst needs to formulate the necessary predictive hypotheses and formulate the appropriate predictive analytic models and methods. These can include multivariate regression, logistic regression, forecasting, nonlinear models, classification trees, neural networks, and so on. Predictive models generally fall under two categories: predictions of values (e.g., future sales) and predictions about classification (e.g., will a loan client default on the loan or not?). We discuss this further in the next section.

However, as Michael Wu of Lithium Technologies (http://www.lithium.com/) has pointed out:

> The purpose of predictive analytics is NOT to tell you what will happen in the future. It cannot do that. In fact, no analytics can do that. Predictive analytics can only forecast what might happen in the future, because all predictive analytics are probabilistic in nature.

Prescriptive Analytics. *Prescriptive analytics* uses optimization and simulation algorithms to advise on possible outcomes and answer: "What should we do?" Prescriptive analytics may also be used to assess different courses of action and their likely outcomes.

Prescriptive analytics uses a combination of modeling approaches such as business rules, machine learning, and computational modeling procedures, among others, which can be run against historical and transactional data, real-time data feeds. You then identify and develop decision and optimization models incorporating domain knowledge to develop possible outcomes and rank them in some way so that the best possible outcome, given the inputs, emerges from the mix of outcomes. The prescriptive model can also suggest or recommend one or more possible courses of action, ranking them according to some set of constraints.

But one must remember what some called the *MD^3 rule → Models Don't Make Decisions, Managers Do!!!* That is, unless the prescriptive models are implemented in an automated application (e.g., to recommend products to buy based on prior purchase history), prescriptive models should be viewed as aids to human decision making. Humans should make their own decisions, aided by prescriptive models, but the prescribed decisions should not be followed blindly.

Modeling Types

There are several modeling types, but they can be succinctly classified as either numerical or categorical. *Numerical models* are generally based on correlation and associations of the various variables relevant for the analysis. Numerical models for descriptive analytics include things such as descriptive statistics, clustering analysis, factor analysis through principal components, correlation analysis, and analysis of frequencies and distributions, among others.

Numerical models for prescriptive analytics include things such as multiple regression analysis, regression trees, structural equation models, and neural networks, among others. *Categorical models* are generally based on classifications in the data. For example, students in a university data file may be classified as freshmen, sophomore, junior, senior, and graduate. Such categorizations do not lend themselves to quantitative analysis. But these categories are often transformed into numerical values or comparison groups where quantitative methods can be applied.

Categorical models for descriptive analytics include things such as analysis of variance (ANOVA) in which the means and variances of two or more categories are compared and chi-square tests in which category counts can be analyzed and compared, among others. Categorical models for predictive analytics are referred to as "classification" models, and they include things such as logistic regression models, in which the predicted variable is categorical (e.g., default on loan versus no default) but is converted into a quantitative value for analysis (e.g., no default = 0; default = 1); classification trees, in which the data is partitioned using predictor variables and the categories that fall in each partition are analyzed and further subpartitioned.

Analysis Approach

The analysis approach is based on the nature of the data being analyzed and it can be structured, unstructured, visual, or mixed. Structured data is data that can be easily organized in tables with columns, such that each column contains data of the same type (e.g., numerical value with 2 decimal points; text with up to 16 characters; dates in dd/mm/yyyy format). Structured data generally contains numerical data, which can be analyzed with statistical methods, and categorical data, which can be classified with numerical values.

Unstructured data refers to data that does not have many restrictions in what it can contain, except for the type of data. For example, text data is considered unstructured, but it is restricted to text only. Similarly, video data is also unstructured, but it can only contain video footage. There are essentially two ways to analyze unstructured data: using unstructured data analysis methods or developing some structured metadata from the unstructured data and then using structured data analysis methods.

More and more software tools and methods are being introduced to analyze unstructured data. For example, there is an abundance of text mining tools that can process millions of pages, uncovering patterns in the data by identifying themes. Because words can have various meanings, there is an abundance of synonym files customized to particular industries. Unstructured data can often be processed to give it some structure for analysis. For example, a data set with movies will generally have some metadata associated with it with information like actors, producer, director, release year, duration, and minutes on screen for each actor, movie segments with timeline pointers, and so on. Similarly, text data can have things like word counts, frequency of usage for words, or a combination of words of interest.

A picture is worth 1,000 words. *Visual analytics* is the science of analytical reasoning facilitated by interactive visual interfaces. It is also a set of techniques for visualizing information to facilitate human decision-making. The field of visual analytics is maturing rapidly with very sophisticated tools like Tableau and SAP's Lumira coming to market all the time. For example, IBM has an advanced visualization website (http://www-01.ibm.com/software/analytics/many-eyes/), which provides a number of very clever web-based visualization tools for free.

Statistical Learning. Machine learning is often thought of as the science of getting computers to act without being explicitly programmed. For example, getting an e-mail system to identify spam is an application of machine learning. In more simple terms, machine learning or statistical learning refers to the ability for computer models to learn patterns from the data. Models are "trained" with the data, and as new data comes in the models can learn the new patterns. In very simple terms, running a regression model and estimating coefficients that explain a predictor variable is a form of statistical learning. As new data is collected the regression models can be re-estimated to produce new coefficients. In essence, every time we develop an analytic model we are applying machine-learning methods.

Developing coefficients that can explain relationships in the data is referred to as "learning." Such learning can be of two types: unsupervised and supervised. Unsupervised learning refers to machine learning methods in which the outcome is not known or there is no specific goal for this outcome. For example, when an analyst looks at sales data and discovers surprisingly that beer and diapers sell well together, he or she is employing unsupervised learning methods. Examples of unsupervised learning methods include things like cluster analysis, classification algorithms, correlation analysis, and multidimensional scaling, and are most typically employed in descriptive analytics. In supervised learning the outcome of interest is specified and there is a specific goal for the analysis outcomes. For example, when an analyst is trying to predict which advertising approaches lead to increased sales, he or she is employing supervised learning methods. Examples of supervised learning methods include things such as regression analysis, regression trees, classification trees, and neural networks and are most typically employed in predictive analytics.

Step 4—Analysis

We are often asked by non-analytical professionals to describe what analytics is. Inevitably, when we answer this question using an INFORMS definition—that is, to extract meaning from the data for decision making—most people then ask, what is new? Or they make comments like, "but we have been doing this for years" and they are right. The work analysts do in the analysis step is no different than the work that statisticians, data miners, and decision modelers have been doing for years. What distinguishes analytics from these other fields are two main things: (1) analytics is more than analysis and it includes all aspects of the life cycle described in this chapter; and (2) analytics incorporates a confluence of various fields that were previously viewed as different, including statistics, mathematical modeling, quantitative sciences, data mining, software programming (e.g., computational statistics), database and Big Data, management sciences, and decision sciences, among many others. But all activities in the life cycle have a single purpose, which is to support the analysis that will answer the analytics question at hand.

Analysis is often performed by humans executing statistical, data mining, or machine learning algorithms with increasingly more sophisticated tools. But the analysis is sometimes automated. For example, a predictive model designed to catch spam e-mail is running in the background by some mail server feature, without human intervention. Or a company like Amazon may be running predictive models in the background to make customized purchase recommendations to customers. Humans intervene in these cases to tune the parameters of the models from time to time, but the analytics models are executed automatically by a software program.

One important aspect in analysis is the selection of the appropriate analytic tools. Tools can be proprietary and expensive, but there are an increasing number of free open source software (OSS) systems. It is becoming difficult to select the appropriate tools because of the overwhelming number of them in the market. We present some of the tools in common use today in Chapter 4.

Step 5—Reporting

This step involves writing up the results, conclusions, and recommendations for the audience interested in the respective analytics question. Document your conclusions and recommendations, remembering to present these findings at a level and format that your end users will find understandable. It is important to note that analytics is a quantitative and highly technical discipline and, therefore, many high-level managers, clients, and stakeholders may not have the background to understand statistical output. A well-articulated report that contains the main findings, conclusions, and recommendations will go long ways.

Analytics reports are often accompanied by attractive visuals and graphics. But we often get confused or overwhelmed when the report contains an overwhelming amount of graphics without proper explanation. To maximize the impact and the business value of the analytics report, we recommend a few important guidelines in Table 3.4.

Table 3.4 Selected guidelines for analytic reporting

Reports should have a text narrative that briefly describe the analytics question, hypotheses, data sources, and methods employed
The report should contain a well-written and well-articulated explanation of key findings, and whether the hypotheses were supported or not.
Supported hypotheses should have a brief commentary and unsupported hypotheses should have an intuitive rational or explanation for why the expected results were not found, or why they had effects opposite to the predicted direction.
No graphic or visual exhibit should be unexplained and they should all be referenced and introduced in the main text—isolated visual exhibits that do not contribute or complement the narrative are generally useless and even distracting.
All visual exhibits should have all the necessary information in the exhibit to understand it—the most common omission in our experience is an explanation of the vertical or horizontal axes in the graph.

Analytic Modeling Methods and Approaches

There are a wide variety of analytical modeling methods and approaches available to the analyst when analyzing Big Data. It is not possible to identify and describe all of them, so we will highlight the most commonly used methods below.

Quantitative Analysis

Qualitative analytics approaches are primarily statistical or other mathematically based. These techniques include linear regression analysis to analyze continuous dependent variables; logistic regression to analyze binary or categorical dependent variables, classification trees, such as decision trees; correlation; data reduction; and so on. Other techniques include:

- Associations: correlation among variables, analysis of variance, regression models, which variables co-vary with which?—For example, how much does annual income increase with each year of additional university education?
- Classification (and probability estimation): in which class does a case belong (predicting the probability that a new case will fall in a given class); chi-square analysis, logistic regression models—for example, patient tested positive (or negative) for

a disease. What are the probabilities of testing positive for a disease?

- Others: clustering, similarity matching, co-occurrence grouping, profiling, link (strength) prediction, data reduction (factor analysis), causal modeling, and so on.

Qualitative Analysis

Qualitative analysis commonly refers to relatively unstructured, nondirective discussions or interviews (such as focus groups, depth interviews, and ethnography or observation) to explore a topic. It can use subjective judgment based on non quantifiable information.

Visual Analytics

Visual analytics is driving new ways of presenting data and information to the user. Wong and Thomas (2004) defined it as follow: "Visual analytics is the formation of abstract visual metaphors in combination with a human information discourse (interaction) that enables detection of the expected and discovery of the unexpected within massive, dynamically changing information spaces." In 2005, a scientific panel defined it as "the science of analytical reasoning facilitated by interactive visual interfaces" (Thomas and Cook 2005).

It is important to note that while we are describing visual analytics as a specific analysis approach, it actually permeates all types of analysis. For example, descriptive, predictive, and prescriptive analytics results are often accompanied by powerful graphics that convey results very clearly. A picture is worth 1,000 words. Visual analytics is the science of analytical reasoning facilitated by interactive visual interfaces. It is also a set of techniques for visualizing information to facilitate human decision-making, which is also called *visual analytics*. The field of visual analytics is maturing rapidly with very sophisticated tools like Tableau and SAP's Lumira hitting the market all the time. For example, IBM has an advanced visualization website (http://www-01.ibm.com/software/analytics/many-eyes/), which provides a number of very clever web-based visualization tools for free.

Visual analytics is often viewed as providing visual renditions of the available data and analytic model results (e.g., pie charts, scatterplots, and social network diagrams). But the true power of visual analytics is in actually employing visual methods to present easy interpretation of predictions and prescriptions, not just descriptions of data, due to the sophistication of the eye–brain interaction. As humans, we can quickly grasp information presented graphically as opposed to information presented in text or columns of numbers.

Visualization is a very old and mature science. All analytical tools have features to prepare charts, box plots, distribution curves, pie charts, and line graphs with trends. So this is not new. What is new is the level of sophistication that is constantly being added to existing tools and new ones arriving to the market all the time. For example, statistical analysis software (SAS) has a full suite of visual analytics tools. Other tools like Tableau specialize specifically on visualization, but like many other visual analytics tools, it offers the capability of running statistics and other quantitative methods from within the tool. For example, Tableau allows users to load and run R scripts to estimate statistical models like multiple regressions, which can then be visualized in many ways by the tool.

Visualization has a long tradition for data representation. Nothing illustrates the point better than a well-prepared slide show full of interesting graphics. But visual analytics has taken this to the next step by also providing ways to conduct the actual analysis visually. For example, visual analysis has a long tradition in social network analysis. Tools like Krackplot, NetDraw (free), and Pagent have been around for several years to illustrate how actors in a social network interconnect. For example, identifying the most central actor in a network can often be done much easier visually than statistically.

Similarly, popular sites providing interesting and intuitive visual analysis websites are rapidly appearing, including https://infogr.am/, http://www-01.ibm.com/software/analytics/many-eyes/, and http://www.informationisbeautiful.net/. There are also many scientific and prominent conferences devoted to visualization, not just visual analytics, including http://ieeevis.org/ and http://s2015.siggraph.org/.

Statistical Machine Learning (Supervised Learning)

Statistical machine learning refers to analytics models that learn from the data as the data changes. Commonly called *supervised learning*, these techniques partition the data into training and testing sets:

- The training data is used to build the models by associating predictors (or rules) with outcomes (e.g., spam filtering)
- The testing data is used to test the models by evaluating if the models predicted new data outcomes correctly
- With supervised learning a specific target is specified—for example, which loan clients are more likely to default on their loan?

Unsupervised Learning

As opposed to supervised learning, unsupervised learning does not typically partition the data. Rather, then data is treated as a whole and the learning approach is to explore the data set looking for patterns and trends. The learning analysis program may specify criteria for what is sought, but there is no external feedback provided to the program about how it is doing. It either converges to a solution based on the exit criteria or it diverges and finds no significant patterns or trends.

Some Emerging Analytics

New analytics are being developed continually and are moving beyond the simple statistical approaches in data mining. New infrastructure has provided an ability to process information from sources that were previously the focus of research. New analytics are based on advanced mathematical, engineering, and computer science techniques that incorporate models and data from social, demographic, political, and other scientific disciplines.

Social analytics often focus on intangible and qualitative phenomena with varying parameters and interdisciplinary contexts. Sometimes, data are not directly measureable, but must be determined through proxy variables that can be imprecise and uncertain. Data may be non-numeric, thus requiring reduction and encoding to numeric values, but may lose precision and information as a result.

Global business problems are a mix of physical and social problems, but are strongly influenced by social phenomena. Hence, a mix of physical and social analytics must be developed and applied to model and understand business operations in the global environment. Kaisler and Cioffi-Revilla (2007) reviewed a set of analytical methods and classified them into 17 classes. Kaisler (2012) extended that analysis to the problem of intelligence analysis. Appendix A briefly describes these classes.

Sentiment Analysis

Sentiment analysis mines data streams to determine trending issues. It attempts to determine demographic, social, and customer sentiments regarding issues and how they relate to products. A hefty dose of psychology accompanied by behavioral models is the current approach.

There is a gold mine of information being collected by service organizations—both customer-focused and across the web. One use of this data is to gauge the collective consciousness of the respective populations. On the one hand, an organization can assemble and assess a variety of sentiments based on customer orders, complaints, comments, and even blogs devoted to specific products or organizations. Online opinion has become a powerful force that can make or break a product in the marketplace or an organization's reputation through information or disinformation. On the other hand, an organization can assess different types of trends based on a broad collection across the web.

Sentiment analysis is an emerging field that attempts to analyze and measure human emotions and convert them into symbolic and quantitative data. In either case above, organizations monitor news articles, online forums (blogs, chat rooms, twitter feeds, etc.), and social networking sites for trends in opinions about their products and services or topics in the news that may affect their business decisions or operations. Sentiment analysis is one example of the field of social analytics.

Generally, sentiment analysis uses raw text as its input. Sentiment analysis works by classifying the polarity of a given text—whether tweet, blog entry, or document, either in part or in full. The simplest algorithms work by scanning keywords in the text to categorize a statement as negative, neutral, or positive, based on a simple binary analysis. For example, "enjoyed" is good, but "miserable" is bad. The individual scores for words, phrases, sentences, and documents—whatever units you use—are then combined to form an overall sentiment for the set of data that you have. You can interpret the final score based on the domain you are working in. For example, if you sampled sportswriters' tweets regarding athletes at the London Olympics as one of us did, you can get a "sentiment" of who the preferred athletes were in several of the major sports.

However, not all opinions are equally important. Some pundits carry more weight because of their stature in the business community, the number of times they have ordered products or services from the company, or their popularity, for example, as measured by the number of followers. A tweet by Lady GaGa will have much more impact than a person who has used a product or service just a few times. Thus, organizations must also develop profiles of commenters and rank their importance to generate weights when assessing comments.

As an example, a restaurant gathers tweets about its menu and service. It identifies tweeters who have large numbers of followers and, perhaps, the number of times that have actually visited the restaurant and the number of different selections made. Tweeters with greater numbers in each case would be given greater importance in assessing their sentiments. And, the number of retweets would measure the level of engagement with their followers and the degree of their influence. Using this technique the restaurant could identify opinion shapers of importance to them, assign them higher weights to calculate a more accurate indicator of sentiment, and engage them to correct negative impressions and to generate more positive impressions.

There are many challenges in applying sentiment analysis to a selection of text. Most tools use simple assignments to individual words. For example, the words "sinful" and "decadent" are often positive sentiments when applied to a chocolate confection obtained from a bakery, but have negative connotations in other instances. Most algorithms cannot handle irony, slang, jargon, or idioms—leading to erroneous polarity assignments than can skew the sentiment assigned to a text. Reliable sentiment analysis is going to require many linguistic shades of gray.

It should not replace opinion polling or other structured opinion assessment techniques, but it can complement their results due to the law of large numbers. In sentiment analysis, one does not have accurate control over the sample space, so the results are often hard to verify and validate. At best, they can show trends in real-time whereas other techniques often take days to produce results.

Geospatial Analytics

Geospatial analytics (or *location analytics*) expressly integrates geographic and demographic information into the analysis process. It applies statistical processing to data associated with geographical and demographical information, such as maps, census areas, mineral deposits, and so on. Many of the complex problems in business, intelligence, social science, and government have a geographic component that is often best expressed through visual analytics.

Geospatial analytics often incorporates a temporal dimension into the analysis process to complement the spatial dimension. For example, a retailer may want to analyze year-over-year sales per store to determine the best performing stores. With these results, the retailer may take corrective actions at underperforming stores based on demographic analysis, perhaps even closing stores that do not have the ability to increase sales. Similarly, health care organizations can use spatial and temporal analysis to track and predict the spread of disease. Such techniques were used to determine where to commit resources in the recent Ebola outbreak in Africa.

Unstructured Text Processing

With over 80 percent of the world's knowledge residing in unstructured text, whether on paper or in electronic form, the capability to process it to extract data and information is essential. Typical methods rely on statistical or symbolic natural language processing (NLP), and, sometimes, a hybrid of the two. Neither type is new as NLP research has been ongoing for over four decades—supported by Defense Advanced Research Projects Agency (DARPA), National Science Foundation (NSF), and academia. Within the past decade or so, many commercial firms have discovered that developing tools for processing unstructured text is a viable market.

One of the major issues with NLP is ambiguity. For example, in English, many nouns and verbs have several close synonyms. The word *strike* has over 30 common meanings. For example, there are more than 45,000 people in the United States named "John Smith." *Entity resolution* is the process of resolving a name to an explicit individual. In a series of articles, John Talburt (2009–2011) identified a hierarchy of five methods for entity resolution ranging from simple deterministic matching to asserted or knowledge-based matching. As one ascends the hierarchy, the complexity of the processing required increases. Some of the symbolic classes are more flexible in using ambiguous data.

Resolving ambiguous data at higher levels requires domain knowledge to make decisions about the cause of ambiguity within the data. Assuming default values when encountering ambiguous data can lead to erroneous results. However, analysis and reasoning, as at the highest level of Talburt's hierarchy, may require intensive computation that affects performance. The tradeoff becomes end-to-end processing time for large quantities of data versus the fidelity and quality of individual data items.

Image, Video, and Audio Processing

With more data and information being recorded in images and video streams due to the ubiquity of cell phones and tablets with embedded cameras, a major percentage of the world's knowledge is becoming represented by image and video data. Granted a lot of it may be currently frivolous, but it is becoming a major influence given the success of services such as FaceBook, YouTube, SnapChat, InstaGram, Pinterest, Tumblr, and other social media.

Extract information from imagery and full motion video is very much a major research problem, but one that has significant potential for providing data for social and behavioral models that can lead to trends and customer preferences. Many organizations have been exploring and researching this for years but this is still an extremely difficult problem to solve computationally—although the human mind, brain, and eye do it extremely well. While unstructured text remains the largest repository of Big Data, imagery and video are a strong second.

Another very difficult problem is how to extract information from audio recordings such as telephone conversations and the audio tracks accompanying full motion video recordings. Most current audio processing systems do not process free range speech, but rely on phrases chosen from distinct vocabularies.

Edge and Location-Specific Analytics

Because moving large amounts of data on the order of petabytes can be both time- and bandwidth-consuming, an emerging discipline is edge analytics, which also incorporates location-specific analytics. It may not be physically possible or economically feasible to store all the data streaming in at the endpoints of an organization's data and information network. Moreover, data perishability may be a significant factor if the lifetime of the data is less than the transport time from the endpoints to where it might be processed. One of the maxims of enterprise architecture is to place the processing (at least, the initial stages) close to where the data is collected following, if possible, the 80-20 rule.

Edge analytics is focused on moving analytics to the frontier of the domain. For example, many cameras contain image processing functions that can be done right in the camera before you download the images. It is estimated that there is one camera for every 11 individuals in the United Kingdom. London is replete with cameras—some obvious and some not. With Google Glass beginning to shape the future of body-worn video cameras, significant issues in privacy, processing, and ownership of data are going to emerge that have yet to be addressed. Key questions include: What will we do with all this video? How will we process it? How will we store it (Satyanarayanan et al. 2015)?

Network Analytics

Social network analysis methods and tools have been around for many years. With the explosion of social media in the last several years, social network analytics has increased substantially in popularity. But social network analysis methods can be used for other types of network analysis that are not necessarily social.

Network analysis derives from the field of "graph theory" and it has a rich tradition of quantitative and visual methods. A network can be represented mathematically as a table called "sociomatrix" by listing all members of the network in rows and also in columns. The cells in the sociomatrix contain numerical values that measure some relation of interest between the row member and the column member. The same network can be represented visually in "sociograms" by depicting each member as a node in a graph and then adding connecting lines representing the relationship between nodes (i.e., cells in the socio matrix). These are called "1-mode" networks because rows and columns represent the same thing (e.g., friends on a social network).

However, networks can also be "2-mode" in which the rows contain the members and the columns contain some affiliation that members have. For example, if one were to list all the students in a university, one per row, and create a column for each course taught by the university, a 2-mode network can be created by entering a 1 in the respective cell indicating if the student took that course or 0 if he or she did not. The interesting thing about 2-mode networks is that they can be de-composed into two 1-mode networks (e.g., student by student; and course by course), which could help, for example, determine how students cluster together around fields of education, or how courses cluster together based on student enrollments. A full discussion of network analytics is beyond the scope of this book, but suffice it to say that there is a very rich and abundant set of quantitative and visual tools and methods to analyze dyadic relationships in networks.

Cognitive Analytics

Recently, a new type of analytics—*cognitive analytics*—exemplified by IBM's Watson, has emerged. Cognitive analytics exemplify the potential for machines to actually learn from experience inspired by how the human brain processes information, draws conclusions, and codifies instincts and experience into learning. It often uses concepts from deep machine learning, which focus on how the human brain learns and attempts to translate those techniques into software mechanisms.

IBM has created its System G Cognitive Analytics Toolkit (http://systemg.research.ibm.com/cognitiveanalytics.html), which builds upon its decades of research in machine learning to provide a wide range of tools to detect human's emotion and perception on text, images, or videos. IBM's visual sentiment and recognition tools can detect visual objects, such as faces, in images or video and predict the feelings expressed. Their text emotion tool uses supervised learning to classify unstructured text into 1 of 12 categories.

IBM's Watson, a Jeopardy winner, is being converted to a general-purpose tool and applied to many different disciplines. A new product, Watson Analytics, "is like a data scientist in a box," according to Marc Altshuller, vice president of IBM Watson Analytics (http://searchdata-management.techtarget.com/news/2240238506/IBM-works-to-deliver-on-Watsons-cognitive-computing-promise). It is a subset of the analytics embedded in Watson that focuses on analytical discovery.

Key Challenges for Analytics

BDA is clearly here to stay. However, BDA is still in an emerging state in many respects. Perhaps the biggest challenge for success in Big Data is lack of analytical professionals and managers who understand and can make decisions based on Big Data insights.

For example, industry, academia, and government currently suffer from a lack of analytical talent. Training, growing, and organizing Big Data professionals within an organization will be one of the key challenges to successful BDA initiatives going forward. We will discuss these and other organizational and people issues in Chapter 5.

Another key challenge in BDA includes the continued development and deployment of analytic tools that support business managers, analysts, and other "non" data scientist individuals to efficiently and successfully analyze outcomes. As mentioned earlier, the vast amount of effort in a typical analytics initiative is in the extracting, moving, cleaning, and preparing the data, not actually analyzing it. As Big Data source become larger, more complex,

Table 3.5 Selected big data analytics challenges

Quantity versus quality: What data are required to satisfy a given value proposition? At what precision and accuracy?
Tracking data and information provenance from data generation through data preparation and processing to derived data and information
Performing data validation and verification, including accuracy, precision, and reliability
Coping with sampling biases and heterogeneity
Using more diverse data, not just more data (Feinleb 2012)
Working with and integrating data having different formats and structures
Developing scalable algorithms to exploit current and emerging parallel and distributed architectures
Developing methods for visualizing massive data
Ensuring data sharing with security and integrity
Enabling data and information discovery
Understanding that approximate results given limited computational capacity are better than no results
Is more data better than better algorithms?

and more numerous, approaches to addressing this problem will become critical.

Table 3.5 presents other selected BDA challenges (Kaisler et al. 2013).

Finally, do not disregard traditional analytics—traditional and big analytics will live beside each other for years to come.

Analytical Resource and Tool Survey

Analytical tools are packaged applications that allow users to begin working with data without having to do any programming. This will be a very brief survey of some of the different types of tools available. Managers need some familiarity with these tools even if they do not fully understand how they work in order to be able to discuss cogently with their IT support staff.

The available analytics resources and tools are expanding at an explosive rate. It is not possible, in a document this size, to list them all. Following is a list of resources and tools that the authors have used.

Although some of the authors have had experience with these packages, we are not endorsing them over other tools.

Commercial Packages

There are numerous commercial packages that now incorporate Big Data analytic capabilities—too many to identify all of them here. Although some of the authors have had experience with some of these packages, we are *not* endorsing any of them. Table 3.6 presents a few of these.

Open Source Packages

There are many OSS packages that are widely used in Big Data processing and analysis—far too many to describe here. There are advantages and disadvantages to using OSS. Generally, the source code is available for you to download and modify if you want, but then you diverge from the main code line. The farther you diverge, the harder it becomes to maintain. Unless you actively participate in the development of open source code, you are dependent upon the community members for their vision of what the software system should be and do and their schedule for upgrades and updates.

Often, there is a lack of (good) documentation for OSS, so it can take longer to understand how it works and how to use it. Moreover, to build the infrastructure and subsystems you need for your business environment, you may have to integrate several different OSS programs together, which with the lack of documentation can take a while. Again, together these software systems may not do exactly what you want and so your infrastructure may be a hodge-podge of pieced-together software systems.

A few of the more popular OSS packages in use today are presented in Table 3.7. While these packages are largely self-contained and have large libraries associated with them, they can be difficult to use if they do not do exactly what you need.

R, a free, open source, object-oriented programming language, specifically designed for computational statistics, is very popular with the social science community and is becoming one of the main tools for rapid prototyping of analytics. It is expected to surpass many commercial statistical

Table 3.6 Selected commercial packages

Package and website	Brief description
SAS http://www.sas.com/en_us/home.html	SAS is a software suite developed by SAS Institute for advanced analytics, BI, data management, and predictive analytics. See also: SAS Enterprise Guide http://support.sas.com/software/products/guide/index.html SAS Enterprise Miner http://www.sas.com/en_us/software/analytics/enterprise-miner.html
IBM SPSS http://www-01.ibm.com/software/analytics/spss	SPSS started out as a suite of statistical processing techniques, but has significantly expanded into the BDA domain. See also: SPSS Modeler http://www-01.ibm.com/software/analytics/spss/products/modeler/
Mathematica http://www.wolfram.com/mathematica/	A computational software program used in many scientific, engineering, mathematical, and computing fields, based on symbolic mathematics. It is an integrated system of mathematical techniques accessed by the Wolfram programming language that allows users to express problems, display results, and so on.
Mathwork's MATLAB http://www.mathworks.com/	MATLAB (Matrix Laboratory) is a multi paradigm numerical computing environment and associated programming language intended primarily for numerical computing. Since many of the modern analytics require numerical computing, MATLAB provides a foundation for developing a rich analytical framework. It is one of the most widely used computing systems with well over a million users. It interfaces with programs written in C/C++, Java, FORTRAN, and Python.
RapidMiner https://rapidminer.com/	RapidMiner can be used as either a standalone application for data analysis or as a data mining engine for integration into an organization's systems. It provides procedures for data loading and transformation, data preprocessing, visualization, modeling, evaluation, and deployment. The Waikato Environment for Knowledge Analysis (WEKA) library is included in the package.

(*Continued*)

Table 3.6 Selected commercial packages (Continued)

Package and website	Brief description
XLMiner http://www.solver.com/xlminer-platform	A friendly data mining tool. The advantage of a data mining tool like XLMiner (which also has statistics functionality) is that it runs as an add-on to MS Excel, so it is convenient to further develop spreadsheet models with the analysis results.
SAP Lumira http://go.sap.com/product/analytics/lumira.html	SAP Lumira is data visualization software that allows users to access, transform, and visualize data. A 30-day trial version of the standard edition is available.
Tableau http://www.tableausoftware.com	Tableau in particular is increasing in popularity because visual analysis and presentations have strong appeal among managers.
IBM Many Eyes http://www-01.ibm.com/software/analytics/many-eyes/	IBM's Many Eyes is a visualization package that allows users to gain insights from data without having to do extensive programming or have deep technical expertise.

Table 3.7 Selected open source packages

Package and Website	Brief description
Project R www.r-project.org	R is a free software environment for statistical computing and graphics. It compiles and runs on a wide variety of UNIX platforms, Windows, and MacOS.
WEKA http://www.cs.waikato.ac.nz/ml/weka/	WEKA is a widely used suite of machine-learning software, developed at the University of Waikato, New Zealand. It contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization. It is widely used in industry, academia, and government to implement statistical machine-learning applications for data mining and predictive analytics.
Octave http://www.gnu.org/software/octave/	GNU Octave is a high-level interpreted language, primarily intended for numerical computations, including linear and nonlinear problems.
KNIME (Konstanz Information Miner) http://www.predictiveanalyticstoday.com/knime/	KNIME Desktop is a graphical workbench for data access, data transformation, predictive analytics, visualization, and reporting. KNIME uses a modular pipelining architecture to integrate multiple techniques from machine learning and data mining accessible through a graphical interface. Other variants are available for servers and clusters.
Scikit-learn http://scikit-learn.org/dev/index.html NumPy http://www.numpy.org/ SciPy http://www.scipy.org/index.html	A simple, easy to use tool for data mining and data analysis written in Python, a scripting language, for use in scientific analysis and programming. It incorporates NumPy and SciPy—two Python packages for numerical and scientific computing.
Apache Mahout http://mahout.apache.org/	Apache Mahout is an alternative to the WEKA library that provides scalable machine-learning algorithms for filtering, clustering, and classification. Many other data analytics packages use Mahout as a core subsystem.

programs in its adoption. There are a number of very good reasons for this popularity, including (1) R is a special implementation of an older language called S, optimized for mathematical operations, matrix algebra, and statistics, which makes it ideally suited to write computational programs; (2) R can be used with a development tool called R Studio, which is a very friendly user interface with four panes that allow the analyst to work directly in the R console or in an R script, but also view things like variables, data sets, help screens, results, and so on; (3) R is easy to learn—novice analysts begin by typing simple interactive commands in the R console; (4) R has a data object called "data frame," which is credited for some of the popularity of R because it makes it very easy to work with data in tables; (5) R has literally thousands of free packages in public libraries, which users can install, activate, and use instantly with minimal programming; and (6) R has a feature called "views," which makes it easy to install a collection of all the necessary packages for a particular type of analysis (e.g., econometrics) in one scoop.

CHAPTER 4

Big Data Infrastructure—A Technical Architecture Overview

Four elements composed of processing capability, storage capacity, data transport bandwidth, and visualization capability provided by systems and analytical software techniques constitute the basic infrastructure of Big Data from our perspective. We will address each of these components in this chapter.

First, we view data processing as having two basic paradigms: batch and stream processing. Batch processing has high-latency, whereas stream processing analyzes small amounts of data as they arrive. It has low-latency and, depending on the arrival rate, volume can mount up very quickly. If you try to process a terabyte or more of data all at once, you will not be able to do it in less than a second with batch processing. On the other hand, smaller amounts of data can be processed very fast—even on the fly.

As Big Data scales upwardly to exabyte and much larger volumes, it is clear that single processors, and even small multiprocessors, cannot provide the computational power to process all the data. Large multiprocessor systems have evolved—as grid architectures and cloud-based systems—to handle the large volumes of data. Having powerful computers providing trillions of instructions per second is not enough. The computer system must therefore be balanced across processing capability, and both the second and third components—storage capacity and data transport bandwidth—to ensure that Big Data can be processed in a time interval consonant with the time to decision and utilization of the extracted and derived information.

Additionally, a Big Data processing system must incorporate visualization techniques to provide the user with the ability to understand and navigate through the data and the resulting information derived from the data by analytics. These four elements, along with the systems and analytical software suites, constitute the basic infrastructure of a Big Data computing capability.

Data and Information Processing

Data processing infrastructure has evolved through several generations since the first mainframes were developed in the 1950s. The most recent manifestations have been threefold: (1) cluster computing, (2) cloud computing, and (3) processing stacks. Cluster computing and cloud computing are focused on scaling the computational infrastructure as an organization's needs evolve. Processing stacks provide open source software (OSS) frameworks for developing applications to support a business data analytics capability. In addition, an organization must decide on a suite of programming systems, tools, and languages in order to develop custom applications compatible with the analytic suites that it may purchase or obtain through OSS.

Big Data success will ultimately depend on a scalable and extensible architecture and foundation for data, information, and analytics. This foundation must support the acquisition, storage, computational processing, and visualization of Big Data and the delivery of results to the clients and decision-makers.

Service-Oriented Architecture

Service-oriented architecture (SOA) is a paradigm for designing distributed, usually interactive, systems. An SOA is essentially a collection of services running on one or more hardware-software platforms. These services communicate with each other through established protocols, by which we say the services are *interoperable*. The communication can involve either simple data passing or it could involve multiple services coordinating some activity. The services are connected to each other through software mechanisms supported by the software infrastructure.

SOAs evolved from transaction processing systems as a general software architecture. A *service* is a self-contained software unit that performs one or a few functions. Here, by service, we mean the software module that implements the service that was previously defined in "Defining Services" section. It is designed and implemented to ensure that the service can exchange information with any other service in the network without human interaction and without the need to make changes to the underlying program itself. Thus, services are usually autonomous, platform-independent, software modules that can be described, published, discovered, and loosely coupled within a single platform or across multiple platforms. Services adhere to well-defined protocols for constructing and parsing messages using description metadata.

Web Services

One implementation of SOA is known as *web services* because they are delivered through the web. The advantages of web services are interoperability, functional encapsulation and abstraction, loose coupling, reusability, and composability. Because communication between two web service modules is through HTML or eXtended Markup Language (XML), the communication is independent of any particular messaging system. The message definition is embedded in the message so that each receiver, knowing how to parse HTML/XML, can readily understand the message contents. This allows any service to *interoperate* with any other services without human intervention and thus provides a capability to *compose* multiple services to implement complex business operations. Services can be *reused* by many other different services without having to implement a variation for each pair of business transaction interactions. *Functional encapsulation and abstraction* means that functions performed on the client and server sides are independent of each other. Through *loose coupling*, in which the client sends a message and sometime later, the server receives it, allows the client and server to operate independently of each other, and more importantly to reside separately on geographically and physically independent platforms. Web services are built on a number of components as described in Table 4.1.

Table 4.1 Typical web service components

Component	Brief description
Web browser	A software module that displays web pages encoded in HTML and provides interactive functionality through the modules below.
Javascript https://angularjs.org/ https://nodejs.org/	A web page and applet scripting language that is similar to, but not Java. Many variations of Javascript exist, including AngularJS, NodeJS
Ajax https://netbeans.org/kb/docs/web/ajax-quickstart.html	This (asynchronous JavaScript and XML) is a method for building interactive Web applications that process user requests immediately. Ajax combines several programming tools including JavaScript, dynamic HTML, XML, cascading style sheets, the Document Object Model and the Microsoft object XMLHttpRequest. Ajax is often embedded in a jQuery framework.
jQuery http://jQuery.com	jQuery is a small, fast Javascript library for developing client-side applications within browsers to send and retrieve data to back-end applications. It is designed to navigate a document, select DOM elements, create animations, handle events, and develop Ajax applications.
Jersey https://jersey.java.net/	Jersey is a Java framework for developing RESTful web services that connect a browser-based client to applications code on a server. Jersey implements the JAX-RS API and extends it to provide programmers with additional features and utilities to simplify RESTful service development.
Glassfish https://glassfish.java.net/	An open-source application server for the Java EE platform that supports Enterprise JavaBeans, JPA, JavaServer Faces, Java Message System, Remote Method Invocation, JavaServer Pages, servlets, and so on. It is built on the Apache Felix implementation of the OSGI framework.

Cluster Computing

Cluster computing is an outgrowth of the distributed processing architectures of the 1980s but achieved its major impetus from the high performance computing community as it was applied to very large-scale scientific processing requiring trillions of computational cycles. Cluster machines can be connected through fast local area networks, but are sometimes geographically distributed. Each node runs its own instance of an operating system. The machines in a cluster may be homogeneous or heterogeneous.

As with parallel processors and cloud computing, effective and efficient use of cluster systems requires careful attention to software architecture and distribution of work across multiple machines. *Middleware* is software that sits atop the operating systems and allows users to "see" the cluster of machines as essentially a single, multi node machine. One common approach is to use Beowulf clusters built in commodity hardware and OSS modules.

Cloud Computing

Cloud computing is a maturing technology in which an IT user does not have to physically access, control (operate), or own any computing infrastructure other than, perhaps, workstations, routers, and switches, and, more recently, mobile client devices. Rather, the user "rents or leases" computational resources (time, bandwidth, storage, etc.) in part or whole from some external entity. The resources are accessed and managed through logical and electronic means. A *cloud architecture* can be physically visualized as the arrangement of large to massive numbers of computers in distributed data centers to deliver applications and services via a utility model. In a true physical sense, many servers may actually function on a high capacity blade in a single data center.

Rather than providing the user with a permanent server to connect to when application execution is required, cloud computing provides "virtualized servers" chosen from a pool of servers at one of the available data centers. A user's request for execution of a web application is directed to one of the available servers that have the required operating environment, tools, and application locally installed. Within a data center, almost any application can be run on any server. The user knows neither the physical server nor, in many cases, where it is physically located, that is, it is *locationally irrelevant*.

Confusion still exists about the nature of cloud computing. Gartner asserts that a key characteristic is that it is "massively scalable" (Desisto, Plummer, and Smith 2008). Originally, cloud computing was proposed as a solution to deliver large-scale computing resources to the scientific community for individual users who could not afford to make the huge investments in permanent infrastructure or specialized tools, or could not lease needed infrastructure and computing services. It evolved, rapidly, into a medium of storage and computation for Internet users that offers economies of scale in several areas. Within the past 10 years, a plethora of applications based on cloud computing have emerged including various e-mail services (HotMail, Gmail, etc.), personal photo storage (Flickr), social networking sites (Facebook, MySpace), or instant communication (Skype Chat, Twitter).

While there are public cloud service providers (Amazon, IBM, Microsoft, Apple, Google, to name a few) that have received the majority of attention, large corporations are beginning to develop "private" clouds to host their own applications in order to protect their corporate data and proprietary applications while still capturing significant economies of scale in hardware, software, or support services.

Types of Cloud Computing

Clouds can be classified as public, private, or hybrid. A public cloud is a set of services provided by a vendor to any customer generally without restrictions. Public clouds rely on the service provider for security services, depending on the type of implementation. A private cloud is provided by an organization solely for the use of its employees, and sometimes for its suppliers. Private clouds are protected behind an organization's firewalls and security mechanisms. A hybrid cloud is distributed across both public and private cloud services.

Many individuals are using cloud computing without realizing that social media sites such as Facebook, Pinterest, Tumblr, and Gmail all use cloud computing infrastructure to support performance and scalability. Many organizations use a hybrid approach where publicly available information is stored in the public cloud while proprietary and protected information is stored in a private cloud.

Implementations of Cloud Computing

The original perspective on cloud computing was defined by the National Institute for Science and Technology (NIST) as *software-as-a-service* (SaaS), *platform-as-a-service* (PaaS), or *infrastructure-as-a-service* (IaaS)

Table 4.2 Linthicum's perspectives on cloud computing

Category	Admin	Client	Example
Storage-as-a-service (SaaS)	Limited control	Access only	Amazon S3
Database-as-a-service (DBaaS)	DB management	Access only	Microsoft SSDS
Information-as-a-service (INaaS)	DB management	Access only	Many
Process-as-a-service (PRaaS)	Limited control	Access only	Appian Anywhere
Application-as-a-service (AaaS) (software-as-a-service)	Total control	Limited tailoring	SalesForce.com Google Docs Gmail
Platform-as-a-service (PaaS)	Total control	Limited programmability	Google App Engine
Integration-as-a-service (IaaS)	No control (except VM)	Total control (except VM)	Amazon SQS
Security-as-a-service (SECaaS)	Limited control	Access only	Ping Identity
Management/governance-as-a-service (MGaaS)	Limited control	Access only	Xen, Elastra
Testing-as-a-service (TaaS)	Limited control	Access only	SOASTA

(Mello and Grance 2011). As cloud computing concepts have evolved, additional perspectives have emerged. Linthicum (2009) identified those presented in Table 4.2.

Access to Data in a Cloud

There are three generally accepted methods for access data stored in a cloud: file-based, block-based, and web-based. A *file-based method* allows users to treat cloud storage as essentially an almost infinite capacity disk. Users store and retrieve files from the cloud as units. File-based systems may use a network file system, common Internet file system, or other file management system layered on top of a standard operating system.

A *block-based method* allows the user a finer granularity for managing very large files where the time to retrieve a whole file might be very long. For example, retrieving a multi-petabyte file might take hours. But, if only a subset of data is required, partitioning the file into blocks and storing and retrieving the relevant blocks yields substantially better performance. These two methods just treat the data as a group of bytes without any concern for the content of the file.

A different approach is to use *web-based methods* to access data through REST services or web servers. Here data may be stored as web pages or as semantically annotated data via XML or as linked data via resource description framework (RDF). Using these methods, a user can store or retrieve explicit data items or sets of related data.

The granularity of storage and retrieval is from high-to-low for file-based to block-based to web-based methods. However, the performance is also high-to-low as well because as the size of the data subset becomes smaller, more computation is required at the cloud computing infrastructure to locate the data units for storage or retrieval.

Moving to the Cloud

Cloud computing is an enabling technology for creating target platforms on which to host services. It is not a mechanism for specifying the explicit technologies to be used in implementing a cloud to support a particular business environment.

Today, given the popularity of cloud computing, the emergence of robust system software and application frameworks, and the significant pressure on business operations of IT infrastructure costs, many organizations are making a decision whether to move to the cloud or not. Many organizations decide to experiment with cloud computing, to implement a small low risk application to assess cloud computing. The availability of web-accessible cloud computing environments makes it easy to try cloud computing (Kaisler, Money, and Cohen 2012).

Processing Stacks

The advent of OSS has introduced a new approach to designing and developing software for commercial and business use—the so-called stack approach. A stack is a vertical hierarchy of different software modules that provide a suite of services for a comprehensive computing infrastructure. Usually, no additional software is required to provide an environment for developing and supporting applications. Several stacks have come into widespread use—originating the scientific community, but now adopted by the business community as well.

There are many extent stacks developed by and for specific communities of developers and users. Four of the most popular are described in the following sections.

The LAMP Stack

The LAMP stack (Linux-Apache-MySQL-Perl/PHP/Python) consists of four components:

- Linux operating system: Linux, an open source operating system, provides the framework in which the other components operate. The major versions are Debian, Ubuntu, Centos, Red Hat, and Fedora. These distributions generally provide a package system that includes a complete LAMP stack.
- Apache web server: While Apache remains the primary web server, other web servers such as Tomcat and Nginx have also been incorporated into the stack.
- MySQL or Maria database management system: Other DBMSs, including the MongoDB NoSQL DBMS, have been integrated into the LAMP stack.
- PHP, Perl, or Python programming systems: Scripting systems that allow relative non programmers to write some small programs on the server side to manage data and do simple analysis.

The structure of the LAMP stack is depicted in Figure 4.1.

The Leap Stack

The *LEAP stack* is a stack for cloud-based solution infrastructure that consists of four elements:

- The **L**inux operating system
- **E**ucalyptus, a free and open-source computer software for building Amazon Web Services compatible private and hybrid cloud computing environment

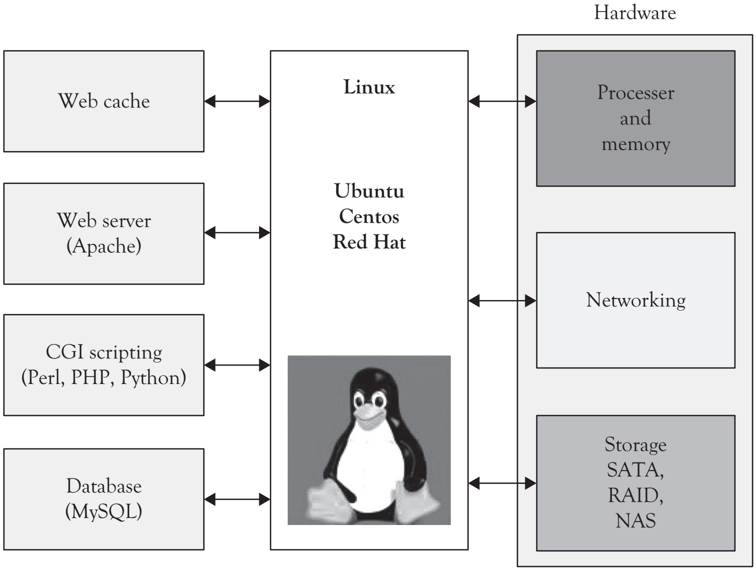


Figure 4.1 The LAMP stack

- **A**ppScale, a platform that automatically deploys and scales unmodified Google App Engine applications over public and private cloud systems
- **P**ython programming language

MAMP/WAMP Stacks

The *MAMP/WAMP stacks* are composed of four elements, of which the lower three layers are common to both stacks. The difference is that these stacks are based on either the Macintosh operating systems or the Windows operating system.

- **M**acOS or **W**indows operating system
- **A**pache Web Server
- **M**ySQL database management system
- **P**HP, **P**erl, or **P**ython programming systems

These two stacks operate in a manner similar to the LAMP stack, but are comprised of different components.

Real-Time Big Data Analytics Stack

The *real-time big data analytics* (RTBDA) stack is an emerging stack that focuses on predictive analytics (Barlow 2013). Figure 4.2 depicts a modified adaptation from Barlow.

In the data layer, data is provided from multiple sources including both streaming and persistent sources. The analytic layer contains the different analytics for processing the data. It also contains the local data mart (a subset of a data warehouse typically focused on a small portion of the total data stored in a warehouse) where processed data is forward deployed that is about to be processed. Depending on the context and the problems to be worked, not all data may be forward deployed, but only data necessary for immediate processing.

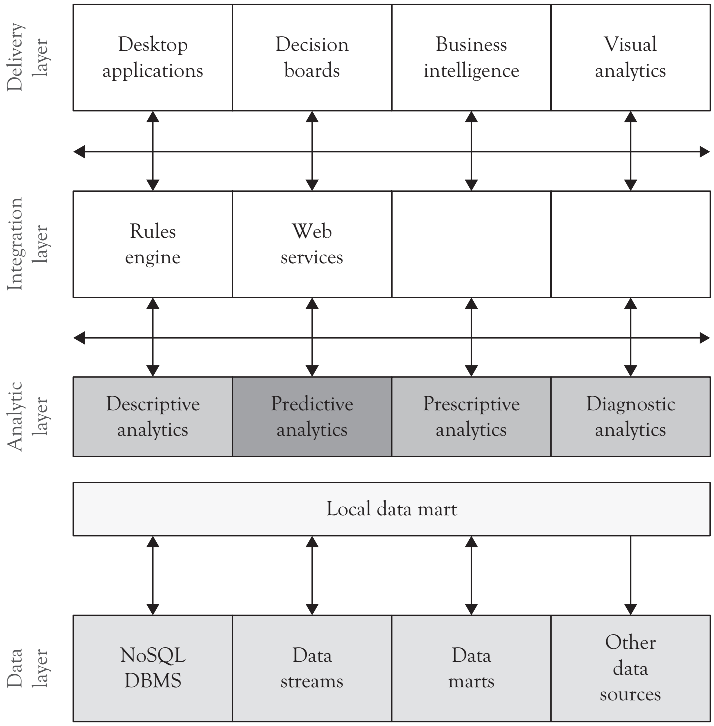


Figure 4.2 RTBDA stack

Source: Adapted from Barlow (2013).

The integration layer merges the different processed data stream into a single useful block of data or information. Finally, in the delivery layer, processed data and information for decision making and control is presented. Different types of high-level analytics most closely associated with the actual business decisions to be made are located in this layer. In addition, applications or modules that transmit data to other locations for use may be located here.

Fedora Repository Stack

The *Fedora repository stack*, initially developed at the University of Virginia but now an open community project, provides a structure for storing and archiving data sets. Figure 4.3 depicts the structure of the Fedora repository stack.

> Fedora is a robust, modular, open source repository system for the management and dissemination of digital content. It is especially suited for digital libraries and archives, both for access and

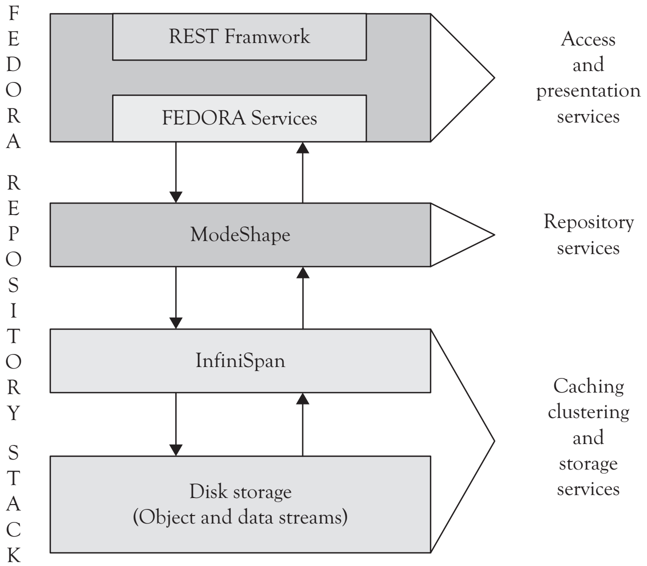

FEDORA REPOSITORY STACK
REST Framwork
FEDORA Services
Access and presentation services
ModeShape
Repository services
InfiniSpan
Disk storage (Object and data streams)
Caching clustering and storage services

Figure 4.3 Fedora repository stack

Source: Adapted from Fedora website.

> preservation. It is also used to provide specialized access to very large and complex digital collections of historic and cultural materials as well as scientific data. Fedora has a worldwide installed user base that includes academic and cultural heritage organizations, universities, research institutions, university libraries, national libraries, and government agencies. The Fedora community is supported by the stewardship of the DuraSpace organization. (https://wiki.duraspace.org/display/FF/Fedora+Repository+Home)

The rest framework provides access to the repository services from web services that can be implemented in a number of different web browsers. Currently, it uses the Drupal content management system running within your favorite web browser. Fedora services provide the essential set of services for archiving, management, and dissemination of data sets. Repository services manage the individual data sets and documents stored in the repository. The caching, clustering, and storage services interact with the underlying operating system to manage physical disks within the repository computer system.

Analytical Frameworks

An *analytical framework* is a software application framework that provides a structure for hosting applications and providing runtime services to them. Application frameworks allow the application developers to focus on the problem to be solved rather than the challenges of delivering a robust infrastructure to support web applications or distributed processing/parallel processing (Kaisler 2005). Several analytical frameworks are briefly described in Table 4.3.

The allure of hardware replication and system expandability as represented by cloud computing along with the MapReduce and MPI parallel programming systems offers one solution to solving these infrastructure challenges by utilizing a distributed approach. Even with this creative approach, significant performance degradation can still occur because of the need for communication between the nodes and movement of data between the nodes (Eaton et al. 2012).

Table 4.3 Selected analytical frameworks

System/website	Brief description
Hadoop MapReduce http://hadoop.apache.org/	MapReduce, an idea first originated by Google, is both a programming model and a runtime environment for distributing data and supporting parallel programming on commodity hardware to overcome the performance challenges of Big Data processing.
Pig http://pig.apache.org/	Pig is a platform for analyzing large data sets that consists of a high-level language for expressing data analysis programs, coupled with infrastructure for evaluating these programs. It is a subset of the functionality contained in SQL, but adapted to NoSQL databases. Pig queries are amenable to parallelization which makes it an easy to learn and use query language for large data sets.
Open Service Gateway Initiative (OSGI) http://www.osgi.org/Main/HomePage	OSGI is a specification for a distributed framework of composed and many different reusable components. It enables components to hide their implementations from other components while communicating through *services*, which are objects that are specifically shared between components. It is designed and specified by the OSGI Alliance, a worldwide consortium of technology innovators dedicated to open source implementations that enable the modular assembly of software built with Java technology
Message Passing Interface (MPI) http://www.open-mpi.org/	Developed at Argonne National Laboratories, MPI is a software platform and set of libraries for enabling communication and coordination among a distributed set of processes via a message passing paradigm. It supports multiple 32- and 64-bit operating systems and several programming languages.
KNIME https://www.knime.org/	KNIME is an open source platform for data-driven analytics that includes multiple components for data access, data transformation, analysis and data mining, visualization, and deployment.

Programming Systems

Aside from the standard suite of programming and scripting languages that are used today (such as Java, Python/Perl, PHP, C/C++/C#, VSBASIC, etc.), numerous programming systems have been developed and used that provide support for the development of analytic applications.

One class of programming systems is interactive development environments (IDEs) for supporting software development. The two leading IDEs are Eclipse and Netbeans. BlueJ, developed by the

Table 4.4 Selected development systems

System/website	Brief description
Python https://www.python.org/	An interpreted software language, which has become very popular for analytics because it is object oriented and simple, and facilitates rapid programming. It has an extensive library of contributed packages that are applicable to many different types of analytic problems.
Eclipse https://www.eclipse.org/	Eclipse is an integrated development environment (IDE), originally developed by IBM, but then converted to an open source community project. It supports a wide variety of programming languages including Ada, C/C++, PHP, Python, Ruby, FORTRAN, and Javascript among others.
Netbeans https://netbeans.org/	Netbeans is an IDE originally developed for Java by Sun Microsystems, but then converted to an open source community project. It recently has been extended to support PHP, Ruby, C, Javascript, and HTML. Versions of Netbeans run on the major operating systems.
BlueJ http://www.bluej.org	A simple IDE geared toward academic environments. However, if your staff has no experience with IDEs previously, this tool is easy to use, provides very good graphical support for understanding the structure of the application, and good support for debugging applications. Supports Java only.

University of Kent and supported by Sun Microsystems, is an IDE primarily used in academic and small team environments. Table 4.4 presents some popular open source systems that have been used by many organizations.

Martin Fowler (2011) has observed that many large-scale Big Data analytics (BDA) systems exhibit polyglot programming. Polyglot programming was defined by Neal Ford (2006) as the development of applications in a mix of programming languages. While as recently as a decade ago most applications were written in a single programming language, the influence of the model-view-controller paradigm that underlies web development has led to the use of different languages for different components of the system. Each component has different performance requirements, structural representations, and functional processing needs. Thus, it makes sense to select a programming language most suitable for each of these needs.

Data Operations Flow

In order to put the data to use—or find the value in the data, it is not enough to have Big Data that one can query. A business owner needs to be able to extract meaningful, useful, and usable information for strategic, tactical, and operational decision-making in the business environment. An organization cannot begin analyzing raw Big Data right away for several reasons. It needs to be cleansed, perhaps formatted for specific analytics, even transformed or translated to different scales to yield useful results. This problem has been characterized as the ETL (extract, transform, load) problem.

Data operations flow is the sequence of actions from acquisition through storage and analysis, eventually to visualization and delivery. Table 4.5 briefly describes some of the operations to be performed on Big Data.

Data Acquisition

With Big Data, it is often easier to get data into the system than it is to get it out (Jacobs 2009). Data entry and storage can be handled with processes currently used for relational databases up to the terabyte range, albeit with some performance degradation. However, for petabytes and beyond new techniques are required, especially when the data is streaming into the system.

Many projects demand "real-time" acquisition of data. But, what they do not understand is that real-time online algorithms are constrained by time and space limitations. If you allow the amount of data to be unbound, then these algorithms no longer function as real-time algorithms. Before specifying that data should be acquired in "real time," we need to determine which processing steps are time-bound and which are not. This implies that designing a real-time system does not mean every stage must be real-time, but only selected stages. In general, this approach can work well, but we must be careful about scaling. All systems will break down when the scale or volume exceeds a threshold maximum velocity and volume (plus a fudge factor) for which the system was designed. Over engineering a real-time system is a good idea because

Table 4.5 Big Data operations

Operation	Description
Acquiring/capturing	*Data acquisition* is the first step in utilizing Big Data. Because Big Data exists in many different forms and is available from many different sources, you will need many different acquisition methods. Many of these may already be built into your business operations systems.
Management	*Data curation* is the active management of data throughout its life cycle. It includes extracting, moving, cleaning, and preparing the data.
Cleansing	*Data cleansing* includes verifying that the data are valid within the problem parameters and removing outliers from the data set. However, there is a risk that the phenomena may not be fully understood when outliers are removed.
Storing	*Data storing* is a key function, but involves critical decisions about whether to move or not move data from the point of origin to the point of storage or the point of processing; how much data to keep (all or some); whether any data is perishable and must be analyzed in near real-time; and how much data must be online versus off-line—among other aspects.
Movement	*Data movement* is a critical problem when the volume of data exceeds the size of the applications that will be processing it. Moving a gigabyte of data across a 100 Mbit/s Ethernet takes more than 80 seconds depending on the network protocols and the handling of data at either end of the transfer route. Data movement should be minimized as much as possible. Moving programs to the data locations is a more efficient use of network bandwidth, but the tradeoff is more powerful servers at the data storage sites.
Analyzing/processing	In *data analysis*, there are basically two paradigms: batch and stream processing. Batch processing has high-latency, whereas stream processing analyzes small amounts of data as they arrive. It has low-latency and, depending on the arrival rate, the volume can mount up very quickly. The section "Data and Information Processing" has described some of the analytics and processing approaches and challenges for Big Data.

bursts are often unpredictable and create queuing theory problems with unpredictable queue lengths and waiting times for processing or transport of data. Table 4.6 presents a sample of the many challenges in data acquisition. Table 4.7 presents some data acquisition tools that address many of these issues.

Table 4.6 Some challenges for Big Data acquisition and storage

Has quality control been maintained throughout the acquisition of the data? If not, are quality control parameters clearly identified for different data subsets?
What validation and verification procedures were used during data acquisition?
How to handle data lost during the data acquisition process? What if it is due to acceptance and storage problems?
Guaranteeing reliability and accessibility of data given the axiom of maintaining one copy of data, not many (Feinleb 2012).
How does one drive the acquisition of high-value data?
A corollary to the above: How does one predict what data is needed to plan, design, and engineer systems to acquire it?
How does one decide what data to keep, e.g., what data is relevant, if it cannot all be kept?
How does one access and manage highly distributed data sources?

Table 4.7 Selected data acquisition tools

Tool/website	Brief description
Storm http://storm.apache.org/	Storm is an open-source event processing system developed by BackType, but then transferred to Apache after acquisition by Twitter. It has low latency and has been used by many organizations to perform stream processing.
IBM InfoSphere Streams http://www-01.ibm.com/software/data/infosphere/	Streams are a platform and execution environment for stream processing analytics, which supports the diversity of streams and stream processing techniques.
Apache Spark Streaming https://spark.apache.org/streaming/	Apache Spark is a fast and general-purpose cluster computing system overlaid on commodity hardware. It provides high-level APIs in Java, Scala, Python, and R and an optimized engine that supports general execution graphs. It also supports a rich set of higher-level tools including Spark SQL for SQL and structured data processing, MLlib for machine learning, GraphX for graph processing, and Spark Streaming.

You also need to determine what is "real-time." It can depend on the context in which it is used: options trading, missiles flying, driving down the highway at 75 mph. Typically, what we mean by "real-time" is that the system can respond to data as it is received without necessarily storing it in a database first. This means that you can process the data as it arrives as opposed to storing it and processing it later. In effect, you can respond to events as they are happening and situations as they are unfolding.

Input Processes

Data can be acquired in batches or in streams. Batches may be small, medium, or large, but as they aggregate, and the frequency with which they are acquired increases, they can amount to Big Data very easily. On the other hand, streaming data arrives continuously and, often, at high speed.

Streaming is the "continuous" arrival of data from a data source. Arrival may be periodic, but at fixed intervals with fixed or variable amounts of data. More often, streaming is assumed to be a continuous flow of data that is sampled to acquire data. In the worst case, sampling acquires every piece of data, but often it involves selecting only certain types of data or sampling on a periodic basis. A good example is social media where posts are made at varying intervals depending on the poster. In general, the flow of information appears continuous, but is really aperiodic.

Unlike a data set, a *data source* has no beginning and no end. One begins collecting data and continues to do so until one has enough data or runs out of patience or money or both. Stream processing is the act of running analytics against the data as it becomes *initially available* to the organization. The data streams in with varied speed, frequency, volume, and complexity. The data stream may dynamically change in two ways: (1) the data formats change, necessitating changes in the way the analytics process the data, or (2) the data itself changes necessitating different analytics to process it. A complicating factor is the implicit assumption that the data streams are well-behaved and that the data arrive more or less in order. In reality, data streams are not so well-behaved and often experience disruptions and mixed-in data, possibly unrelated, to the primary data of interest.

Data Growth versus Data Expansion

Most organizations expect their data to grow over their lifetime as the organization increases its services, its business and business partners and clients, its projects and facilities, and its employees. Few businesses adequately consider data expansion, which occurs when the data records grow in richness, when they evolve over time with additional information as new techniques, processes, and information demands evolve. Most data is time-varying—the same data items can be collected over and over with different values based on a timestamp. Much of this time-stamped data is required for retrospective analysis—particularly that which is used in estimative and predictive analytics.

As the volume of data grows, the "big" may morph from the scale of the data warehouse to the amount of data that can be processed in a given interval, say 24 hours, using the current computational infrastructure. Gaining insight into the problem being analyzed is often more important than processing all of the data. Time-to-information is critical when one considers (near) real-time processes that generate near-continuous data, such as radio frequency identifiers (RFIDs—used to read electronic data wirelessly, such as with EZPass tags) and other types of sensors. An organization must determine how much data is enough in setting its processing interval because this will drive the processing system architecture, the characteristics of the computational engines, and the algorithm structure and implementation.

Output Processes

A major issue in Big Data system design is the output processes. Jacobs (2009) summarized the issue very succinctly—"… it's easier to get the data in than out." However, the tools designed for transaction processing that add, update, search for, and retrieve small to large amounts of data are not capable of extracting the huge volumes and cannot be executed in seconds or even in a few minutes.

How to access very large quantities of semi- or unstructured data, and how to convey this data to visualization systems for assessment and managerial or executive decision making is a continuing problem. It is clear the problem may neither be solved by dimensional modeling and online analytical processing, which may be slow or have limited functionality, nor by simply reading all the data into memory and then reading it out again, although recent in-memory data analysis systems have demonstrated an ability to analyze terabytes of data in a large memory.

Technical considerations that must be factored into the design include the ratio of the speed of sequential disk reads to the speed of random memory access. The current technology shows that random access to memory is 150,000 times slower than sequential access. Joined tables, an assumed requirement of associating large volumes of disparate but somehow related data, perhaps by observations over time alone, will come at further huge performance costs (Jacobs 2009).

Data Curation

Once acquired, most data is not ready for immediate use. It must be cleansed to remove inaccuracies, transformed to canonical forms, and even restructured to make it easier to manage and access. A critical aspect is identifying and associating metadata to the data sets in order to make them searchable and to understand what is contained therein.

Data curation includes extracting, moving, cleaning, and preparing the data, which includes the process of data enrichment. The University of Illinois Graduate School of Library and Information Science defined data curation as the active management of data throughout its useful life cycle (CLIR 2012)—a more robust handling of data. It is all about improving the quality of the data in hand. Curation is an important process because data is often acquired with varying degrees of reliability, accuracy, precision, and veracity. One aspect of curation is to canonicalize the data to a common set of units or to base concepts. Poor data quality can skew or invalidate results produced by the best algorithms. However, data curation tools cannot replace human analysis and understanding. However, with Big Data, humans beings cannot do it all (or even a small part of it), and so we must trust the tools to assist in assessing and improving data quality. Table 4.8 presents some elements of the digital curation process.

Data Cleansing

Data cleansing is the process of detecting and correcting or removing data from a data set. It includes verifying that the data are valid within the problem parameters and removing outlying data while risking that the phenomena may not be fully understood. The objective of data cleansing is to ensure a consistent data set that will allow some degree or measure of trust in the results generated from the data set. Table 4.9 presents some criteria for data cleansing.

Table 4.8 Selected elements of digital curation

Know your data—both what you acquire and what you generate. Develop an acquisition plan, a storage plan, a management plan, and a retirement plan. Authenticate the data with metadata so its provenance is established.
Creating data—for data you generate, associate the appropriate metadata with it.
Determine accessibility and use. Identify the classes of users of the data, what they can and cannot use, and what the accessibility constraints on the data are and how they will be enforced. Check periodically to ensure that the data is accessible and uncorrupted.
Establish selection criteria—you may not be able to keep all of the data that you acquire. Establish selection criteria that determine what data you will keep based on its relevancy to the organization's mission and operation. One criterion to consider is perishability—how useful the data is or will be after a given time.
Disposal plan and operations—it is not simply a matter of erasing digital data or throwing paper or other media in a recycling bin. Some data is sensitive and must be properly discarded—shredding of paper and other media, overwriting data stores with zeroes, and so on. Periodically appraise you data to determine it is still relevant and useful for your organization's mission and operations.
Ingest the data—including data curation, cleaning, and transforming prior to storing it. It may be desirable or necessary to transform the data into a different digital format before storing it—based on its intended use or efficiency of storage.
Preserving the data—ensuring the data is available, protected, and backed up so if the primary store is lost, the data will be recoverable.
Retention of data—are there legal requirements for retaining the data? For a fixed period? For a long time?
Access to the data—is access to the data limited by law? For example, HIPAA and the Privacy Act (amended) restrict access to personal data through personal identifiable information (PII). (Data that in itself or combined with other information can identify, contact, or locate a specific person, or an individual in a context.)

Data correction may include correcting typographical errors created during data collection. Or, it may be correcting values against a list of known entities or a range of valid values. For example, checking postal zip codes against the names of towns and ensuring the correct zip code is associated with the address.

Data Storage

As data science has become more important in the business analytics community, the data can often be aggregated and linked in different ways. Sharing and archiving have become important aspects of being able to

Table 4.9 Some criteria for data cleansing

Validity: A measure of the conformance of data to the criteria set established for useful data within the organization. Each data set should have a set of validity criteria and a mechanism for establishing the validity of each element. If the data set is refreshed—either periodically or continually—then the validity criteria should be checked on the same schedule.
Retention: Although data has been removed from the data set, should it be retained? For example, there may be legal requirements to retain the original data set even though the cleansed data set is used in further processing.
Duplicate removal: Removal of duplicates is often performed through data correction or outright removal. However, there may not be enough information available to determine whether duplicate removal should occur. The risk is loss of information.
Streaming data: Cleansing of streaming data is problematic for at least two reasons. First, decision about correction, transformation, and duplicate removal have to be made within a short time frame and, often, with insufficient information. Second, if the data re-occurs, a question arises as to whether the earlier cleansing was the right decision.
User perspective: Different users may have different perspectives on how data should be cleansed for a particular problem. For example, the question of "how many employees do we have" should result in different answers depending on the person asking the question. This implies that different versions of the data set must be maintained to satisfy different user perspectives.
Data homogenization: Sometimes, data must be homogenized to eliminate bias due to data collection methodologies and variability in instruments—whether physical or textual, such as opinion survey instruments. Homogenization is a transformation of a data to a base value with scaling to ensure the all data represents the same phenomena.

use this data efficiently. Whether Big Data exists as a very large number of small data files, a few very large data sets, or something in between, a key challenge is organizing this data for access, analysis, updating, and visualization.

This is a fundamental change in the way many businesses manage their data—a cultural change, where silos still often prevail among divisions. This problem is not unique to businesses, however, as government and science also experience the same challenges. Although the Federal government and a few state governments have moved toward increased sharing of data, including the Federal government's www.data.gov website, it is a slow process to overcome institutional and individual barriers.

To archive data for future use, one needs to develop a management plan, organize it, and physically store it in both an accessible repository as well as a backup repository. Data management is briefly described in the next section. Here, we will be concerned about technology for storing and archiving data.

The capacity of disk drives seems to be doubling about every 18 months due to new techniques, such as helical recording, that lead to higher density platters. However, disk rotational speed has changed little over the past 20 years, so the bottleneck has shifted—even with large disk caches—from disk drive capacity to getting data on and off the disk. In particular, as National Academy of Science (2013) noted, if a 100-TByte disk requires mostly random-access, it was not possible to do so in any reasonable time.

Data storage is usually divided into three structures: flat files, relational databases, and NoSQL databases. The first two are well-known structures that have been discussed extensively elsewhere.

Flat Files

Flat files are the oldest type of data storage structure. Basically, they are a set of records that has a structure determined by an external program. One of the most common types of flat files are "csv" (comma-separated value) files where an arbitrary sequence of tokens (e.g., words or numbers or punctuation other than commas) are separated by commas. The interpretation of the meaning of the sequence of tokens is provided by one or more applications programs that read and process one record at a time. CSV files are often used to exchange data between applications, including across distributed systems, because they are text files.

Relational Database Management Systems

Relational database management systems (RDBMSs) have been available both commercially and as OSS since the early 1980s. As data sets have grown in volume, relational DBMSs have provided enhanced functionality to deal with the increased volumes of data.

RDBMSs implement a table model consisting of rows with many fields. Each table is described by a primary key consisting of one or more fields. One way to think of relational databases is to view a table as a spreadsheet where the column headers are the field names and the rows are the records. Then, a database with multiple tables is like a set of spreadsheets where each spreadsheet represents a different table. This model can implement complex relationships as entries in the field of one table can be entries in the field of another table thus linking the tables together.

Much has been written about relational databases, and hence we will not address them further here. The Further Reading section contains a number of references on relational databases.

NoSQL Databases

NoSQL is a database model that differs from the traditional relational database and flat file models. It implements a simplified data model that trades off speed of access and large volume scalability for complexity in representing and managing detailed structural relationships. Typical implementations use either key-value, graph, or document representations. The term "NoSQL" is often interpreted to mean "Not Only SQL." Yet, NoSQL databases can support a subset of the standard SQL-like queries.

NoSQL Properties. NoSQL databases provide three properties, known by the acronym BASE:

- Basically available: Since a NoSQL database is often distributed across multiple servers, parts of it may not always be consistent. However, the goal is to ensure that most of it is available all of the time.
- Soft state: The state of the system may change over time, even without input. This is because of the eventual consistency model.
- Eventual consistency: Given a reasonably long duration over which no changes are sent, all updates will propagate throughout the system. This means that some copies of the data distributed across multiple servers may be inconsistent.

Types of NoSQL Databases. NoSQL databases utilize different structures to provide different levels of storage flexibility and performance. You should select a database technology that fits both the natural representation of the data and the way that you primarily expect to access the data. Table 4.10 describes the four major types of NoSQL database structures.

Table 4.10 NoSQL database types

Type	Brief description
Column	Column-family databases store data in column families as rows that have many columns associated with a row key. Each column can be thought of as a set of rows in a relational database. The difference is that various rows do not have to have the same columns, and columns can be added to any row at any time without having to add it to other rows. Column stores do not need indexes for fast access to the database contents.
Key-value	Key-value stores are the simplest NoSQL data stores to use from an access perspective. A client can either get the value for the key, put a value for a key, or delete a key from the data store. The value is a blob of data that the data store just stores, without caring or knowing what is inside; it is the responsibility of the application to understand what was stored. Key-value stores generally have very good performance and are easily scaled.
Document	A document is the unit of data stored in a document store such as MongoDB. The document can be represented in many ways, but XML, JSON, and BSON are typical representations. Documents are self-describing tree-structures that can consist of maps, collections, and scalar values. The documents stored are similar to each other but do not have to be exactly the same. The documents are stored in the value part of a key-value store.
Graph	Graph databases allow you to store entities and relationships between these entities. Entities are also known as nodes, which have properties. Relations are known as edges that can have properties. Edges have directional significance; nodes are organized by relationships which allow you to find interesting patterns between the nodes. The graph model represents data as RDF triples of the form <subject, predicate, object> tuple. Complex relationships can be represented in this format, including network and mesh data structures that features multiple named edges between named nodes.

Selected Popular NoSQL Databases

The ease of use and popularity of the NoSQL model has spawned a rich competitive environment in developing NoSQL database systems. Table 4.11 presents selected NoSQL database systems that are widely used across the Big Data community.

Table 4.11 Selected NoSQL database systems

Database/website	Brief description
HBase http://hbase.apache.org/	HBase is an open-source, nonrelational database that stores data in columns. It runs on top of the Hadoop distributed file system (HDFS), which provides fault-tolerance in storing large volumes of data. Typically, it is used with Zookeeper, a system for coordinating distributed applications, and Hive, a data warehouse infrastructure (George 2011).
MongoDB https://www.mongodb.org/	MongoDB is a distributed document-oriented database, which stores JSON—like documents with dynamic schemas.
SCIDB http://www.scidb.org/	SCIDB is an array database management system. Arrays are the natural way to organize, store, and retrieve ordered or multifaceted data. Data extraction is relatively easy—by selecting any two dimensions you extract a matrix. Array processing is very scalable parallel processors that can yield substantial performance gains over other types of databases.
Accumulo http://accumulo.apache.org/	Accumulo is a sorted, distributed key-value store which embeds security labels at the column level. Data with varying security requirements can be stored in the same table, and access to individual columns depends on the user's permissions. It was developed by a government agency, but then released as open source.
Cassandra http://cassandra.apache.org/	Cassandra is an open-source distributed database system that was designed to handle large volumes of data distributed across multiple servers with no single point of failure. Created at Facebook, Cassandra has emerged as a useful hybrid of a column-oriented database with a key-value store. Rows in the database are partitioned into tables each of whose first component is a primary key (Hewitt 2010).
Neo4J http://neo4j.com/	Neo4J is an open-source graph database which implements a data model of nodes and edges—each described by attributes.

A larger listing of NoSQL databases of various type may be found at http://nosql-database.org/.

Advantages and Disadvantages of NoSQL Databases

As with RDBMSs, NoSQL databases have both advantages and disadvantages. The primary advantage of NoSQL databases is claimed to be scalability into the petabyte range. Although several RDBMSs claim this ability, it remains to be seen whether performance scales with data volume. A few of the advantages and disadvantages are described in Tables 4.12 and 4.13 (Harrison 2010).

Table 4.12 Advantages of NoSQL databases

Advantage	Brief description
Elastic scaling	NoSQL databases allow organizations to scale out (across distributed servers) rather than scaling up (with more expensive and faster hardware). Scaling out in cloud environments is relatively easy, often transparent, and allows organizations to adjust their data storage to fit their current requirements.
Big Data	As databases grow to volumes of petabytes and beyond, transaction processing times increase exponentially. By distributing transactions across multiple servers, processing time can be kept to near-linear increases.
Reducing the DBA role	DBA expertise is often expensive. The simple models used by NoSQL databases usually require only a basic understanding of data structures. Moreover, their inherent reliability, maintainability, and performance reduce the need for many highly trained database administrators.
Economies of scale	NoSQL databases tend to use clusters of low-cost commodity hardware coupled with mechanisms to ensure high reliability and fault tolerance. Thus, the cost per terabyte or transaction/second is often significantly much less than for a conventional RDBMS.
Flexible data models	The simple and flexible data models allow data to be easily described and changed. And, databases can be expanded through new fields and columns without affecting the rest of the database. NoSQL databases are often said to be “schema-free” in that data is represented as simple JSON structures that may be rapidly searched through the application programming interface (API).

Table 4.13 Disadvantages of NoSQL databases

Disadvantage	Brief description
Maturity	Some NoSQL DBMSs are relatively mature, although all are less than 10 years old. Some data models are still evolving; the reliability and functionality of the software is still evolving.
Support	Most NoSQL DBMSs are open-source projects meaning there are many anonymous contributors to source code that continues to evolve at varying rates.
Analytics support	All NoSQL DBMSs provide a Java API, but this may be difficult to use without significant Java programming experience. The emergence of SQL-like query languages such as Pig and Hive can provide easier access to data, but at the loss of some rich SQL functionality.
Administration	NoSQL databases are NOT zero-administration databases. Organizations must still retain expertise to describe the data within the data model supported by the DBMS. Most NoSQL DBMSs require considerable expertise to install and operate within a distributed server environment and require considerable infrastructure expertise to support their operation.
Data model expertise	Even with the simple models provided by NoSQL DBMSs, expertise is still required to describe the external data structures in the internal data model representation and tune the structures to achieve high-performance.

Data Ownership

Data ownership presents a critical and ongoing challenge, particularly in the social media arena. While petabytes of social media data reside on the servers of Facebook, MySpace, and Twitter, it is not really owned by them (although they may contend so because of residency). Certainly, the "owners" of the pages or accounts believe they own the data. This dichotomy will have to be resolved in court. If your organization uses a lot of social media data that it extracts from these sites, you should be concerned about the eventual outcome of these court cases. Kaisler, Money, and Cohen (2012) addressed this issue with respect to cloud computing as well as other legal aspects that we will not delve into here.

With ownership comes a modicum of responsibility for ensuring its accuracy. This may not be required of individuals, but almost certainly is so of businesses and public organizations. However, enforcement of

Table 4.14 Some Big Data ownership challenges

When does the validity of (publicly available) data expire?
If data validity is expired, should the data be removed from public-facing websites or data sets?
Where and how do we archive expired data? Should we archive it?
Who has responsibility for the fidelity and accuracy of the data? Or, is it a case of user beware?

such an assumption (much less a policy) is extremely difficult. Simple user agreements will not suffice since no social media purveyor has the resources to check every data item on its servers. Table 4.14 presents some ownership challenges.

With the advent of numerous social media sites, there is a trend in BDA toward mixing of first-party, reasonably verified data, with public and third-party external data, which has largely not been validated and verified by any formal methodology. The addition of unverified data compromises the fidelity of the data set, may introduce non relevant entities, and may lead to erroneous linkages among entities. As a result, the accuracy of conclusions drawn from processing this mixed data varies widely.

Unlike the collection of data by manual methods, where rigorous protocols are/were often followed in order to ensure accuracy and validity, digital data collection is much more relaxed. The richness of digital data representation prohibits a bespoke methodology for data collection. Data qualification often focuses more on missing data or outliers than trying to validate every item. Data is often very fine-grained such as clickstream or metering data. Given the volume, it is impractical to validate every data item: new approaches to data qualification and validation are needed.

Going forward, data and information provenance will become a critical issue. JASON has noted (2008) that "there is no universally accepted way to store raw data, reduced data, and … the code and parameter choices that produced the data." Further, they note: "We are unaware of any robust, open source, platform-independent solution to this problem." As far as we know, this remains true today. To summarize, there is *no* perfect Big Data management solution yet. This represents an important gap in the research literature on Big Data that needs to be filled.

Data Management

Data management will, perhaps, be the most difficult problem to address with Big Data. This problem first surfaced a decade ago in the UK eScience initiatives where data was distributed geographically and "owned" and "managed" by multiple entities. Resolving issues of access, metadata, utilization, updating, governance, and reference (in publications) have proven to be major stumbling blocks. Within a large organization, data is often distributed among different business operations and "owned" and "managed" by those business operations. Without a centralized data dictionary and uniform management practices, conflicts, including inconsistencies, representational issues, data duplication among others, will arise between one or more business operations.

If we consider that data and information are both sources of business strategies, the raw material of business operations and the basis for metrics for assessing how well the organization is doing, then we begin to understand how critical it is for it to be managed well. "Managed well" is best simply described as the following: *delivering the right information to the right place (and people) at the right time in the right form at the right cost*. To do so requires a strong, sound architecture, good processes, and effective project management. It is hard to get all three elements correct.

A good analogue is to think of an *information supply chain*. We need to ensure that information flows smoothly through the business from multiple sources through multiple points of processing and analysis to multiple storage facilities and, then, out to multiple points of consumption. As with other supply chains, an effective information supply chain can enhance customer satisfaction, better support analytics to plan and manage business operations, and facilitate regulatory compliance as required. It can also reduce the need for redundant data and leverage infrastructure and personnel resources to improve the cost–benefit ratio.

Data management is different from database management. It is concerned more with organizing the data in a logical, coherent fashion with appropriate naming conventions and managing the metadata associated with the data. However, data management also has the function of distributing large data sets across multiple, distributed processors such as cloud-based or geographically distributed systems.

The reason for distributing data is that single systems cannot often process the large volumes of data represented by Big Data, for example, terabytes or petabytes of data. Thus, to achieve reasonable performance, the data must be distributed across multiple processors—each of which can process a portion of the data in a reasonable amount of time. There are two approaches to distributing data:

- *Sharding*: Sharding distributes subsets of the data set across multiple servers, so each server acts as the single source for a subset of data.
- *Replication*: Replication copies data across multiple servers, so each bit of data can be found in multiple places. Replication comes in two forms:
  - *Master-slave replication* makes one node the authoritative copy that handles writes while slave nodes synchronize with the master and handle reads. Thus, multiple concurrent reads can occur across a set of slave nodes. Since reads are more frequent than writes, substantial performance gains can be achieved.
  - *Peer-to-peer replication* allows writes to any node. The nodes coordinate to synchronize their copies of the data usually as background tasks. The disadvantages are that for short periods of time different nodes may have different versions of the data.

Master-slave replication reduces the chance of conflicts occurring on updates. Peer-to-peer replication avoids loading all writes onto a single server, thus attempting to eliminate a single point of failure. A system may use either or both techniques. Some databases shard the data and also replicate it based on a user-specified replication factor.

Some data management issues that you should consider are presented in Table 4.15 (Kaisler et al. 2013).

Data Enrichment

Data enrichment is the process of augmenting collected raw data or processed data with existing data or domain knowledge to enhance the

Table 4.15 Some management challenges for Big Data

Are the data characteristics sufficiently documented in the metadata? If not, how difficult is it to find them?
If all the data cannot be stored, how does one filter/censor/process data to store only the most relevant data?
Can ETL (extract, transform, load) be performed on all data without resorting to external mass storage?
How much data enrichment must be performed before the data can be analyzed for a specific problem?
Determining the amount of enrichment to perform on acquired data such that it does not skew or perturb the results from the original data.
How does one handle outliers and uncorrelated data?
Where is the tradeoff between the integration of diverse data (multimedia, text, and web) versus more complex analytics on multiple data sets?
What visualization techniques will help to understand the extent and diversity of the data?

analytic process. Few, if any, business problems represent entirely new domains or memoryless processes. Thus, extant domain knowledge can be used to initialize the problem solving process and to guide and enhance the analysis process. Domain knowledge can be used to enrich the data representation during and after the analysis process. Transformation through enrichment can add the necessary domain knowledge to enable analytics to describe and predict patterns and, possibly, prescribe one or more courses of action.

Data enrichment is directly influenced by the V: data veracity. About 80 percent of the work to analyze data is about preparing the data for analysis. This percentage will vary substantially depending on the quality of the data. According to Davenport (2006), the most important factor for using sophisticated analytics is the availability of high-quality data. This requires that data be precise (within context), accurate, and reliable. Accuracy of data will influence the quality of the analytic's output, per the old axiom "garbage in, garbage out." Precision will influence the exactness of the analysis and, in numerical analysis, minimize the error field. Reliability—how well we trust the data to reflect the true situation—will influence the believability and value of the results. Ensuring veracity is a result of good data acquisition and curation practices.

An example of data enrichment that resonates with us is the following advertisement from AT&T in the early 1980s. Years ago, there was an AT&T commercial on TV about selling telephone service to a trucking company. The AT&T representative approaches the general manager of the trucking company who says "Let's talk about telephones." The AT&T representative says, "No, let's talk about trucking, then we will talk about telephones." This vignette captures the essence of the problem: Understanding the domain is essential to being able to solve problems in the domain! The AT&T representative knew about telephones, but needed to understand the trucking company manager's perspective in order to understand how AT&T's services could impact and enhance his business.

Data Movement

The *velocity* dimension of Big Data often refers to just the acceptance of incoming data and the ability to store it without losing any of it. However, velocity also refers to the ability to retrieve data from a data store and the ability to transmit it from one place to another, such as from where it is stored to where it is to be processed. Google, for example, wants their data to be transmitted as fast as possible. Indeed, they establish internal constraints on their systems to make sure they give up on certain approaches very quickly if they are going to take more time than a given threshold.

Network communications speed and bandwidth have not kept up with either disk capacity or processor performance. Of the 3Es—exabytes, exaflops, and exabits—only the first two seem attainable within the next 10 years, and only with very large multiprocessor systems. The National Academy of Science (2013) has noted that data volumes on the order of petabytes mean that the data cannot be moved to where the computing is; instead, the analytical processes must be brought to the data.

Data Retrieval

In order to use Big Data to assist in solving business operations problems, we need to find the relevant data that applies to the business problem at hand. Identifying the relevant data is a search problem where we generally wish to extract a subset of data that satisfy some criteria associated with the problem at hand.

Quality versus Quantity

An emerging challenge for Big Data users is "quantity vs. quality." As users acquire and have access to more data (quantity), they often want even more. For some users, the acquisition of data has become an addiction. Perhaps, because they believe that with enough data, they will be able to perfectly explain whatever phenomenon they are interested in.

Conversely, a Big Data user may focus on quality which means not having all the data available, but having a (very) large quantity of high quality data that can be used to draw precise and high-valued conclusions. Table 4.16 identifies a few issues that must be resolved.

Another way of looking at this problem is: what is the level of precision you require to solve business problems? For example, trend analysis may not require the precision that traditional DB systems provide, but which require massive processing in a Big Data environment.

Data Retrieval Tools

The tools designed for transaction processing that add, update, search for, and retrieve small to large amounts of data are not capable of extracting the huge volumes typically associated with Big Data and cannot be executed in seconds or a few minutes.

Some tools for data retrieval are briefly described in Table 4.17.

Table 4.16 Some quantity and quality challenges

How do we decide which data is irrelevant versus selecting the most relevant data?
How do we ensure that all data of a given type is reliable and accurate? Or, maybe just approximately accurate?
How much data is enough to make an estimate or prediction of the specific probability and accuracy of a given event?
How do we assess the "value" of data in decision making? Is more necessarily better?

Table 4.17 Selected data retrieval tools

Tool	Brief description
Drill http://drill.apache.org/	Apache Drill is a low latency distributed query engine for large-scale data sets, including structured and semi-structured/nested data. It is the open source version of Google's Dremel. A version of Drill has been pre-installed in Hadoop's MapR sandbox to facilitate experimentation with drillbits, the components that receive and execute user's queries.
Flume http://flume.apache.org/	Flume is a distributed, reliable, and available service for efficiently collecting, aggregating, and moving large amounts of log data. It implements that a streaming data flow model is robust and fault tolerant with tunable reliability mechanisms and many failover and recovery mechanisms.

Table 4.18 Some data value challenges

For a given problem domain, what is the minimum data volume required for descriptive, estimative, predictive, and prescriptive analytics and decision modeling with a specified accuracy?
For a given data velocity, how do we update our data volume to ensure continued accuracy and support (near) real-time processing?
For a given problem domain, what constitutes an analytic science for non-numerical data?
"What if we know everything?"—What do we do next?

The Value of "Some Data" versus "All Data"

Not all data are created equal; some data are more valuable than other data—temporally, spatially, contextually, and so on. Previously, storage limitations required data filtering and deciding what data to keep. Historically, we converted what we could and threw the rest away (figuratively, and often, literally).

With Big Data and our enhanced analytical capabilities, the trend is toward keeping everything with the assumption that analytical significance will emerge over time. However, at any point in time the amount of data we need to analyze for specific decisions represents only a very small fraction of all the data available in a data source and most data will go un-analyzed. Table 4.18 presents some data value challenges.

New techniques for converting latent, unstructured text, image, or audio information into numerical indicators to make them computationally tractable are required in order to improve the efficiency of large-scale processing. However, such transformations must retain the diversity of values associated with words and phrases in text or features in image or audio files.

Data Visualization

Data visualization involves the creation and study of the visual representation of data. It is a set of techniques used to communicate data or information by encoding it as visual objects (e.g., points, lines, or bars) contained in graphics.

Table 4.19 describes the common types of data visualization methods, adapted from http://vadl.cc.gatech.edu/taxonomy/. A comprehensive set

Table 4.19 Common data visualization types

Type	Description
1D/linear	A list of data items, usually organized by a single feature. Useful for textual interfaces, although frequently used in simple pages within web browsers.
2D/planar	A set of data items organized into a 2D representations such as matrices, maps, charts, and plots.
3D/volumetric	A set of data displayed in a 3S space such as a 3D plot, 3D models such as wire models, surface and volume models, and animated computer simulations of different phenomena.
Temporal	A set of data is organized and displayed according to its temporal attributes using techniques such as timelines, time series, Gantt charts, stream graphs, rose charts, and Scatter plots where time is one of the axes. Tools include MS Excel, the R Time Series library, and Google Charts.
Multi-dimensional	A set of data is organized according to multiple attributes taken one at a time in which the data set is partitioned into groups. Techniques include pie charts, histograms, tag clouds, bubble charts, and tree maps, among other techniques.
Tree/hierarchy charts	A set of data is partitioned into sets and subsets where the subsets are more explicitly descriptive based on the attributes of the items. Techniques include trees, dendrograms, tree maps, and partition charts.
Networks	A set of items is displayed as a partially connected graph where the nodes are connected by edges labeled by attributes shared by the nodes. Techniques include subway/tube maps, node-link diagrams, and alluvial diagrams.

of visualization techniques is available at http://www.visualliteracy.org/periodic_table/periodic_table.html. A lack of space prohibits detailed descriptions, but further reading contains a number of references that will allow you to explore the different techniques.

Communications and Networking

As little as three decades ago, our means of communication between individuals, individuals and computers, and computer to computer was very limited. The advent of Ethernet as a simple-to-implement network and the TCP/IP stack as a communication protocol unleashed a diversity of technology and applications. Continued development of Ethernet devices led to wireless communications, Wi-Fi, cellphones, tablets, and a plethora of mobile device applications.

Network communications speed and bandwidth have not kept up with either disk capacity or processor performance. Of the 3Es—exabytes, exaflops, and exabits—only the first two seem attainable within the next 10 years. The National Academy of Science (2013) has noted that data volumes on the order of petabytes mean that the data cannot be moved to where the computing is in a reasonable amount of time; instead, the analytical processes must be brought to the data. There seem to be physical limitations on moving data at an exabit per second over current physical transmission media. If data is compressed prior to transmission, it may be possible to achieve apparent exabit per second bandwidth.

Over the past two decades, as wireless technology has evolved, transmission rates have continuously increased. These increases, accompanied by the shrinkage of the cell phone and the emergence of other wireless devices—tablets, watches, Google glasses, and so on—has resulted in a new discipline—mobility science—which focuses on research and development in mobile devices and applications. Indeed, some of the fastest growing conferences have to do with mobile technology, devices, and applications.

Computer and Information Security

Today's threat environment is extremely challenging, but none more so than in the online world. The rapid growth in mobile applications and the increasing web-hopping by Internet users exposes them to a myriad of threats. As the number of websites increases, so do the opportunities for computer and information security threats.

The threat environment has evolved rapidly. In the 1990s very few spam messages were sent. By 2014, it was estimated by Spam Laws (http://www.spamlaws.com/) that over 14.5 billion spam messages were sent per day, which comprised over 45 percent of the total e-mails transmitted across the Internet. As IPv6, the expanded Internet routing protocol, is implemented across the Internet, approximately 4 billion unique IP addresses will become available. This has created an enormous opportunity for cybercriminals to use to attack and exploit user sites. Defensive cyber technology applies to all sites, not just Big Data sites.

Cybercriminals have become extremely sophisticated in their methods of attack, exploitation, and exfiltration. They use modern software engineering techniques and methodologies to design, develop, and deploy their software. Polymorphic software can evolve and propagate in thousands of ways, much of which is undetectable by traditional methods. The newest types of complex and sophisticated threats (advanced persistent threats—APTs) are beyond the scope of this book. The attackers of today are not limited in their attack platforms or targets. In today's environment, *multiplatform* means that mobile devices are also highly susceptible to attack. The volume of data, the increasing frequency of attacks and diversity of targets, and the variety of threat software pose a Big Data challenge to organizations and individuals alike.

To defend themselves, organizations must gather, analyze, and assess more data than ever before at more locations within the organization infrastructure. This means dedicating significantly greater resources to ensure the viability of the organization in the face of this aggressive and evolving threat environment. It is becoming blatantly clear that off-the-shelf solutions have not and cannot address these problems. Scaling up resources for defense must be managed intelligently, because just identifying each threat and fixing it as it is found is economically not viable, and simply not technically possible. The bad guys are too creative and highly skilled!

Most organizations in academia, government, and industry are expected to do much more to protect themselves against the deluge of cyberattacks that are exponentially increasing. Today's organizations are adding security operation centers to meet these growing threats. In medieval times, one could notice when the Huns were attacking the city gates or someone had stolen half the grain supply. Today, cybercriminals can steal the data while leaving the original data in place. In many cases, an organization does not know it has been attacked until its own data is used against it.

Security software organizations must not only collect, analyze, detect, and develop responses, but they must also develop tools to predict how software will morph, where attacks might occur and where from, and how frequently they might occur.

Firewalls

A primary component of a security architecture is the *firewall*, which is a network security device that inspects and controls incoming and outgoing network traffic. Inspection and control is based on a set of rules defined by the organization. For example, many organizations prohibit the acceptance of traffic from known phishing scans and prohibit access to known websites that host malware. In effect, the firewall is intended to establish a barrier between a trusted network, such as a company's internal network, and an external untrusted network, such as the Internet. Firewalls may be implemented in hardware, software, or a combination of the two—although it is generally accepted that hardware-based firewalls may be more secure. Most personal computer operating systems offer software-based firewalls.

Firewalls can operate at many levels—as packet filters for raw network traffic, as state-based systems that operate at the message level, to application level systems that handle protocol-based applications such as browsers and file transfer programs.

Intrusion Detection System

An *intrusion detection system* (IDS) is hardware or software that monitors network traffic for malicious activities, policy violations, and attempts to bypass security mechanisms to gain access to an organization's internal computer systems. Some IDSs also attempt to prevent intrusions through various mechanisms. Generally, an IDS will identify possible incidents of intrusion, log the information, and possibly send alerts to system administrators, ensuring the organizational security policies are heeded.

Passive systems usually log the incident and send alerts. Reactive systems will automatically attempt to prevent an intrusion by rejecting traffic, or even shutting down the host system to prevent attack and corruption. IDS may, like firewalls, look outward, but they also look inward at the operation of organizational systems for anomalous behavior based on known models of operation and communications between those systems.

It is impossible to detect every possible intrusion and even harder to prevent some of the intrusions that take or are taking place without analyzing them. Often, prevention is difficult because of the time required to identify and analyze the attack. Detection and prevention become harder as system usage grows and more computational effort must be focused on detecting anomalous behavior through comparison and prevention based on understanding the nature and target of the attack and having adequate methods for preventing or defeating it.

A good reference for intrusion detection is the NIST 800-94 Publication: A Guide to Intrusion Detection and Prevention Systems, 2007, which is available at http://csrc.nist.gov/publications/nistpubs/800-94/SP800-94.pdf.

Virtual Private Network

A *virtual private network* (VPN) extends a private network across a public network, such as the Internet. In effect, the VPN rides on top of the public network. A VPN is implemented as a point-to-point connection between two nodes in which the communications are encrypted at each end before transmission and decrypted at the receiving end. Two commonly used VPN techniques are openVPN (https://openvpn.net/) and ipSec (http://ipsec-tools.sourceforge.net/).

Security Management Policies and Practices

Security management is the identification, classification, and protection of an organization's information and IT assets. Both computer systems and data must be managed and protected. The objective of security management is to prevent the loss of information, compromise of IT assets, and financial impact on an organization due to corruption or exfiltration of critical data, or denial of service to internal and external clients. To do so, an organization must identify, categorize as to impact and likelihood, and develop methods for prevention and mitigation of security threats.

Security management is implemented through a set of security policies developed and mandated by the organization. A security policy is a specification for achieving some level of security in an IT system, such as, for example, a password policy. Security policies may be machine- or human-focused. In either case, they prescribe the behavior of the entity to which they are addressed. Policies are typically decomposed into subpolicies which then specify specific operational constraints on the use and operation of the IT systems and access to organizational data.

Compliance with Security and Privacy Laws and Regulations

Security and privacy laws regulate how we deal with personal data in the United States. Organizations must ensure that they are in compliance with such laws and regulations or face stiff penalties. In some domains (e.g., health), the laws and regulations are much stricter than in others. Moreover, there is an emerging concern that as data accumulates about an individual or group of individuals, the aggregation of such data will allow the determination of properties of the individual through reasoning that were not explicitly represented in the original data itself.

In certain domains, such as social media and health information, as more data is accumulated about individuals, there is a fear that certain organizations will know too much about individuals. For example, data collected in electronic health record systems in accordance with HIPAA/HITECH provisions is already raising concerns about violations of one's privacy. International Data Corporation (IDC) coined the term "digital shadow" to reflect the amount of data concerning an individual that has been collected, organized, and perhaps analyzed to form an aggregate "picture" of the individual. It is the information about you that is much greater than the information you create and release about yourself. A key problem is how much of this information—either original or derived—do we want to remain private?

Perhaps the biggest threat to personal security is the unregulated accumulation of data by numerous social media companies. This data represents a severe security concern, especially when many individuals so willingly surrender such information. Questions of accuracy, dissemination, expiration, and access abound. For example, the State of Maryland became the first state to prohibit—by law—employers asking for Facebook and other social media passwords during employment interviews and afterward.

Clearly, some Big Data must be secured with respect to privacy and security laws and regulations. IDC suggested five levels of increasing security (Gantz and Reinsel 2011): privacy, compliance-driven, custodial, confidential, and lockdown. The recent spate of international hackings and exfiltration of customer data from multiple large corporations in the United States demonstrates the critical nature of the problem and highlights the risk facing all businesses and organizations. Your corporation should take the initiative to review the Big Data assets and its structure and properties to determine the risk to your corporation if it is exposed. The IDC classifications are a good start absent a formal standard developed by the IT community. Table 4.20 presents some of the data compliance challenges.

Table 4.20 Some Big Data compliance challenges

What rules and regulations should exist regarding combining data from multiple sources about individuals into a single repository?
Do compliance laws (such as HIPAA) apply to the entire data warehouse or just to those parts containing relevant data?
What rules and regulations should exist for prohibiting the collection and storage of data about individuals—either centralized or distributed?
Should an aggregation of data be secured at a higher level than its constituent elements?
Given IDC's security categorization, what percentage of data should reside in each category? What mechanisms will allow data to move between categories?

The Buy or Own Dilemma?

The primary issue facing any organization considering cloud computing as an infrastructure for providing computing services can be categorized as the "buy or build dilemma." Basically, does an organization buy its cloud computing services from a public vendor or does it develop its own in-house cloud computing services. These are obviously the extremes. Many organizations have opted for a hybrid approach.

From a buy perspective, there are many issues that an organization should consider before deciding to wholly commit to an external, third-party supplier of cloud computing services. Kaisler and Money (2010, 2011) and Kaisler, Money, and Cohen (2012) have described some of the issues that should be considered. Table 4.21 presents a summary of some of these issues.

Table 4.21 Selected issues in buying cloud computing services

If the organization moves to a competing service provider, can you take your data with you?
Do you lose access (and control and ownership) of your data if you fail to pay your bill?
What level of control over your data do you retain, for example, the ability to delete data that you no longer want?
If your data is subpoenaed by a government agency, who surrenders the data (e.g., who is the target of the subpoena)?
If a customer's information/data resides in the cloud, does this violate privacy law?
How does an organization determine that a cloud computing service provider is meeting the security standards it espouses?
What legal and financial provisions are made for violations of security and privacy laws on the part of the cloud computing service provider?
Will users be able to access their data and applications without hindrance from the provider, third parties, or the government?

CHAPTER 5

Building an Effective Big Data Organization

> *"The key to future success is to build an organization that can innovate with data," said Rod Morris, former senior vice-president of marketing and operations at Opower, a cloud-services firm that supplies SaaS applications to the energy and utilities industries. "To disperse the data, democratize the tools and unleash the creativity of individual employees."*
>
> —http://blogs.clicksoftware.com/clickipedia/three-steps-to-building-a-better-big-data-culture/

Good organizational design and practices are essential for information systems success, but it is not enough. Having great technologies are also necessary for success, but they are insufficient. Having skilled personnel is key to success, but good people alone will not cut it. The key to effective information systems lies in the integration of all three. It is not only important to have great organizational practices, technology, and people, but it is also important that these three components integrate and complement each other well. This is no different for Big Data organizations. For example, Tom Davenport proposed these same three key elements of analytical capability—organization, human, and technology, suggesting that building an effective Big Data organization requires organizational practices, information technology, and talent working in harmony, just like a symphony.

In this chapter we look at some of the key components that an organization should have in place to build an enterprise level Big Data program. We discussed the Big Data analytics (BDA) life cycle process model in Chapter 3, but what are the key practices and structures an organization needs to succeed in its BDA efforts? We attempt to answer these questions in this chapter. Based on the many one-on-one interviews we have had with analytics professionals, countless workshops with academics and practitioners in this area, and survey studies we have conducted, we have identified these key aspects to build an effective Big Data organization, or to transform an organization into one.

- Organizational Design and Practices
- People
- Talent Identification and Acquisition
- BDA Teamwork

We discuss these key aspects in more detail in the next few sections. Before we begin, it is important to note that the shift to a service-oriented economy from a product-oriented economy has intensified over the past 20 years. Increased globalization arising from the Web and telecommunication technologies means that established markets and advanced economies must now compete with emerging economies, such as India, China, and, more recently, Africa. The widespread availability and access to open source and free tools like R and Big Data cloud services means that any small group of individuals with great talent and education in any part of the globe can be a formidable competitor in this field. At a time in which analytic talent is so scarce, professionals in every country are waking up to this opportunity. Professionals and organizations will therefore need to be nimble and train and re-train themselves on best practices in this area to remain competitive.

A recent Bain and Company survey (Pearson and Wegener 2013) found that the use of BDA can give an organization a clear competitive advantage over its competition. An examination of more than 400 large companies found that those with the most advanced analytics were significantly outperforming their competition. These organizations were twice as likely to be in the top quartile of financial performance within their industries and were also more likely to use data very frequently when making decisions, while making decisions faster than their competitors and much more likely to successfully execute these decisions. Moreover, at the same time the window to respond to business needs is always shrinking. Organizations should establish their organization structure for an analytics and start the design and launch of a new analytics organization with a basic overall organizational architecture and approach to ensure that all of the roles, skills, and capabilities are in place from the beginning.

Organizational Design and Practices

Organizational design and management practices include a number of areas that have a significant effect on the successful use of analytics efforts within an organization. These design approaches and practices include enterprise and other architectural designs, an AnBoK to provide guidance to analytic practitioners, governance to help management and ensure that the analytics efforts are aligned with business needs, and finally a culture and level of maturity that provides a positive environment for analytics success.

Enterprise, Domain, and Application Architecture

Circular A130 of the Office of Management and Budget describes enterprise architecture (EA) as "the explicit description and documentation of the current and desired relationships among business and management processes and information technology," and it provides the "blueprint" to build information systems in an organization in a coordinated manner. While it is not the objective in this chapter to promote architecture practices, we lean on EA principles to make the point that, like any other information system implementation, BDA requires careful architecting from the beginning.

The data needed for effective analytics can be either gathered from external sources or produced internally. In either case, an architecture needs to be in place to store, manage, and access this data. But data management alone is insufficient and we also need to pay attention to business processes, applications, and the technology infrastructure. Several EA frameworks have been proposed—for example, Zachman, Federal EA Framework, and The Open Group Architectural Framework—representing the various views of the architecture.

As depicted in Figure 5.1, one thing all frameworks agree on is on EA's four key players: business process, information, applications, and

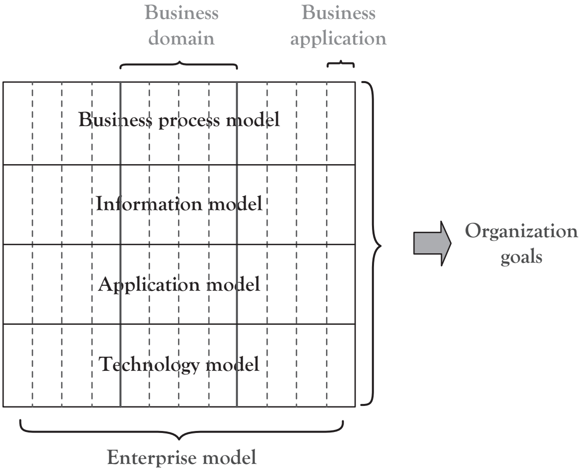


Figure 5.1 Enterprise, domain, and application architecture

technology infrastructure. An organization's EA is often broken down into business domains (e.g., divisions, functions, and geographical regions), which often have their own domain architecture, and there are multiple applications running within each of these domains. In the end, all EA layers, domains, and applications must work in unison to support the organizational goals. Again, this is no different for BDA.

The point we are trying to make is that any organization that is serious about implementing effective BDA practices needs to think way beyond talent and tools. It becomes an organizational design issue. Business processes, whether they are enterprise, domain, or application processes, need to incorporate Big Data and analytic thinking into them. It was sufficient in the past to conceptualize business processes to fulfill the transactional needs of the organization, but this is no longer sufficient. For example, some of the most advanced Internet marketing analytics practice collect terabytes of data, gathering all the customer shopping and clicking behavior online, which can later be mined to develop customized marketing strategies. However, the necessary business processes to gather, store, retrieve, and analyze the necessary data must be in place.

The same is true for the information model. Many online shopping sites were not interested in collecting shopping cart data in the early years and most of this data was stored locally in cookies in the users' computers.

All the companies needed were to be able to move shopping cart contents to the checkout application to process the sale transaction. Not any more, today companies are tracking all aspects of shopping cart behavior (e.g., the timing when items are added, modified, or deleted; the elapsed time between a shopping cart addition and a product purchase; the likelihood of purchasing once a product is added to the shopping cart; the amount of time a buyer spends in the product specifications before buying; the most effective colors and graphics to motivate a buyer to make a purchase, etc.).

The application architecture also needs to be designed with BDA in mind. If the analysis is done by humans, maybe all that is necessary is to have applications that produce the necessary data and the tools to analyze it. But more progressive organizations do analytics on the fly as the data arrives. An online shopping application that makes product recommendations based on social filtering (i.e., "people like you who bought this item also bought this other item") cannot afford to wait for human intervention to make these recommendations, so they embed analytic applications within their transactional systems.

Finally, the technology architecture to support BDA has to be in place. An organization committed to collecting lots of data for analytics must have the necessary infrastructure to manage these data effectively. Again, if the analysis is done by humans, perhaps all that is needed is large storage data capacity and data warehousing facilities, where the data can be easily accessed and downloaded for analysis. If the data needs to be analyzed in real time, then the organization will need to invest in Big Data facilities so that the analysis can be run directly in the Big Data environment (i.e., true BDA).

Analytics Body of Knowledge Focus

While it is common knowledge that analytical talent is in high demand and short supply, the necessary knowledge to do effective analytic work is not just confined to the analytic and quantitative domains. Functional knowledge domain is also necessary to do effective analytical work. It is important to note that some organizations are reporting that the number of years of experience to become an expert in a knowledge domain (e.g., financial analysis) is dramatically shrinking, because younger less experienced analysts can access the data and derive similar intelligence from the data than a former domain expert without the data analysis skills. Nevertheless, decisions cannot be made in a vacuum and, given that analytics permeates all aspects of organizational work, it is necessary for analysts to have some degree of functional domain knowledge. In order to better understand the knowledge that analytics professionals need to possess, we developed a body of knowledge framework for analytics (AnBoK). We have used this body of knowledge successfully to analyze the educational market in BDA and design a master's of science based on it.

Implementing a successful enterprise organizational analytics program can be a challenging set of activities. When we speak of a holistic approach to BDA within an organization, there are knowledge areas to address when implementing the program. We developed our AnBoK, illustrated in Figure 5.2, from information gathered from numerous interviews, workshops, roundtables, conference discussions, surveys, and interviews of analytics professionals and academics. This framework is composed of four layers: foundations, analytics core, functional domains, and management. We now discuss each of these in more detail.

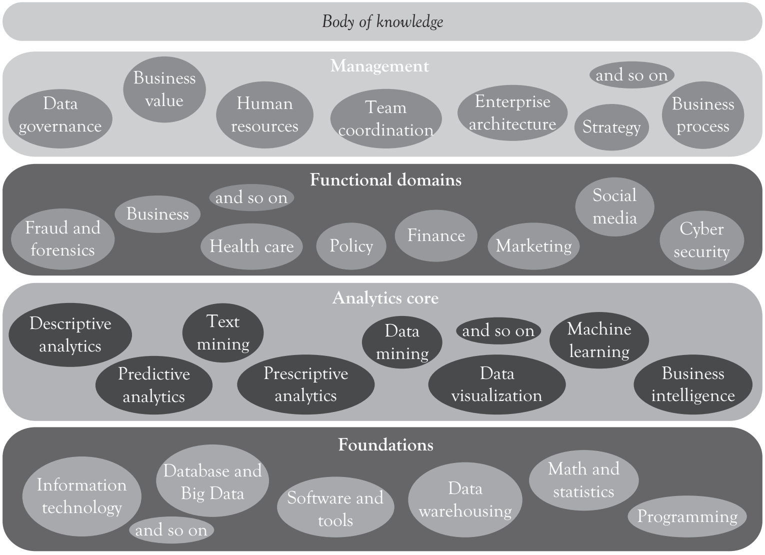


Figure 5.2 Analytics body of knowledge

Foundations Layer

These are the basic skills and infrastructure needed to operate and support a robust organizational analytics program. BDA is at the intersection of several disciplines, including computer science, software programming, database querying, mathematics, statistics, operations research, management science, and information systems, among others. These are many of the core information technology tools and infrastructure discussed in Chapter 4. While a full discussion of the basic foundations of analytics is beyond the scope of this chapter, we highlight important foundations.

Database, Big Data, and Data Warehousing. The analytics team will need access to the data as large amounts of raw data may exist in a multitude of various formats: structured (relational) and unstructured (tweets, Web blogs, text documents, etc.). For structured data management, knowledge of tools and data platforms is necessary to manage high-volume structured data (for instance, clickstream data or machine or sensor data). For unstructured data management, there is a focus on managing the explosion in data volumes due to a large extent to sources such as social media, Web transaction data, RFID and other sensor data, videos, pictures, machine generated and even text data from customer support logs, and so on. Tools and technologies are needed to manage, analyze, and make sense of this data to build understanding and to correlate with other forms of the structured data. Big Data is either downloaded to large data warehouse for analysis (i.e., traditional analytics) or analyzed directly in the Big Data environment (i.e., BDA). Either way, any organization dealing with the 5 Vs of Big Data needs to have the talent to work with and analyze data in these environments.

Information Technology, Software, Programming, and Tools. To effectively capture value from BDA, organizations need to integrate information technology, software, and programming tools into the analytics architecture and process. These technology decisions to utilize need careful analysis since they can have a long-term impact on an organization's ability to integrate BDA as well as inform what skills need to be developed and how are they best developed. In many cases, pilots and prototyping initiatives are needed to test out and experiment with new tools and architectures.

Math and Statistics. The organization will need individuals that have foundation skills in statistics and math. To be able to perform advanced analytics on Big Data, the analyst will need to understand such statistical concepts as correlation, multivariate regression, as well as the ability to model and view data from different perspectives for use in predictive and prescriptive modeling.

Software and Tools. As we mentioned earlier, Big Data can be harvested from a variety of sources. Organizations need to identify and deploy tools to gather, process, and visualize the information in useful and effective ways. Technologies include high-capacity storage repositories, modern databases, and specialized Big Data applications that can be used to reduce, consolidate Big Data, as well as find patterns and make sense of the data. Compounding these challenges, the analytics technology tool space is rapidly changing. New tools and software are being introduced at a rapid rate. Choices need to be made based on analytics. Organizations need to address some key technology decisions including the following:

- Which technology and tools are needed to support our analytics architecture based on our business goals and objectives?
- Which platforms are best for integrating these new tools, as well as with any existing technologies?
- Are open source or proprietary solutions the right fit for your organization?

The analysts may need access to such advanced analytical tools, such as Hadoop, NoSQL, SAS and so on, as well as programming skills particularly in languages such as R and Python. Regardless the specific technologies of choice of the day, a critical component of the AnBoK is the ability to manipulate data using software, so the need for skilled programmers will not go away. For example, given that Big Data may come from a variety of sources and in a variety of formats, merging structured and unstructured Big Data requires specialized, advanced knowledge, tools and technologies to prepare the data for analysis in analytical models.

Technologies such as in-database and in-memory analytics provide capabilities to process large data sets for analysis at near real-time speeds and to combine the analytics environment within, for example, structured data management tools. Additionally, an organization will need to develop an approach for master data management and consistently support an analytics culture at all levels of the organization.

Finally, the representation of data and results are an important component of Big Data architecture. Visualization tools allow the analysis to be more clearly understood and insights discovered. Visualization tools and technologies for quick drill down and analysis are now available and need to be integrated into analytics architectures, like the one illustrated in Figure 5.3. Finally, the organization will need to address how to integrate these analytic technologies to provide value to the organization, by aligning to and supporting the business decision-making process.

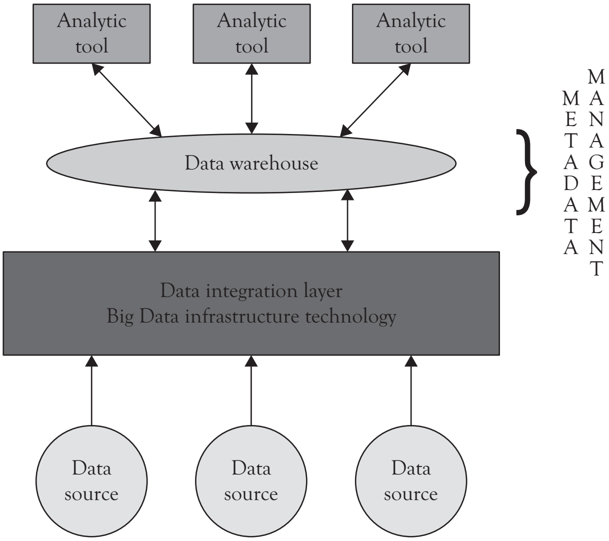


Figure 5.3 Logical analytics integration architecture

Analytics Core

These are the disciplines that often come to mind when we read or hear about BDA and it includes the typical topics found in most analytics books, including things like descriptive analytics, predictive analytics, machine learning, data mining, and business intelligence, among others. As discussed in Chapter 3, analytics approaches can be partitioned in three broad areas: descriptive, predictive, and prescriptive. The integrated role these approaches play in organization analytics approach is outlined in Table 5.1.

Functional Domains

Analytics is never performed in isolation, but in the context of a specific functional domain like health care, marketing, finance, policy analysis, fraud detection, accounting forensics, cyber security analysis, and sustainability, among many others. Just about every discipline can benefit from analytics. So the basic question is: what makes the best analyst? Is it a data scientist who develops some knowledge of a given functional domain through experience? Or is it a functional worker (e.g., investment broker) who gets additional education in analytics? We argue that functional professionals with deep understanding of and expertise in their functional domains make up the best analyst because they have a stronger foundation to formulate analytics questions and interpret and explain results.

Management

To be successful at the organizational level, the analytics efforts must be appropriately supported by management functions and processes.

Table 5.1 The role various analytics play in data-driven decision making

Past and present—descriptive analytics	The future—predictive and prescriptive analytics
Create awareness and determine that a decision needs to be made	Identify likely outcomes—predictive
Report on the results of action	Identify the best course of action—prescriptive
Understand the scope and context of the decision	

Analytics efforts are not likely to yield results of strategic value if the organization's long-term strategy is not aligned with analytics. Similarly, no analytics project will be successful unless the various component activities are appropriately coordinated and staffed. Finding analytics talent is difficult these days, so human resources plays a critical role. It is not uncommon to hear complaints from Chief Information Officers (CIOs) that they need to pay very high salaries for analytic talent who end up leaving the firm before one year has elapsed.

Data governance is another key issue. Data is the most valuable asset in modern organizations today, beyond their personnel. No insurance policy that we have heard of provides coverage for data losses. Hence, all aspects of data ownership, sharing, access, security, and location need to be covered by governance policies.

Governance

BDA governance can be thought of as the approach that analytic-based organizations use to define, prioritize, and track analytic initiatives as well as to manage different types and categories of data related to analytics. BDA governance strives to ensure that organizations have reliable and consistent data to perform analytics and make management decisions.

Without a corporate wide BDA governance framework in place, the risks an organization faces include, but are not limited to:

- Missing opportunities due to incorrect or insufficient utilization of analytics.
- Damage to the organizations' reputation due to illegal or unethical application of analytics (e.g., privacy and security).
- Making bad decisions based on analytics and data is not often fully understood by the decision makers.
- The legal and economic consequences of not complying with regulatory requirements.
- Finally, cost or quality issues due to misalignment of analytical-based performance measurement across the organization.

Thoughtful and appropriate governance helps to encourage actions in support of business objectives and preventing unnecessary efforts. General governance attributes include defining and assigning decision rights within the organization (e.g., with senior management or within areas of competency, ensuring that analytics efforts are aligned with business objectives and processes). Determining the priority alignment of analytical initiatives to business objectives and processes requires both portfolio management and governance. The governance approach should be documented and communicated with analytic Big Data policies and procedures: For analytic initiatives, measures should be defined to ensure and track the effort and the accomplishment of a business goal or decision.

Key tasks in implementing BDA governance include:

- Delegating analytics authority, budget, and responsibility to the appropriate unit or department.
- Defining and taking responsibility for the corporate analytics metrics framework and then ensuring alignment with the business' goals and strategy.
- Reviewing, prioritizing, and selecting BDA initiatives and deciding on corporate level investment budget for BDA.
- Tracking and enforcing metrics and indicators to assess return on Big Data investments.
- Developing performance indicators to identify policy, compliance, and regulatory adherence.

In starting to build an analytics governance framework, consider the goals and objectives for the various forms and types of Big Data initiatives shown, for example, in Table 5.2.

Big Data Analytics Culture

In implementing a successful analytics program, organizations need to focus on a number of key analytical and technology practices broadly and be business-centric rather than IT-centric. The organizations need to plan and quickly adapt to changing roles and skill sets, and they need to balance business and IT skills and resources, while building partnership between the business stakeholders and the information technology group. With the core technology and skills in place throughout the organization,

Table 5.2 Big data initiatives type

Information portal	Business analytics	Analytics laboratory	Operational
Descriptive	Descriptive	Predictive	Prescription
Focus on data models and information governance aspects	Focus on analytic models and business analyst self-service	Focus on prediction models	Focus on decision models and integration in business processes
Efficiency Integrity Security	Consistency Alignment Relevance	Transparency Privacy Quality	Viability Fidelity Appropriateness

the organization needs to institutionalize the importance and value of data-driven decision making to the company at large. A culture that embraces data and facts and applies them to everyday business should be grown. The challenge requires both top-down and bottom-up approaches to ensure everyone gets on board with the new data-driven paradigm.

When the organization has institutionalized the management, skills, tools, and technology necessary to drive an analytics-based organization, they will be prepared to innovate with the Big Data and continually find new ways to leverage their Big Data analytical abilities. Data-driven organizations need to value analytics as much as instinct when making decisions, and they need to view information as an asset and understand how to measure and communicate the value of BDA to all the stakeholders. The organization's culture should grow and change as people use the technology to take informed actions. When taking an organization wide approach to analytics, key technology practices should be implemented, including developing and executing an EA strategy, and work to prevent or eliminate silos of capability as well as implement an integrated analytics toolset and delivery platform.

Big Data Analytics Maturity

As the field of BDA matures, frameworks will emerge to evaluate the analytic maturity of organizations. The Software Engineering Institute developed the infamous *Capability Maturity Model* (CMM) establishing five levels of software organization maturity based on whether the company employed certain key processes needed to build software effectively and efficiently with repeatable patterns of success. Today, the CMM is the de-facto standard of software maturity.

Similarly, Jeanne Ross and colleagues developed a four-level CMM for EA, starting with stove piped silos at level 1 all the way to a modular architecture at level 4. CMMs are very good because they help gauge the level of sophistication and reliability organizations have in the particular domain the model is designed to measure. For example, bids for contracts often require the software companies to be at CMM levels 4 or 5 (Ross, Weill, and Robertson 2006, p. 69).

Naturally, various CMMs are likely to be proposed for BDA. Such a model, if effectively specified, would be a wonderful development for this growing field. To the best of our knowledge, such a model has not been effectively developed to date. Some models have been proposed. For example, Tom Davenport (2007) proposed five stages of development for analytics capabilities as depicted in Table 5.3.

These stages of analytics development are very useful, but we argue that more needs to be done to define a nuanced CMM for analytics. Perhaps the five stages described earlier are adequate, but it is our opinion that the stages need more precise definition of specific capabilities or key processes to reach the particular stages or levels. Examples of possible capabilities or key processes include things such as data governance, having a chief data officer

Table 5.3 Five stages of analytic capabilities

Stage	Description
Prerequisites to analytical capabilities	A good transaction data environment and operations exist.
Prove-it detour	Most organizations progress to stage two directly, but some need this detour to prove their analytical capabilities.
Analytical aspirations	This stage occurs when analytics gains executive support and first major analytics projects are launched.
Analytical company	This stage is achieved when the organization has demonstrated world-class analytical capabilities at the enterprise level.
Analytical competitors	This stage occurs when analytics go beyond being a capability to become strategic for competitive advantage.

(CDO), widespread data access, staff trained in analytics and data science, standardized corporate tools for analysis, Big Data storage capabilities, programming staff, having an analytics department with key roles fully staffed, having functional domain experts available to the analytics team, and so on. The formulation of a BDA CMM is beyond the scope of this chapter, but we hope that this discussion illustrates the importance of developing one.

People

As with most endeavors in life, people are perhaps the most critical aspect for the success of an analytics-based organization. While roles and skills will vary depending on the organization (e.g., size, business domain, specific needs etc.), a number of key roles and related skills needed in Big Data analytical organizations are highlighted in Table 5.4 and discussed.

Chief Data Officer

While typically not formally on the analytics team, the CDO plays a critical role to ensure that the data needed by the analytics team is available, accurate, consistent, and reliable. The CDO is a senior executive who

Table 5.4 Key Big Data team roles

Role	Description
Chief data officer	The chief data officer (CDO) is a senior executive who bears responsibility for the firm's enterprise wide data and information strategy and governance.
Chief analytic officer	The chief analytics officer (CAO) is the senior manager responsible for the analysis of data within an organization.
Analytically informed business managers	But all managers and business analysts need to understand and have an appreciation for what analytics can and cannot do.
Data analyst	A data analyst's job is to capture business requirements, analyze, and model supporting data and use it to help companies make better business decisions.
Data scientist	Data scientists are analytical data experts who have the technical skills to solve complex, analytical problems.
Big Data technologist	The Big Data technologist architects, implements, and administers the BDA technology and tools within the Big Data ecosystem.

bears responsibility for the firm's enterprise wide data and information strategy and governance.

The CDO's role will combine accountability and responsibility for information protection and privacy, information governance and information, data quality and data life cycle management, along with the exploitation of data assets to create business value.

Chief Analytic Officer

While the CDO is responsible for the collection, storage, and management of data, the *Chief Analytics Officer* (CAO) focuses on providing input into operational decisions on the basis of the analysis. The CAO is the senior manager responsible for the analysis of data within an organization. The CAO addresses creating real business value through data analytics and promoting the organization's data-driven culture. As such, the CAO requires experience in statistical analysis and marketing, finance, or operations. The CAO oversees the overall organization's analytic approach and helps to insure that the analytic efforts are prioritized at the strategic level and the analytic initiatives align with business goals and objectives, with the analytic results informing business decision makers.

Analytically Informed Business Managers

While it is not reasonable, nor desirable, in most cases, that the vast majority of business managers and analysts are deeply trained in the knowledge of analytics, it is very important that they need to learn how to analyze data and become a savvy consumer of analytics information. Managers and business analysts need to understand and have an appreciation for what analytics can and cannot do as well as how to identify business needs that can be addressed with analytics and communicate these business needs to the analytics team.

Data Analyst

The *data analyst's* job is to capture business requirements, analyze and model supporting data, and use it to help companies make better business decisions. The data analyst will be responsible for gathering, modeling, analyzing, and reporting data for a wide range of sources. The data analyst will need the business knowledge to understand the business problem and determine what data is needed to model the problem. They also need knowledge of statistical and analytical techniques to be able to analyze the data as well as business insight to be able to interpret analysis results in business terms.

Data Scientist

Data scientists are analytical data experts who have the technical skills to solve complex, analytical problems utilizing, in many cases, advanced statistics, math, and programming languages. They are part mathematician, part computer scientist, and part analyst.

They normally have a strong foundation in such disciplines as computer science and applications, modeling, statistics, analytics, and math. Data scientists have advanced training in multivariate statistics, artificial intelligence, machine learning, mathematical programming, and simulation to perform descriptive, predictive, and prescriptive analytics. Data scientists often hold PhD degrees. Big Data organizations typically will need to augment existing analytical staffs with data scientists, who will have a higher level of technical capabilities, as well as the ability to manipulate Big Data technologies. These capabilities might include natural language processing and text mining skills, experience with video and image manipulation. Data scientists also have the ability to code in scripting languages such as Python, Pig, and Hive.

Big Data Technologist—Infrastructure and Tools

The *Big Data technologist* architects, implements, and administers the BDA technology and tools within the Big Data ecosystem. These technologies vary, but can include Hadoop, MapReduce, databases (both NOSQL and SQL), in memory databases, data warehouses, analytical modeling and visualization tools, and so on. This role can also be responsible for the extracting, transforming, and loading of data from various source systems into the analytical platform/data staging areas.

Talent Identification and Acquisition—Staffing the Analytics Operation

Many large organizations already have data analysts even if they do not have a formal BDA program. As you start a BDA program, you must consider where to obtain the skilled resources to staff it. Four primary means for obtaining skilled team members are: (1) hiring trained data and analytics scientists, (2) developing an in-house training program, (3) partnering with a college/university to provide a formal training program, (4) utilizing commercial educational programs, and (5) contracting out the BDA function. Each approach has its advantages and disadvantages.

Most importantly it needs people that have the skills in the analytic techniques and methods that best address the business problem.

Hiring Trained Staff

This is one of the quickest ways to organize and staff a BDA program. The number of trained and skilled data and analytic scientists is still rather small, although college and university programs are increasing and will eventually provide a large pool of potential employees within the next few years. Of course, we understand that you cannot wait and must have a BDA staff right now.

Harris, Murphy, and Vaisman (2013) analyzed corporate requirements for BDA staff and the corresponding skills of people who billed themselves as "data scientists." They found that many employers have unrealistic expectations of potential staff. Harris et al. called this the "rock star syndrome"—the desire to have one person who can do it all—the *Data Creatives*. These individuals can do it all—extract data, integrate it, create compelling visualization mechanisms, perform statistical and analytical analyses, and build the necessary tools when good ones do not exist. Good Luck! These people are rare and are most likely hackers rather than disciplined practitioners.

Harris et al. identified three other categories of data scientists. The *Data Business people* are focused on the organization and how data projects can yield profit or provide better service to clients. Most have a technical undergraduate degree and an advanced degree (such as an MBA), are highly disciplined in their approach, and have experience in contracting or consulting work. Another group is the *Data Developers*, who are usually focused on managing the data—preparing it, organizing it, transforming it, and storing it. They do not necessarily analyze it as much as make it ready for analysis or derive data that can be used in analysis. The final group is the *Data Researchers*, who generally have deep academic training—PhDs and technical publications. They are focused on developing new methods of analysis to understand complex processes in business or other fields.

So, which ones do you need, or—put another way—how do you build a staff given these categories of data scientists? It depends on what you are trying to do. Our suggestion is to start off with a few people from the Data Business people and a few from the Data Developers. This provides you with a staff to manage your data and a staff to begin developing solutions to benefit the business operations and decision making. As your organization gains experience with BDA, you can move into more complex business and technical decision processes and you can add a few Data Researchers to develop advanced and alternative analytical techniques. Finally, to handle ad hoc, critical problems, you may decide to hire one or two people from the Data Creatives category.

Communication, Organizational Skills, and Knowledge

We have and will continue to discuss in this chapter the need for individuals that have a sound foundation of skills that include analytical and Big Data modeling, tools, and technology. However, while some analytical personnel will focus narrowly on specific Big Data and analytics approaches, there will be a need for more broad-based individuals who match technical data knowledge with great business, communication, and presentation skills.

Analytic teams should have the capability to identify unexpected and critical analytical insights via an understanding of business needs. The analytics professional also needs to be able to effectively converse in the language of business with the users to determine the best analytical approaches to address the specific business user's requirements. Additionally, they must be able to explain and communicate these insights to the business users and executives.

In Walmart (Marr 2015), for example, analytics and Big Data are now integrated into every vertical within the company. Every new analytics team member participates in the Walmart Analytics Rotation Program and spends time in different business departments to understand how they operate and how analytics can be used across the various organizations. We discuss typical analytics organizational structures later in this chapter.

INFORMS (https://www.informs.org/Community/Analytics), a professional society that focuses on promoting the use of data-driven analytics and fact-based decision making in practice, has an analytics certification called Certified Analytics Professional (CAP). When creating the CAP, INFORMS (Nestler et al. 2012) discussed "T" shaped analytics knowledge versus "I" shaped knowledge. A person with "T" expertise understands, at a high level, a wide breadth of knowledge surrounding their discipline, with in-depth knowledge of one or more narrower areas, whereas a person with "I" shaped expertise tends to focus exclusively on narrower, but a deep set of skills. Organizations should be thoughtful as they staff their analytics program to include a mix of individuals with both "T" and "I" shaped expertise.

The INFORMS CAP certification requires, among other things, passing a certification test, having at least a BA/BS degree, multiple years of analytical work experience, and being effective using soft skills. For more information on the INFORMS CAP, see https://www.certifiedanalytics.org/.

In-House Training

Another approach, which will take some time, is to develop an in-house training program. One advantage is that you can select people with the apparent aptitude and background, who may already be business analysts and have considerable business domain knowledge, to enhance their skills and capabilities and value to the organization. Most often, an organization will contract with an external organization to train their people in a select set of skills directly relevant to the immediate business needs.

At a recent roundtable of C-level executives (Private Sector IT Assembly 2015, Panel on Global Enterprise, Big Data Trends, Analytics and Insights), one experienced panelist expressed concern about the difficulty of attracting highly trained and qualified data scientists. The observation this colleague made was that these data scientists came at a very high price tag and often did not last more than one year with the organization. Data scientists are in very high demand these days and they are very difficult to pin down. The same colleague indicated that their organization (a very large consulting firm) had more success training in-house personnel.

Unless an employee has been highly educated in both, analytics and a functional domain, it is likely that they will only have expertise in one or the other. Hence, one approach is to train quantitative and technical staff in the specific functional domains of interest. Another approach is to train managers and functional domain experts in BDA. This panelist indicated that the best analytics person they had was a functional domain employee who had some basic foundation of statistics, who became a leading analyst after receiving the necessary training.

College and University Education Programs

This approach takes an academic approach to developing a BDA capability by allowing existing staff to become "credentialed" by earning a certificate or Master's degree in Big Data. At the same time, it provides an organization with a conduit for examining and hiring current and graduating students to build a BDA staff. Two advantages of this approach are that it can provide a pipeline for future hires and access to professors with expertise in a variety of domains and analytic disciplines.

Because of the widespread belief that there is an ever widening gap between supply and demand of both, deep analytical data scientists and managers with analytical skills, schools have embarked in a gold-rush-like race to create appealing degrees in data science, analytics, and Big Data. One thing that is important to note is that BDA and traditional analytics can be classified as STEM programs. STEM stands for Science, Technology, Engineering and Math, and there is a big national push to increase STEM education in the United States. Furthermore, international students enrolled in STEM education programs can get an additional 17 months of practical training visas. We strongly encourage readers with an interest in analytics education to seek STEM approved programs. For more information visit:

http://www.dhs.gov/news/2012/05/11/dhs-announces-expanded-list-stem-degree-programs

With the help of an external consulting organization, American University conducted a comprehensive study of about 40 different programs in Big Data and analytics. Our conclusion is that there is rapid growth of educational programs catering to this market and those seeking further college education in this area need to be careful about the programs they select.

We see a lot of confusion among applicants, which is not unexpected, given the fact that BDA is still an emerging field. In our attempt to make some sense out of the jungle of possibilities for education in analytics, we classified Big Data and analytics programs using our AnBoK framework into the categories presented in Table 5.5.

Numerous programs exist in business analytics. American University's new program in the Kogod School of Business is an exemplar of the type of business analytics program that will evolve in the future. American University has implemented a novel one-year Master of Science in Analytics by partnering with various departments in the business school and other schools across campus. The program includes 9 credits of foundational courses, 9 credits of analytics core courses, and 12 credits of functional specialization courses from the various departments and schools participating in the degree. For example, if a student wishes to specialize on accounting forensics analytics, they can take the required foundational and core analytic courses and combine that with 12 credits of accounting forensics courses. In addition, the program has a capstone practicum in which students work on a real project with real data, and no lectures.

Commercial Education Opportunities

Commercial training organizations, tool vendors, and open source organizations provide many possible training opportunities. The list is endless and is evolving very quickly. Following is a short sampling of training sites that provide Big Data training options. Table 5.6 lists a few commercial education opportunities. We should caution you to thoroughly investigate the backgrounds of the teachers and lecturers in a commercial firm before committing to them to educate your staff.

Table 5.5 Big Data and analytics programs

Category	Description
Data science programs	These programs center around the first two layers of the AnBoK. Their focus is on developing fundamental skills in quantitative sciences and statistics, followed by a solid foundation of topics in the analytics core layer. It is *doubtful* that any school can really prepare a true data scientist in a short one-year program. Data scientists often have doctoral degrees in quantitative fields; they are in very short supply and high demand.
Analytics programs	These programs differ in key aspects from traditional data science programs. Like with data science programs, they focus on the first two layers of the AnBoK framework, but the coursework focuses more on analytics than Big Data, whereas data science tends to go deeper into Big Data topics.
Business analytics programs	While this may seem to be a subtle semantic differentiation, the market appears to be viewing "analytics" programs as deeply analytical and "business analytics" programs as ones that aim to support functional domains, focusing on managers that need analytical skills and consumers of analytics reports. As such, these programs focus on the first three layers of the AnBoK—foundational, analytics core, and functional domain. For example, some schools have very effective programs specializing in specific functional areas, like marketing analytics, health care analytics, and cybersecurity analytics, among others.
Analytics programs—highly technical and foundational programs	These programs focus on education around the foundational layer of the AnBoK. These programs are typically found in computer science or mathematics schools. They tend to focus on basic mathematics, statistics, quantitative analysis, software programming, and computing infrastructures.
Tool-based programs	These programs focus on providing students the skills to use popular analytics and business intelligence tools like RapidMiner, XLMiner, and Tableau, among others. These are typically short certification programs and we caution the reader to study these programs carefully. A power tool in the hands of a novice can do more damage than good and we encourage readers to look into more foundational programs.
Theoretical versus practical approaches	Some schools focus on teaching students the fundamentals and theories associated with BDA, whereas others follow a more practical approach. In our experience, we find that students trained with ample opportunity to put into practice what they have learned—through practicums or internships—become the most effective analysts who are ready to hit the ground running upon graduation. We call this "experiential learning."

Table 5.6 Selected commercial training opportunities

Commercial training	Description
Teradata University Network (http://www.teradatauniversitynetwork.com/	Free, web-based portal that provides teaching and learning tools. The resources and content support everything from introduction to IT at the undergraduate level to graduate and executive level courses in Big Data and analytics.
Big Data University (http://bigdatauniversity.com/)	Online courses, developed by experienced professionals and teachers that are mostly free on Big Data topics.
Lynda (http://www.lynda.com/)	A wide selection of online video on a multiple of topics including, but not limited to, Big Data and analytic methods and tools.
SAS (https://support.sas.com/training/)	Provides a variety of training course and on SAS software and associated analytics skills.
Hortonworks (http://hortonworks.com/training/)	Provides training designed by Hadoop experts. Scenario-based training courses are available in-classroom or online.
EMC (https://education.emc.com/guest/campaign/data_science.aspx	Curriculum-based training and approach to the techniques and tools required for BDA.

Outsourcing the Big Data Analytics Function

The final approach is to contract out the BDA capability to a third-party firm already established in BDA. This approach is usually more expensive, but can often be implemented relatively quickly. It allows the BDA staff to grow and/or shrink as the organization's need vary. Moreover, it can often provide a quicker route to introduce new analytic techniques into the BDA process.

Big Data Analytics Teamwork

> The sexy job in the next ten years will be statisticians … The ability to take data—to be able to understand it, to process it, to extract value from it, to visualize it, to communicate it—that's going to be a hugely important skill. (Varian, *Mckinsey Quarterly*, 2009)

The analytics teams will oversee the design, build, test, and deployment of analytical models, reports, and data visual representations.

Creating a successful analytics team requires both the right people and the right culture. Applying BDA within an organization is not a solitary process. While data and analytic scientists may work on individual projects, most BDA projects are a team effort. Collaboration serves to reinforce the joint understanding of the domain and the types of problems the team is trying to solve.

BDA, with its need for many skill sets, is normally done in multi disciplinary teams. There are multiple organizational structures for locating and managing this team. For example, how are analytic teams set up across the organization, how do they interact with the business units, and how are they compensated? How are projects assigned and tracked? We outline several possible organizational analytics structures in the following text.

Distributed Analytics—Analytics within the Business Unit

In a distributed analytics structure, as depicted in Figure 5.4, analysts sit in one or more business units, and there is no enterprise wide analytics perspective. This is a typical structure within less mature analytically based organizations. This structure has the advantage that it allows a quick start-up to provide the ability to prioritize, test out, and refine different "Pilot" approaches before implementing them enterprise wide. The disadvantage is the only the business unit with the analytics gets direct benefit from the efforts.

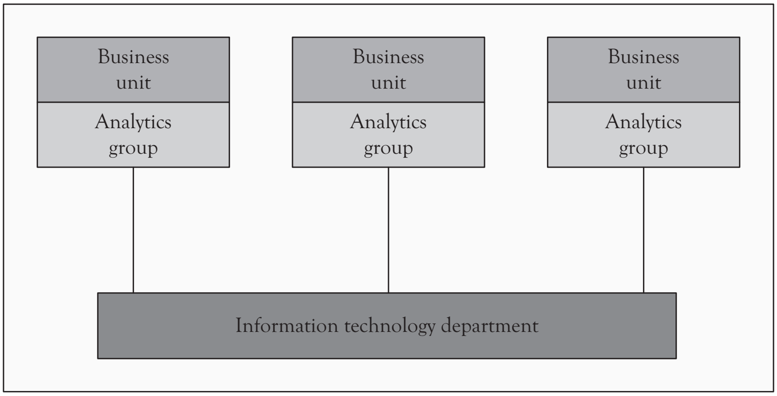


Figure 5.4 Distributed analytics group

Centralized Analytics Group

A centralized analytics structure, as depicted in Figure 5.5, is one in which the analytics team is organized as a single centralized unit and sets the analytical direction for the organization. The team can consult to individual business units for analytic services. Analytic services can be deployed strategically. Analytical insights and knowledge across business units can be centralized and shared.

Analytics Group within the IT Department

An analytics group within the IT department structure, as depicted in Figure 5.6, is one where the analytics group resides within the IT department. This structure is normally very technology focused and while it can be very technically advanced, the challenge within this structure is a lack of communication and understanding of the business stakeholders and business needs.

Distributed Analytics Groups within the IT and Business Units

Another analytics organizational structure having distributed analytics groups within the IT department and business units is depicted in

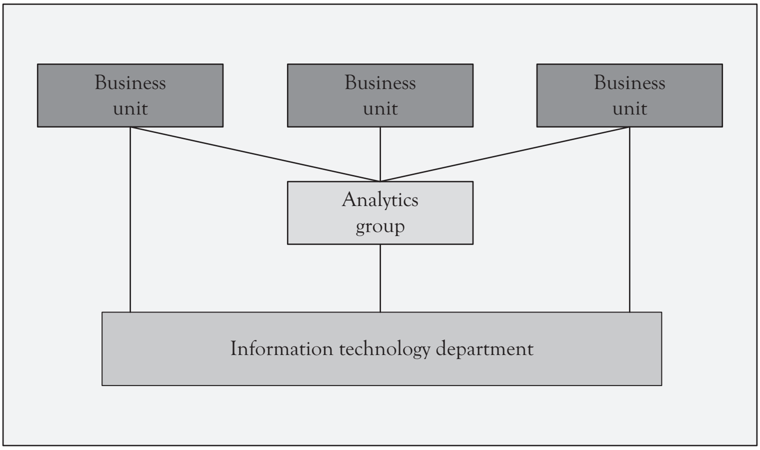


Figure 5.5 Centralized analytics group

Figure 5.7. The IT-based group can focus on the Big Data technology challenges, while the business unit focuses on the use and application of analytics. The danger with this approach is that the individual groups are or may be uncoordinated in these analytic approaches and there is no overall enterprise governance.

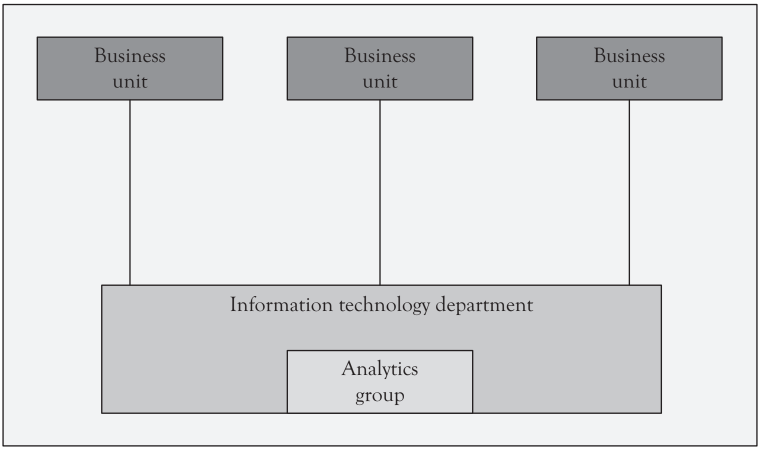


Figure 5.6 Analytics group within the IT department

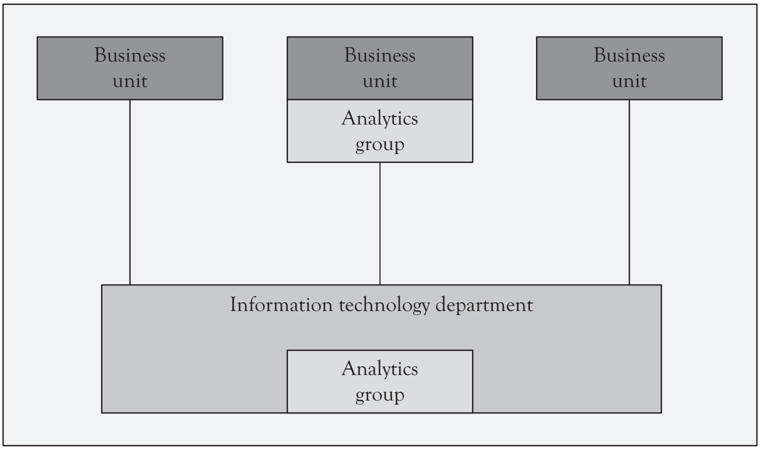


Figure 5.7 Distributed analytics groups within the IT and business units

CHAPTER 6

Issues and Challenges in Big Data and Analytics

In every discipline, there are issues and challenges that managers and decision makers must deal with in the normal course of their work. With Big Data, these issues and challenges arise not only from technology and its use, but also from the scale of the data and the use of insights and results generated by analyzing the data.

First and foremost, analytics reside on a continuum as depicted in Figure 6.1. Few problems actually reside at either end of the continuum; most require some combination of analytics to extract the information and derive the actionable intelligence necessary for decision making.

Finding the Needle in the Haystack

At one end of the continuum is the problem we characterize as *finding the needle in the haystack*. With all the data available to us, how do we find the key event or events, or the patterns characterizing key events that indicate(s) a critical situational change? If we have an initial description of the type of event or pattern we are looking for, this problem could devolve into a massive search problem. More often, we are not sure what the structure or type of the event is—we are trying to not only describe the event characteristics, but also find if the event is actually present in the data set(s). This discovery process is iterative and often involves redesign of the analytics pipeline as one better understands the data and how it will support the decision making.

One problem affecting this process, which also affects the next challenge, is how to integrate data from different sources in order to describe the event or event sequence. Another problem is how to decide what the "right" answer is. Sometimes, the needle is a complex structure

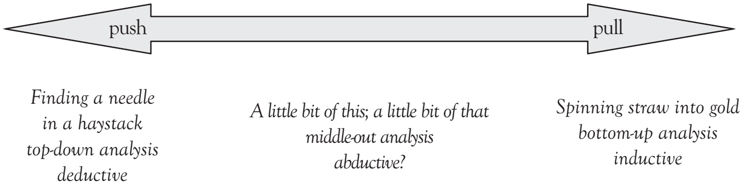


Figure 6.1 The Big Data analytics continuum

Source: Kaisler et al. (2015).

that must be assembled and an instance given a set of values before testing to see if it is an answer to the problem. Often, the needle is not growing as fast as the haystack, which increases the complexity of the search.

Even with large-scale parallelism, brute force processing may not succeed. As Borne (2013) notes: "Absence of evidence (so far) is not evidence of absence!" Success is all in asking the right question with the right parameters. Consider the problem of finding the few packets in a large corporate network among the tens of billions flowing through every day to find the ones that carry a virus or malware that might initiate a cyberattack.

If the haystack is a complex, multidimensional graph with multiple edges linking the nodes, traditional DataBase Management Systems (DBMSs) will not perform efficiently. Graph traversal problems often grow exponentially and will overwhelm most computer systems. Recognizing that a subgraph is a solution to a problem is one of the most difficult problems to work in mathematics. And, this difficulty carries over to the field of business analytics where we have complex graph structures representing connections among disparate trading partners, markets, suppliers, and so on.

Spinning Straw into Gold

This challenge, a play on the Rumpelstiltskin fairy tale, focuses on processing a large set of discrete data points into high-valued data. We call it *spinning straw into gold*. It attempts to describe a situation in a generalized form such that predictions for future events and prescriptions can be made about how to deal with those events through mitigation, whether passive or via active intervention. The objective is to identify one or more patterns that characterize the behavior of the system. However, these patterns may be initially unknown or ambiguously defined. These patterns may also be morphing over time, as in wicked problems. Some have characterized this problem as "dross into gold," but this seems too pejorative. All data have value to someone, but not all data have value to everyone.

With Big Data, the problem is not interpretation, but *sensemaking*—the dual process of trying to fit data to a frame or model and of fitting a frame around the data. Neither data nor frame comes first. They evolve concurrently as one understands events and their context, for example, develops a situational awareness that will help decision makers to respond to known but unexpected or unknown situations. Sensemaking requires deductive, abductive, and inductive reasoning that require domain knowledge (Weick 1995).

Managing Data Provenance

Much of Big Data is raw data collected from original sources or cleansed to ensure accuracy, precision, and validity. Once prepared, it is processed to yield actionable intelligence for decision making. Often, the data will progress through multiple analytics and processing stages until it reaches a stage where it can be used for decision making. A key problem with many analytic efforts is explaining how the analysis process reached the results that it did—its provenance, particularly when analytics involve complex algorithms and many iterations. Sometimes, it is not the result that is important, but how the result is reached, for example, what were the intermediate results that were strung together to form a chain of events or logical chain of reasoning that led to a critical event or conclusion.

Going forward, Big Data and information provenance will become a critical issue. Buneman and Davidson (2010) noted that provenance—the tracing of the evolution of data through transformation, translation, and enrichment—is largely a manual operation. Automated collection requires extensive changes to existing data processing systems and significant changes to the software architectures of future data processing systems. Further, the collection of provenance increases the amount of data that must be stored and managed. As data becomes information and proceeds through the processing steps to yield actionable information, the new data may match or exceed the volume of the original data. As a result, mechanisms for capturing provenance need to be developed to store it in association with both original and derived information, and to be able to backtrack through the provenance to determine how specific information was created during the analytic process.

Since applications that process data will undergo evolution, copies of application code and the associated parameters and other data sources must be stored with each generation of derived information to ensure the ability to recover how the derived information was generated. Very few, if any, applications now collect provenance, and this provenance cannot now be re-created because the original data may have been or is likely to have been discarded.

Scaling Beyond Petabytes

Turning Big Data into meaningful insight is a key problem. Hilbert and Lopez (2011, 2012) analyzed the world's technological capacity to store, communicate, and compute information. They concluded, in 2007, that the world was capable of storing 2.9×10^{20} optimally compressed bytes, communicate about 2×10^{21} bytes, and carry out about 6.4×10^{18} instructions per second. However, annual growth in computing capacity was about 58 percent, in bidirectional communication about 28 percent, and in globally stored information at 23 percent. They noted that while general-purpose computers had comprised 41 percent of the world's computers in 1986, this had shrunk to 3 percent by 2007 due to the advent of microcontrollers, GPUs, and embedded processors. Using this latter number, one can conclude that the world will not be able to keep up with the increase in available information.

A major challenge for Big Data is scaling, which has two aspects. As Hilbert and Lopez noted, there is not or may not be enough computational power to analyze all the data that is collected. Second, "as data sets grow as fast or faster than, Moore's Law, they are growing at least as fast as computing power increases" (National Academy of Science 2013). This limits analytical techniques that can scale linearly with the number of data items, N, or, at most, N log N. Thus, there is a need for new near-linear or nonlinear computational techniques that can address these challenges. Alternatively, new techniques for filtering or censoring data to yield tractable subsets will need to be developed.

While cloud computing seems to be the current infrastructure architecture for processing Big Data, scalability remains an issue. Cloud computing may be the culmination of the commodity-processor, commodity-disk, or Linux-based virtualized infrastructure architecture. New infrastructure architectures will need to emerge and become viable to reach beyond the 3Es—exabytes, exabits, exaflops—threshold for data volumes.

Ethical Challenges in Using Big Data

Big Data has emerged as a potentially powerful tool for understanding social phenomena based on massive quantities of information about people, things, and locations. Many individuals and groups argue that the use of Big Data without considering the social, economic, and personal issues raise significant security and privacy issues that represent ethical challenges that must be addressed by government, commercial, and academic organizations.

Data mining has been transformed by Big Data availability. Very public data—collected by businesses, organizations, and government; stored in social media websites—can reveal much about the public and private lives of citizens and corporations. Many Americans accept that purchases and activities are tracked through websites, mobile apps, and customer loyalty programs to name a few. How this data is used has initiated major debates about data mining in the U.S. Congress and in the European Union regarding whether data mining infringes on personal privacy rights. Cellphone companies track their subscribers everywhere in order to be able to switch calls among cell towers nearly transparently. How is tracking Wi-Fi signals from laptops, PDAs, Google glasses, or even wristwatches different from what the cell phone companies do?

Social media providers and web-based vendors are using the data stored in their systems in ways that challenge their presumed role as neutral storage sites. For example, how far should vendors, such as Amazon or Facebook, be allowed to utilize personal data or transactions, such as recommending to one's friends books, articles, or other merchandise that an individual bought or even just looked at? The use (and abuse?) of such data presents researchers, politicians, and users with an ethical dilemma arising from a set of ethical challenges. Anonymizing data from social media may prevent personal data from unauthorized release. Table 6.1 presents some challenges to privacy and security.

Table 6.1 Some Big Data challenges to privacy and security

Should personal data be used without the permission of individual owners, such as copying publicly available data about individuals and organizations? A critical legal determination must be made as to whether an individual or group owns data about itself that has been publicly made available.
Should an individual or organization be required to inform individuals or organizations of the use of their personal public data—for example, if it is extracted from social media sites?
Should an individual or organization be required to respect a person's privacy even on publicly accessible websites such as Facebook, MySpace, YouTube, Reddit, Pinterest, and so on?
Should an individual or organization be required to check the accuracy of what is posted on a publicly available website before using it?
What penalties should be imposed for the misuse of publicly available data for criminal purposes? Should they be harsher for private data obtained fraudulently?
How can we prevent the use of Big Data to harass or abuse culturally sensitive groups? Such as fostering racial prejudices or gender stereotypes?
What penalties should be imposed for failing to correct errors in acquired data and derived results even after proper notification?
Can we prevent the dissemination of data and derived results that are not in the best interest of the public to know?
Will Big Data be used to target and harass protesters who are exercising their right to protest (as guaranteed by the U.S. Constitution)?
How much and what types of data collected by the government or by private industry (such as the social media providers) should be made publicly available?
What rules and regulations should exist regarding combining data from multiple sources about individuals into a single repository?
Do compliance laws (such as HIPAA) apply to an entire data warehouse or just to those parts containing relevant health care data?
What rules and regulations (should) exist for prohibiting the collection and storage of data about individuals—either centralized or distributed?

Real-Time Streaming Challenges

Initially, Big Data focused on slowly varying static repositories, for example, the amount of data added to the repository was a fraction of the total volume of data in the repository. This meant that as the repository grew, it was still manageable. The advent of high-speed streaming processors and analytic software that operates in near real-time has increased the flow of data into an organization's repositories and exacerbated the management problem.

Streaming, however, raises an interesting problem for management. Feinleb (2012) has noted that the faster one analyzes data, the greater its predictive value. Near real-time processing of data streaming into an organization allows them to detect trends within the OODA (observe, orient, decide, act) loop of both customers and managers. The ability to detect and react to trends and changing situations can yield a faster decision-making capability which begets the need for more data to stay ahead of the situation. However, this seems to be a race between data volume, data velocity, data variety, and computational capability. With data volume seemingly expanding exponentially, and computational capacity increasing, but decelerating, new techniques will be required to extract the same or more amounts of information to support the increased pace of decision making.

Big Data Perishability

A problem related to streaming and to the collection and storage of Big Data as a whole is *data perishability*. Data has a lifetime. Some of it is long term, such as well-established facts in the physical and social sciences, but some of it may be measured in mere nanoseconds, such as financial transactions in computerized trading on Wall Street.

Transient data, if not processed within the time required to use it for decision making, must often be discarded to make room for newer data. Perishability puts the emphasis on insight, not retention. Historically, most enterprises chose to keep data for as long as possible and as cheaply as possible. But, for Big Data, ideas and policies regarding the duration and cost of retention must be revaluated because it is often just not feasible to

Table 6.2 Some data value challenges

For a given problem domain, what is the minimum data volume required for descriptive, estimative, predictive, and prescriptive analytics and decision modeling with a specified accuracy?
For a given data velocity, how do we update our data volume to ensure continued accuracy and support (near) real-time processing?
For a given problem domain, what constitutes an analytic science for non-numerical data?
"What if we know everything?"—What do we do next?

store the information for later processing. Table 6.2 suggests a few issues that you might want to consider. Also, see the section "Storage: What If You Can't Keep Everything?" that addresses the data storage issue.

The Role of Advanced Analytics

Modern businesses require a suite of appropriate analytical tools—both quantitative and qualitative—that go beyond data mining, because of issues with scalability, parallelizability, and numeric versus symbolic representation that may well affect analytic utility and the analytical results.

Numerous analytical methods have been applied to problems in business and the physical and social sciences. Advanced analytic tools, coupled with Big Data, comprised of diverse data sets, and domain knowledge, provide a means for solving complex problems. Up until 30 years ago, simple business models often sufficed for international business. The advent of globalization, brought on partly due to advances in digital technology, massive amounts of information available at our fingertips that coupled with a rapidly changing, even chaotic, international political environment, have up-ended these models. Globalization has increased the diversity and uncertainty in outcomes when complex systems such as financial flows and markets, regional economies and political systems, and transnational threats involving multiple actors are in constant flux.

Advanced analytics is the application of multiple analytic methods that address the diversity of Big Data—structured or unstructured—to provide estimative results and to yield *actionable* descriptive, predictive, and prescriptive results. We cannot delve deeply into advanced analytics here, but the paper by Kaisler et al. (2014) addresses some of the issues and challenges associated with advanced analytics. Appendix A presents a brief description of the analytics classes we consider to be examples of advanced analytics.

The Internet of Things

Much has been written about the Internet of Things (IoT), where every device and, even someday, every human might be wired into the global network. IoT promises instant communication and access to information, but also a deluge of information unlike anything we have seen so far in the Big Data arena.

All of that data might have to be processed. Note that we say might because it is unclear now and will be for another decade or so how much of that information will be really useful. That means that we may be collecting data that we do not really need, perhaps storing it for some period of time, and then eventually discarding it.

The major challenge we face with IoT today is that machine-generated data will exceed the amount of data created by humans. We do not seem to be there yet, but it seems certain that the day cannot be far off. More sensors are being introduced into products, such as bicycles, coffee pots, washing machines, and thermostats that we use in everyday living. The data collected by these sensors are used to create other data which begets even more data and so on. Another example is Twitter tweets, where many original tweets are retweeted automatically based on profiles set up by receiving users. All retweets are stored someplace. In fact, it is likely that many retweets are retweeted in a cascading effect that distributes copies around the world.

Storage: What If You Cannot Keep Everything?

The quantity of data has exploded each time we have invented a new storage medium. What is different about the most recent explosion—due largely to social media—is that there has been no new storage medium. Moreover, data is being created by everyone and everything (e.g., devices, sensors, appliances, etc.)—not just, as heretofore, by professionals such as scientist, journalists, writers, and so on. Data is being accumulated at an accelerating rate in the scientific, business, and social media arenas. Social media accumulation results—in large part—from the growth in people using social media, but also the duplication of data such as retweeting of tweets and repinning of pictures on Pinterest.

Current disk technology—still the preferred storage medium—has increased to four terabytes per disk. So, one petabyte requires 250 disks and one exabyte would require about 250,000 disks to store raw data—independent of any indexing and redundancy requirements (which are large and significant). Even if one exabyte could be processed on a computer system, it would not be able to attach all of the disks necessary to store the data. Hilbert and Lopez (2011) calculated that the current sum of humanity's knowledge encompasses about 295 EBytes, but much of this still resides in the text on paper.

Assume a Radio Electronics Television Manufacturers Association (RETMA) rack takes up about 10 square feet of space to hold the disks and processors. One of these rack costs about $100,000 and has a rough capacity of about 1.2 PBytes. Hence, to store the total of humanity's knowledge, reduced to bits and bytes, would require somewhat over 250,000 racks costing about $250 billion and approximately 2.5 million square feet of space, not counting cooling and power equipment, infrastructure, and access space.

It is clear that the solution is not one large data center, but many, many data centers such as Amazon, Google, and Microsoft, among other firms, are creating in order to store the data that they host. And, these data centers need to be managed by cloud computing technology in order to ensure access to that data (EMC 2014).

Bandwidth: Getting the Data from Here to There

We need to differentiate between bandwidth and data transfer. Bandwidth is a measurement of the data transfer rate over a network. Your data will load faster if you have more bandwidth. On the other hand, data transfer is the amount of data transferred from a site, which will depend on the number of requests to the site and the file sizes. Bandwidth is a raw number while data transfer rate is a number composed of many implications of processing and overhead, error handling, and infrastructure support.

If we assume that a computer system could attach to a disk farm that could store an exabyte of data, we now face the problem of getting the data from the disks to the processor(s). Let us assume a 100 Gbit/s (or about 12.5 GBytes/s) communications network with an 80 percent sustainable transfer rate yielding a bandwidth of 10 Gbytes/s. It would take about 1×10^5 seconds to transfer a petabyte (or about 27.7 hours) and 1×10^8 seconds (or about 27,777 hours) to transfer one exabyte.

1 TByte = 1,000 GBytes, so 1×10^2 seconds to transfer
1 PByte = 1,000 TBytes, so 100,000 seconds to transfer = 1×10^5 seconds
1 EByte = 1,000 PBytes, so 100,000,000 seconds to transfer = 1×10^8 seconds
With 3600 s/hour, 1 EByte will take about 28×10^3 hours

Of course, this makes some simplifying assumptions such as just raw speed and no overhead associated with managing the data transfer.

The basic limitation is physical: data can only travel so fast through a wire, even if superconducting, or through an optical fiber. At present, the upper limit seems to be in the 10s to 100s of Gbits/s, with the next target about 400 Gbits/s. 1 Tbit/s (one terabit per second) could be just around the corner, but actually seems to be much farther in the future, although it has been demonstrated under experimental conditions (Brodkin 2012; Cisco 2015a, 2015b).

As the amount of data to be transferred continues to exponentially increase while the single conduit bandwidth is limited, the only solution—in the near term—will be to use multiple conduits between sites, with the attendant overhead costs, in order to transfer increasing amounts of data. Costs will vary with technology, distance, provider, and many other considerations. However, for large organizations or those dependent on real-time data transfers of Big Data, the cost will have to be borne until the technology catches up.

CHAPTER 7

Conclusion

Capturing the Value of Big Data Projects

It is about adding *value.*

This book has defined Big Data, described its important analytical characteristics, and illustrated some of the many ways managers and executives can utilize Big Data to support business operations and improve service delivery. The Big Data picture painted thus is quite complex. But simply understanding the Big Data picture is not the goal of an effective decision maker. Effective decision makers and executives are constantly and clearly focusing attention and concentration on the prize: *applying* the Big Data findings and *obtaining strategic and tactical results* by capturing the promise of value in the data and applying this to improve services. The methods used to extract understandable and usable information that provide organizations with the knowledge to make highly improved operational, tactical, and strategic decisions must be combined with both a sound understanding of methods with technical and methodological names, and clear business strategies shaping how the knowledge derived from the Big Data can be effectively employed.

The key for a manager is not in simply being able to describe the characteristics of these tools, or in wildly and randomly applying the tools to all the data owned or available to an organization. It is in deeply appreciating the business and decisional impacts of the characteristics of the Big Data. Two examples will show why the Big Data analysis may be so useful or even critical to success of a service, product, or business.

The reports and research documenting the value of Big Data have been arriving for several years. Bain research surveys of executives of more than 400 companies with more than a billion dollars of revenues have shown that companies that are good at applying analytics and changing their business due to analytic findings are twice as likely to be in the highest quarter of the financial performers in their industry, are three times more likely to implement the executive decision as they desire, and are five times as fast as making decisions (Wegener and Sinha 2013).

Examples of value can be very specific. The Bain research cites Nest as an example of a company that not only applies remote control thermostat technology to control the environment of a home through the Web, but additionally uses crowdsourced intelligence to determine when and how the home's thermostats are set and changed. These data are then associated with other factors such as the home's local environmental conditions, weather, physical location, and construction type to assist in determining the setting to create a more pleasing living environment inside the home (Wegener and Sinha 2013).

Where is the value being found by using Big Data and data analytics? Early leaders may be financial services, technology, and health care. Examples include mail-order pharmacies that found increases in service calls associated with specific refill windows for prescriptions. Deep analysis showed that customers had variable dosages. The pharmacy responded with periodic predictive calls inquiring how many pills customers had remaining and reduced time-consuming customer service calls and emergency refills (Wegener and Sinha 2013). Examples are beginning to proliferate—with call centers routing executive request, more important clients, or premium consumers based upon telephone numbers. In the airline industry, these data can be combined with flight status data that can possibly predict why one is calling (flight diversion or delay), and used to deliver a more rapid update without taking time to receive, record, analyze, and identify a response.

Where can value be expected to be realized? Look for Big Data to make information (of many types—customers, products, inventories, complaints, and so on) transparent and usable at much higher frequency. Digitized transactional data is more available and precise thus exposing relationships to outcome variables showing true costs of performance (product variability, personal absences, traffic blockages, etc.). Value is realized by acting on the enormous opportunities to experiment with clients, alternate deliveries, service variations, and product improvements. One can then compare success rates thereby making managers and executives into scientists who can realistically apply a more scientific or analytical method to seek out and pinpoint the key actions and reasons behind sales increase or decrease or behaviors finely tuned likes and dislikes of customers (Manyika et al. 2011). Value is thus found in multiple areas of businesses and in numerous processes—through the tailoring of services, managerial decision-making, and the rapid design and development of generations of products and services.

The more generic lists of possible potential business benefits include timely insights from the vast amounts of data; real-time monitoring and forecasting of actions and activities that effect performance; ability to interchange the data tools used such as SAP HANA, SAP Sybase®, SAP Intelligence Analysis for Public Sector application by Palantir, Kapow®, Hadoop; improvement through including data of different types and increasing velocity; and better validation and verification, better risk recognition, assessment, and decision making.

A few common characteristics of Big Data insights include the following.

The first example is one of swatting—where in current "practice" someone may report a murder, kidnapping, shooting, or some other crime that requires a swat team response. The team is then directed to an innocent's location, where doors are broken in, and the "swat rescue" is imposed by an unknowing police but well-intentioned rescue force under the pressure of potentially life-saving consequences. What occurs is "swatting, " a more and more prevalent Internet-aided trick on the police in which cybercriminals report a criminal hostage situation or shooter on the loose threat with the goal of unleashing a SWAT response on an unprepared and unknowing individual.

How does this relate to Big Data? First, careful analysis of the data may identify the perpetrators. But further analysis may enable one to examine and understand the linkages in the network, and how to utilize the connectivity and speed of the functional communication networks that "share" the information or report in real time. Swatting incidents do not even have to be real to illustrate how communications mapping can demonstrate the power of Big Data analysis. A recent viral Internet video depicts a 15-year-old boy being convicted of this crime and sentenced to a long jail term for his actions. Many believe the "conviction story" is true, but careful analysis of the data shows it is false. (No one was convicted of the crime—it did not even occur.) However, the Big Data will map the rapid dissemination and communication of this story. When this mapping is combined with survey and Big Data analysis, the power or reach of the story is clearly demonstrated.

A second similar example is of a non melting ice-cream sandwich. Several videos show that in 80° sunny temperatures, the ice-cream sandwich does not melt or change its shape. It is embarrassing, and the video obviously is intended to reflect negatively on this particular treat. The Big Data impact is derived from the transmission of this message, tracking of views, and viral nature of the communication generated. The communication paths themselves, wide dissemination of the "event" or message through channels, methods by which the messages are "picked-up" and passed, and immediacy with which the social media communicate the real or supposedly real event are useful for planning communication events and developing marketing programs. The networks of communication paths can be mapped, the impact can easily be seen in supporting social communication and video views, and then used to project an attached or implanted service message, product capability, or to defuse (or accentuate) a problem event or situation.

Finally, there is value in the data beyond that than can be captured within the processes of the organization. But there are no official guidelines for assessing data value because data is not a real asset like a factory or tangible cash. The traffic in information is large and generates revenue, but standard methods for valuing data must be developed. The issue is great. The *Wall Street Journal (WSJ)* reports that supermarket operator Kroger Co. has more than 2,600 stores where it tracks the purchases of 55 million loyalty-card members. Data are analyzed for trends and then, through a joint venture, sold to the store vendors. The consumer-products makers purchase this information and are able to thereby adjust products and marketing to match consumer choices, likes, and dislikes. It is estimated that Kroger receives $100 million a year from such data sales (Monga 2014).

Metrics: What Are They and How Should They Be Used?

Performance assessment is a critical foundation of any agreement, deal, or transactional relationship established between parties involved in a service arrangement. And managers, executives, and organizations are all judged by performance. The judgment process requires that measures or metrics be set that will aid in establishing the value of a service, and if the service is effective. The metric may further aid in setting behaviors and actions and in assessing tactics and strategies, thus determining managerial and executive decisions and actions. Metrics can be constructed to evaluate the "value" by determining degrees of performance and compliance, by encouraging future performance improvements, by demonstrating increases in effectiveness and efficiency, and by providing managers with mechanisms to determine the levels of effective internal controls.

Determining Value

Determining the value that can be or is derived from analysis of Big Data is a critical problem from many perspectives. The manager may realistically focus on quantitatively determining the impact of Big Data on the customer's overall satisfaction (with a single purchase or service, or a reoccurring use), with the performance and productivity of the organization, and with the impact on the organization's workforce. Questions that executives and managers must ask are the keys to assessing value. Do the data show that the outcomes are in full agreement with the mission and goals of the organization? Is the product or service less costly to produce, or of higher quality? Does the service meet all (or more) of the needs and requirements of a customer? Does the service exhibit higher quality characteristics? Are inventory reduced or are inventory turns increased? Are corporate promises being met (explicit or implied)? Are services being delivered in a more well-timed frame (thus reducing backlogs or shortening wait times for customers) and increasing the number of customer positive communications to others?

Measurement Costs of Big Data

Of course, there are other organization investment metrics that can be assessed that will also demand that costs of producing the benefit be fully assessed. This calculation requires that one begin measuring the *value* delivered to businesses from Big Data and its corresponding analysis by defining the expenses associated with producing analytics, and including the infrastructure components required to structure, cleanse, analyze, and present the Big Data to decision makers. The costs are openly assessed in their initial collection. They are composed of the costs of tools (software, computation, management, and storage), costs to acquire data (either by purchase or direct collection), analysis costs (personnel), and visualization costs for preparing the data for consumption by managers and executives and to associate, support, or relate the data to the decisions to be made.

Using Organization Performance Metrics to Explain Big Data Findings

Assessing Big Data's value is a far more difficult problem when the analysis must deliver explanations as to why end measures of organizational performance previously identified are being obtained. The Big Data and analytic results must be correlated with increased sales, customer growth, increased size of purchases, profits, faster inventory turns, and the list goes on and on. The explanatory process requires that performance metrics for the organization be developed by involving the employees who are directly accountable for the effort to be assessed. They are knowledgeable about the processes and activities involved in the work. The steps in the value assessment process include:

- Establishing the essential work activities and customer requirements
- Aligning work outcomes to customer requirements
- Setting measures for the critical work processes or critical results
- Establishing work goals, criterions, or benchmarks

Goals may be set at multiple levels, for end objectives related to the mission or function, direct targets that match divisions of responsibility within primary functions, and end metrics designed to link to specific improvements characterizing advances for each criteria.

These must be quantifiable and concrete targets based on individual expected work results.

With careful organizational planning, the measures themselves will be clear and focused, not subject to misunderstanding, quantified and compared to other data through statistical analysis, reasonable and credible, and collected under normalized or expected conditions. Measures must also appropriately fall within the organization's constraints and be actionable within a time determined frame. Measures may take various forms—and be seen as trends or rate-of-change (over time) data including views set against standards and benchmarks—or be fixed with baselines to be established at some future point, or determined incrementally via milestones.

A question still remains—when are the organizational metrics good? Are they useful or important for the organization's success? Although they may be done well, following all appropriate applied rules and procedures, a manager still wonders—is the analytical data good? There are no silver bullets guaranteeing how the quality of the metrics can be assured. Answering key questions can help to make this determination. The straightforward questions listed below will point the way.

First, ask oneself about the content: Was it objectively measured? Was there a clear statement of the goal or result? Did the metric support customer service requirements and meet all compliance requirements? Did it focus on effectiveness and efficiency of the system being measured? Secondly, ask about the properties: Does the metric allow for meaningful trend or statistical analysis, meet any industry or other external standards, possess incremental milestones, and does it include qualitative criteria? Finally, how does the metric fit the organization objectives: Are the metrics challenging but at the same time attainable? Are assumptions and definitions specified for excellent performance? Are the employees who must enact the activity being measured fully involved in the development of this metric? Has the metric been jointly fixed by you and your service customer?

Table 7.1 Metric quality assessment table

Assessment question	Stronger	Weaker
Measurement approach	Objective	Subjective
Clarity	High	Ambiguous
Directly related to customer service	Yes	No
Focus on effectiveness, efficiency	Yes	Yes
Properties	Display as trend	Isolated, on time
Meet external standards	Yes	No
Incremental data	Yes	No
Qualitative	Yes	No
Attainable		No
Clear assumptions	Yes	No
Employee full involvement	Yes	No
Service customer agreement	Yes	No

Table 7.1 can be used as a starting point for assessing the development and tailoring of good metrics that are specific to an organization.

The end results are that the organization must deliver services to customers on time and deliver increased revenue or a protected market position when the data and measures are associated. Targets achieved must be quality services with increased buyer satisfaction, minimized or reduced rework or contract responsibility, possess a high return on investment (ROI), and ensure that the organizational capacity in resources and equipment is present to deliver the services.

Analytic Metrics

Beyond assessing the value that Big Data brings to organizational performance and decision making, a corresponding question is the need for metrics to measure the usefulness of analytics and analytic tools. There are few metrics for assessing tools or assessing analytical techniques and mapping outputs (of tools) to outcomes (of analytical approaches). Specialized metrics for specific domains will be required to assess analytical approaches and take account of how domain knowledge is used to derive actionable information and intelligence. Additionally, new verification and validation techniques will be required to support the assessment of tool outputs and analytical method outcomes where the results may be Big Data in their own right.

Going Forward: What Should a Manager Do?

The basis for applying Big Data analysis will require that decision makers and executives develop comprehensive measures of performance associated with organization transaction activity. With today's technology, this can be extended to a sweeping instruction—*measure everything*, or at least establish a defined transaction baseline for beginning states of the outcomes and activities you will assess. But do it before you begin collecting and analyzing the data! However, there is no universal standard approach to establishing baselines. The collection and observing tools and the range and complexity of information made available can vary greatly among service providers, services, and customers. This enormous variability ultimately thwarts a generic or simple set of metrics from being applied for spotting the trends and findings that will be important in the analyses. The steps a manager and executive are to follow are straightforward.

Set Baselines

In the unassuming language of services, a performance baseline is composed of metrics that can define acceptable or standard operational conditions of a service. Obviously, the baseline performance will be used as a comparison to identify changes that indicate anything from a problem to an opportunity. Baselines may in themselves provide early warnings regarding service demands, capacity limitations, and upgrades in input requirements. Aligning baselines with targets or other objects can aid in staying within parameters or discerning difficult service spaces that may not meet standards or compliance goals.

Improve and Expand Classification Analysis

Decision makers and executives are familiar with the concept of grouping. Similarities in clients, services, requirements, demands, locations, and huge numbers of other characteristics are used to envisage participation that the client for a service is a member of a group or class, and therefore will find the offered service attractive. The class or classification membership is a predictor and therefore hugely important.

Predicting and explaining responses is the desired outcome. Our methods of developing classification are robust. A simple example of the use of a familiar classification tree is a good example of the benefits of classification. If we want to devise a system for approaching customers we may use a variety of tools—income, location, age, previous purchases, and interests—to classify them according to purchase potential. Thus, someone who files income taxes (at a level) does not live in a desert, is over the age of 17, has previously purchased products related to sailing, and subscribes to a sailing magazine may be a candidate for a sailboat purchase. Categorization processes and procedures can be more positive (e.g., categorizing plants and coins), but the principles for obtaining classifications and benefits are the same.

A decision maker or analyst does not need to know in advance what categories may be important. One can permit the data or values of the customers or attributes that were measured to describe the grouping or the structure of the metrics representing the customers included in our data. This permits the metric to aid in determining what groups of observations are all near each other and from that likeness find either other similarities or data to aid in reasoning why they may be together. Thus, it is not sufficient to just realize that there are natural groupings; one must still determine why the customers are grouped by an observation (or multiple values) of the metrics that we have. Simply said, the data we collect, and metrics describing the service or customer, will do some of our analytical work for us.

However, the metrics and data do not do it all. There is a critical role for the business knowledgeable expert, manager, and executive. They must apply their organizational, service, and business knowledge. Managers must contribute knowledge that is complex and constructed from a deep understanding of an overall market, market segmentation, and opportunities or limitations that can vary by geography, culture, technology, economic and social trends that could impact service performance, and customer requirement. They must also understand the competition's strengths and weaknesses, and plans. Finally, managers must have an essential understanding of their envisioned customers in detail. Ask what the customer does, wants, values, and how current or similar services are now being employed to deliver value.

This complex understanding can be considered to be domain knowledge. And a manager or executive does not simply collect or acquire this understanding of the meaning of the Big Data but interprets, tests, predicts, and assembles all or portions of the Big Data to create a coherent depiction of the situation from the Big Data that are available for the analysis.

As an example, consider the expanding field of biology and the importance of Big Data. It is another useful Big Data domain with significant current research and great promise, but a deep technical, chemical, and biological understanding of this domain is essential for obtaining value from the analysis of Big Data. This domain covers the analysis of metabolites and biological molecules that are not as predicted or expected in terms of a location or quantity and the assessment of metabolites, macromolecules, proteins, DNA, and large molecule drugs in biological systems. Big Data analytical methods are available for measurable evaluation in biopharmaceutics and clinical pharmacology research. The data are used to assess the sample preparation of drugs in a biological matrix and to analyze the selectivity, specificity, limit of detection, lower limit of quantitation, linearity, range, accuracy, precision, recovery, stability, ruggedness, and robustness of liquid chromatographic methods. Researchers can then use Big Data tools to understand and report in studies on the pharmacokinetic (PK), toxicokinetic, bioavailability, and bioequivalence of a drug (Tiwari and Tiwari 2010).

It is readily apparent that this domain requires highly specialized knowledge to obtain value from these Big Data. Without such knowledge how can one possibly know or appreciate the relevance of these data? The executive, managers, and analysts of the future will have far more metrics available, but will need the classifications, knowledge, and insight to interpret and explain the associations. Thus, managers must invest in classification tools and expect more and better analysis as the groupings and findings fall out of the metrics and data. However, there is still an essential role for the manager!

Expect to Merge and Mix Data Types

The era of Big Data will see more and more to mixed data types. The events and activities of the world are experienced in conjunction with many other activities. Thus, the combined data may contain many types of information and classifications. Businesses and organizations have experienced this problem for many years. But the size and complexity of the data now make this problem more acute.

Mixed data can be best understood when one thinks of various categories. Some may be categories such as different types of service. Others may be continuous such as measurements showing sizes or quantities that can increase or decrease. Decision makers may expect that data will simultaneously combine data that have both categories and sizes—many of each at the same time. Differentiating and analyzing why metrics are obtained under each situation, and how changes will happen in the future, is the role of the manager and decision maker.

Examples will show that we can have an item or event categorized (as a yes or no, present or not present property); this could be combined with data that are frequency counts of how many times something happened. Medical situations can have these characteristics where patients are categorized by age, height, weight, test results, and previous occurrences of the event. Under these circumstances, the data can simply be counted, continuous, ranked, or as present or not present. Decision makers will need to be prepared for the volume of data, number of tools, results, and mixtures of data. Undeniably, all will increase as Big Data increases.

Executives: Listen and Watch Carefully for the Breakthroughs and Success Stories

It is difficult to envision the types of changes that will be encountered because of the growth of Big Data stimulated by the attraction, swiftness, and unpredicted impacts of social communications. The Big Data sets will also be added to and combined with the numerous sensor-driven data collections and ubiquity of enabled devices.

Many stories and articles of Big Data findings, applications, and successes will stimulate ideas and analytical approaches. Know and watch the competition, customers, services, and related industries for changes and breakthroughs via cross-industry pollination. A dramatic example of the use of mixed data detected structural changes on two publicly available social network data sets: Tactical Officer Education Program e-mail Network and open source Al-Qaeda communications network. Significant changes in both networks were detected with Al-Qaeda network structurally changing prior to the attacks of September 11 (McCulloh et al. 2008). Changes of this nature, that really make a service difference, will likely be exponential rather than incremental and similar to the black-swan event. One cannot be fully prepared, but do not be surprised.

What Is the Big Data Future?

All predictions point to a very bright future for Big Data! Obviously, executives and managers will encounter a great proliferation of new analysis tools, and trials of cleverly crafted organizational applications of this technology. Further, there are many educational programs churning out well trained and prepared students of the Big Data technology. They will have an enormous business impact. Organizations will hire and then quickly learn to strategically respond to the market with even more inventive uses of tools and combinations of data. Thus, there is no time to waste in applying the poser that can be drawn from Big Data to service delivery. What seems to be approaching, or may already be upon us, is a storm of changes (or an "arms race" like proliferation of technology) driven by Big Data.

This future is supported by a look at past predictions that help us understand the rapidity of the Big Data driven changes upon us. For a number of years, research firms and consultants predicted the rise and value that could be derived from Big Data and analytics. For example, McKinsey studied Big Data in five domains—health care in the United States, the public sector in Europe, retail in the United States, and manufacturing and personal-location data globally. Their findings lead to the prediction that Big Data could generate value in each area (Manyika et al. 2011).

With the growing attention on Big Data, the Pew Research Center began assessing it as a topic of interest in 2012 when the National Science Foundation and National Institutes of Health began seeking new methods, infrastructure, and approaches to find scientific and technological means for working with large data sets to accelerate progress in science and engineering, associated education, and training. The report described how Big Data is already having impacts with descriptions of spelling changes in Google search queries attributed to previous queries on the same subject that employed different spelling—a finding that enables an ability to identify search trends that may predict economic and public health trends (Anderson and Rainie 2012).

Other 2012 example uses include calls about "unusual activity" based on assessments of transaction anomalies in buyer behavior that may be evidence in indicators of fraud, and subscriber movie recommendations based upon profiles and previous actions of users. But there were still questions in the minds of managers and executives regarding effectiveness. Pew also reported that a 2012 survey of chief marketing officers report that more than half of the respondents said they then lacked the tools to mine true customer insights, and 58 percent lacked the skills and technology to perform analytics on marketing data (Anderson and Rainie 2012).

Today, as shown in this work, we can see that many predictions of the growth and uses of Big Data were correct. But what of the future for technology associated with Big Data, the public's opinion and acceptance of the collection and uses of Big Data, and specific evolvement of business applications and uses?

In the area of analytics and technology, it is predicted that Hadoop will address many different scenarios and applications that require a mix of batch, real-time, streaming, and interactive scenarios (Taft 2014). However, limitations of the Hadoop MapReduce paradigm are already leading to new programming and analytics paradigms, such as Spark, Storm, and so on. From a hardware perspective, systems will use photonics for interconnections, memristers for huge amounts of storage that can be placed near the processing engines, and optimized computer and analytic processors for specialized tasks (Shaw 2014).

In the public space dealing with acceptance and understanding at large, the Big Data explosion has generated societal issues related to privacy, security, intellectual property, and liability. This may well lead to public policy and regulation beyond those current in place. Further, the value of the data may begin to be formalized as organizations learn and conduct transactions that determine in the marketplace what all the information is worth. There may be implied or explicit guidelines for assessing data value because data becomes recognized as an asset and can be equated to cash as standard methods for valuing data are developed. (And with the costs to society, perhaps we can expect new forms of taxation of this real, but difficult to measure, asset.)

From a business perspective, companies will continue to invent new models, search for associations in the data, and rapidly integrate information from multiple data sources, purchased from secondary sources and third parties. The sources and elements that will be added will include all forms of social media, volumes of sales and commercial data, location data, and accessible financial data. It is further predicted that users (subject matter experts) will create their own 360° views of data; the users will choose the data sources, analysis tool and technique, and the visualization—and will then be able to repeat this analytic as an experiment (as the gaming industry does today); organizational data will be structured and then connected to human interaction data; will improve our understanding of workers—state, mood, goals, preferences—and will recommend how to assist the human doing the work; and the tools will look for patterns in the data and will offer predictions of events for the users to asses and address (Shaw 2014).

Our Conclusion

Big Data has been a physical science problem for many years but has now become the "new" business and social science frontier. The amount of information and knowledge that can be extracted from the digital universe is continuing to expand as users come up with new ways to message and process data. Moreover, it has become clear that "more data is not just more data," but that "more data is different," for example, more diversity may produce better results.

Big Data is just the beginning of the problem. Technology evolution and placement guarantee that in a few years more data will be available in a year than has been collected since the dawn of man. If Facebook and Twitter are producing, collectively, around 50 gigabytes of data per day, and tripling every year, within a few years (perhaps three to five) we are indeed facing the challenge of *Big Data becoming really Big Data*.

There are numerous issues and challenges associated with Big Data usage. Technical scaling challenges will need to be solved as the volume of data approaches the 3Es—exabytes, exabits, exaflops—thresholds. However, the complexity of data—its complex semantics and structures, the context required to understand the data, and the need to track the provenance of derived data back to the original data—pose much larger problems.

The importance of Big Data to business and organizations—both profit and non profit—is indicated by the increasing number of Big Data startups each year. And more companies are engaging in Big Data initiatives either in-house or through consultants. Increased demand is putting pressure on the pool of data and analytic scientists. Training for data and analytic scientists will need to increase in both academic and corporate and organizational venues, with in-house training possibly becoming the training of first resort, at least for large organizations.

We conclude that the impact will be great, and the Big Data business itself will continue expanding. Research and Markets have predicted the evolution of Big Data technologies, markets, and outlays (expenditures) over the next five years—from 2015 to 2020. Using a triple-scenario analysis, including "steady state," the emergence of new Big Data analytical tools, and the rise of new analytical tools that can replace Big Data analysis; their prediction is that the market will almost double in the coming five years, from nearly $39 billion in 2015 to more than $76 billion in 2020 (PRNewswire: Dublin 2014).

Our conclusion: Big Data is here, evolving, and swiftly becoming a major force that must be understood and effectively employed.

APPENDIX A

Methods-Based Analytics Taxonomy

Business analytics has recently focused on statistical analysis, data mining, and statistical machine learning methods to analyze Big Data. These are mathematical methods that work with numerical data or symbolic data encoded as numerical data, but they do not capture the domain material that reflects critical factors that affect business decisions, such as culture; political environment; gender, age, and ethnic preferences; geographic and demographic features; and so on.

There are a multitude of other methods for analyzing data and information. Table A.1 presents a taxonomy of analytics based on classes of methods. This taxonomy was first developed by Stephen Kaisler and Claudio Cioffi-Revilla (Kaisler and Cioffi-Revilla 2007). Each class represents many different methods—each with its own advantages and disadvantages.

Table A.1 Taxonomy of analytic classes

Analytic class	Description
Indications and warning (NS)	Assess situations to determine if any event is likely to occur. The goals are to give a probability of occurrence and describe the characteristics of the event. For example, analysis and prediction of rare earth supply chain disruption due to ROC policies.
Dynamical systems (N)	Usually comprised of a set of differential or difference equations of low dimensionality representing known competing forces. A problem is how to handle missing forces. For example, applying sociodynamic principles to patterns of recruitment for terrorism (Choucri et al. 2006).
(Hidden) Markov models (N)	(H)MMs represent the problem space as a set of discrete states, fully linked with associated weights, and an associated set of probabilities. Input data drives the transitions between states. For example, numerous applications to automatic speech recognition (Rabiner 1989).
Event data analysis (NS)	Examine an event stream to identify patterns of events leading to particular outcomes. A problem is how to handle missing or exogenous events due to incomplete models. For example, modeling political decisions in the face of uncertain information (Cioffi-Revilla 1998).
Econometric and sociometric models (N)	E and S models are large-scale aggregate models of actors in economic or social contexts. For example, Paul Collier's work on civil war and rebellion and their motivations (Collier and Hoeffler 2000; Collier, Hoeffler, and Soderbom 2001).
Regression models (N)	Regression methods, for example, logistics regression, use recurrence equations to relate past results to future values.
Reduction techniques (N)	Mathematical methods that reduce data dimensionality to identify essential variables, including principle components analysis and singular value decomposition.
Game theory models (NS)	Apply rational decision models to 2- or n-person situations exhibiting both local and strategic interdependence in cooperative or competitive situations. For example, the many applications of the "Tragedy of the Common Good" (Hardin 1968).

Expected utility models (N)	Compute the value of certain outcomes for individual and collective choices, for example, Bayesian decision theory. For example, analyzing public lottery contests such as MegaMillions.
Control theory models (N)	Apply (non)linear and optimal control theory principles to modeling interactions among individuals and organizations. For example, the qualitative application of Nisbett's work on cultural differences among different societies (Nisbett 2003).
Survival models (N)	Compute the hazard rate or intensity function of a process to determine its lifetime. For example, reliability models for equipment operation or the spread of rumors among a population.
Evolutionary computation (NS)	Apply evolutionary models, such as genetic algorithms, to determine feasible alternatives.
State transition models (NS)	Model interactions between entities as transitions between discrete states over time. For example, cellular automata and Petri nets.
Graph and mesh models (NS)	Model entities linked through one or more relations. For example, Renfro and Deckro's analysis of influence in the Iranian Government (Renfro and Deckro 2003).
Agent-based simulation (NS)	Apply multi agent system models to simulate human and social dynamics in complex environments. For example, the STRADS geopolitical simulation systems (Oresky et al. 1990).
Field theory models (NS)	Apply spatial field models of interactions among entities to extend graph analytics and agent-based simulations.
Logic systems (S)	Use of logical formulae and systems to represent and solve qualitative problems, including deductive, abductive, and inductive techniques. For example, the application of constraint solvers to support dynamic taint analysis in program understanding (Zhang 2008).

Legend: N = numerical data, S = symbolic data

References

Anderson, J., and L. Rainie. 2012. "Main Findings: Influence of Big Data in 2020." Pew Research Center. http://www.pewinternet.org/2012/07/20/main-findings-influence-of-big-data-in-2020/

Barlow, M. 2013. *Real-Time Big Data Analytics: Emerging Architecture*. Sebastopol, CA: O'Reilly Media.

Barth, P., and R. Bean. 2012. "Who's Really Using Big Data?" *Harvard Business Review Blog Network*, September 12.

Borne, K. 2013. "Statistical Truiisms in the Age of Big Data." Retrieved December 2013 from http://www.statisticsviews.com/details/feature/4911381/Statistical-Truisms-in-the-Age-of-Big-Data.html

Boyd, D.F. 1950. "Applying the Group Chart for X and R." *Industrial Quality Control* 7, no. 3, pp. 22–25.

Brodkin, J. 2012. "Bandwidth Explosion: As Internet Use Soars, Can Bottlenecks Be Averted?" Ars Technica. http://arstechnica.com/business/2012/05/bandwidth-explosion-as-internet-use-soars-can-bottlenecks-be-averted/

Brynjolfsson, E. December 3, 2010. International Institute for Analytics.

Buneman, P., and S.B. Davidson. 2010. *Data Provenance—The Foundation of Quality Data*. Pittsburgh, PA: Software Engineering Institute, Carnegie-Mellon University.

Chakrabarti, D., and C. Faloutsos. 2012. *Graph Mining: Laws, Tools and Case Studies*. San Rafael, CA: Morgan and Claypool Publishers.

Choucri, N., C. Electris, D. Goldsmith, D. Mistree, S.E. Madnick, J.B. Morrison, M.D. Siegel, M. Sweitzer-Hamilton. January 2006. "Understanding and Modeling State Stability: Exploiting System Dynamics." *MIT Sloan Research Papers*, no. 4574–06.

Cioffi-Revilla, C. 1998. *Politics and Uncertainty: Theory, Models and Applications*. Cambridge and New York: Cambridge University Press.

Cisco, Inc. 2015a. "Cisco Visual Networking Index: Forecast and Methodology 2014–2019 White Paper." http://www.cisco.com/c/en/us/solutions/collateral/service-provider/ip-ngn-ip-next-generation-network/white_paper_c11-481360.html

Cisco. Inc. 2015b. "The Zettabyte Era: Trends and Analysis." http://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/VNI_Hyperconnectivity_WP.html

Clifford, S. 2012. "Shopper Alert: Price May Drop for You Alone." *The New York Times*. http://www.nytimes.com/2012/08/10/business/supermarkets-try-customizing-prices-for-shoppers.html

Collier, P., and A. Hoeffler. 2000. "Greed and Grievance in Civil War." Policy Research Working Paper No. 2355. Washington, DC: World Bank.

Collier, P., A. Hoeffler, and M. Soderbom. 2001. "On the Duration of Civil War, Volume 1." Policy Research Working Paper No. 2681. Washington, DC: World Bank.

Cook, D., and L. Holder, eds. 2006. *Mining Graph Data.* New York: John Wiley.

Council on Library and Information Resources (CLIR). 2012. "Data Curation". Retrieved November 17, 2013 from http://www.clir.org/initiatives-partnerships/data-curation

Cruz, J.A., and D.S. Wishart. 2006. "Applications of Machine Learning in Cancer Prediction and Prognosis." *Cancer Informatics* 2, 59–77. http://www.ncbi.nlm.nih.gov/pubmed/19458758

Davenport, T.H. January 2006. "Competing on Analytics." *Harvard Business Review* 84, no. 1, pp. 98–107.

Davenport, T.H. 2007. *Competing on Analytics*. Cambridge, MA: Harvard Business School Publishing Corporation.

Deng, H., G.C. Runger, and E. Tuv. 2012. "Systems Monitoring with Real Time Contrasts." *Journal of Quality Technology* 44, no. 1, pp. 9–27.

Desisto, R.P., D.C. Plummer, and D.M. Smith. 2008. *Tutorial for Understanding the Relationship Between Cloud Computing and SaaS*. Gartner, G00156152.

Eaton, C., D. Deroos, T. Deutsch, G. Lapis, and P. Zikopoulos. 2012. *Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data*. New York, NY: McGraw Hill.

EMC. Inc. 2014. "Managing Storage: Trends, Challenges, and Options (2013–2014)." https://education.emc.com/content/_common/docs/articles/Managing_Storage_Trends_Challenges_and_Options_2013_2014.pdf

Feinleb, D. 2012. "Big Data Trends." Retrieved November 10, 2013 from http://thebigdatagroup.com/

Ferraty, F. and Y. Romain. 2011. *The Oxford Handbook of Functional Data Analysis*. Oxford Handbooks. Oxford, New York: Oxford University Press.

Ford, N. 2006. "Polyglot Programming." http://memeagora.blogspot.com/2006/12/polyglot-programming.html

Fowler, M. 2011. "Polyglot Persistence." http://martinfowler.com/bliki/PolyglotPersistence.html

Gantz, J., and E. Reinsel. 2011. "Extracting Value from Chaos." IDC's Digital Universe Study, sponsored by EMC. http://www.emc.com/collateral/analyst-reports/idc-extracting-value-from-chaos-ar.pdf

George, L. 2011. *HBase: The Definitive Guide.* 1st ed. Sebastopol, CA: O'Reilly Media.

Hardin, G. 1968. "The Tragedy of the Commons." *Science* 162, no. 3859, pp. 1243–48.

Harris, H.D., S.P. Murphy, and M. Vaisman. 2013. *Analyzing the Analyzers*. Sebastopol, CA: O'Reilly Media.

Harrison, G. 2010. "10 Things You Should Know About NoSQL Databases." Tech Republic. http://b2b.cbsimg.net/downloads/Gilbert/dl_10_things_nosql.pdf

Hewitt, E. 2010. *Cassandra: The Definitive Guide*. 1st ed. Sebastopol, CA: O'Reilly Media.

Hilbert, M., and P. Lopez. 2011. "The World's Technological Capacity to Store, Communicate, and Compute Information." *Science* 332, no. 6025, pp. 60–65.

Hilbert, M., and P. Lopez. 2012. "How to Measure the World's Technological Capacity to Communicate, Store and Compute Information, Part I: Results and Scope." *International Journal of Communication* 6, pp. 956–79.

Hill, K. 2012. "How Target Figured Out a Teen Girl Was Pregnant Before Her Father Did." *Forbes*. http://www.forbes.com/sites/kashmirhill/2012/02/16/how-target-figured-out-a-teen-girl-was-pregnant-before-her-father-did/

Hwang, W., G. Runger, and E. Tuv. 2007. "Multivariate Statistical Process Control with Artificial Contrasts." *IIE Transactions* 39, no. 6, pp. 659–69.

IDC (International Data Corporation). December, 2012. "Big Data in 2020." http://www.emc.com/leadership/digital-universe/2012iview/big-data-2020.htm

Jacobs, A. 2009 "Pathologies of Big Data." *Communications of the ACM* 52, no. 8, pp. 36–44.

JASON. 2008. Data Analysis Challenges. The Mitre Corporation, McLean, VA, JSR-08-142.

Jirasettapong, P., and N. Rojanarowan. 2011. "A Guideline to Select Control Charts for Multiple Stream Processes Control." *Engineering Journal* 15, no. 3, pp. 1–14.

Johnston, L. 2013. "Defining the 'Big' in Big Data." Retrieved from November 5, 2013. http://blogs.loc.gov/digitalpreservation/2012/05/defining-the-big-in-big-data/

Kaisler, S. March 2005. Software Paradigms. New York: John Wiley & Sons.

Kaisler, S., and C. Cioffi-Revilla. 2007. "Quantitative and Computational Social Science." *Tutorial Presented at 45th Hawaii International Conference on System Sciences*. Wailea, HI.

Kaisler, S., and W. Money. 2010. *Dynamic Service Migration in a Cloud Architecture*. England: ARCS 2010 Workshop, Said Business School, University of Oxford.

Kaisler, S., and W. Money. January 8, 2011. "Service Migration in a Cloud Computing Architecture." *44th Hawaii International Conference on System Sciences*. Poipu, Kauai, HI.

Kaisler, S. 2012. “Advanced Analytics”, CATALYST Technical Report, AFRL technical Report (based on work by S. Kaisler and C. Cioffi-Revilla (George Mason University), *Quantitative and Computational Social Sciences Tutorial, 40th Hawaii International Conference on System Sciences*. Waikoloa, HI, 2007.

Kaisler, S., W. Money, and S.J. Cohen. January 4–7, 2012. “A Decision Framework for Cloud Computing.” *45th Hawaii International Conference on System Sciences*. Grand Wailea, Maui, HI.

Kaisler, S., F. Armour, A. Espinosa, and W. Money. 2013. “Big Data: Issues and Challenges Moving Forward.” *46th Hawaii International Conference on System Sciences*. Maui, HI: IEEE Computer Society.

Kaisler, S., F. Armour, A. Espinosa, and W. Money. 2014. “Advanced Analytics: Issues and Challenges in the Global Environment.” *47th Hawaii International Conference on System Sciences*. Hilton Waikoloa, Big Island, HI: IEEE Computer Society.

Kaisler, S., F. Armour, A. Espinosa, and W. Money. 2015. “Introduction to Big Data.” *Tutorials presented at 48th Hawaii International Conference on System Sciences*. Poipu, Kauai, HI: IEEE Computer Society.

Laney, D. 2001. “3D Data Management: Controlling Data Volume, Velocity and Variety.” Retrieved October 30, 2103 from http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf

Linthicum, D. 2009. “Defining the Cloud Computing Framework: Refining the Concept.” *Cloud Computing Journal*. http://cloudcomputing.sys-con.com/node/811519

Maglio, P.P., and J. Spohrer. 2008. “Fundamentals of Service Science.” *Journal of the Academy of Marketing Science* 36, no. 1.

Manyika, J., M. Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh, and A. Hung Byers. May 2011. “Big data: The Next Frontier for Innovation, Competition, and Productivity.” McKinsey Global Institute. http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation

Marr, B. 2015. “Walmart: The Big Data Skills Crisis and Recruiting Analytics Talent.” *Forbes Magazine*. http://www.forbes.com/sites/bernardmarr/2015/07/06/walmart-the-big-data-skills-crisis-and-recruiting-analytics-talent/2/

McAfee, A., and E. Brynjolfsson. 2012. “Big Data: The Management Revolution.” *Harvard Business Review*, Cambridge, MA.

McCulloh, I., M. Webb, J. Graham, K. Carley, and D.B. Horn. 2008. Technical Report 1235 Change Detection in Social Networks. United States Army Research Institute for the Behavioral and Social Sciences. Arlington, VA.

Megahed, F.M., W.H. Woodall, and J.A. Camelio. 2011. “A Review and Perspective on Control Charting with Image Data.” *Journal of Quality Technology* 43, no. 2, pp. 83–98.

Mello, P., and T. Grance. 2011. The NIST Definition of Cloud Computing (Draft), SP800-145. Gaithersburg, MD: National Institute of Standards and Technology.

Monga, V. 2014. The Big Mystery: What's Big Data Really Worth? The Wall Street Journal, October 12. http://www.wsj.com/articles/whats-all-that-data-worth-1413157156

National Academy of Science. 2013. *Frontiers in Massive Data Analysis*. WA: National Academies Press. Retrieved November 1, 2013 from http://www.nap.edu

Nestler S., J. Levis, B. Klimack, and M. Rappa. 2012. "The Shape of Analytics Certification." *INFORMS OR/MS Today* 39, no. 1. https://www.informs.org/ORMS-Today/Public-Articles/February-Volume-39-Number-1/The-shape-of-analytics-certification

Nisbett, R. 2003. *Geography of Thought: How Asians and Westerners Think Differently … and Why*. New York, NY: Simon and Schuster.

Oresky, C., A. Clarkson, D.B. Lenat, and S. Kaisler. 1990. "Strategic Automated Discovery System (STRADS)." In *Knowledge Based Simulation: Methodology and Application*, eds. P. Fishwick and D. Modjeski, 223–60. New York: Springer-Verlag.

Pearson, T., and R. Wegener. 2013. Big Data: The Organizational Challenge. Bain & Company, San Francisco.

Peters, B. 2014. "Big Data's Fading Bloom." http://www.forbes.com/sites/bradpeters/2014/02/28/big-datas-fading-bloom/

Pricewaterhouse Coopers, Inc. 2009. "12th Annual Global CEO Survey, Redefining Success." http://www.pwc.ch/user_content/editor/files/publ_corp/pwc_12th_annual_global_ceo_survey_e.pdf

PRNewswire: Dublin. 2014. "The Future of Big Data Analytics—Global Market and Technologies Forecast—2015–2020." Research and Markets. http://www.prnewswire.com/news-releases/the-future-of-big-data-analytics---global-market-and-technologies-forecast---2015-2020-275637471.html

Provost, F., and T. Fawcett. 2103. "Data Science and Its Relationship to Big Data and Data-Driven Decision Making." *Big Data* 1, no. 1, pp. 51–59.

Rabiner, L.R. 1989. "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition." *Proceedings of the IEEE* 77, no. 2, pp. 257–86.

Renfro, R.S. II, and R.F. Deckro. 2003. "A Flow Model Social Network Analysis of the Iranian Government." *Military Operations Research* 8, no. 1, pp. 5–16.

Ross, J.W., P. Weill, and D.C. Robertson. 2006. *Enterprise Architecture as Strategy: Creating a Foundation for Business Execution*. Cambridge, MA: Harvard Business School Press.

Satyanarayanan, M., P. Simoens, Y. Xiao, P. Pillai, Z. Chen, K. Ha, W. Hu, and B. Amos. April–June 2015. "Edge Analytics in the Internet of Things." *IEEE Pervasive Computing* 14, no. 2, pp. 24–31.

Shaw, M. 2014. "Big data 2020: What the Future of Analytics Means for the Enterprise." http://www.slideshare.net/hpsoftwaresolutions/big-data-2020 whatthefutureofanalyticsmeansfortheenterprise

Spam Laws. 2014. "Spam Statistics and Facts." http://www.spamlaws.com/spam-stats.html

Taft, D.K. 2014. "Hadoop 2020: The Future of Big Data in the Enterprise." *eWeek*, December 2. http://www.eweek.com/database/slideshows/hadoop-2020-the-future-of-big-data-in-the-enterprise.html

Talburt, J. 2009–2011. "Reference Linking Methods." *Identity Resolution Daily*. Retrieved November 3, 2013 from http://identityresolutiondaily.com/ (Site no longer exists)

Thomas, J.J., and K.A. Cook, eds. 2005. Illuminating the Path—the Research and Development Agenda for Visual Analytics. *IEEE Computer Society Visual Graphics Technical Committee.*

Tiwari, G., and R. Tiwari. 2010. "Bioanalytical Method Validation: An Updated Review." *Pharmaceutical Methods* 1, no. 1, pp. 25–38. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3658022/

Turner, D., M. Schroeck, and R. Shockley. 2012. *Analytics: The Real World Use of Big Data*. A Collaborative Research Study by the IBM Institute for Business Value and the Said Business School at the University of Oxford.

Varian, H. January 2009. "Hal Varian on How the Web Challenges Managers." *Mckinsey Quarterly*. http://www.mckinseyquarterly.com/Hal_Varian_on_how_the_Web_challenges_managers_2286

Wegener, R., and V. Sinha. 2013. "The Value of Big Data: How Analytics Differentiates Winners." Bain & Company. http://www.bain.com/publications/articles/the-value-of-big-data.aspx

Weick, K.E. 1995. *Sensemaking in Organizations*. Thousand Oaks, CA: Sage Publications.

Wikipedia. 2013. "Big Data." Retrieved October 24, 2013 from http://en.wikipedia.org/wiki/Big_data/

Wixom, B., and H. Watson. 2010. "The BI-based Organization." *International Journal of Business Intelligence Research* 1, no. 1, pp. 13–28.

Wong, P.C., and J. Thomas. 2004. "Visual Analytics." *IEEE Computer Graphics and Applications* 24, no. 5, pp. 20–21.

Zhang, H., S.L. Albin, S. Wagner, D.A. Nolet, and S. Gupta. 2010. "Determining Statistical Process Control Baseline Periods in Long Historical Data Streams." *Journal of Quality Technology* 42, no. 1, pp. 21–35.

Zhang, Y. 2008. *Constraint Solver Techniques for Implementing Precise and Scalable Static Program Analysis*. Technical University of Denmark, IMM-PHD-2008-211.

Further Reading

Ayres, I. 2007. *Supercrunchers*. New York: Bantam Books.

Box-Steffensmeier, J.M., and B.S. Jones. 2004. *Event History Modeling*. Cambridge, UK: The Press Syndicate of the University of Cambridge.

Clarkson, A. 1981. *Towards Effective Strategic Analysis*. Boulder, CO: Westview Press.

Davenport, T.H., and J.G. Harris. 2007. *Competing on Analytics: The New Science of Winning*. Cambridge: Harvard Business School Publishing Corporation.

Dunteman, G.H. 1989. *Principal Components Analysis*. Thousand Oaks, CA: Sage Publications.

Forrester, J.W. 1968. *Principles of Systems*. Cambridge, MA: Wright-Allen Press.

Forrester, J.W. 1973. *World Dynamics*. Cambridge, MA: Wright-Allen Press.

Gilbert, N., and K. Troitzsch. 2005. *Simulation for the Social Scientist*. 2nd ed. Buckingham and Philadelphia: Open University Press.

Goldberg, D.E. 1989. *Genetic Algorithms in Search, Optimization and Machine Learning*. Reading, MA: Addison Wesley.

Grabo, C. 2010. *Handbook of Warning Intelligence: Assessing the Threat to National Security*. Lanham, MD: Scarecrow Press.

Hilbe, J.M. 2009. *Logistic Regression Models*. Boca Raton, FL: Chapman and Hall/CRC Press.

Peterson, J.L. 1977. "Petri Nets." *ACM Computing Surveys* 9, no. 3, pp. 223–52.

Philpott, S. 2010. "Advanced Analytics: Unlocking the Power of Insight." http://ibmtelconewsletter.files.wordpress.com/2010/04/advanced-analytics.pdf

Rausand, M., and A. Hoyland. 2004. *System Reliability Theory: Models, Statistical Methods, and Applications*. Hoboken, NJ: John Wiley and Sons.

Resilience Alliance. 2007. "Assessing Resilience in Social-Ecological Systems: A Scientist's Workbook." http://www.resalliance.org/3871.php

Rosvall, M., and C.T. Bergstrom. 2010. "Mapping Change in Large Networks." *PLoS One* 5, no. 1, e8694.

Soares, S., T. Deutsch, S. Hanna, and P. Malik. 2012. Big Data Governance: A Framework to Assess Maturity. *IBM Data Magazine*. Retrieved on February 1, 2014 from http://ibmdatamag.com/2012/04/big-data-governance-a-framework-to-assess-maturity/

Scott, J. 2000. *Social Network Analysis: A Handbook*. London: Sage Publications.

Tufte, E.R. 2001. *The Visual Display of Quantitative Information*. 2nd ed. Cheshire, CT: Graphics Press.

Walter, C. 2005. “Insights: Kryder’s Law.” *Scientific American*. Retrieved November 3, 2013 from http://www.scientificamerican.com/article.cfm?id=kryders-law

Webb, J., and T. O’Brien. 2013. Big Data Now. Sebastopol, CA: O’Reilly Media.

Weld, W.E. 1959. *How to Chart: Facts from Figures with Graphs*. Norwood, MA: Codex Book Company, Inc.

Welling, M. 2011. “A First Encounter with Machine Learning.” https://www.ics.uci.edu/~welling/teaching/ICS273Afall11/IntroMLBook.pdf

Wills, G. 2012. *Visualizing Time*. New York: Springer-Verlag.

Glossary

Advanced analytic: a set of analytics integrated through an analytic architecture to solve a complex problem.

Analytics: the process of transforming data into insight for the purpose of making decisions.

Analytic architecture: a software architecture or application framework designed to solve a set of problems within a complex domain.

Application programming interface (API): a set of functions, procedures, or methods implemented in an application, which allows programs to invoke the functionality therein.

Big Data: the volume of data just beyond our capacity to process it efficiently by traditional data base tools and methods.

Business analytics: the application of analytics specifically in the sphere of business, for example, to include marketing analytics, CRM analytics, operations analytics, and so on.

Data mining: a term often used interchangeably with analytics, but, in fact, is a subset of the realm of analytics. Data mining is usually based on statistical methods, but includes variants such as text mining, web mining, and so on.

Geospatial analytics: the integration of geographic and demographic information into the analysis process.

HTML: HypertText Markup language is a specification for describing web pages in a structured way that facilitates their presentation independent of the web browser.

Machine learning: a discipline in which a program is designed in such a way that it can improve its behavior by processing data, examining the results, and deducing or inferring changes to program parameters that result in improvement in some figure of merit.

Platform: A hardware or software system for performing computations, processing data, storing data and information, and visualizing data and information. An example of a hardware platform is a Dell or Apple laptop or server, while an example of a software platform is Apple's MacOS or Microsoft's Windows 2013 Server operating system.

Population imbalance: in very large data sets, events of interest occur relatively infrequently.

RTBDA: real-time Big Data analytics.

Visual analytics: the science of analytical reasoning facilitated by interactive visual interfaces.

Web service: a method of communication between two computing systems linked through a network, but, typically, a software system supporting machine-to-machine interaction.

XML: E**x**tensible **M**arkup **L**anguage is a specification for describing the fields and values of a document in a structured form.

Index

OTHER TITLES IN OUR SERVICE SYSTEMS AND INNOVATIONS IN BUSINESS AND SOCIETY COLLECTION

Jim Spohrer, IBM and Haluk Demirkan, Arizona State University, Editors

- *Business Engineering and Service Design with Applications for Health Care Institutions* by Oscar Barros
- *Achieving Service Excellence: Maximizing Enterprise Performance Through Innovation and Technology* by Carl M. Chang
- *Service and Service Systems: Provider Challenges and Directions in Unsettled Times* by Steve Baron, Philip Hunter-Jones, and Gary Warnaby
- *Service Thinking: The Seven Principles to Discover Innovative Opportunities* by Hunter Hastings and Jeff Saperstein
- *Profiting From Services and Solutions: What Product-Centric Firms Need to Know* by Valarie A. Zeithaml, Stephen W. Brown and Mary Jo Bitner
- *People, Processes, Services, and Things: Using Services Innovation to Enable the Internet of Everything* by Hazim Dahir, Bil Dry, and Carlos Pignataro
- *Service Design and Delivery: How Design Thinking Can Innovate Business and Add Value to Society* by Toshiaki Kurokawa
- *All Services, All the Time: How Business Services Serve Your Business* by Doug McDavid
- *Modeling Service Systems* by Ralph D. Badinelli

Obtaining Value from Big Data for Service Delivery

Stephen H. Kaisler • Frank Armour • J. Alberto Espinosa • William H. Money

Big data is an emerging phenomenon that has enormous implications and impacts upon business strategy, profitability, and process improvements. All service systems generate big data these days, especially human-centered service systems. It has been characterized as the collection, analysis and use of data characterized by the five Vs: volume, velocity, variety, veracity, and value (of data).

This booklet will help middle, senior, and executive managers to understand what big data is; how to recognize, collect, process, and analyze it; how to store and manage it; how to obtain useful information from it; and how to assess its contribution to operational, tactical, and strategic decision-making in service-oriented organizations.

Dr. Stephen H. Kaisler is a senior scientist at a small business focused on machine learning, big data and advanced analytics, natural language processing, and enterprise architecture. He earned a DSc in computer science at George Washington where he is an adjunct professor of engineering in computer. He is co-chair for the Enterprise Architecture minitrack and primary co-chair of the Big Data and Analytics minitrack at HICSS.

Dr. Frank Armour is an assistant professor of information technology at the Kogod School of Business, American University and program director for the MS in analytics degree. He earned a PhD at the Volgenau School of Engineering at George Mason University. He is also an independent senior IT consultant.

Dr. J. Alberto Espinosa is a professor information technology at American University. He earned PhD in information systems from Carnegie Mellon University and his work is published in leading journals, including *Management Science* and the *Journal of Management Information Systems*. He has 20 years of experience designing and implementing systems worldwide including two masters programs on analytics at American University.

Dr. William H. Money is an associate professor at the School of Business Administration, The Citadel. He earned his PhD in organizational behavior/systems engineering at Northwestern University. His publications and recent research interests focus on collaborative solutions to complex business problems; business process engineering and analytics; and information system development, collaboration, and workflow tools and methodologies.

Service Systems and Innovations in Business and Society Collection

Jim Spohrer and Haluk Demirkan, *Editors*

ISBN: 978-1-63157-222-7
90000
9 781631 572227

BEP BUSINESS EXPERT PRESS