CHAPTER

13

Big Data and Analytics

LEARNING OBJECTIVES

- Learn what Big Data is and how it is changing the world of analytics
- Understand the motivation for and business drivers of Big Data analytics
- Become familiar with the wide range of enabling technologies for Big Data analytics
- Learn about Hadoop, MapReduce, and NoSQL as they relate to Big Data analytics
- Understand the role of and capabilities/skills for data scientist as a new analytics profession
- Compare and contrast the complementary uses of data warehousing and Big Data
- Become familiar with the vendors of Big Data tools and services
- Understand the need for and appreciate the capabilities of stream analytics
- Learn about the applications of stream analytics

Big Data, which means many things to many people, is not a new technological fad. It is a business priority that has the potential to profoundly change the competitive landscape in today's globally integrated economy. In addition to providing innovative solutions to enduring business challenges, Big Data and analytics instigate new ways to transform processes, organizations, entire industries, and even society all together. Yet extensive media coverage makes it hard to distinguish hype from reality. This chapter aims to provide a comprehensive coverage of Big Data, its enabling technologies, and related analytics concepts to help understand the capabilities and limitations of this emerging paradigm. The chapter starts with the definition and related concepts of Big Data, followed by the technical details of the enabling technologies, including Hadoop, MapReduce, and NoSQL. After describing "data scientist" as a new, fashionable organizational role/job, we provide a comparative analysis between data warehousing and Big

Data analytics. The last part of the chapter is dedicated to stream analytics, which is one of the most promising value propositions of Big Data analytics. This chapter contains the following sections:

13.1 OPENING VIGNETTE: Big Data Meets Big Science at CERN

The European Organization for Nuclear Research, known as CERN (which is derived from the acronym for the French "Conseil Européen pour la Recherche Nucléaire"), is playing a leading role in fundamental studies of physics. It has been instrumental in many key global innovations and breakthrough discoveries in theoretical physics and today operates the world's largest particle physics laboratory, home to the Large Hadron Collider (LHC) nestled under the mountains between Switzerland and France. Founded in 1954, CERN, one of Europe's first joint ventures, now has 20 member European states. At the beginning, their research primarily concentrated on understanding the inside of the atom, hence, the word "nuclear" in its name.

At CERN physicists and engineers are probing the fundamental structure of the universe. They use the world's largest and the most sophisticated scientific instruments to study the basic constituents of matter—the fundamental particles. These instruments include purpose-built particle accelerators and detectors. Accelerators boost the beams of particles to very high energies before the beams are forced to collide with each other or with stationary targets. Detectors observe and record the results of these collisions, which are happening at or near the speed of light. This process provides the physicists with clues about how the particles interact, and provides insights into the fundamental laws of nature. The LHC and its various experiments have received media attention following the discovery of a new particle strongly suspected to be the elusive Higgs Boson—an elementary particle initially theorized in 1964 and tentatively confirmed at CERN on March 14, 2013. This discovery has been called "monumental" because it appears to confirm the existence of the Higgs field, which is pivotal to the major theories within particle physics.

THE DATA CHALLENGE

Forty million times per second, particles collide within the LHC, each collision generating particles that often decay in complex ways into even more particles. Precise electronic circuits all around LHC record the passage of each particle via a detector as a series of electronic signals, and send the data to the CERN Data Centre (DC) for recording and digital reconstruction. The digitized summary of data is recorded as a "collision event." Physicists must sift through the 15 petabytes or so of digitized summary data produced annually to determine if the collisions have thrown up any interesting physics. Despite

the state-of-the-art instrumentation and computing infrastructure, CERN does not have the capacity to process all of the data that it generates, and therefore relies on numerous other research centers all around the world to access and process the data.

The Compact Muon Solenoid (CMS) is one of the two general-purpose particle physics detectors operated at the LHC. It is designed to explore the frontiers of physics and provide physicists with the ability to look at the conditions presented in the early stages of our universe. More than 3,000 physicists from 183 institutions representing 38 countries are involved in the design, construction, and maintenance of the experiments. An experiment of this magnitude requires an enormously complex distributed computing and data management system. CMS spans more than a hundred data centers in a three-tier model and generates around 10 petabytes (PB) of summary data each year in real data, simulated data, and metadata. This information is stored and retrieved from relational and nonrelational data sources, such as relational databases, document databases, blogs, wikis, file systems, and customized applications.

At this scale, the information discovery within a heterogeneous, distributed environment becomes an important ingredient of successful data analysis. The data and associated metadata are produced in variety of forms and digital formats. Users (within CERN and scientists all around the world) want to be able to query different services (at dispersed data servers and at different locations) and combine data/information from these varied sources. However, this vast and complex collection of data means they don't necessarily know where to find the right information or have the domain knowledge to extract and merge/combine this data.

SOLUTION

To overcome this Big Data hurdle, CMS's data management and workflow management (DMWM) created the Data Aggregation System (DAS), built on MongoDB (a Big Data management infrastructure) to provide the ability to search and aggregate information across this complex data landscape. Data and metadata for CMS come from many different sources and are distributed in a variety of digital formats. It is organized and managed by constantly evolving software using both relational and nonrelational data sources. The DAS provides a layer on top of the existing data sources that allows researchers and other staff to query data via free text-based queries, and then aggregates the results from across distributed providers—while preserving their integrity, security policy, and data formats. The DAS then represents that data in defined format.

"The choice of an existing relational database was ruled out for several reasons—namely, we didn't require any transactions and data persistency in DAS, and as such can't have a pre-defined schema. Also the dynamic typing of stored metadata objects was one of the requirements. Amongst other reasons, those arguments forced us to look for alternative IT solutions," explained Valentin Kuznetsov, a research associate from Cornell University who works at CMS.

"We considered a number of different options, including file-based and in-memory caches, as well as key-value databases, but ultimately decided that a document database would best suit our needs. After evaluating several applications, we chose MongoDB, due to its support of dynamic queries, full indexes, including inner objects and embedded arrays, as well as auto-sharding."

ACCESSING THE DATA VIA FREE-FORM QUERIES

All DAS queries can be expressed in a free text-based form, either as a set of keywords or key-value pairs, where a pair can represent a condition. Users can query the system using a simple, SQL-like language, which is then transformed into the MongoDB query syntax, which is itself a JSON record. "Due to the schema-less nature of the underlying

MongoDB back-end, we are able to store DAS records of any arbitrary structure, regardless of whether it's a dictionary, lists, key-value pairs, etc. Therefore, every DAS key has a set of attributes describing its JSON structure," added Kuznetsov.

DATA AGNOSTIC

Given the number of different data sources, types, and providers that DAS connects to, it is imperative that the system itself be data agnostic and allow us to query and aggregate the metadata information in customizable ways. The MongoDB architecture easily integrates with existing data services while preserving their access, security policy, and development cycles. This also provides a simple plug-and-play mechanism that makes it easy to add new data services as they are implemented and configure DAS to connect to specific domains.

CACHING FOR DATA PROVIDERS

As well as providing a way for users to easily access a wide range of data sources in a simple and consistent manner, DAS uses MongoDB as a dynamic cache, collating the information fed back from the data providers—feedback in a variety of formats and file structures. "When a user enters a query, it checks if the MongoDB database has the aggregation the user is asking for and, if it does, returns it; otherwise, the system does the aggregation and saves it to MongoDB," said Kuznetsov. "If the cache does not contain the requested query, the system contacts distributed data providers that could have this information and queries them, gathering their results. It then merges all of the results, doing a sort of 'group by' operation based on predefined identifying keys and inserts the aggregated information into the cache."

The deployment specifics are as follows:

- The CMS DAS currently runs on a single eight-core server that processes all of the queries and caches the aggregated data.
- OS: Scientific Linux
- Server hardware configuration: 8-core CPU, 40GB RAM, 1TB storage (but data set usually around 50–100GB)
- Application Language: Python
- Other database technologies: Aggregates data from a number of different databases including Oracle, PostGreSQL, CouchDB, and MySQL

RESULTS

"DAS is used 24 hours a day, seven days a week, by CMS physicists, data operators, and data managers at research facilities around the world. The average query may resolve into thousands of documents, each a few kilobytes in size. The performance of MongoDB has been outstanding, with a throughput of around 6,000 documents a second for raw cache population," concluded Kuznetsov. "The ability to offer a free text query system that is fast and scalable, with a highly dynamic and scalable cache that is data agnostic, provides an invaluable two-way translation mechanism. DAS helps CMS users to easily find and discover information they need in their research, and it represents one of the many tools that physicists use on a daily basis toward great discoveries. Without help from DAS, information lookup would have taken orders of magnitude longer." As the data collected by the various experiments grows, CMS is looking into horizontally scaling the system with sharding (i.e., distributing a single, logical database system across a cluster of machines) to meet demand. Similarly the team are spreading the word beyond CMS and out to other parts of CERN.

QUESTIONS FOR THE OPENING VIGNETTE

1. What is CERN? Why is it important to the world of science?
2. How does Large Hadron Collider work? What does it produce?
3. What is essence of the data challenge at CERN? How significant is it?
4. What was the solution? How did Big Data address the challenges?
5. What were the results? Do you think the current solution is sufficient?

WHAT WE CAN LEARN FROM THIS VIGNETTE

Big Data is big, and much more. Thanks largely to the technological advances, it is easier to create, capture, store, and analyze very large quantities of data. Most of the Big Data is generated automatically by machines. The opening vignette is an excellent example to this testament. As we have seen, LHC at CERN creates very large volumes of data very fast. The Big Data comes in varied formats and is stored in distributed server systems. Analysis of such a data landscape requires new analytical tools and techniques. Regardless of its size, complexity, and velocity, data need to be made easy to access, query, and analyze if promised value is to be derived from it. CERN uses Big Data technologies to make it easy to analyze vast amount of data created by LHC to scientists all over the world, so that the promise of understanding the fundamental building blocks of the universe is realized. As organizations like CERN hypothesize new means to leverage the value of Big Data, they will continue to invent newer technologies to create and capture even Bigger Data.

Sources: Compiled from N. Heath, "Cern: Where the Big Bang Meets Big Data," TechRepublic, 2012, **techrepublic.com/blog/european-technology/cern-where-the-big-bang-meets-big-data/636** (accessed February 2013); **home.web.cern.ch/about/computing;** and 10gen Customer Case Study, "Big Data at the CERN Project," **10gen.com/customers/cern-cms** (accessed March 2013).

13.2 DEFINITION OF BIG DATA

Using data to understand customers/clients and business operations to sustain (and foster) growth and profitability is an increasingly more challenging task for today's enterprises. As more and more data becomes available in various forms and fashions, timely processing of the data with traditional means becomes impractical. This phenomenon is nowadays called Big Data, which is receiving substantial press coverage and drawing increasing interest from both business users and IT professionals. The result is that Big Data is becoming an overhyped and overused marketing buzzword.

Big Data means different things to people with different backgrounds and interests. Traditionally, the term "Big Data" has been used to describe the massive volumes of data analyzed by huge organizations like Google or research science projects at NASA. But for most businesses, it's a relative term: "Big" depends on an organization's size. The point is more about finding new value within and outside conventional data sources. Pushing the boundaries of data analytics uncovers new insights and opportunities, and "big" depends on where you start and how you proceed. Consider the popular description of Big Data: Big Data exceeds the reach of commonly used hardware environments and/or capabilities of software tools to capture, manage, and process it within a tolerable time span for its user population. Big Data has become a popular term to describe the exponential growth, availability, and use of information, both structured and unstructured. Much has been written on the Big Data trend and how it can serve as the basis for innovation, differentiation, and growth.

Where does the Big Data come from? A simple answer is "everywhere." The sources of data that were ignored because of technical limitations are now being treated like gold mines. Big Data may come from Web logs, RFID, GPS systems, sensor networks, social networks, Internet-based text documents, Internet search indexes, detailed call records, astronomy, atmospheric science, biological, genomics, nuclear physics, biochemical experiments, medical records, scientific research, military surveillance, photography archives, video archives, and large-scale ecommerce practices.

Big Data is not new. What is new is that the definition and the structure of Big Data constantly change. Companies have been storing and analyzing large volumes of data since the advent of the data warehouses in the early 1990s. While terabytes used to be synonymous with Big Data warehouses, now it's petabytes, and the rate of growth in data volumes continues to escalate as organizations seek to store and analyze greater levels of transaction details, as well as Web- and machine-generated data, to gain a better understanding of customer behavior and business drivers.

Many (academics and industry analysts/leaders alike) think that "Big Data" is a misnomer. What it says and what it means are not exactly the same. That is, Big Data is not just "big." The sheer volume of the data is only one of many characteristics that are often associated with Big Data, such as variety, velocity, veracity, variability, and value proposition, among others.

The Vs That Define Big Data

Big Data is typically defined by three "V"s: volume, variety, velocity. In addition to these three, we see some of the leading Big Data solution providers adding other Vs, such as veracity (IBM), variability (SAS), and value proposition.

VOLUME Volume is obviously the most common trait of Big Data. Many factors contributed to the exponential increase in data volume, such as transaction-based data stored through the years, text data constantly streaming in from social media, increasing amounts of sensor data being collected, automatically generated RFID and GPS data, and so forth. In the past, excessive data volume created storage issues, both technical and financial. But with today's advanced technologies coupled with decreasing storage costs, these issues are no longer significant; instead, other issues emerge, including how to determine relevance amidst the large volumes of data and how to create value from data that is deemed to be relevant.

As mentioned before, big is a relative term. It changes over time and is perceived differently by different organizations. With the staggering increase in data volume, even the naming of the next Big Data echelon has been a challenge. The highest mass of data that used to be called petabytes (PB) has left its place to zettabytes (ZB), which is a trillion gigabytes (GB) or a billion terabytes (TB). Technology Insights 13.1 provides an overview of the size and naming of Big Data volumes.

TECHNOLOGY INSIGHTS 13.1 The Data Size Is Getting Big, Bigger, and Bigger

The measure of data size is having a hard time keeping up with new names. We all know kilobyte (KB, which is 1,000 bytes), megabyte (MB, which is 1,000,000 bytes), gigabyte (GB, which is 1,000,000,000 bytes), and terabyte (TB, which is 1,000,000,000,000 bytes). Beyond that, the names given to data sizes are relatively new to most of us. The following table shows what comes after terabyte and beyond.

Name	Symbol	Value
Kilobyte	kB	10^3
Megabyte	MB	10^6
Gigabyte	GB	10^9
Terabyte	TB	10^{12}
Petabyte	PB	10^{15}
Exabyte	EB	10^{18}
Zettabyte	ZB	10^{21}
Yottabyte	YB	10^{24}
Brontobyte*	BB	10^{27}
Gegobyte*	GeB	10^{30}

*Not an official SI (International System of Units) name/symbol, yet.

Consider that an exabyte of data is created on the Internet each day, which equates to 250 million DVDs' worth of information. And the idea of even larger amounts of data—a zettabyte—isn't too far off when it comes to the amount of info traversing the Web in any one year. In fact, industry experts are already estimating that we will see a 1.3 zettabytes of traffic annually over the Internet by 2016—and soon enough, we might start talking about even bigger volumes. When referring to yottabytes, some of the Big Data scientists often wonder about how much data the NSA or FBI have on people altogether. Put in terms of DVDs, a yottabyte would require 250 trillion of them. A brontobyte, which is not an official SI prefix but is apparently recognized by some people in the measurement community, is a 1 followed by 27 zeros. Size of such magnitude can be used to describe the amount of sensor data that we will get from the Internet in the next decade, if not sooner. A gegobyte is 10 to the power of 30. With respect to where the Big Data comes from, consider the following:

- The CERN Large Hadron Collider generates 1 petabyte per second.
- Sensors from a Boeing jet engine create 20 terabytes of data every hour.
- 500 terabytes of new data per day are ingested in Facebook databases.
- On YouTube, 72 hours of video are uploaded per minute, translating to a terabyte every 4 minutes.
- The proposed Square Kilometer Array telescope (the world's proposed biggest telescope) will generate an exabyte of data per day.

Sources: S. Higginbotham, "As Data Gets Bigger, What Comes After a Yottabyte?" 2012, **gigaom.com/2012/10/30/as-data-gets-bigger-what-comes-after-a-yottabyte** (accessed March 2013); and **en.wikipedia.org/wiki/Petabyte** (accessed March 2013).

From a short historical perspective, in 2009 the world had about 0.8ZB of data; in 2010, it exceeded the 1ZB mark; at the end of 2011, the number was 1.8ZB. Six or seven years from now, the number is estimated to be 35ZB (IBM, 2013). Though this number is astonishing in size, so are the challenges and opportunities that come with it.

VARIETY Data today comes in all types of formats—ranging from traditional databases to hierarchical data stores created by the end users and OLAP systems, to text documents, e-mail, XML, meter-collected, sensor-captured data, to video, audio, and stock ticker data. By some estimates, 80 to 85 percent of all organizations' data is in some sort of unstructured or semistructured format (a format that is not suitable for traditional

database schemas). But there is no denying its value, and hence it must be included in the analyses to support decision making.

VELOCITY According to Gartner, velocity means both how fast data is being produced and how fast the data must be processed (i.e., captured, stored, and analyzed) to meet the need or demand. **RFID** tags, automated sensors, GPS devices, and smart meters are driving an increasing need to deal with torrents of data in near–real time. Velocity is perhaps the most overlooked characteristic of Big Data. Reacting quickly enough to deal with velocity is a challenge to most organizations. For the time-sensitive environments, the opportunity cost clock of the data starts ticking the moment the data is created. As the time passes, the value proposition of the data degrades, and eventually becomes worthless. Whether the subject matter is the health of a patient, the well-being of a traffic system, or the health of an investment portfolio, accessing the data and reacting faster to the circumstances will always create more advantageous outcomes.

In the Big Data storm that we are witnessing now, almost everyone is fixated on at-rest analytics, using optimized software and hardware systems to mine large quantities of variant data sources. Although this is critically important and highly valuable, there is another class of analytics driven from the velocity nature of Big Data, called "data stream analytics" or "in-motion analytics," which is mostly overlooked. If done correctly, data stream analytics can be as valuable, and in some business environments more valuable, than at-rest analytics. Later in this chapter we will cover this topic in more detail.

VERACITY Veracity is a term that is being used as the fourth "V" to describe Big Data by IBM. It refers to the conformity to facts: accuracy, quality, truthfulness, or trustworthiness of the data. Tools and techniques are often used to handle Big Data's veracity by transforming the data into quality and trustworthy insights.

VARIABILITY In addition to the increasing velocities and varieties of data, data flows can be highly inconsistent, with periodic peaks. Is something big trending in the social media? Perhaps there is a high-profile IPO looming. Maybe swimming with pigs in the Bahamas is suddenly the must-do vacation activity. Daily, seasonal, and event-triggered peak data loads can be challenging to manage—especially with social media involved.

VALUE PROPOSITION The excitement around Big Data is its value proposition. A preconceived notion about "big" data is that it contains (or has a greater potential to contain) more patterns and interesting anomalies than "small" data. Thus, by analyzing large and feature rich data, organizations can gain greater business value that they may not have otherwise. While users can detect the patterns in small data sets using simple statistical and machine-learning methods or ad hoc query and reporting tools, Big Data means "big" analytics. Big analytics means greater insight and better decisions, something that every organization needs nowadays.

Since the exact definition of Big Data is still a matter of ongoing discussion in academic and industrial circles, it is likely that more characteristics (perhaps more Vs) are likely to be added to this list. Regardless of what happens, the importance and value proposition of Big Data are here to stay. Figure 13.1 shows a conceptual architecture where big data (at the left side of the figure) is converted to business insight through the use of a combination of advanced analytics and delivered to a variety of different users/roles for faster/better decision making.

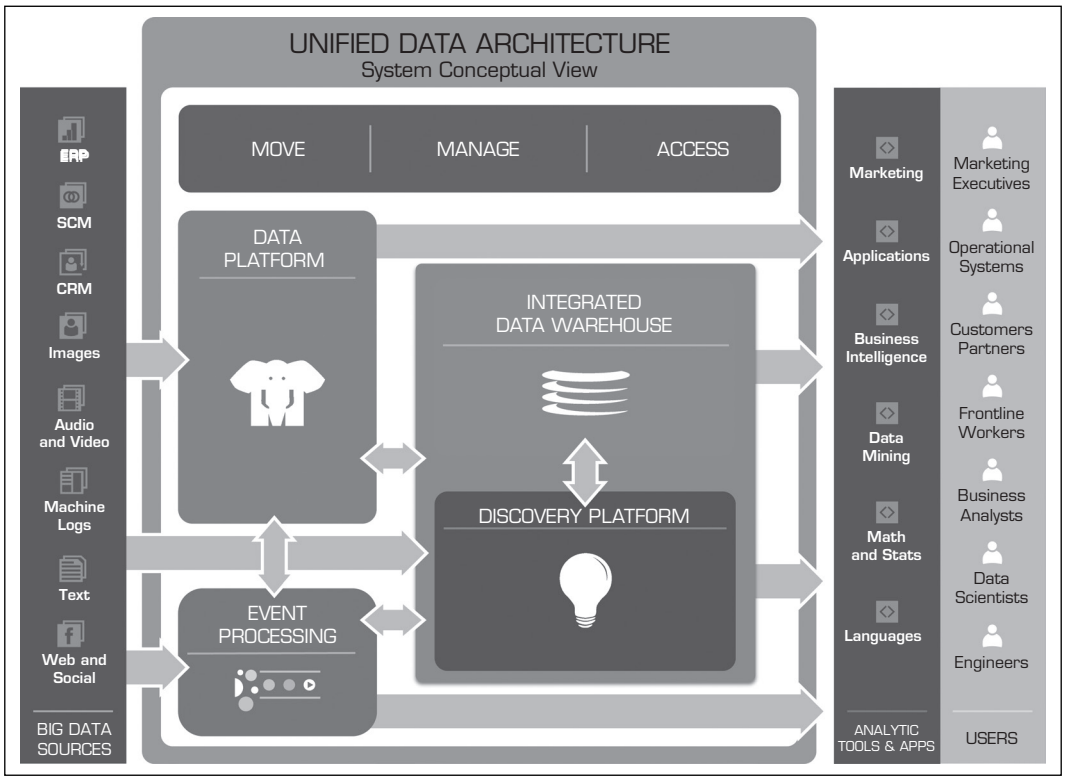

FIGURE 13.1 A High-Level Conceptual Architecture for Big Data Solutions. (*Source:* AsterData—a Teradata Company)

Application Case 13.1 shows the creative use of Big Data analytics in the ever-so-popular social media industry.

Application Case 13.1

Big Data Analytics Helps Luxottica Improvement Its Marketing Effectiveness

Based in Mason, Ohio, Luxottica Retail North America (Luxottica) is a wholly owned retail arm of Milan-based Luxottica Group S.p.A, the world's largest designer, manufacturer, distributer and seller of luxury and sports eyewear. Employing more than 65,000 people worldwide, the company reported net sales of EUR6.2 billion in 2011.

Problem - Disconnected Customer Data

Nearly 100 million customers purchase eight house brands from Luxottica through the company's numerous websites and retail chain stores. The big data captured from those customer interactions (in the form of transactions, click streams, product reviews, and social media postings) constitutes a massive source of business intelligence for potential product, marketing, and sales opportunities.

Luxottica, however, outsourced both data storage and promotional campaign development and management, leading to a disconnect between data analytics and marketing execution. The outsource model hampered access to current, actionable data, limiting its marketing value and the analytic value of the IBM PureData System for Analytics appliance that Luxottica used for a small segment of its business.

Luxottica's competitive posture and strategic growth initiatives were compromised for lack of an individualized view of its customers and an inability to act decisively and consistently on the different types of information generated by each retail channel. Luxottica needed to be able to exploit all data regardless of source or which internal or external application it resided on. Likewise, the company's marketing team wanted more control over

promotional campaigns, including the capacity to gauge campaign effectiveness.

Solution - Fine-tuned Marketing

To integrate all data from its multiple internal and external application sources and gain visibility into its customers, Luxottica deployed the Customer Intelligence Appliance (CIA) from IBM Business Partner Aginity LLC.

CIA is an integrated set of adaptable software, hardware, and embedded analytics built on the IBM PureData System for Analytics solution. The combined technologies help Luxottica highly segment customer behavior and provide a platform and smart database for marketing execution systems, such as campaign management, e-mail services and other forms of direct marketing.

IBM® PureData™ for Analytics, which is powered by Netezza data warehousing technology, is one of the leading data appliances for large-scale, real-time analytics. Because of its innovative data storage mechanisms and massively parallel processing capabilities, it simplifies and optimizes performance of data services for analytic applications, enabling very complex algorithms to run in minutes, not hours or days, rapidly delivering invaluable insight to decision makers when they need it.

The IBM and Aginity platform provides Luxottica with unprecedented visibility into a class of customer that is of particular interest to the company: the omni-channel customer. This customer purchases merchandise both online and in-store and tends to shop and spend more than web-only or in-store customers.

"We've equipped their team with tools to gain a 360-degree view of their most profitable sales channel, the omni-channel customers, and individualize the way they market to them," says Ted Westerheide, chief architect for Aginity. "With the Customer Intelligence Appliance and PureData System for Analytics platform, Luxottica is a learning organization, connecting to customer data across multiple channels and improving marketing initiatives from campaign to campaign."

Benefits

Successful implementation of such an advanced big data analytics solution brings about numerous business benefits. In the case of Luxottica, the top three benefits were:

- Anticipates a 10 percent improvement in marketing effectiveness
- Identifies the highest-value customers out of nearly 100 million
- Targets individual customers based on unique preferences and histories

Questions for Discussion

1. What does Big Data mean to Luxottica?
2. What were their main challenges?
3. What was the proposed solution, and the obtained results?

Source: IBM Customer Case, "Luxottica anticipates 10 percent improvement in marketing effectiveness" **http://www-01.ibm.com/software/success/cssdb.nsf/CS/KPES-9BNNKV?OpenDocument&Site=default&cty=en_us** (accessed October 2013).

SECTION 13.2 REVIEW QUESTIONS

1. Why is Big Data important? What has changed to put it in the center of the analytics world?
2. How do you define Big Data? Why is it difficult to define?
3. Out of the Vs that are used to define Big Data, in your opinion, which one is the most important? Why?
4. What do you think the future of Big Data will be like? Will it leave its popularity to something else? If so, what will it be?

13.3 FUNDAMENTALS OF BIG DATA ANALYTICS

Big Data by itself, regardless of the size, type, or speed, is worthless unless business users do something with it that delivers value to their organizations. That's where "big" analytics comes into the picture. Although organizations have always run reports and

dashboards against data warehouses, most have not opened these repositories to in-depth on-demand exploration. This is partly because analysis tools are too complex for the average user but also because the repositories often do not contain all the data needed by the power user. But this is about to change (and had already changed for some) in a dramatic fashion, thanks to the new Big Data analytics paradigm.

With the value proposition, Big Data also brought about big challenges for organizations. The traditional means for capturing, storing, and analyzing data are not capable of dealing with Big Data effectively and efficiently. Therefore, new breeds of technologies need to be developed (or purchased/hired/outsourced) to take on the Big Data challenge. Before making such an investment, organizations should justify the means. Here are some questions that may help shed light on this situation. If any of the following statements are true, then you need to seriously consider embarking on a Big Data journey.

- You can't process the amount of data that you want to because of the limitations posed by your current platform or environment.
- You want to involve new/contemporary data sources (e.g., social media, RFID, sensory, Web, GPS, textual data) into your analytics platform, but you can't because it does not comply with the data storage schema-defined rows and columns without sacrificing fidelity or the richness of the new data.
- You need to (or want to) integrate data as quickly as possible to be current on your analysis.
- You want to work with a schema-on-demand (as opposed to the predetermined schema used in RDBMS) data storage paradigm because the nature of the new data may not be known, or there may not be enough time to determine it and develop a schema for it.
- The data is arriving so fast at your organization's doorstep that your traditional analytics platform cannot handle it.

As is the case with any other large IT investment, the success in **Big Data analytics** depends on a number of factors. Figure 13.2 shows a graphical depiction of the most critical success factors (Watson 2012).

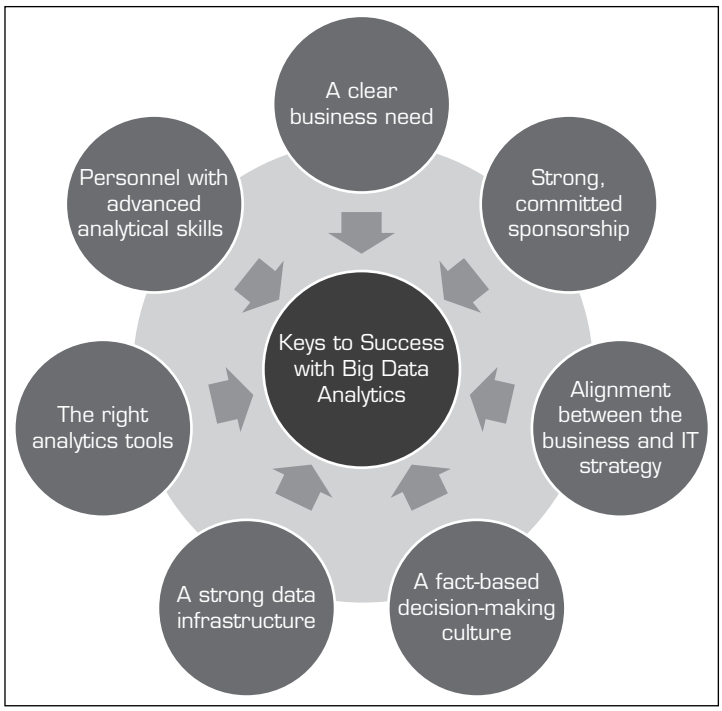

FIGURE 13.2 Critical Success Factors for Big Data Analytics. (*Source:* AsterData—a Teradata Company)

Following are the most critical success factors for Big Data analytics (Watson et al., 2012):

1. ***A clear business need (alignment with the vision and the strategy)***. Business investments ought to be made for the good of the business, not for the sake of mere technology advancements. Therefore the main driver for Big Data analytics should be the needs of the business at any level—strategic, tactical, and operations.
2. ***Strong, committed sponsorship (executive champion).*** It is a well-known fact that if you don't have strong, committed executive sponsorship, it is difficult (if not impossible) to succeed. If the scope is a single or a few analytical applications, the sponsorship can be at the departmental level. However, if the target is enterprise-wide organizational transformation, which is often the case for Big Data initiatives, sponsorship needs to be at the highest levels and organization-wide.
3. ***Alignment between the business and IT strategy.*** It is essential to make sure that the analytics work is always supporting the business strategy, and not other way around. Analytics should play the enabling role in successful execution of the business strategy.
4. ***A fact-based decision making culture.*** In a fact-based decision-making culture, the numbers rather than intuition, gut feeling, or supposition drive decision making. There is also a culture of experimentation to see what works and doesn't. To create a fact-based decision-making culture, senior management needs to:
   - Recognize that some people can't or won't adjust
   - Be a vocal supporter
   - Stress that outdated methods must be discontinued
   - Ask to see what analytics went into decisions
   - Link incentives and compensation to desired behaviors
5. ***A strong data infrastructure.*** Data warehouses have provided the data infrastructure for analytics. This infrastructure is changing and being enhanced in the Big Data era with new technologies. Success requires marrying the old with the new for a holistic infrastructure that works synergistically.

As the size and the complexity increase, the need for more efficient analytical systems is also increasing. In order to keep up with the computational needs of Big Data, a number of new and innovative computational techniques and platforms have been developed. These techniques are collectively called *high-performance computing,* which includes the following:

- ***In-memory analytics:*** Solves complex problems in near–real time with highly accurate insights by allowing analytical computations and Big Data to be processed in-memory and distributed across a dedicated set of nodes.
- ***In-database analytics:*** Speeds time to insights and enables better data governance by performing data integration and analytic functions inside the database so you won't have to move or convert data repeatedly.
- ***Grid computing:*** Promotes efficiency, lower cost, and better performance by processing jobs in a shared, centrally managed pool of IT resources.
- ***Appliances:*** Bringing together hardware and software in a physical unit that is not only fast but also scalable on an as-needed basis.

Computational requirement is just a small part of the list of challenges that Big Data imposes upon today's enterprises. Following is a list of challenges that are found by business executives to have a significant impact on successful implementation of Big Data analytics. When considering Big Data projects and architecture, being mindful of these challenges could make the journey to analytics competency a less stressful one.

- ***Data volume:*** The ability to capture, store, and process the huge volume of data at an acceptable speed so that the latest information is available to decision makers when they need it.
- ***Data integration:*** The ability to combine data that is not similar in structure or source and to do so quickly and at reasonable cost.
- ***Processing capabilities:*** The ability to process the data quickly, as it is captured. The traditional way of collecting and then processing the data may not work. In many situations data needs to be analyzed as soon as it is captured to leverage the most value (this is called *stream analytics*, which will be covered later in this chapter).
- ***Data governance:*** The ability to keep up with the security, privacy, ownership, and quality issues of Big Data. As the volume, variety (format and source), and velocity of data change, so should the capabilities of governance practices.
- ***Skills availability:*** Big Data is being harnessed with new tools and is being looked at in different ways. There is a shortage of people (often called data scientists, covered later in this chapter) with the skills to do the job.
- ***Solution cost:*** Since Big Data has opened up a world of possible business improvements, there is a great deal of experimentation and discovery taking place to determine the patterns that matter and the insights that turn to value. To ensure a positive ROI on a Big Data project, therefore, it is crucial to reduce the cost of the solutions used to find that value.

Though challenges are real, so is the value proposition of Big Data analytics. Anything that you can do as business analytics leaders to help prove the value of new data sources to the business will move your organization beyond experimenting and exploring Big Data into adapting and embracing it as a differentiator. There is nothing wrong with exploration, but ultimately the value comes from putting those insights into action.

Business Problems Addressed by Big Data Analytics

The top business problems addressed by Big Data overall are process efficiency and cost reduction as well as enhancing customer experience, but different priorities emerge when it is looked at by industry. Process efficiency and cost reduction are common business problems that can be addressed by analyzing Big Data, which are perhaps among the top-ranked problems that can be addressed with Big Data analytics for the manufacturing, government, energy and utilities, communications and media, transport, and healthcare sectors. Enhanced customer experience may be at the top of the list of problems addressed by insurance companies and retailers. Risk management usually is at the top of the list for companies in banking and education. Here is a list of problems that can be addressed using Big Data analytics:

- Process efficiency and cost reduction
- Brand management
- Revenue maximization, cross-selling, and up-selling
- Enhanced customer experience
- Churn identification, customer recruiting
- Improved customer service
- Identifying new products and market opportunities
- Risk management
- Regulatory compliance
- Enhanced security capabilities

Application Case 13.2 illustrates an excellent example in the banking industry, where disparate data sources are integrated into a Big Data infrastructure to achieve a single source of the truth.

Application Case 13.2

Top 5 Investment Bank Achieves Single Source of Truth

The Bank's highly respected derivatives team is responsible for over one-third of the world's total derivatives trades. Their derivatives practice has a global footprint with teams that support credit, interest rate, and equity derivatives in every region of the world. The Bank has earned numerous industry awards and is recognized for its product innovations.

Challenge

With its significant derivatives exposure the Bank's management recognized the importance of having a real-time global view of its positions. The existing system, based on a relational database, was comprised of multiple installations around the world. Due to the gradual expansions to accommodate the increasing data volume varieties, the legacy system was not fast enough to respond to growing business needs and requirements. It was unable to deliver real-time alerts to manage market and counterparty credit positions in the desired timeframe.

Solution

The Bank built a derivatives trade store based on the MarkLogic (a Big Data analytics solution provider) Server, replacing the incumbent technologies. Replacing the 20 disparate batch-processing servers with a single operational trade store enabled the Bank to know its market and credit counterparty positions in real time, providing the ability to act quickly to mitigate risk. The accuracy and completeness of the data allowed the Bank and its regulators to confidently rely on the metrics and stress test results it reports.

The selection process included upgrading existing Oracle and Sybase technology. Meeting all the new regulatory requirements was also a major factor in the decision as the Bank looked to maximize its investment. After the Bank's careful investigation, the choice was clear—only MarkLogic could meet both needs plus provide better performance, scalability, faster development for future requirements and implementation, and a much lower total cost of ownership (TCO). Figure 13.3 illustrates the transformation from the old fragmented systems to the new unified system.

Results

MarkLogic was selected because existing systems would not provide the sub-second updating and analysis response times needed to effectively

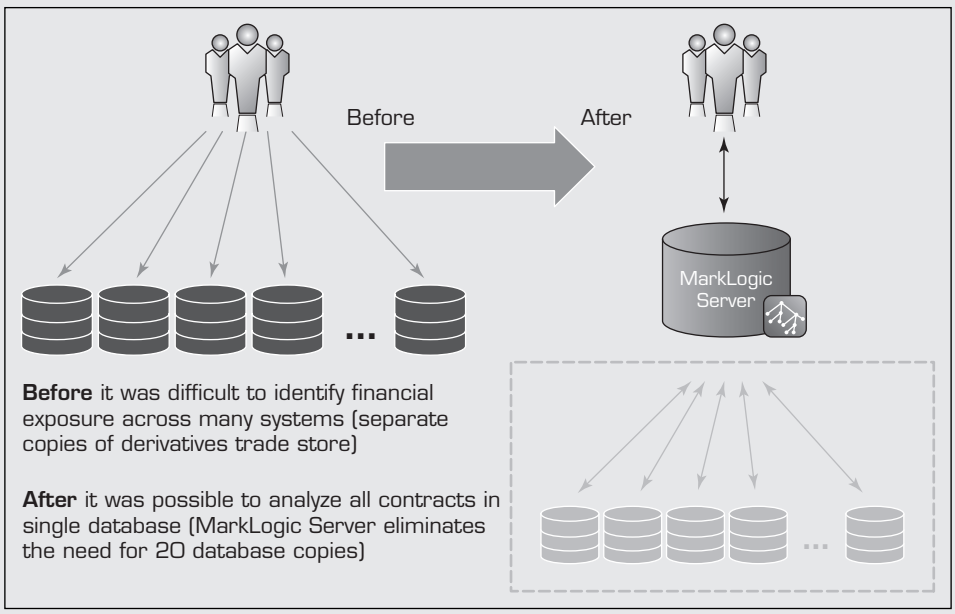

FIGURE 13.3 Moving from Many Old Systems to a Unified New System. *Source:* MarkLogic.

(*Continued*)

Application Case 13.2 (Continued)

manage a derivatives trade book that represents nearly one-third of the global market. Trade data is now aggregated accurately across the Bank's entire derivatives portfolio, allowing risk management stakeholders to know the true enterprise risk profile, to conduct predictive analyses using accurate data, and to adopt a forward-looking approach. Not only are hundreds of thousands of dollars of technology costs saved each year, but the Bank does not need to add resources to meet regulators' escalating demands for more transparency and stress-testing frequency. Here are the highlights from the obtained results:

- An alerting feature keeps users appraised of up-to-the-minute market and counterparty credit changes so they can take appropriate actions.
- Derivatives are stored and traded in a single MarkLogic system requiring no downtime for maintenance, a significant competitive advantage.
- Complex changes can be made in hours versus days, weeks, and even months needed by competitors.
- Replacing Oracle and Sybase significantly reduced operations costs: one system versus 20, one database administrator instead of up to 10, and lower costs per trade.

Next Steps

The successful implementation and performance of the new system resulted in the Bank's examination of other areas where it could extract more value from its Big Data—structured, unstructured, and/or poly-structured. Two applications are under active discussion. Its equity research business sees an opportunity to significantly boost revenue with a platform that provides real-time research, repurposing, and content delivery. The Bank also sees the power of centralizing customer data to improve onboarding, increase cross-sell opportunities, and support know your customer requirements.

Questions for Discussion

1. How can Big Data benefit large-scale trading banks?
2. How did MarkLogic infrastructure help ease the leveraging of Big Data?
3. What were the challenges, the proposed solution, and the obtained results?

Source: MarkLogic, Customer Success Story, **marklogic.com/resources/top-5-derivatives-trading-bank-achieves-single-source-of-truth** (accessed March 2013).

SECTION 13.3 REVIEW QUESTIONS

1. What is Big Data analytics? How does it differ from regular analytics?
2. What are the critical success factors for Big Data analytics?
3. What are the big challenges that one should be mindful of when considering implementation of Big Data analytics?
4. What are the common business problems addressed by Big Data analytics?

13.4 BIG DATA TECHNOLOGIES

There are a number of technologies for processing and analyzing Big Data, but most have some common characteristics (Kelly 2012). Namely, they take advantage of commodity hardware to enable scale-out, parallel processing techniques; employ nonrelational data storage capabilities in order to process unstructured and semistructured data; and apply advanced analytics and data visualization technology to Big Data to convey insights to end users. There are three Big Data technologies that stand out, that most believe will transform the business analytics and data management markets: MapReduce, Hadoop, and NoSQL.

MapReduce

MapReduce is a technique popularized by Google that distributes the processing of very large multi-structured data files across a large cluster of machines. High performance is achieved by breaking the processing into small units of work that can be run in parallel across the hundreds, potentially thousands, of nodes in the cluster. To quote the seminal paper on MapReduce:

"MapReduce is a programming model and an associated implementation for processing and generating large data sets. Programs written in this functional style are automatically parallelized and executed on a large cluster of commodity machines. This allows programmers without any experience with parallel and distributed systems to easily utilize the resources of a large distributed system" (Dean and Ghemawat, 2004).

The key point to note from this quote is that MapReduce is a programming model, not a programming language, that is, it is designed to be used by programmers, rather than business users. The easiest way to describe how MapReduce works is through the use of an example—see the geometric shape counter in Figure 13.4.

The input to the MapReduce process in Figure 13.4 is a set of geometric shapes. The objective is to count the number of geometric shapes of each type (diamond, circle, square, star, and triangle). The programmer in this example is responsible for coding the map and reducing programs; the remainder of the processing is handled by the software system implementing the MapReduce programming model.

The MapReduce system first reads the input file and splits it into multiple pieces. In this example, there are two splits, but in a real-life scenario, the number of splits would typically be much higher. These splits are then processed by multiple map programs

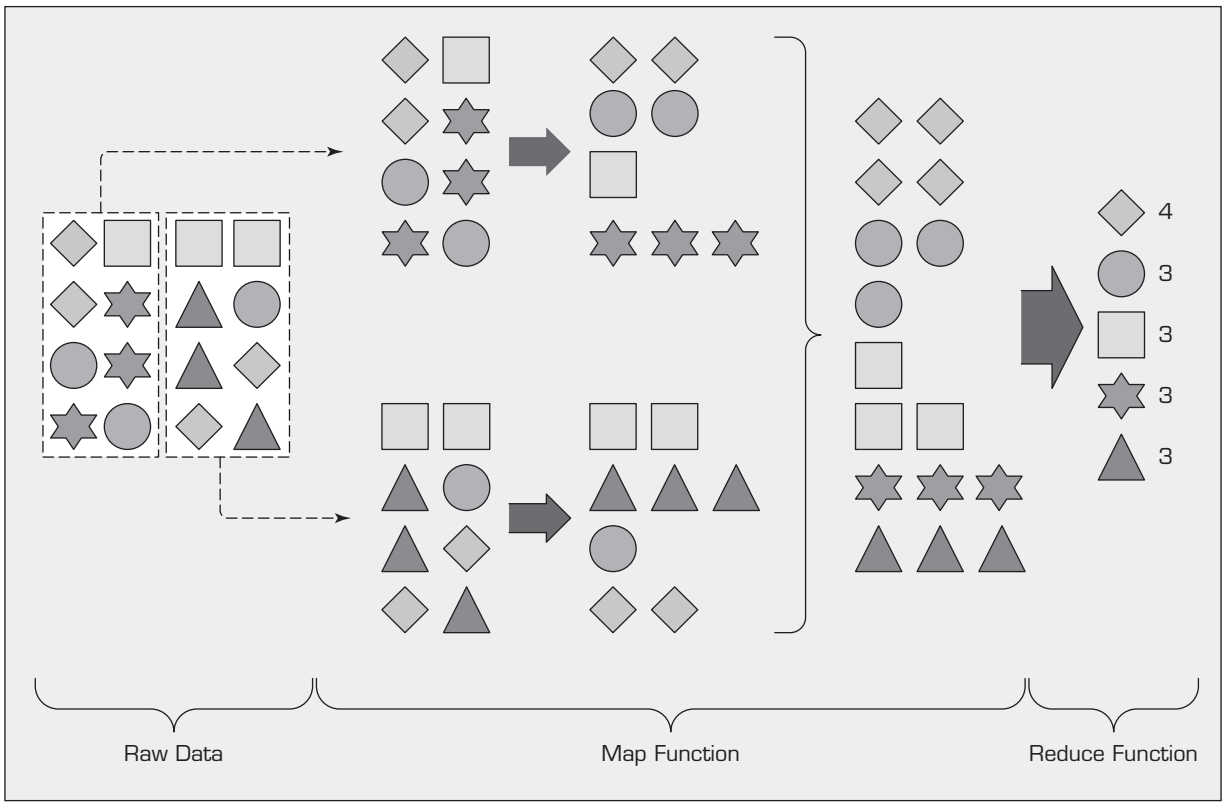

FIGURE 13.4 A Graphical Depiction of the MapReduce Process.

running in parallel on the nodes of the cluster. The role of each map program in this case is to group the data in a split by the type of geometric shape. The MapReduce system then takes the output from each map program and merges (shuffles/sorts) the results for input to reduce the program, which calculates the sum of the number of different types of geometric shapes. In this example, only one copy of the reduce program is used, but there may be more in practice. To optimize performance, programmers can provide their own shuffle/sort program and can also deploy a combiner that combines local map output files to reduce the number of output files that have to be remotely accessed across the cluster by the shuffle/sort step.

Why Use MapReduce?

MapReduce aids organizations in processing and analyzing large volumes of multi-structured data. Application examples include indexing and search, graph analysis, text analysis, machine learning, data transformation, and so forth. These types of applications are often difficult to implement using the standard SQL employed by relational DBMSs.

The procedural nature of MapReduce makes it easily understood by skilled programmers. It also has the advantage that developers do not have to be concerned with implementing parallel computing—this is handled transparently by the system. Although MapReduce is designed for programmers, non-programmers can exploit the value of prebuilt MapReduce applications and function libraries. Both commercial and open source MapReduce libraries are available that provide a wide range of analytic capabilities. Apache Mahout, for example, is an open source machine-learning library of "algorithms for clustering, classification, and batch-based collaborative filtering" that are implemented using MapReduce.

Hadoop

Source: Hadoop. Used with permission.

Hadoop is an open source framework for processing, storing, and analyzing massive amounts of distributed, unstructured data. Originally created by Doug Cutting at Yahoo!, Hadoop was inspired by MapReduce, a user-defined function developed by Google in the early 2000s for indexing the Web. It was designed to handle petabytes and exabytes of data distributed over multiple nodes in parallel. Hadoop clusters run on inexpensive commodity hardware so projects can scale-out without breaking the bank. Hadoop is now a project of the Apache Software Foundation, where hundreds of contributors continuously improve the core technology. Fundamental concept: Rather than banging away at one, huge block of data with a single machine, Hadoop breaks up Big Data into multiple parts so each part can be processed and analyzed at the same time.

How Does Hadoop Work?

A client accesses unstructured and semistructured data from sources including log files, social media feeds, and internal data stores. It breaks the data up into "parts," which are then loaded into a file system made up of multiple nodes running on commodity hardware. The default file store in Hadoop is the **Hadoop Distributed File System (HDFS)**. File systems such as HDFS are adept at storing large volumes of unstructured and semistructured data as they do not require data to be organized into relational rows and columns. Each "part" is replicated multiple times and loaded into the file system, so that if a node fails, another node has a copy of the data contained on the failed node.

A Name Node acts as facilitator, communicating back to the client information such as which nodes are available, where in the cluster certain data resides, and which nodes have failed.

Once the data is loaded into the cluster, it is ready to be analyzed via the MapReduce framework. The client submits a "Map" job—usually a query written in Java—to one of the nodes in the cluster known as the Job Tracker. The Job Tracker refers to the Name Node to determine which data it needs to access to complete the job and where in the cluster that data is located. Once determined, the Job Tracker submits the query to the relevant nodes. Rather than bringing all the data back into a central location for processing, processing then occurs at each node simultaneously, or in parallel. This is an essential characteristic of Hadoop.

When each node has finished processing its given job, it stores the results. The client initiates a "Reduce" job through the Job Tracker in which results of the map phase stored locally on individual nodes are aggregated to determine the "answer" to the original query, and then loaded onto another node in the cluster. The client accesses these results, which can then be loaded into one of a number of analytic environments for analysis. The MapReduce job has now been completed.

Once the MapReduce phase is complete, the processed data is ready for further analysis by data scientists and others with advanced data analytics skills. **Data scientists** can manipulate and analyze the data using any of a number of tools for any number of uses, including searching for hidden insights and patterns or use as the foundation for building user-facing analytic applications. The data can also be modeled and transferred from Hadoop clusters into existing relational databases, data warehouses, and other traditional IT systems for further analysis and/or to support transactional processing.

Hadoop Technical Components

A Hadoop "stack" is made up of a number of components, which include:

- ***Hadoop Distributed File System (HDFS):*** The default storage layer in any given Hadoop cluster
- ***Name Node:*** The node in a Hadoop cluster that provides the client information on where in the cluster particular data is stored and if any nodes fail
- ***Secondary Node:*** A backup to the Name Node, it periodically replicates and stores data from the Name Node should it fail
- ***Job Tracker:*** The node in a Hadoop cluster that initiates and coordinates MapReduce jobs, or the processing of the data
- ***Slave Nodes:*** The grunts of any Hadoop cluster, slave nodes store data and take direction to process it from the Job Tracker

In addition to these components, the Hadoop ecosystem is made up of a number of complimentary sub-projects. NoSQL data stores like Cassandra and HBase are also used to store the results of MapReduce jobs in Hadoop. In addition to Java, some MapReduce jobs and other Hadoop functions are written in Pig, an open source language designed specifically for Hadoop. Hive is an open source data warehouse originally developed by Facebook that allows for analytic modeling within Hadoop. Here are the most commonly referenced sub-projects for Hadoop.

HIVE **Hive** is a Hadoop-based data warehousing–like framework originally developed by Facebook. It allows users to write queries in an SQL-like language called HiveQL, which are then converted to MapReduce. This allows SQL programmers with no MapReduce experience to use the warehouse and makes it easier to integrate with business intelligence and visualization tools such as MicroStrategy, Tableau, Revolutions Analytics, and so forth.

PIG **Pig** is a Hadoop-based query language developed by Yahoo!. It is relatively easy to learn and is adept at very deep, very long data pipelines (a limitation of SQL.)

HBASE HBase is a nonrelational database that allows for low-latency, quick lookups in Hadoop. It adds transactional capabilities to Hadoop, allowing users to conduct updates, inserts, and deletes. eBay and Facebook use HBase heavily.

FLUME Flume is a framework for populating Hadoop with data. Agents are populated throughout one's IT infrastructure—inside Web servers, application servers, and mobile devices, for example—to collect data and integrate it into Hadoop.

OOZIE Oozie is a workflow processing system that lets users define a series of jobs written in multiple languages—such as Map Reduce, Pig, and Hive—and then intelligently link them to one another. Oozie allows users to specify, for example, that a particular query is only to be initiated after specified previous jobs on which it relies for data are completed.

AMBARI Ambari is a Web-based set of tools for deploying, administering, and monitoring Apache Hadoop clusters. Its development is being led by engineers from Hortonworks, which include Ambari in its Hortonworks Data Platform.

AVRO Avro is a data serialization system that allows for encoding the schema of Hadoop files. It is adept at parsing data and performing removed procedure calls.

MAHOUT Mahout is a data mining library. It takes the most popular data mining algorithms for performing clustering, regression testing, and statistical modeling and implements them using the MapReduce model.

SQOOP Sqoop is a connectivity tool for moving data from non-Hadoop data stores—such as relational databases and data warehouses—into Hadoop. It allows users to specify the target location inside of Hadoop and instructs Sqoop to move data from Oracle, Teradata, or other relational databases to the target.

HCATALOG HCatalog is a centralized metadata management and sharing service for Apache Hadoop. It allows for a unified view of all data in Hadoop clusters and allows diverse tools, including Pig and Hive, to process any data elements without needing to know physically where in the cluster the data is stored.

Hadoop: The Pros and Cons

The main benefit of Hadoop is that it allows enterprises to process and analyze large volumes of unstructured and semistructured data, heretofore inaccessible to them, in a cost- and time-effective manner. Because Hadoop clusters can scale to petabytes and even exabytes of data, enterprises no longer must rely on sample data sets but can process and analyze *all* relevant data. Data scientists can apply an iterative approach to analysis, continually refining and testing queries to uncover previously unknown insights. It is also inexpensive to get started with Hadoop. Developers can download the Apache Hadoop distribution for free and begin experimenting with Hadoop in less than a day.

The downside to Hadoop and its myriad components is that they are immature and still developing. As with any young, raw technology, implementing and managing

Hadoop clusters and performing advanced analytics on large volumes of unstructured data require significant expertise, skill, and training. Unfortunately, there is currently a dearth of Hadoop developers and data scientists available, making it impractical for many enterprises to maintain and take advantage of complex Hadoop clusters. Further, as Hadoop's myriad components are improved upon by the community and new components are created, there is, as with any immature open source technology/approach, a risk of forking. Finally, Hadoop is a batch-oriented framework, meaning it does not support real-time data processing and analysis.

The good news is that some of the brightest minds in IT are contributing to the Apache Hadoop project, and a new generation of Hadoop developers and data scientists is coming of age. As a result, the technology is advancing rapidly, becoming both more powerful and easier to implement and manage. An ecosystems of vendors, both Hadoop-focused start-ups like Cloudera and Hortonworks and well-worn IT stalwarts like IBM and Microsoft, are working to offer commercial, enterprise-ready Hadoop distributions, tools, and services to make deploying and managing the technology a practical reality for the traditional enterprise. Other bleeding-edge start-ups are working to perfect NoSQL (Not Only SQL) data stores capable of delivering near– real-time insights in conjunction with Hadoop. Technology Insights 13.2 provides a few facts to clarify some misconceptions about Hadoop.

TECHNOLOGY INSIGHTS 13.2 A Few Demystifying Facts About Hadoop

Although Hadoop and related technologies have been around for more than 5 years now, most people still have several misconceptions about Hadoop and related technologies such as MapReduce and Hive. The following list of 10 facts intends to clarify what Hadoop is and does relative to BI, as well as in which business and technology situations Hadoop-based BI, data warehousing, and analytics can be useful (Russom, 2013).

Fact #1. Hadoop consists of multiple products. We talk about Hadoop as if it's one monolithic thing, whereas it's actually a family of open source products and technologies overseen by the Apache Software Foundation (ASF). (Some Hadoop products are also available via vendor distributions; more on that later.)

The Apache Hadoop library includes (in BI priority order) the Hadoop Distributed File System (HDFS), MapReduce, Hive, Hbase, Pig, Zookeeper, Flume, Sqoop, Oozie, Hue, and so on. You can combine these in various ways, but HDFS and MapReduce (perhaps with Hbase and Hive) constitute a useful technology stack for applications in BI, DW, and analytics.

Fact #2. Hadoop is open source but available from vendors, too. Apache Hadoop's open source software library is available from ASF at apache.org. For users desiring a more enterprise-ready package, a few vendors now offer Hadoop distributions that include additional administrative tools and technical support.

Fact #3. Hadoop is an ecosystem, not a single product. In addition to products from Apache, the extended Hadoop ecosystem includes a growing list of vendor products that integrate with or expand Hadoop technologies. One minute on your favorite search engine will reveal these.

Fact #4. HDFS is a file system, not a database management system (DBMS). Hadoop is primarily a distributed file system and lacks capabilities we'd associate with a DBMS, such as indexing, random access to data, and support for SQL. That's okay, because HDFS does things DBMSs cannot do.

Fact #5. Hive resembles SQL but is not standard SQL. Many of us are handcuffed to SQL because we know it well and our tools demand it. People who know SQL can quickly learn to hand code Hive, but that doesn't solve compatibility issues with SQL-based tools. TDWI feels that over time, Hadoop products will support standard SQL, so this issue will soon be moot.

Fact #6. Hadoop and MapReduce are related but don't require each other. Developers at Google developed MapReduce before HDFS existed, and some variations of MapReduce work with a variety of storage technologies, including HDFS, other file systems, and some DBMSs.

Fact #7. MapReduce provides control for analytics, not analytics per se. MapReduce is a general-purpose execution engine that handles the complexities of network communication, parallel programming, and fault tolerance for any kind of application that you can hand code—not just analytics.

Fact #8. Hadoop is about data diversity, not just data volume. Theoretically, HDFS can manage the storage and access of any data type as long as you can put the data in a file and copy that file into HDFS. As outrageously simplistic as that sounds, it's largely true, and it's exactly what brings many users to Apache HDFS.

Fact #9. Hadoop complements a DW; it's rarely a replacement. Most organizations have designed their DW for structured, relational data, which makes it difficult to wring BI value from unstructured and semistructured data. Hadoop promises to complement DWs by handling the multi-structured data types most DWs can't.

Fact #10. Hadoop enables many types of analytics, not just Web analytics. Hadoop gets a lot of press about how Internet companies use it for analyzing Web logs and other Web data, but other use cases exist. For example, consider the Big Data coming from sensory devices, such as robotics in manufacturing, RFID in retail, or grid monitoring in utilities. Older analytic applications that need large data samples—such as customer-base segmentation, fraud detection, and risk analysis—can benefit from the additional Big Data managed by Hadoop. Likewise, Hadoop's additional data can expand 360-degree views to create a more complete and granular view.

NoSQL

A related new style of database called **NoSQL** (Not Only SQL) has emerged to, like Hadoop, process large volumes of multi-structured data. However, whereas Hadoop is adept at supporting large-scale, batch-style historical analysis, NoSQL databases are aimed, for the most part (though there are some important exceptions), at serving up discrete data stored among large volumes of multi-structured data to end-user and automated Big Data applications. This capability is sorely lacking from relational database technology, which simply can't maintain needed application performance levels at Big Data scale.

In some cases, NoSQL and Hadoop work in conjunction. The aforementioned HBase, for example, is a popular NoSQL database modeled after Google BigTable that is often deployed on top of HDFS, the Hadoop Distributed File System, to provide low-latency, quick lookups in Hadoop. The downside of most NoSQL databases today is that they trade ACID (atomicity, consistency, isolation, durability) compliance for performance and scalability. Many also lack mature management and monitoring tools. Both these shortcomings are in the process of being overcome by both the open source NoSQL communities and a handful of vendors that are attempting to commercialize the various NoSQL databases. NoSQL databases currently available include HBase, Cassandra, MongoDB, Accumulo, Riak, CouchDB, and DynamoDB, among others. Application Case 13.3 shows the use of NoSQL databases at eBay.

Application Case 13.3

eBay's Big Data Solution

eBay is the world's largest online marketplace, enabling the buying and selling of practically anything. Founded in 1995, eBay connects a diverse and passionate community of individual buyers and sellers, as well as small businesses. eBay's collective impact on e-commerce is staggering: In 2012, the total value of goods sold on eBay was $75.4 billion. eBay currently serves over 112 million active users and has 400+ million items for sale.

The Challenge: Supporting Data at Extreme Scale

One of the keys to eBay's extraordinary success is its ability to turn the enormous volumes of data it generates into useful insights that its customers can glean directly from the pages they frequent. To accommodate eBay's explosive data growth—its data centers perform billions of reads and writes each day—and the increasing demand to process data at blistering speeds, eBay needed a solution that did not have the typical bottlenecks, scalability issues, and transactional constraints associated with common relational database approaches. The company also needed to perform rapid analysis on a broad assortment of the structured and unstructured data it captured.

The Solution: Integrated Real-Time Data and Analytics

Its Big Data requirements brought eBay to NoSQL technologies, specifically Apache Cassandra and DataStax Enterprise. Along with Cassandra and its high-velocity data capabilities, eBay was also drawn to the integrated Apache Hadoop analytics that come with DataStax Enterprise. The solution incorporates a scale-out architecture that enables eBay to deploy multiple DataStax Enterprise clusters across several different data centers using commodity hardware. The end result is that eBay is now able to more cost effectively process massive amounts of data at very high speeds, at very high velocities, and achieve far more than they were able to with the higher cost propriety system they had been using. Currently, eBay is managing a sizable portion of its data center needs—250TBs+ of storage—in Apache Cassandra and DataStax Enterprise clusters.

Additional technical factors that played a role in eBay's decision to deploy DataStax Enterprise so widely include the solution's linear scalability, high availability with no single point of failure, and outstanding write performance.

Handling Diverse Use Cases

eBay employs DataStax Enterprise for many different use cases. The following examples illustrate some of the ways the company is able to meet its Big Data needs with the extremely fast data handling and analytics capabilities the solution provides. Naturally, eBay experiences huge amounts of write traffic, which the Cassandra implementation in DataStax Enterprise handles more efficiently than any other RDBMS or NoSQL solution. eBay currently sees 6 billion+ writes per day across multiple Cassandra clusters and 5 billion+ reads (mostly offline) per day as well.

One use case supported by DataStax Enterprise involves quantifying the social data eBay displays on its product pages. The Cassandra distribution in DataStax Enterprise stores all the information needed to provide counts for "like," "own," and "want" data on eBay product pages. It also provides the same data for the eBay "Your Favorites" page that contains all the items a user likes, owns, or wants, with Cassandra serving up the entire "Your Favorites" page. eBay provides this data through Cassandra's scalable counters feature.

Load balancing and application availability are important aspects to this particular use case. The DataStax Enterprise solution gave eBay architects the flexibility they needed to design a system that enables any user request to go to any data center, with each data center having a single DataStax Enterprise cluster spanning those centers. This design feature helps balance the incoming user load and eliminates any possible threat to application downtime. In addition to the line of business data powering the Web pages its customers visit, eBay is also able to perform high-speed analysis with the ability to maintain a separate data center running

(*Continued*)

Application Case 13.3 (Continued)

Hadoop nodes of the same DataStax Enterprise ring (see Figure 13.5).

Another use case involves the Hunch (an eBay sister company) "taste graph" for eBay users and items, which provides custom recommendations based on user interests. eBay's Web site is essentially a graph between all users and the items for sale. All events (bid, buy, sell, and list) are captured by eBay's systems and stored as a graph in Cassandra. The application sees more than 200 million writes daily and holds more than 40 billion pieces of data.

eBay also uses DataStax Enterprise for many time-series use cases in which processing high-volume, real-time data is a foremost priority. These include mobile notification logging and tracking (every time eBay sends a notification to a mobile phone or device it is logged in Cassandra), fraud detection, SOA request/response payload logging, and RedLaser (another eBay sister company) server logs and analytics.

Across all of these use cases is the common requirement of uptime. eBay is acutely aware of the need to keep their business up and open for business, and DataStax Enterprise plays a key part in that through its support of high availability clusters. "We have to be ready for disaster recovery all the time. It's really great that Cassandra allows for active-active multiple data centers where we can read and write data anywhere, anytime," says eBay architect Jay Patel.

Questions for Discussion

1. Why Big Data is a big deal for eBay?
2. What were the challenges, the proposed solution, and the obtained results?
3. Can you think of other e-commerce businesses that may have Big Data challenges comparable to that of eBay?

Source: DataStax, Customer Case Studies, **datastax.com/resources/casestudies/eBay** (accessed January 2013).

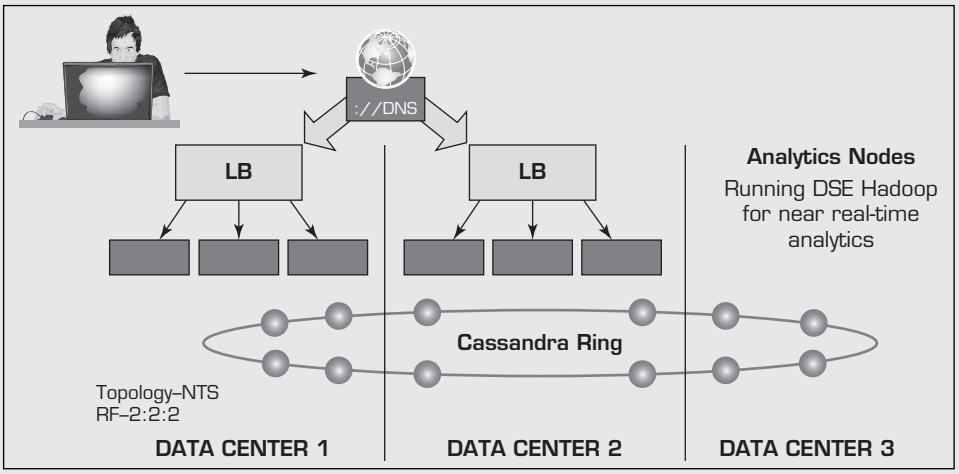

FIGURE 13.5 eBay's Multi–Data-Center Deployment. *Source:* DataStax.

SECTION 13.4 REVIEW QUESTIONS

1. What are the common characteristics of emerging Big Data technologies?
2. What is MapReduce? What does it do? How does it do it?
3. What is Hadoop? How does it work?
4. What are the main Hadopp components? What functions do they perform?
5. What is NoSQL? How does it fit into the Big Data analytics picture?

13.5 DATA SCIENTIST

Data scientist is a role or a job frequently associated with Big Data or data science. In a very short time it has become one of the most sought-out roles in the marketplace. In a recent article published in the October 2012 issue of the *Harvard Business Review,* authors Thomas H. Davenport and D. J. Patil called data scientist "The Sexiest Job of the 21st Century." In that article they specified data scientists' most basic, universal skill as the ability to write code (in the latest Big Data languages and platforms). Although this may be less true in the near future, when many more people will have the title "data scientist" on their business cards, at this time it seems to be the most fundamental skill required from data scientists. A more enduring skill will be the need for data scientists to communicate in a language that all their stakeholders understand—and to demonstrate the special skills involved in storytelling with data, whether verbally, visually, or—ideally—both (Davenport and Patil, 2012).

Data scientists use a combination of their business and technical skills to *investigate* Big Data looking for ways to improve current business analytics practices (from descriptive to predictive and prescriptive) and hence to improve decisions for new business opportunities. One of the biggest differences between a data scientist and a business intelligence user—such as a business analyst—is that a data scientist investigates and looks for new possibilities, while a BI user analyzes existing business situations and operations.

One of the dominant traits expected from data scientists is an intense curiosity—a desire to go beneath the surface of a problem, find the questions at its heart, and distill them into a very clear set of hypotheses that can be tested. This often entails the associative thinking that characterizes the most creative scientists in any field. For example, we know of a data scientist studying a fraud problem who realized that it was analogous to a type of DNA sequencing problem (Davenport and Patil, 2012). By bringing together those disparate worlds, he and his team were able to craft a solution that dramatically reduced fraud losses.

Where Do Data Scientists Come From?

Although there still is disagreement about the use of "science" in the name, it is becoming less of a controversial issue. Real scientists use tools made by other scientists, or make them if they don't exist, as a means to expand knowledge. That is exactly what data scientists are expected to do. Experimental physicists, for example, have to design equipment, gather data, and conduct multiple experiments to discover knowledge and communicate their results. Even though they may not be wearing white coats, and may not be living in a sterile lab environment, that is exactly what data scientists do: use creative tools and techniques to turn data into actionable information for others to use for better decision making.

There is no consensus on what educational background a data scientist has to have. The usual suspects like Master of Science (or Ph.D.) in Computer Science, MIS, Industrial Engineering, or the newly popularized postgraduate analytics degrees may be necessary but not sufficient to call someone a data scientist. One of the most sought-out characteristics of a data scientist is expertise in both technical and business application domains. In that sense, it somewhat resembles to the professional engineer (PE) or project management professional (PMP) roles, where experience is valued as much as (if not more than) the technical skills and educational background. It would not be a huge surprise to see within the next few years a certification specifically designed for data scientists (perhaps called "Data Science Professional" or "DSP," for short).

Because it is a profession for a field that is still being defined, many of its practices are still experimental and far from being standardized; companies are overly sensitive about the experience dimension of data scientist. As the profession matures, and

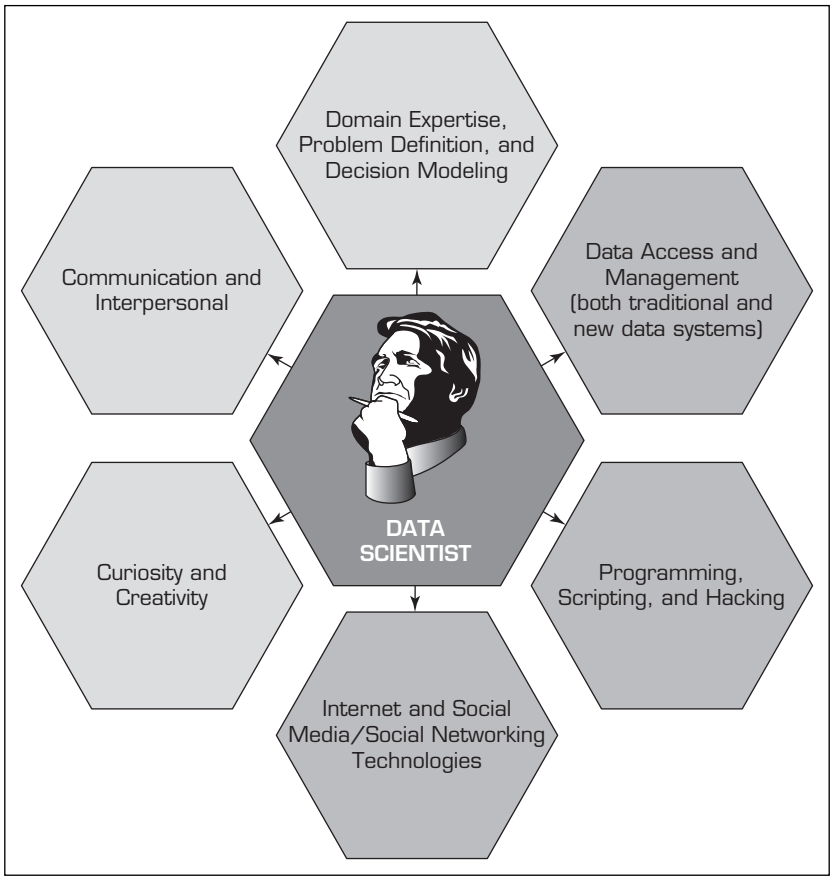

FIGURE 13.6 Skills That Define a Data Scientist.

practices are standardized, experience will be less of an issue when defining a data scientist. Nowadays, companies looking for people who have extensive experience in working with complex data have had good luck recruiting among those with educational and work backgrounds in the physical or social sciences. Some of the best and brightest data scientists have been Ph.D.s in esoteric fields like ecology and systems biology (Davenport and Patil, 2012). Even though there is no consensus on where data scientists come from, there is a common understanding of what skills and qualities they are expected to possess. Figure 13.6 shows a high-level graphical illustration of these skills.

Data scientists are expected to have soft skills such as creativity, curiosity, communication/interpersonal, domain expertise, problem definition, and managerial (shown with light background hexagons on the left side of the figure) as well as sound technical skills such as data manipulation, programming/hacking/scripting, and Internet and social media/networking technologies (shown with darker background hexagons on the right side of the figure). Technology Insights 13.3 is about a typical job advertisement for a data scientist.

TECHNOLOGY INSIGHTS 13.3 A Typical Job Post for Data Scientists

[Some company] is seeking a Data Scientist to join our Big Data Analytics team. Individuals in this role are expected to be comfortable working as a software engineer and a quantitative researcher. The ideal candidate will have a keen interest in the study of an online social network and a passion for identifying and answering questions that help us build the best products.

Responsibilities

- Work closely with a product engineering team to identify and answer important product questions
- Answer product questions by using appropriate statistical techniques on available data
- Communicate findings to product managers and engineers
- Drive the collection of new data and the refinement of existing data sources
- Analyze and interpret the results of product experiments
- Develop best practices for instrumentation and experimentation and communicate those to product engineering teams

Requirements

- M.S. or Ph.D. in a relevant technical field, or 4+ years of experience in a relevant role
- Extensive experience solving analytical problems using quantitative approaches
- Comfort with manipulating and analyzing complex, high-volume, high-dimensionality data from varying sources
- A strong passion for empirical research and for answering hard questions with data
- A flexible analytic approach that allows for results at varying levels of precision
- Ability to communicate complex quantitative analysis in a clear, precise, and actionable manner
- Fluency with at least one scripting language such as Python or PHP
- Familiarity with relational databases and SQL
- Expert knowledge of an analysis tool such as R, Matlab, or SAS
- Experience working with large data sets, experience working with distributed computing tools a plus (Map/Reduce, Hadoop, Hive, etc.)

People with this range of skills are rare, which explains why data scientists are in short supply. Because of the high demand for these relatively fewer individuals, the starting salaries for data scientists are well above six figures, and for ones with ample experience and specific domain expertise, salaries are pushing near seven figures. For most organizations, rather than looking for individuals with all these capabilities, it will be necessary instead to build a team of people that collectively have these skills. Here are some recent anecdotes about data scientists:

- Data scientists turn Big Data into big value, delivering products that delight users and insight that informs business decisions.
- A data scientist is not only proficient to work with data, but also appreciates data itself as an invaluable asset.
- By 2020 there will be 4.5 million new data scientist jobs, of which only one-third will be filled because of the lack of available personnel.
- Today's data scientists are the quants of the financial markets of the 1980s.

Data scientists are not limited to high-tech Internet companies. Many of the companies that do not have much Internet presence are also interested in highly qualified Big Data analytics professionals. For instance, as described in the End-of-Chapter Application Case, Volvo is leveraging data scientists to turn data that comes from its corporate transaction databases and from sensors (placed in its cars) into actionable insight. An interesting area where we have seen the use of data scientists in the recent past is in politics. Application Case 13.4 describes the use of Big Data analytics in the world of politics and presidential elections.

Application Case 13.4

Big Data and Analytics in Politics

One of the application areas where Big Data and analytics promise to make a big difference is arguably the field of politics. Experiences from the recent presidential elections illustrated the power of Big Data and analytics to acquire and energize millions of volunteers (in the form of a modern-era grassroots movement) to not only raise hundreds of millions of dollars for the election campaign but to optimally organize and mobilize potential voters to get out and vote in large numbers, as well. Clearly, the 2008 and 2012 presidential elections made a mark on the political arena with the creative use of Big Data and analytics to improve chances of winning. Figure 13.7 illustrates a graphical depiction of the analytical process of converting a wide variety of data into the ingredients for winning an election.

As Figure 13.7 illustrates, data is the source of information; the richer and deeper it is, the better and more relevant the insights. The main characteristics of Big Data, namely volume, variety, and velocity (the three Vs), readily apply to the kind of data that is used for political campaigns. In addition to the structured data (e.g., detailed records of previous campaigns, census data, market research, and poll data) vast volumes and a variety of **social media** (e.g., tweets at Twitter, Facebook wall posts, blog posts) and Web data (Web pages, news articles, newsgroups) are used to learn more about voters and obtain deeper insights to enforce or change their opinions. Often, the search and browsing histories of individuals are captured and made available to customers (political analysts) who can use such data for better insight and behavioral targeting. If done correctly, Big Data and analytics can provide invaluable information to manage political campaigns better than ever before.

From predicting election outcomes to targeting potential voters and donors, Big Data and analytics have a lot to offer to modern-day election campaigns. In fact, they have changed the way presidential election campaigns are run. In the 2008 and 2012 presidential elections, the major political parties (Republican and Democratic) employed social media and data-driven analytics for a more effective and efficient campaign, but as many agree, the Democrats clearly had the competitive advantage (Issenberg, 2012). Obama's 2012 data and analytics-driven operation was far more sophisticated and more efficient than its much-heralded 2008 process, which was primarily social media driven. In the 2012

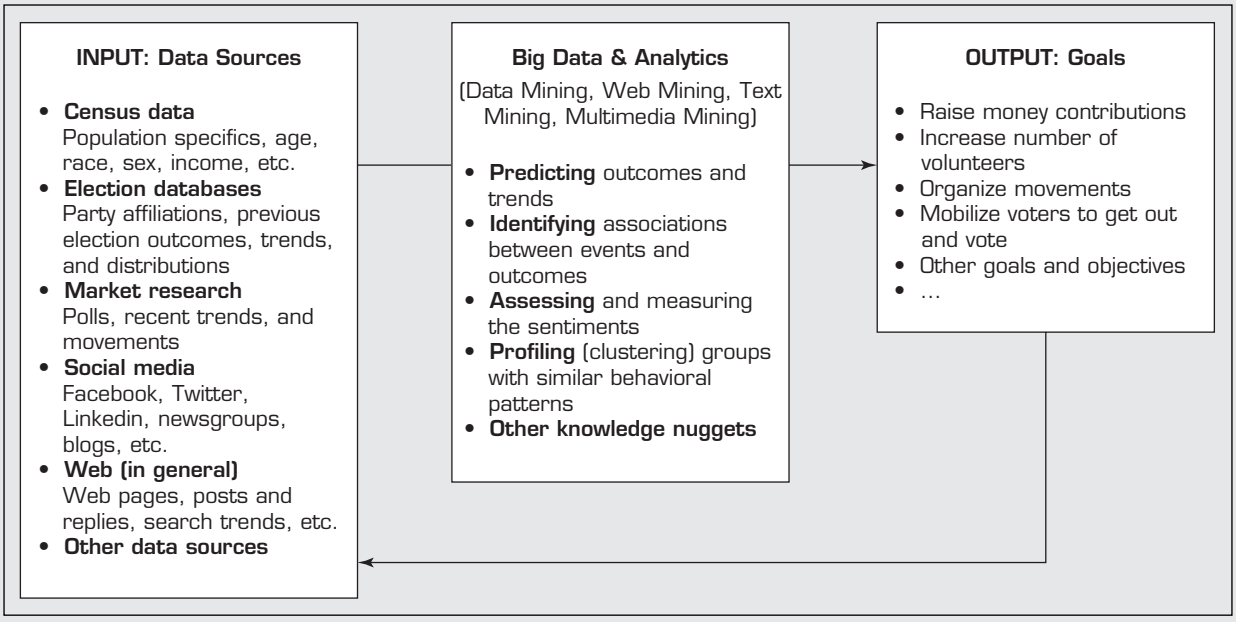

FIGURE 13.7 Leveraging Big Data and Analytics for Political Campaigns.

campaign, hundreds of analysts applied advanced analytics on very large and diverse data sources to pinpoint exactly who to target, for what reason, with what message, on a continuous basis. Compared to 2008, they had more expertise, hardware, software, data (e.g., Facebook and Twitter were orders of magnitude bigger in 2012 than they had been in 2008), and computational resources to go over and beyond what they had accomplished previously (Shen, 2013). Before the 2012 election, in June of last year, a *Politico* reporter claimed that Obama had a data advantage and went on to say that the depth and breadth of the campaign's digital operation, from political and demographic data mining to voter sentiment and behavioral analysis, reached beyond anything politics had ever seen (Romano, 2012).

According to Shen, the real winner of the 2012 elections was analytics (Shen, 2013). While most people, including the so-called political experts (who often rely on gut feelings and experiences), thought the 2012 presidential election would be very close, a number of analysts, based on their data-driven analytical models, predicted that Obama would win easily with close to 99 percent certainty. For example, Nate Silver at FiveThirtyEight, a popular political blog published by *The New York Times,* predicted not only that Obama would win but also by exactly how much he would win. Simon Jackman, professor of political science at Stanford University, accurately predicted that Obama would win 332 electoral votes and that North Carolina and Indiana—the only two states that Obama won in 2008—would fall to Romney.

In short, Big Data and analytics have become a critical part of political campaigns. The usage and expertise gap between the party lines may disappear, but the importance of analytical capabilities will continue to evolve for the foreseeable future.

Questions for Discussion

1. What is the role of analytics and Big Data in modern-day politics?
2. Do you think Big Data Analytics could change the outcome of an election?
3. What do you think are the challenges, the potential solution, and the probable results of the use of Big Data Analytics in politics?

Sources: Compiled from G. Shen, "Big Data, Analytics and Elections," INFORMS' *Analytics Magazine,* January–February 2013; L. Romano, "Obama's Data Advantage," *Politico,* June 9, 2012; M. Scherer, "Inside the Secret World of the Data Crunchers Who Helped Obama Win," *Time,* November 7, 2012; S. Issenberg, "Obama Does It Better" (from "Victory Lab: The New Science of Winning Campaigns"), *Slate,* October 29, 2012; and D. A. Samuelson, "Analytics: Key to Obama's Victory," INFORMS' *ORMS Today,* February 2013 Issue, pp. 20–24.

SECTION 13.5 REVIEW QUESTIONS

1. Who is a data scientist? What makes them so much in demand?
2. What are the common characteristics of data scientists? Which one is the most important?
3. Where do data scientists come from? What educational backgrounds do they have?
4. What do you think is the path to becoming a great data scientist?

13.6 BIG DATA AND DATA WAREHOUSING

There is doubt that the emergence of Big Data has changed and will continue to change data warehousing in a significant way. Until recently, enterprise data warehouses were the centerpiece of all decision support technologies. Now, they have to share the spotlight with the newcomer, Big Data. The question that is popping up everywhere is whether Big Data and its enabling technologies such as Hadoop will replace data warehousing and its core technology relational data base management systems (RDBMS). Are we witnessing a data warehouse versus Big Data challenge (or from the technology standpoint, Hadoop versus RDBMS)? In this section we will explain why these questions have no basis—and at least justify that such an either-or choice is not the reflection of the reality at this point in time.

In the last decade or so, we have seen a significant improvement in the area of computer-based decision support systems, which can largely be credited to data warehousing and technological advancements in both software and hardware to capture, store, and analyze data. As the size of the data increased, so did the capabilities of data warehouses. Some of these data warehousing advances included massively parallel processing (moving from one or few to many parallel processors), storage area networks (easily scalable storage solutions), solid-state storage, in-database processing, in-memory processing, and columnar (column oriented) databases, just to name a few. These advancements helped keep the increasing size of data under control, while effectively serving analytics needs of the decision makers. What has changed the landscape in recent years is the variety and complexity of data, which made data warehouses incapable of keeping up. It is not the volume of the structured data but the variety and the velocity that forced the world of IT to develop a new paradigm, which we now call "Big Data." Now that we have these two paradigms, data warehousing and Big Data, seemingly competing for the same job—turning data into actionable information—which one will prevail? Is this a fair question to ask? Or are we missing the big picture? In this section, we try to shed some light on this intriguing question.

As has been the case for many previous technology innovations, hype about Big Data and its enabling technologies like Hadoop and MapReduce is rampant. Both non-practitioners as well as practitioners are overwhelmed by diverse opinions. According to Awadallah and Graham (2012), people are missing the point in claiming that Hadoop replaces relational databases and is becoming the new data warehouse. It is easy to see where these claims originate since both Hadoop and data warehouse systems can run in parallel, scale up to enormous data volumes, and have shared-nothing architectures. At a conceptual level, it is easy to think they are interchangeable. The reality is that they are not, and the differences between the two overwhelm the similarities. If they are not interchangeable, then how do we decide when to deploy Hadoop and when to use a data warehouse?

Use Case(s) for Hadoop

As we have covered earlier in this chapter, Hadoop is the result of new developments in computer and storage grid technologies. Using commodity hardware as a foundation, Hadoop provides a layer of software that spans the entire grid, turning it into a single system. Consequently, some major differentiators are obvious in this architecture:

- Hadoop is the repository and refinery for raw data.
- Hadoop is a powerful, economical, and active archive.

Thus, Hadoop sits at both ends of the large-scale data life cycle—first when raw data is born, and finally when data is retiring, but is still occasionally needed.

1. ***Hadoop as the repository and refinery.*** As volumes of Big Data arrive from sources such as sensors, machines, social media, and clickstream interactions, the first step is to capture all the data reliably and cost effectively. When data volumes are huge, the traditional single-server strategy does not work for long. Pouring the data into the Hadoop Distributed File System (HDFS) gives architects much needed flexibility. Not only can they capture hundreds of terabytes in a day, but they can also adjust the Hadoop configuration up or down to meet surges and lulls in data ingestion. This is accomplished at the lowest possible cost per gigabyte due to open source economics and leveraging commodity hardware.

   Since the data is stored on local storage instead of SANs, Hadoop data access is often much faster, and it does not clog the network with terabytes of data movement. Once the raw data is captured, Hadoop is used to refine it. Hadoop can act

as a parallel "ETL engine on steroids," leveraging handwritten or commercial data transformation technologies. Many of these raw data transformations require the unraveling of complex free-form data into structured formats. This is particularly true with clickstreams (or Web logs) and complex sensor data formats. Consequently, a programmer needs to tease the wheat from the chaff, identifying the valuable signal in the noise.

2. ***Hadoop as the active archive.*** In a 2003 interview with ACM, Jim Gray claimed that hard disks can be treated as tape. While it may take many more years for magnetic tape archives to be retired, today some portions of tape workloads are already being redirected to Hadoop clusters. This shift is occurring for two fundamental reasons. First, while it may appear inexpensive to store data on tape, the true cost comes with the difficulty of retrieval. Not only is the data stored offline, requiring hours if not days to restore, but tape cartridges themselves are also prone to degradation over time, making data loss a reality and forcing companies to factor in those costs. To make matters worse, tape formats change every couple of years, requiring organizations to either perform massive data migrations to the newest tape format or risk the inability to restore data from obsolete tapes.

   Second, it has been shown that there is value in keeping historical data online and accessible. As in the clickstream example, keeping raw data on a spinning disk for a longer duration makes it easy for companies to revisit data when the context changes and new constraints need to be applied. Searching thousands of disks with Hadoop is dramatically faster and easier than spinning through hundreds of magnetic tapes. Additionally, as disk densities continue to double every 18 months, it becomes economically feasible for organizations to hold many years' worth of raw or refined data in HDFS. Thus, the Hadoop storage grid is useful in both the preprocessing of raw data and the long-term storage of data. It's a true "active archive" since it not only stores and protects the data, but also enables users to quickly, easily, and perpetually derive value from it.

Use Case(s) for Data Warehousing

After nearly 30 years of investment, refinement, and growth, the list of features available in a data warehouse is quite staggering. Built upon relational database technology using schemas and integrating business intelligence (BI) tools, the major differences in this architecture are:

- Data warehouse performance
- Integrated data that provides business value
- Interactive BI tools for end users

1. ***Data warehouse performance.*** Basic indexing, found in open source databases, such as MySQL or Postgres, is a standard feature used to improve query response times or enforce constraints on data. More advanced forms such as materialized views, aggregate join indexes, cube indexes, and sparse join indexes enable numerous performance gains in data warehouses. However, the most important performance enhancement to date is the cost-based optimizer. The optimizer examines incoming SQL and considers multiple plans for executing each query as fast as possible. It achieves this by comparing the SQL request to the database design and extensive data statistics that help identify the best combination of execution steps. In essence, the optimizer is like having a genius programmer examine every query and tune it for the best performance. Lacking an optimizer or data demographic statistics, a query that could run in minutes may take hours, even with many indexes.

For this reason, database vendors are constantly adding new index types, partitioning, statistics, and optimizer features. For the past 30 years, every software release has been a performance release.

2. ***Integrating data that provides business value.*** At the heart of any data warehouse is the promise to answer essential business questions. Integrated data is the unique foundation required to achieve this goal. Pulling data from multiple subject areas and numerous applications into one repository is the *raison d'être* for data warehouses. Data model designers and ETL architects armed with metadata, data-cleansing tools, and patience must rationalize data formats, source systems, and semantic meaning of the data to make it understandable and trustworthy. This creates a common vocabulary within the corporation so that critical concepts such as "customer," "end of month," or "price elasticity" are uniformly measured and understood. Nowhere else in the entire IT data center is data collected, cleaned, and integrated as it is in the data warehouse.
3. ***Interactive BI tools.*** BI tools such as MicroStrategy, Tableau, IBM Cognos, and others provide business users with direct access to data warehouse insights. First, the business user can create reports and complex analysis quickly and easily using these tools. As a result, there is a trend in many data warehouse sites toward end-user self-service. Business users can easily demand more reports than IT has staffing to provide. More important than self-service, however, is that the users become intimately familiar with the data. They can run a report, discover they missed a metric or filter, make an adjustment, and run their report again all within minutes. This process results in significant changes in business users' understanding the business and their decision-making process. First, users stop asking trivial questions and start asking more complex strategic questions. Generally, the more complex and strategic the report, the more revenue and cost savings the user captures. This leads to some users becoming "power users" in a company. These individuals become wizards at teasing business value from the data and supplying valuable strategic information to the executive staff. Every data warehouse has anywhere from two to 20 power users.

The Gray Areas (Any One of the Two Would Do the Job)

Even though there are several areas that differentiate one from the other, there are also gray areas where the data warehouse and Hadoop cannot be clearly discerned. In these areas either tool could be the right solution—either doing an equally good or a not-so-good job on the task at hand. Choosing the one over the other depends on the requirements and the preferences of the organization. In many cases, Hadoop and the data warehouse work together in an information supply chain, and just as often, one tool is better for a specific workload (Awadallah and Graham, 2012). Table 13.1 illustrates the preferred platform (one versus the other, or equally likely) under a number of commonly observed requirements.

Coexistence of Hadoop and Data Warehouse

There are several possible scenarios under which using a combination of Hadoop and relational DBMS-based data warehousing technologies makes more sense. Here are some of those scenarios (White, 2012):

1. ***Use Hadoop for storing and archiving multi-structured data.*** A connector to a relational DBMS can then be used to extract required data from Hadoop for analysis by the relational DBMS. If the relational DBMS supports MapReduce functions, these functions can be used to do the extraction. The Aster-Hadoop adaptor,

TABLE 13.1 When to Use Which Platform—Hadoop Versus DW

Requirement	Data Warehouse	Hadoop
Low latency, interactive reports, and OLAP	☑	
ANSI 2003 SQL compliance is required	☑	☑
Preprocessing or exploration of raw unstructured data		☑
Online archives alternative to tape	☑	
High-quality cleansed and consistent data	☑	☑
100s to 1,000s of concurrent users	☑	☑
Discover unknown relationships in the data		☑
Parallel complex process logic	☑	☑
CPU intense analysis	☑	
System, users, and data governance		☑
Many flexible programming languages running in parallel		☑
Unrestricted, ungoverned sandbox explorations		☑
Analysis of provisional data	☑	
Extensive security and regulatory compliance	☑	☑

for example, uses SQL-MapReduce functions to provide fast, two-way data loading between HDFS and the Aster Database. Data loaded into the Aster Database can then be analyzed using both SQL and MapReduce.

2. ***Use Hadoop for filtering, transforming, and/or consolidating multi-structured data.*** A connector such as the Aster-Hadoop adaptor can be used to extract the results from Hadoop processing to the relational DBMS for analysis.
3. ***Use Hadoop to analyze large volumes of multi-structured data and publish the analytical results*** to the traditional data warehousing environment, a shared workgroup data store, or a common user interface.
4. ***Use a relational DBMS that provides MapReduce capabilities as an investigative computing platform.*** Data scientists can employ the relational DBMS (the Aster Database system, for example) to analyze a combination of structured data and multi-structured data (loaded from Hadoop) using a mixture of SQL processing and MapReduce analytic functions.
5. ***Use a front-end query tool to access and analyze data*** that is stored in both Hadoop and the relational DBMS.

These scenarios support an environment where the Hadoop and relational DBMS systems are separate from each other and connectivity software is used to exchange data between the two systems (see Figure 13.8). The direction of the industry over the next few years will likely be moving toward more tightly coupled Hadoop and relational DBMS-based data warehouse technologies—software as well as hardware. Such integration provides many benefits, including eliminating the need to install and maintain multiple systems, reducing data movement, providing a single metadata store for application development, and providing a single interface for both business users and analytical tools.

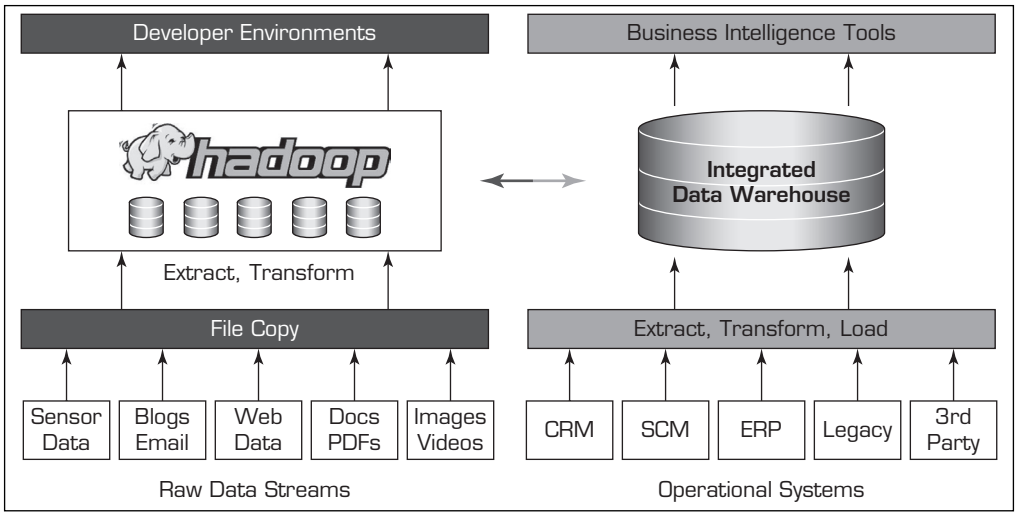

FIGURE 13.8 Coexistence of Hadoop and Data Warehouses. *Source:* Teradata.

SECTION 13.6 REVIEW QUESTIONS

1. What are the challenges facing data warehousing and Big Data? Are we witnessing the end of the data warehousing era? Why or why not?
2. What are the use cases for Big Data and Hadoop?
3. What are the use cases for data warehousing and RDBMS?
4. In what scenarios can Hadoop and RDBMS coexist?

13.7 BIG DATA VENDORS

As a relatively new technology area, the Big Data vendor landscape is developing very rapidly. A number of vendors have developed their own Hadoop distributions, most based on the Apache open source distribution but with various levels of proprietary customization. The clear market leader in terms of distribution seems to be Cloudera (**cloudera.com**), a Silicon Valley start-up with an all-star lineup of Big Data experts, including Hadoop creator Doug Cutting and former Facebook data scientist Jeff Hammerbacher. In addition to distribution, Cloudera offers paid enterprise-level training/services and proprietary Hadoop management software. MapR (**mapr.com**), another Valley start-up, offers its own Hadoop distribution that supplements HDFS with its proprietary NFS for improved performance. EMC Greenplum partnered with MapR to release a partly proprietary Hadoop distribution of its own in May 2011. Hortonworks (**hortonworks.com**), which was spun-out of Yahoo! in summer 2011, released its 100 percent open source Hadoop distribution, called Hortonworks Data Platform, and related support services in November 2011. These are just a few of the many companies (established and start-ups) that are crowding the competitive landscape of tool and service providers for Hadoop technologies.

In the NoSQL world, a number of start-ups are working to deliver commercially supported versions of the various flavors of NoSQL. DataStax, for example, offers a commercial version of Cassandra that includes enterprise support and services, as well as integration with Hadoop and open source enterprise search via Lucene Solr. As mentioned, proprietary data integration vendors, including Informatica, Pervasive Software, and Syncsort, are making inroads into the Big Data market with Hadoop connectors and complementary tools aimed at making it easier for developers to move data around and within Hadoop clusters.

The analytics layer of the Big Data stack is also experiencing significant development. A start-up called Datameer, for example, is developing what it says is an "all-in-one"

business intelligence platform for Hadoop, while data visualization specialist Tableau Software has added Hadoop and Next Generation Data Warehouse connectivity to its product suite. EMC Greenplum, meanwhile, has Chorus, a sort of playground for data scientists where they can mash-up, experiment with, and share large volumes of data for analysis. Other vendors focus on specific analytic use cases, such as ClickFox with its customer experience analytics engine. A number of traditional business intelligence vendors, most notably MicroStrategy, are working to incorporate Big Data analytic and reporting capabilities into their products.

Less progress has been made in the Big Data application space, however. There are few off-the-shelf Big Data applications currently on the market. This void leaves enterprises with the task of developing and building custom Big Data applications with internal or outsourced teams of application developers. There are exceptions. Namely, a start-up called Treasata offers Big-Data-as-a-service applications for the financial services vertical market, and Google makes its internal Big Data analytics application, called BigQuery, available as a service.

Meanwhile, the next-generation data warehouse market has experienced significant consolidation since 2010. Four leading vendors in this space—Netezza, Greenplum, Vertica, and Aster Data—were acquired by IBM, EMC, HP, and Teradata, respectively. Just a handful of niche independent players remain, among them Kognitio and ParAccel. These vendors, by and large, position their products as complementary to Hadoop and NoSQL deployments, providing real-time analytic capabilities on large volumes of structured data.

Mega-vendors Oracle and IBM also play in the Big Data space. IBM's Big Insights platform is based on Apache Hadoop, but includes numerous proprietary modules including the Netezza database, InfoSphere Warehouse, Cognos business intelligence tools, and SPSS data mining capabilities. It also offers IBM InfoSphere Streams, a platform designed for streaming Big Data analysis. Oracle, meanwhile, has embraced the appliance approach to Big Data with its Exadata, Exalogic, and Big Data appliances. Its Big Data appliance incorporates Cloudera's Hadoop distribution with Oracle's NoSQL database and data integration tools. Application Case 6.5 provides an interesting case where Dublin City council used Big Data Analytics to reduce city's traffic congestion.

The cloud is increasingly playing a role in the Big Data market as well. Amazon and Google support Hadoop deployments in their public cloud offerings, Amazon Elastic

Application Case 13.5

Dublin City Council Is Leveraging Big Data to Reduce Traffic Congestion

Employing 6,000 people, Dublin City Council (DCC) delivers housing, water and transport services to 1.2 million citizens across the Irish capital. To keep the city moving, the council's traffic control center (TCC) works together with local transport operators to manage an extensive network of roads, tramways and bus lanes. Using operational data from the TCC, the council's roads and traffic department is responsible for predicting Dublin's future transport requirements, and developing effective strategies to meet them.

Like local governments in many large European cities, DCC has a wide array of technology at its disposal. Sensors such as inductive-loop traffic detectors, rain gauges and closed-circuit television (CCTV) cameras collect data from across Dublin, and each of the city's 1,000 buses transmits a GPS update every 20 seconds.

Tackling Traffic Congestion

In the past, only a small proportion of this Big Data was available to controllers at Dublin's TCC–reducing their ability to identify, anticipate and address the causes of traffic congestion.

As Brendan O'Brien, Head of Technical Services–Roads and Traffic Department at Dublin City Council, explains: "Previously, our TCC systems only offered a narrow window on the overall status of our transport network–for example, controllers could only view the status of individual bus routes. Our legacy systems were also unable to monitor the

(*Continued*)

Application Case 13.5 (Continued)

geospatial location of Dublin's bus fleet, which further complicated the traffic control process." He continues: "Because we couldn't see the 'health' of the whole transport network in real time, it was very difficult to identify traffic congestion in its early stages. This meant that the causes of delays had often moved on by the time our TCC operators were able to select the appropriate CCTV feed–making it hard to determine and mitigate the factors causing congestion."

DCC wanted to ease traffic congestion across Dublin. To achieve this, the council needed to find a way to integrate, process and visualize large amounts of structured and unstructured data from its network of sensor arrays–all in real time.

Becoming a Smarter City

To help develop a smarter approach to traffic control, DCC entered into a research partnership with IBM Research–Ireland. Francesco Calabrese, Research Manager–Smarter Urban Dynamics at IBM Research, comments: "Smarter Cities are cities with the tools to extract actionable insights from massive amounts of constantly changing data, and deliver those insights instantly to decision-makers. At the IBM Smarter Cities Technology Centre in Dublin, our goal is to develop innovative solutions to enable cities like Dublin to support smarter ways of working–delivering a better quality of life for their citizens. "

Today, DCC makes all of its data available to the IBM Smarter Cities Technology Centre in Dublin. Using Big Data analytics technologies, IBM Research is developing new solutions for Smarter Cities, and making the deep insights it discovers available to the council's roads and traffic department.

"From our first discussion with the IBM Research team, we realized that our goals were perfectly aligned," says O'Brien. "Using our data, the IBM Smarter Cities Technology Centre can both drive its own research, and deliver innovative solutions to help us visualize transport data from sensor arrays across the city."

Analyzing the Transport Network

As a first step, IBM integrated geospatial data from buses and data on bus timetables into a central geographic information system. Using IBM InfoSphere Streams and mapping software, IBM researchers created a digital map of the city, overlaid with the real-time positions of Dublin's 1,000 buses. "In the past, our TCC operators could only see the status of individual bus corridors," says O'Brien. "Now, each TCC operator gets a twin-monitor setup–one displaying a dashboard, and the other a real-time map of all buses across the city.

"Using the dashboard screen, operators can drill down to see the number of buses that are on-time or delayed on each route. This information is also displayed visually on the map screen, allowing operators to see the current status of the entire bus network at a glance. Because the interface is so intuitive, our operators can rapidly home in on emerging areas of traffic congestion, and then use CCTV to identify the causes of delays before they move further downstream."

Taking Action to Ease Congestion

By enriching its data with GPS tracking, DCC can produce detailed reports on areas of the network where buses are frequently delayed, and take action to ease congestion. "The IBM Smarter Cities Technology Centre has provided us with a lot of valuable insights," says O'Brien. "For example, the IBM team created trace reports on bus journeys, which showed that at rush hour, some buses were being overtaken by buses that set off later.

"Working with the city's bus operators, we are looking at why the headways are diverging in that way, and what we can do to improve traffic flow at these peak times. Thanks to the work of the IBM team, we can now start answering questions such as: 'Are the bus lane start times correct?', and 'Where do we need to add additional bus lanes and bus-only traffic signals?'"

O'Brien continues: "Over the next two years, we are starting a project team for bus priority measures and road-infrastructure improvements. Without the ability to visualize our transport data, this would not have been possible."

Planning For the Future

Based on the success of the traffic control project for the city's bus fleet, DCC and IBM Research are working together to find ways to further augment traffic control in Dublin. "Our relationship with IBM is quite fluid–we offer them our expertise about

how the city operates, and their researchers use that input to extract valuable insights from our Big Data," says O'Brien. "Currently, the IBM team is working on ways to integrate data from rain and flood gauges into the traffic control solution–alerting controllers to potential hazards presented by extreme weather conditions, and allowing them to take timely action to reduce the impact on road users."

In addition to meteorological data, IBM is investigating the possibility of incorporating data from the under-road sensor network to better understand the impact of private motor vehicles on traffic congestion.

The IBM team is also developing a predictive analytics solution combining data from the city's tram network with electronic docks for the city's free bicycle scheme. This project aims to optimize the distribution of the city's free bicycles according to anticipated demand–ensuring that citizens can seamlessly continue their journey after stepping off a tram.

"Working with IBM Research has allowed us to take a fresh look at our transport strategy," concludes O'Brien. "Thanks to the continuing work of the IBM team, we can see how our transport network is working as a whole–and develop innovative ways to improve it for Dublin's citizens."

Questions for Discussion

1. Is there a strong case to make for large cities to use Big Data Analytics and related information technologies? Identify and discuss examples of what can be done with analytics beyond what is portrayed in this application case.
2. How can a big data analytics help ease the traffic problem in large cities?
3. What were the challenges Dublin City was facing; what were the proposed solution, initial results, and future plans?

Source: IBM Customer Story, "Dublin City Council - Leveraging the leading edge of IBM Smarter Cities research to reduce traffic congestion" **public.dhe.ibm.com/common/ssi/ecm/en/imc14829ieen/IMC14829IEEN.PDF** (accessed October 2013).

MapReduce and Google Compute Engine, respectively, enabling users to easily scale up and scale down clusters as needed. Microsoft abandoned its own internal Big Data platform and will support Hortonworks' Hadoop distribution on its Azure cloud.

As part of its market-sizing efforts, Wikibon (Kelly, 2013) tracked and/or modeled the 2012 Big Data revenue of more than 60 vendors. The list included both Big Data pure-plays—those vendors that derive close to if not all their revenue from the sale of Big Data products and services—and vendors for whom Big Data sales is just one of multiple revenue streams. Table 13.2 shows the top 20 vendors in order of Big Data revenues in 2012, and Figure 13.9 shows the top 10 pure players in the Big Data marketplace.

The services side of the Big Data market is small but growing. The established services providers like Accenture and IBM are just starting to build Big Data practices, while just a few smaller providers focus strictly on Big Data, among them Think Big Analytics. EMC is also investing heavily in Big Data training and services offerings, particularly around data science. Similarly, Hadoop distribution vendors Hortonworks and Cloudera offer a number of training classes aimed at both Hadoop administrators and data scientists.

There are also other vendors approaching Big Data from the visual analytics angle. As Gartner's latest Magic Quadrant indicated, a significant growth in business intelligence and analytics is in visual exploration and visual analytics. Large companies like SAS, SAP, and IBM, along with small but stable companies like Tableau, TIBCO, and QlikView, are making a strong case for high performance analytics built into information visualization platforms. Technology Insights 13.4 provides a few key enablers to succeed with Big Data and visual analytics. SAS is perhaps the one pushing it harder than any other with its recently launched SAS Visual Analytics platform. Using a multitude of computational enhancements, the SAS Visual Analytics platform is capable of turning tens of millions of data records into informational graphics in just a few seconds by using massively parallel processing (MPP) and in-memory computing. Application Case 13.6 is a customer case where the SAS Visual Analytics platform is used for accurate and timely credit decisions.

TABLE 13.2 Top 20 Vendors in Big Data Market

2012 Worldwide Big Data Revenue by Vendor ($US millions)						
Vendor	**Big Data Revenue**	**Total Revenue**	**Big Data Revenue as % of Total Revenue**	**% Big Data Hardware Revenue**	**% Big Data Software Revenue**	**% Big Data Services Revenue**
IBM	$1,352	$103,930	1%	22%	33%	44%
HP	$664	$119,895	1%	34%	29%	38%
Teradata	$435	$2,665	16%	31%	28%	41%
Dell	$425	$59,878	1%	83%	0%	17%
Oracle	$415	$39,463	1%	25%	34%	41%
SAP	$368	$21,707	2%	0%	67%	33%
EMC	$336	$23,570	1%	24%	36%	39%
Cisco Systems	$214	$47,983	0%	80%	0%	20%
Microsoft	$196	$$71,474	0%	0%	67%	33%
Accenture	$194	$29,770	1%	0%	0%	100%
Fusion-io	$190	$439	43%	71%	0%	29%
PwC	$189	$31,500	1%	0%	0%	100%
SAS Institute	$187	$2,954	6%	0%	59%	41%
Splunk	$186	$186	100%	0%	71%	29%
Deloitte	$173	$31,300	1%	0%	0%	100%
Amazon	$170	$56,825	0%	0%	0%	100%
NetApp	$138	$6,454	2%	77%	0%	23%
Hitachi	$130	$112,318	0%	0%	0%	100%
Opera Solutions	$118	$118	100%	0%	0%	100%
Mu Sigma	$114	$114	100%	0%	0%	100%

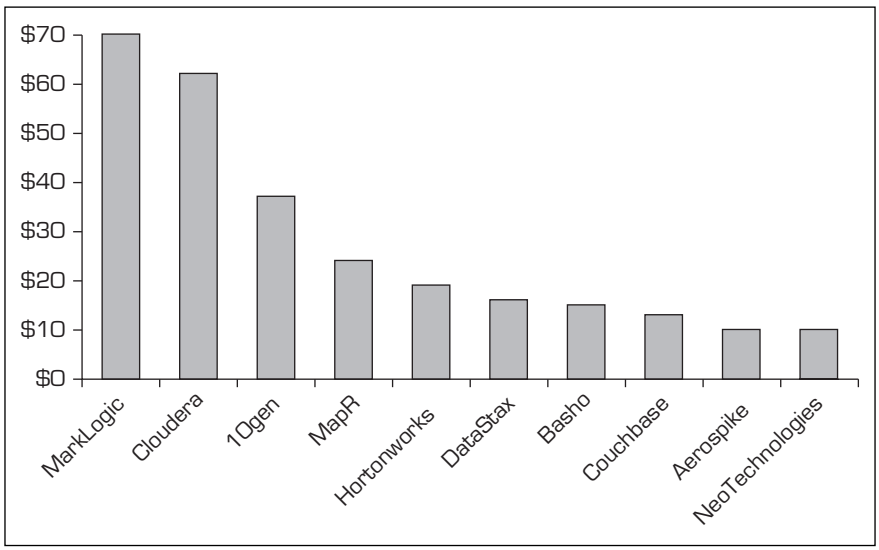

FIGURE 13.9 Top 10 Big Data Vendors with Primary Focus on Hadoop. *Source:* **wikibon.org**.

TECHNOLOGY INSIGHTS 13.4 How to Succeed with Big Data

What a year 2012 was for Big Data! From the White House to your house, it's hard to find an organization or consumer who has less data today than a year ago. Database options proliferate, and business intelligence evolves to a new era of organization-wide analytics. And everything's mobile. Organizations that successfully adapt their data architecture and processes to address the three characteristics of Big Data—volume, variety, and velocity—are improving operational efficiency, growing revenues, and empowering new business models. With all the attention organizations are placing on innovating around data, the rate of change will only increase. So what should companies do to succeed with Big Data? Here are some of the industry testaments:

1. ***Simplify.*** It is hard to keep track of all of the new database vendors, open source projects, and Big Data service providers. It will even be more crowded and complicated in the years ahead. Therefore, there is a need for simplification. It is essential to take a strategic approach by extending your relational and online transaction processing (OLTP) systems to one or more of the new on-premise, hosted, or service-based database options that best reflect the needs of your industry and your organization, and then picking a real-time business intelligence platform that supports direct connections to many databases and file formats. Choosing the best mix of solution alternatives for every project (between connecting live to fast databases and importing data extracts into an in-memory analytics engine to offset the performance of slow or overburdened databases) is critical to the success of any Big Data projects. For instance, eBay's Big Data analytics architecture comprises Teradata (one of the most popular data warehousing companies), Hadoop (most promising solution to Big Data challenge), and Tableau (one of the prolific visual analytics solution providers). eBay employees can visualize insights from more than 52 petabytes of data. eBay uses a visual analytics solution by Tableau to analyze search relevance and quality of the **eBay.com** site; monitor the latest customer feedback and meter sentiments on **eBay.com**; and achieve operational reporting for the data warehouse systems, all of which helped an analytic culture flourish within eBay.
2. ***Coexist.*** Using the strengths of each database platform and enabling them to coexist in your organization's data architecture is essential. There is ample literature that talks about the necessity of maintaining and nurturing the coexistence of traditional data warehouses with the capabilities of new platforms.
3. ***Visualize.*** According to leading analytics research companies like Forrester and Gartner, enterprises find advanced data visualization platforms to be essential tools that enable them to monitor business, find patterns, and take action to avoid threats and snatch opportunities. Visual analytics help organizations uncover trends, relationships, and anomalies by visually shifting through very large quantities of data. A visual analysis experience has certain characteristics. It allows you to do two things at any moment:
   - Instantly change what data you are looking at. This is important because different questions require different data.
   - Instantly change the way you are looking at it. This is important because each view may answer different questions.

   This combination creates the exploratory experience required for anyone to answer questions quickly. In essence, visualization becomes a natural extension of your experimental thought process.
4. ***Empower.*** Big Data and self-service business intelligence go hand in hand, according to Aberdeen Group's recently published "Maximizing the Value of Analytics and Big Data." Organizations with Big Data are over 70 percent more likely than other organizations to have BI/BA projects that are driven primarily by the business community, not by the IT group. Across a range of uses—from tackling new business problems, developing entirely new products and services, finding actionable intelligence in less than an hour, and blending data from disparate sources—Big Data has fired the imagination of what is possible through the application of analytics.
5. ***Integrate.*** Integrating and blending data from disparate sources for your organization is an essential part of Big Data analytics. Organizations that can blend different relational,

semistructured, and raw data sources in real time, without expensive up-front integration costs, will be the ones that get the best value from Big Data. Once integrated and blended, the structure of the data (e.g., spreadsheets, a database, a data warehouse, an open source file system like Hadoop, or all of them at the same time) becomes unimportant; that is, you don't need to know the details of how data is stored to ask and answer questions against it. As we saw in Application Case 13.4, the Obama campaign found a way to integrate social media, technology, e-mail databases, fundraising databases, and consumer market data to create competitive advantage.

6. ***Govern.*** Data governance has always been a challenging issue in IT, and is getting even more puzzling with the advent of Big Data. More than 80 countries have data privacy laws. The European Union (EU) defines seven "safe harbor privacy principles" for the protection of their citizens' personal data. In Singapore, the personal data protection law took effect January 2013. In the United States, Sarbanes-Oxley affects all publicly listed companies, and HIPAA (Health Insurance Portability and Accountability Act) sets national standards in healthcare. The right balance between control and experimentation varies depending on the organization and industry. Use of master data management (MDM) best practices seems to help manage the governance process.
7. ***Evangelize.*** With the backing of one or more executive sponsors, evangelists like yourself can get the ball rolling and instill a virtuous cycle: Tthe more departments in your organization that realize actionable benefits, the more pervasive analytics becomes across your organization. Fast, easy-to-use visual analytics is the key that opens the door to organization-wide analytics adoption and collaboration.

Sources: Compiled from A. Lampitt, "Big Data Visualization: A Big Deal for eBay," *InfoWorld,* December 6, 2012, **infoworld.com/d/big-data/big-data-visualization-big-deal-ebay-208589** (accessed March 2013); Tableau white paper, **cdnlarge.tableausoftware.com/sites/default/files/whitepapers/7-tips-to-succeed-with-big-data-in-2013.pdf** (accessed January 2013).

Application Case 13.6

Creditreform Boosts Credit Rating Quality with Big Data Visual Analytics

Founded as a credit agency in Mainz, Germany, in 1879, Creditreform has grown to now serve more than 163,000 members from 177 offices across Europe and China as one of the leading international providers of business information and receivables management services. Creditreform provides a comprehensive spectrum of integrated credit risk management solutions and services worldwide, provides members with more than 16 million commercial reports a year, and helps them recover billions in outstanding debts.

Challenge

Via its online database, Creditreform makes more than 24 million credit reports from 26 countries in Europe and from China that are available around the clock. Using high-performance solutions Creditreform wants to quickly detect anomalies and relationships within those high data volumes and present results in easy-to-read graphics. Already Germany's top provider of quality business information and debt collection services, Creditreform wants to maintain its leadership and widen its market lead through better and faster analytics.

Solution and the Results

Creditreform decided to use SAS Visual Analytics to simplify the analytics process, so that every Creditreform employee can use the software to make smart decisions without needing extensive training. The new high-performance solution, obtained from one of the business analytics leaders in the market place (SAS Institute), makes Creditreform better at providing the highest quality financial information and credit ratings to its client businesses.

"SAS Visual Analytics makes it faster and easier for our analysts to detect correlations in our business data," said Bernd Bütow, managing director at Creditreform. "That, in turn, improves the quality and forecasting accuracy of our credit ratings."

"Creditreform saw SAS Visual Analytics as a compelling solution," remarked Mona Beck, financial services sales director at SAS Germany. "SAS Visual Analytics advances business analytics by combining Big Data analysis with excellent usability, making it a breeze to represent data graphically. As a company known for providing top-quality information on businesses, Creditreform is a perfect match for the very latest in business analytics technology."

SAS Visual Analytics is a high-performance, in-memory solution for exploring massive amounts of data very quickly. Users can explore all data, execute analytic correlations on billions of rows of data in just minutes or seconds, and visually present results. With SAS Visual Analytics, executives can make quicker, better decisions with instant access, via PC or tablet, to insights based on the latest data. By integrating corporate and consumer data, bank executives gain real-time insights for risk management, customer development, product marketing, and financial management.

Questions for Discussion

1. How did Creditreform boost credit rating quality with Big Data and visual analytics?
2. What were the challenges, proposed solution, and initial results?

Source: SAS, Customer Stories, "With SAS, Creditreform Boosts Credit Rating Quality, Forecasting: SAS Visual Analytics, High-Performance Analytics Speed Decisions, Increase Efficiency," **sas.com/news/preleases/banking-visual-analytics.html** (accessed March 2013).

SECTION 13.7 REVIEW QUESTIONS

1. What is special about the Big Data vendor landscape? Who are the big players?
2. How do you think the Big Data vendor landscape will change in the near future? Why?
3. What is the role of visual analytics in the world of Big Data?

13.8 BIG DATA AND STREAM ANALYTICS

Along with volume and variety, as we have seen earlier in this chapter, one of the key characteristics that define Big Data is velocity, which refers to the speed at which the data is created and streamed into the analytics environment. Organizations are looking for new means to process this streaming data as it comes in to react quickly and accurately to problems and opportunities to please their customers and to gain competitive advantage. In situations where data streams in rapidly and continuously, traditional analytics approaches that work with previously accumulated data (i.e., data at arrest) often either arrive at the wrong decisions because of using too much out-of-context data, or they arrive at the correct decisions but too late to be of any use to the organization. Therefore it is critical for a number of business situations to analyze the data soon after it is created and/or as soon as it is streamed into the analytics system.

The presumption that the vast majority of modern-day businesses are currently living by is that it is important and critical to record every piece of data because it might contain valuable information now or sometime in the near future. However, as long as the number of data sources increases, the "store-everything" approach becomes harder and harder and, in some cases, not even feasible. In fact, despite technological advances, current total storage capacity lags far behind the digital information being generated in the world. Moreover, in the constantly changing business environment, real-time detection of meaningful changes in data as well as of complex pattern variations within a given short time window are essential in order to come up with the actions that better fit with the new environment. These facts become the main triggers for a paradigm that we call *stream analytics*. The stream analytics paradigm was born as an answer to these challenges, namely, the unbounded flows of data that cannot be permanently stored in order to be subsequently analyzed, in a timely and efficient manner, and complex pattern variations that need to be detected and acted upon as soon as they happen.

Stream analytics (also called *data in-motion analytics* and *real-time data analytical,* among others) is a term commonly used for the analytic process of extracting actionable information from continuously flowing/streaming data. A stream can be defined as a continuous sequence of data elements (Zikopoulos et al., 2013). The data elements in a stream are often called *tuples*. In a relational database sense, a tuple is similar to a row of data (a record, an object, an instance). However in the context of semistructured or unstructured data, a tuple is an abstraction that represents a package of data, which can be characterized as a set of attributes for a given object. If a tuple by itself is not sufficiently informative for analysis, a correlation—or other collective relationships among tuples are needed—then a window of data that includes a set of tuples is used. A window of data is a finite number/sequence of tuples, where the windows are continuously updated as new data become available. The size of the window is determined based on the system being analyzed. Stream analytics is becoming increasingly more popular because of two things. First, time-to-action has become an ever decreasing value, and second, we have the technological means to capture and process the data while it is being created.

Some of the most impactful applications of stream analytics were developed in the energy industry, specifically for smart grid (electric power supply chain) systems. The new smart grids are capable of not only real-time creation and processing of multiple streams of data in order to determine optimal power distribution to fulfill real customer needs, but also generating accurate short-term predictions aimed at covering unexpected demand and renewable energy generation peaks. Figure 13.10 shows a depiction of a generic use case for streaming analytics in energy industry (a typical smart grid application). The goal is to accurately predict electricity demand and production in real time by using streaming data that is coming from smart meters, production system sensors, and meteorological models. The ability to predict near future consumption/production trends and detect anomalies in real time can be used to optimize supply decisions (how much to produce, what sources of production to use, optimally adjust production capacities) as well as to adjust smart meters to regulate consumption and favorable energy pricing.

Stream Analytics Versus Perpetual Analytics

The terms "streaming" and "perpetual" probably sound like the same thing to most people, and in many cases they are used synonymously. However, in the context of intelligent systems, there is a difference (Jonas, 2007). Streaming analytics involves applying transaction-level logic to real-time observations. The rules applied to these observations take into account previous observations as long as they occurred in the prescribed window; these windows have some arbitrary size (e.g., last 5 seconds, last 10,000 observations, etc.). **Perpetual analytics**, on the other hand, evaluates every incoming observation against all prior observations, where there is no window size. Recognizing how the new observation relates to all prior observations enables the discovery of real-time insight.

Both streaming and perpetual analytics have their pros and cons, and their respective places in the business analytics world. For example, sometimes transactional volumes are high and the time-to-decision is too short, favoring nonpersistence and small window sizes, which translates into using streaming analytics. However, when the mission is critical and transaction volumes can be managed in real time, then perpetual analytics is a better answer. That way, one can answer questions such as "How does what I just learned relate to what I have known?" "Does this matter?" and "Who needs to know?"

Critical Event Processing

Critical event processing is a method of capturing, tracking, and analyzing streams of data to detect events (out of normal happenings) of certain types that are worthy of the effort. Complex event processing is an application of stream analytics that combines data from multiple sources to infer events or patterns of interest either before they actually

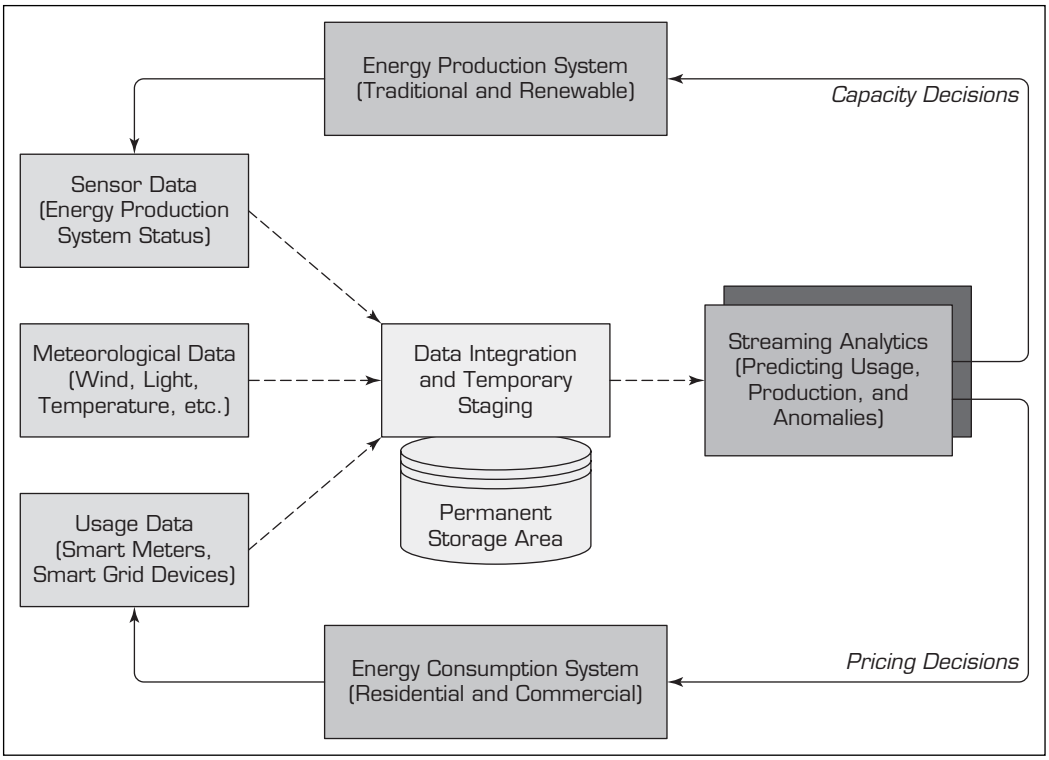

FIGURE 13.10 A Use Case of Streaming Analytics in the Energy Industry.

occur or as soon as they happen. The goal is to take rapid actions to either prevent (or mitigate the negative effects of) these events (e.g., fraud or network intrusion), or in the case of a short window of opportunity, take full advantage of the situation within the allowed time (based on user behavior on a e-commerce site, create promotional offers that they are more likely to respond to).

These critical events may be happening across the various layers of an organization such as sales leads, orders, or customer service calls. Or, more broadly, they may be news items, text messages, social media posts, stock market feeds, traffic reports, weather conditions, or other kinds of anomalies that may have a significant impact on the well-being of the organization. An event may also be defined generically as a "change of state," which may be detected as a measurement exceeding a predefined threshold of time, temperature, or some other value. Even though there is no denying the value proposition of critical event processing, one has to be selective in what to measure, when to measure, and how often to measure. Because of the vast amount of information available about events, which is sometimes referred to as the *event cloud,* there is a possibility of overdoing it, in which case as opposed to helping the organization, it may hurt the operational effectiveness.

Data Stream Mining

Data stream mining, as an enabling technology for stream analytics, is the process of extracting novel patterns and knowledge structures from continuous, rapid data records. As we have seen in the data mining chapter (Chapter 5), traditional data mining methods require the data to be collected and organized in a proper file format, and then processed in a recursive manner to learn the underlying patterns. In contrast, a data stream is a continuous flow of ordered sequence of instances that in many applications of data stream mining can be read/processed only once or a small number of times using limited computing and storage capabilities. Examples of data streams include sensor data, computer network traffic, phone

conversations, ATM transactions, web searches, and financial data. Data stream mining can be considered a subfield of data mining, machine learning, and knowledge discovery.

In many data stream mining applications, the goal is to predict the class or value of new instances in the data stream given some knowledge about the class membership or values of previous instances in the data stream. Specialized machine learning techniques (mostly derivative of traditional machine learning techniques) can be used to learn this prediction task from labeled examples in an automated fashion. An example of such a prediction method is developed by Delen et al. (2005), where they gradually built and refined a decision tree model by using a subset of the data at a time.

SECTION 13.8 REVIEW QUESTIONS

1. What is a stream (in Big Data world)?
2. What are the motivations for stream analytics?
3. What is stream analytics? How does it differ from regular analytics?
4. What is critical event processing? How does it relate to stream analytics?
5. Define data stream mining? What are the additional challenges that are posed?

13.9 APPLICATIONS OF STREAM ANALYTICS

Because of its power to create insight instantly, helping decision makers to be on top of events as they unfold and allowing organizations to address issues before they become problems, the use of streaming analytics is on an exponentially increasing trend. Following are some of the application areas that have already benefited from stream analytics.

e-Commerce

Companies like Amazon and eBay (among many others) are trying to make the most out of the data that they collect while the customer is on their Web sites. Every page visit, every product looked at, every search conducted, and every click made is recorded and analyzed to maximize the value gained from a user's visit. If done quickly, analysis of such a stream of data can turn browsers into buyers and buyers into shopaholics. When we visit an e-commerce Web site, even one where we are not a member, after a few clicks here and there we start to get very interesting product and bundle price offers. Behind the scenes, advanced analytics are crunching the real-time data coming from our clicks, and the clicks of thousands of others, to "understand" what it is that we are interested in (in some cases, even we do not know that) and make the most of that information by creative offerings.

Telecommunications

The volume of data that comes from call detail records (CDR) for telecommunications companies is astounding. Although this information has been used for billing purposes for quite some time now, there is a wealth of knowledge buried deep inside this Big Data that the telecommunications companies are just now realizing to tap. For instance, CDR data can be analyzed to prevent churn by identifying networks of callers, influencers, leaders, and followers within those networks and proactively acting on this information. As we all know, influencers and leaders have the effect of changing the perception of the followers within their network toward the service provider, either positively or negatively. Using social network analysis techniques, telecommunication companies are identifying the leaders and influencers and their network participants to better manage their customer base. In addition to churn analysis, such information can also be used to recruit new members and maximize the value of the existing members.

Continuous stream of data that comes from CDR can be combined with social media data (sentiment analysis) to assess the effectiveness of marketing campaigns. Insight

gained from these data streams can be used to rapidly react to adverse effects (which may lead to loss of customers) or boost the impact of positive effects (which may lead to maximizing purchases of existing customers and recruitment of new customers) observed in these campaigns. Furthermore, the process of gaining insight from CDR can be replicated for data networks using Internet protocol detail records. Since most telecommunications companies provide both of these service types, a holistic optimization of all offerings and marketing campaigns could lead to extraordinary market gains. Application Case 13.7 is an example of how telecommunication companies are using stream analytics to boost customer satisfaction and competitive advantage.

Application Case 13.7

Turning Machine-Generated Streaming Data into Valuable Business Insights

This case study is about one of the largest U.S. telecommunications organizations, which offers a variety of services, including digital voice, high-speed Internet, and cable, to more than 24 million customers. As a subscription-based business, its success depends on its IT infrastructure to deliver a high-quality customer experience. When application failures or network latencies negatively impact the customer experience, they adversely impact company revenue as well. That's why this leading telecommunications organization demands robust and timely information from its operational telemetry to ensure data integrity, stability, application quality, and network efficiency.

Challenges

The environment generates over a billion daily events running on a distributed hardware/software infrastructure supporting millions of cable, online, and interactive media customers. It was overwhelming to even gather and view this data in one place, much less to perform any diagnostics, or hone in on the real-time intelligence that lives in the machine-generated data. Using time-consuming and error-prone traditional search methods, the company's roster of experts would shuffle through mountains of data to uncover issues threatening data integrity, system stability, and applications performance—all necessary components of delivering a quality customer experience.

Solution

In order to bolster operational intelligence, the company selected to work with Splunk, one of the leading analytics service providers in the area of turning machine-generated streaming data into valuable business insights. Here are some of the results.

Application troubleshooting. Before Splunk developers had to ask the operations team to FTP log files to them. And then they waited … sometimes 16+ hours to get the data they needed while the operations teams had to step away from their primary duties to assist the developers. Now, because Splunk aggregates all relevant machine data into one place, developers can be more proactive about troubleshooting code and improving the user experience. When they first deployed Splunk, they started with a simple search for 404 errors. Splunk revealed up to 1,600 404s per second for a particular service. The team identified latencies in a flash player download as the primary blocker, causing viewers to navigate away from the page without viewing any content. Just one search in Splunk has helped to boost video views by 3 percent over the last year. In a business where eyes equal dollars, that's real money to the business. Now when the applications team sees 404s spiking on custom dashboards they've built in Splunk, they can dig in to see what's happening upstream and align appropriate resources to recapture those viewers—and that revenue.

Operations. Splunk's ability to model systems and examine patterns in real time helped the operations team avoid critical downtime. Using Splunk, they spotted the potential for failure in a vendor-provided infrastructure. Modeling the proposed architecture in Splunk, they were able to predict system imbalance

(*Continued*)

Application Case 13.7 (Continued)

and how it might fail based on inability to distribute load. "My team provides guidance to our executives on mission-critical media systems and strategic systems architecture," said Matt Stevens, director of software architecture. "This is just one instance where Splunk paid for itself by helping us avoid deployment of vulnerable systems, which would inevitably result in downtime and upset customers." In day-to-day operations, teams use Splunk to identify and drill into events to identify activity patterns leading to outages. Once they've identified signatures or patterns, they create alerts to proactively avoid future problems.

Compliance. Once seen as a foe, many organizations are looking to compliance mandates as an opportunity to implement best practices in log consolidation and IT systems management. This organization is no different. As Sarbanes-Oxley (SOX) and other compliance mandates evolve, the company uses Splunk to audit its systems, generate scheduled and ad hoc reports, and share information with business executives, auditors, and partners.

Security. When you're a content provider, DNS attacks simply can't be tolerated. By consolidating logs across data centers, the security team has improved the effectiveness of its threat assessments and security monitoring. Dashboards allow analysts to detect system vulnerabilities or attacks on both its content delivery network and critical applications. Trend reports spanning long timeframes also identify recurring threats and known attackers. And alerts for bad actors trigger immediate responses.

Conclusion

No longer does the sheer volume of machine-generated data overwhelm the operations team. The more data that the company's enormous infrastructure generates, the more lurking issues and security threats are revealed. The team even seeks out historical data—going back years—to identify trends and unique patterns. As the discipline of investigating anomalies and creating alerts based on unmasked event signatures spreads throughout the IT organization, the growing knowledge base and awareness fortify the cable provider's ability to deliver continuous quality customer experiences.

Even more valuable than this situational awareness has been the predictive capability gained. When testing a new technology, the decision-making team sees how a solution will work in production—determining the potential for instability by observing reactions to varying loads and traffic patterns. Splunk's predictive analytics capabilities help this leading cable provider make the right decisions, avoiding costly delays and downtime.

Questions for Discussion

1. Why is stream analytics becoming more popular?
2. How did the telecommunication company in this case use stream analytics for better business outcomes? What additional benefits can you foresee?
3. What were the challenges, proposed solution, and initial results?

Source: Splunk, Customer Case Study, **splunk.com/view/SP-CAAAFAD** (accessed March 2013).

Law Enforcement and Cyber Security

Streams of Big Data provide excellent opportunities for improved crime prevention, law enforcement, and enhanced security. They offer unmatched potential when it comes to security applications that can be built in the space, such as real-time situational awareness, multimodal surveillance, cyber-security detection, legal wire taping, video surveillance, and face recognition (Zikopoulos et al., 2013). As an application of information assurance, enterprises can use streaming analytics to detect and prevent network intrusions, cyber attacks, and malicious activities by streaming and analyzing network logs and other Internet activity monitoring resources.

Power Industry

Because of the increasing use of smart meters, the amount of real-time data collected by power utilities is increasing exponentially. Moving from once a month to every 15 minutes (or more frequent), meter read accumulates large quantities of invaluable data for power utilities. These smart meters and other sensors placed all around the power grid are sending information back to the control centers to be analyzed in real time. Such analyses help utility companies to optimize their supply chain decision (e.g., capacity adjustments, distribution network options, real-time buying or selling) based on the up-to-the-minute consumer usage and demand patterns. Additionally, utility companies can integrate weather and other natural condition data into their analytics to optimize power generation from alternative sources (e.g., wind, solar, etc.) and to better forecast energy demand on different geographic granulations. Similar benefits also apply to other utilities such as water and natural gas.

Financial Services

Financial service companies are among the prime examples where analysis of Big Data streams can provide faster and better decisions, competitive advantage, and regulatory oversight. The ability to analyze fast-paced, high volumes of trading data at very low latency across markets and countries offers tremendous advantage to making the split-second buy/sell decisions that potentially translate into big financial gains. In addition to optimal buy/sell decisions, stream analytics can also help financial service companies in real-time trade monitoring to detect fraud and other illegal activities.

Health Sciences

Modern era medical devices (e.g., electrocardiograms and equipment that measures blood pressure, blood oxygen level, blood sugar level, body temperature, and so on) are capable of producing invaluable streaming diagnostic/sensory data at a very fast rate. Harnessing this data and analyzing it in real time offers benefits—the kind that we often call "life and death"—unlike any other field. In addition to helping healthcare companies become more effective and efficient (and hence more competitive and profitable), stream analytics is also improving patient conditions, saving lives.

Many hospital systems all around the world are developing care infrastructures and health systems that are futuristic. These systems aim to take full advantage of what the technology has to offer, and more. Using hardware devices that generate high-resolution data at a very rapid rate, coupled with super-fast computers that can synergistically analyze multiple streams of data, increases the chances of keeping patients safe by quickly detecting anomalies. These systems are meant to help human decision makers make faster and better decisions by being exposed to a multitude of information as soon as it becomes available.

Government

Governments all around the world are trying to find ways to be more efficient (via optimal use of limited resources) and effective (providing the services that people need and want). As the practices for e-government become mainstream, coupled with widespread use and access to social media, very large quantities of data (both structured and unstructured) are at the disposal of government agencies. Proper and timely use of these Big Data streams differentiates proactive and highly efficient agencies from the ones that are still using traditional methods to react to situations as they unfold. Another way in which government agencies can leverage real-time analytics capabilities is to manage natural disasters such as snowstorms, hurricanes, tornados, and wildfires through surveillance of streaming data coming from radars, sensors, and other smart detection devices. They can also use similar approaches to monitor water quality, air quality, and consumption patterns, and detect anomalies before they become significant problems. Yet another area where government

agencies use stream analytics is in traffic management in congested cities. By using the data coming from traffic flow cameras, GPS data coming from commercial vehicles, and traffic sensors embedded in roadways, agencies are able to change traffic light sequences and traffic flow lanes to ease the pain caused by traffic congestion problems.

SECTION 13.9 REVIEW QUESTIONS

1. What are the most fruitful industries for stream analytics?
2. How can stream analytics be used in e-commerce?
3. In addition to what is listed in this section, can you think of other industries and/or application areas where stream analytics can be used?
4. Compared to regular analytics, do you think stream analytics will have more (or fewer) use cases in the era of Big Data analytics? Why?

Chapter Highlights

- Big Data means different things to people with different backgrounds and interests.
- Big Data exceeds the reach of commonly used hardware environments and/or capabilities of software tools to capture, manage, and process it within a tolerable time span.
- Big Data is typically defined by three "V"s: volume, variety, velocity.
- MapReduce is a technique to distribute the processing of very large multi-structured data files across a large cluster of machines.
- Hadoop is an open source framework for processing, storing, and analyzing massive amounts of distributed, unstructured data.
- Hive is a Hadoop-based data warehousing–like framework originally developed by Facebook.
- Pig is a Hadoop-based query language developed by Yahoo!.
- NoSQL, which stands for Not Only SQL, is a new paradigm to store and process large volumes of unstructured, semistructured, and multi-structured data.
- Data scientist is a new role or a job commonly associated with Big Data or data science.
- Big Data and data warehouses are complementary (not competing) analytics technologies.
- As a relatively new area, the Big Data vendor landscape is developing very rapidly.
- Stream analytics is a term commonly used for extracting actionable information from continuously flowing/streaming data sources.
- Perpetual analytics evaluates every incoming observation against all prior observations.
- Critical event processing is a method of capturing, tracking, and analyzing streams of data to detect certain events (out of normal happenings) that are worthy of the effort.
- Data stream mining, as an enabling technology for stream analytics, is the process of extracting novel patterns and knowledge structures from continuous, rapid data records.

Key Terms

Big Data
Big Data analytics
critical event processing
data scientist
data stream mining
Hadoop
Hadoop Distributed File System (HDFS)
Hive
MapReduce
NoSQL
perpetual analytics
Pig
RFID
stream analytics
social media

Questions for Discussion

1. What is Big Data? Why is it important? Where does Big Data come from?
2. What do you think the future of Big Data will be? Will it leave its popularity to something else? If so, what will it be?
3. What is Big Data analytics? How does it differ from regular analytics?
4. What are the critical success factors for Big Data analytics?

5. What are the big challenges that one should be mindful of when considering implementation of Big Data analytics?
6. What are the common business problems addressed by Big Data analytics?
7. Who is a data scientist? What makes them so much in demand?
8. What are the common characteristics of data scientists? Which one is the most important?
9. In the era of Big Data, are we about to witness the end of data warehousing? Why?
10. What are the use cases for Big Data/Hadoop and data warehousing/RDBMS?
11. What is stream analytics? How does it differ from regular analytics?
12. What are the most fruitful industries for stream analytics? What is common to those industries?
13. Compared to regular analytics, do you think stream analytics will have more (or fewer) use cases in the era of Big Data analytics? Why?

Exercises

Teradata University Network (TUN) and Other Hands-On Exercises

1. Go to **teradatauniversitynetwork.com** and search for case studies. Read cases and white papers that talk about Big Data analytics. What is the common theme in those case studies?
2. At **teradatauniversitynetwork.com**, find the SAS Visual Analytics white papers, case studies, and hands-on exercises. Carry out the visual analytics exercises on large data sets and prepare a report to discuss your findings.
3. At **teradatauniversitynetwork.com**, go to the podcasts library. Find podcasts about Big Data analytics. Summarize your findings.
4. Go to **teradatauniversitynetwork.com** and search for BSI videos that talk about Big Data. Review these BSI videos and answer case questions related to them.
5. Go to the **teradata.com** and/or **asterdata.com** Web sites. Find at least three customer case studies on Big Data, and write a report where you discuss the commonalities and differences of these cases.
6. Go to **IBM.com**. Find at least three customer case studies on Big Data, and write a report where you discuss the commonalities and differences of these cases.
7. Go to **cloudera.com.** Find at least three customer case studies on Hadoop implementation, and write a report where you discuss the commonalities and differences of these cases.
8. Go to **MapR.com**. Find at least three customer case studies on Hadoop implementation, and write a report where you discuss the commonalities and differences of these cases.
9. Go to **hortonworks.com**. Find at least three customer case studies on Hadoop implementation, and write a report where you discuss the commonalities and differences of these cases.
10. Go to **marklogic.com**. Find at least three customer case studies on Hadoop implementation, and write a report where you discuss the commonalities and differences of these cases.
11. Go to **youtube.com**. Search for videos on Big Data computing. Watch at least two. Summarize your findings.
12. Go to **google.com/scholar** and search for articles on stream analytics. Find at least three related articles. Read and summarize your findings.
13. Enter **google.com/scholar** and search for articles on data stream mining. Find at least three related articles. Read and summarize your findings.
14. Search the job search sites like **monster.com**, **careerbuilder.com**, and so forth. Find at least five job postings for data scientist. Identify the key characteristics and skills expected from the applicants.
15. Enter **google.com/scholar** and search for articles that talk about Big Data versus data warehousing. Find at least five articles. Read and summarize your findings.

End-of-Chapter Application Case

Discovery Health Turns Big Data into Better Healthcare

Introduction—Business Context

Founded in Johannesburg more than 20 years ago, Discovery now operates throughout the country, with offices in most major cities to support its network of brokers. It employs more than 5,000 people and offers a wide range of health, life and other insurance services.

In the health sector, Discovery prides itself on offering the widest range of health plans in the South African market. As one of the largest health scheme administrators in the country, its is able to keep member contributions as low as possible, making it more affordable to a wider cross-section of the population. On a like-for-like basis, Discovery's plan contributions are as much as 15 percent lower than those of any other South African medical scheme.

Business Challenges

When your health schemes have 2.7 million members, your claims system generates a million new rows of data daily,

and you are using three years of historical data in your analytics environment, how can you identify the key insights that your business and your members' health depend on?

This was the challenge facing Discovery Health, one of South Africa's leading specialist health scheme administrators. To find the needles of vital information in the big data haystack, the company not only needed a sophisticated data-mining and predictive modeling solution, but also an analytics infrastructure with the power to deliver results at the speed of business.

Solutions—Big Data Analytics

By building a new accelerated analytics landscape, Discovery Health is now able to unlock the true potential of its data for the first time. This enables the company to run three years' worth of data for its 2.7 million members through complex statistical models to deliver actionable insights in a matter of minutes. Discovery is constantly developing new analytical applications, and has already seen tangible benefits in areas such as predictive modeling of members' medical needs and fraud detection.

Predicting and preventing health risks

Matthew Zylstra, Actuary, Risk Intelligence Technical Development at Discovery Health, explains: "We can now combine data from our claims system with other sources of information such as pathology results and members' questionnaires to gain more accurate insight into their current and possible future health.

"For example, by looking at previous hospital admissions, we can now predict which of our members are most likely to require procedures such as knee surgery or lower back surgery. By gaining a better overview of members' needs, we can adjust our health plans to serve them more effectively and offer better value."

Lizelle Steenkamp, Divisional Manager, Risk Intelligence Technical Development, adds: "Everything we do is an attempt to lower costs for our members while maintaining or improving the quality of care. The schemes we administer are mutual funds–non-profit organizations–so any surpluses in the plan go back to the members we administer, either through increased reserves or lowered contributions. "One of the most important ways we can simultaneously reduce costs and improve the well-being of our members is to predict and prevent health problems before they need treatment. We are using the results of our predictive modeling to design preventative programs that can help our members stay healthier."

Identifying and eliminating fraud

Estiaan Steenberg, Actuary at Discovery Health, comments: "From an analytical point of view, fraud is often a small intersection between two or more very large data-sets. We now have the tools we need to identify even the tiniest anomalies and trace suspicious transactions back to their source."

For example, Discovery can now compare drug prescriptions collected by pharmacies across the country with healthcare providers' records. If a prescription seems to have been issued by a provider, but the person fulfilling it has not visited that provider recently, it is a strong indicator that the prescription may be fraudulent. "We used to only be able to run this kind of analysis for one pharmacy and one month at a time," says Estiaan Steenberg. "Now we can run 18 months of data from all the pharmacies at once in two minutes. There is no way we could have obtained these results with our old analytics landscape."

Similar techniques can be used to identify coding errors in billing from healthcare providers–for example, if a provider "upcodes" an item to charge Discovery for a more expensive procedure than it actually performed, or "unbundles" the billing for a single procedure into two or more separate (and more expensive) lines. By comparing the billing codes with data on hospital admissions, Discovery is alerted to unusual patterns, and can investigate whenever mistakes or fraudulent activity are suspected.

The Results—Transforming Performance

To achieve this transformation in its analytics capabilities, Discovery worked with BITanium, an IBM Business Partner with deep expertise in operational deployments of advanced analytics technologies. "BITanium has provided fantastic support from so many different angles," says Matthew Zylstra. "Product evaluation and selection, software license management, technical support for developing new models, performance optimization and analyst training are just a few of the areas they have helped us with."

Discovery is an experienced user of IBM SPSS® predictive analytics software, which forms the core of its data-mining and predictive analytics capability. But the most important factor in embedding analytics in day-to-day operational decision-making has been the recent introduction of the IBM PureData™ System for Analytics, powered by Netezza® technology–an appliance that transforms the performance of the predictive models.

"BITanium ran a proof of concept for the solution that rapidly delivered useful results," says Lizelle Steenkamp. "We were impressed with how quickly it was possible to achieve tremendous performance gains." Matthew Zylstra adds: "Our data warehouse is so large that some queries used to take 18 hours or more to process–and they would often crash before delivering results. Now, we see results in a few minutes, which allows us to be more responsive to our customers and thus provide better care."

From an analytics perspective, the speed of the solution gives Discovery more scope to experiment and optimize its models. "We can tweak a model and re-run the analysis in a few minutes," says Matthew Zylstra "This means we can do more development cycles faster–and release new analyses to the business in days rather than weeks."

From a broader business perspective, the combination of SPSS and PureData technologies gives Discovery the ability to put actionable data in the hands of its decision-makers faster. "In sensitive areas such as patient care and fraud investigation, the details are everything," concludes Lizelle Steenkamp. "With the IBM solution, instead of inferring a 'near enough' answer from high-level summaries of data, we can get the right information,

develop the right models, ask the right questions, and provide accurate analyses that meet the precise needs of the business."

Looking to the future, Discovery is also starting to analyze unstructured data, such as text-based surveys and comments from online feedback forms.

About BITanium

BITanium believes that the truth lies in data. Data does not have its own agenda, it does not lie, it is not influenced by promotions or bonuses. Data contains the only accurate representation of what has and is actually happening within a business. BITanium also believes that one of the few remaining differentiators between mediocrity and excellence is how a company uses its data.

BITanium is passionate about using technology and mathematics to find patterns and relationships in data. These patterns provide insight and knowledge about problems, transforming them into opportunities. To learn more about services and solutions from BITanium, please visit bitanium.co.za.

About IBM Business Analytics

IBM Business Analytics software delivers data-driven insights that help organizations work smarter and outperform their peers. This comprehensive portfolio includes solutions for business intelligence, predictive analytics and decision management, performance management, and risk management. Business Analytics solutions enable companies to identify and visualize trends and patterns in areas, such as customer analytics, that can have a profound effect on business performance. They can compare scenarios, anticipate potential threats and opportunities, better plan, budget and forecast resources, balance risks against expected returns and work to meet regulatory requirements. By making analytics widely available, organizations can align tactical and strategic decision-making to achieve business goals. For more information, you may visit ibm.com/business-analytics.

Questions for the End-of-Chapter Application Case

1. How big is Big Data for Discovery Health?
2. What big data sources did Discovery Health use for their analytic solutions?
3. What were the main data/analytics challenges Discovery Health was facing?
4. What were the main solutions they have produced?
5. What were the initial results/benefits? What do you think will be the future of Big Data analytics at Discovery?

Source: **IBM Customer Story, "Discovery Health turns big data into better healthcare"** public.dhe.ibm.com/common/ssi/ecm/en/ytc03619zaen/YTC03619ZAEN.PDF (accessed October 2013).

References

Awadallah, A., and D. Graham. (2012). "Hadoop and the Data Warehouse: When to Use Which." White paper by Cloudera and Teradata. **teradata.com/white-papers/Hadoop-and-the-Data-Warehouse-When-to-Use-Which** (accessed March 2013).

Davenport, T. H., and D. J. Patil. (2012, October). "Data Scientist." *Harvard Business Review,* pp. 70–76.

Dean, J., and S. Ghemawat. (2004). "MapReduce: Simplified Data Processing on Large Clusters." **research.google.com/archive/mapreduce.html** (accessed March 2013).

Delen, D., M. Kletke, and J. Kim. (2005). "A Scalable Classification Algorithm for Very Large Datasets." *Journal of Information and Knowledge Management,* Vol. 4, No. 2, pp. 83–94.

Ericsson. (2012). "Proof of Concept for Applying Stream Analytics to Utilities." Ericsson Labs, Research Topics, **labs.ericsson.com/blog/proof-of-concept-for-applying-stream-analytics-to-utilities** (accessed March 2013).

Issenberg, S. (2012, October 29). "Obama Does It Better" (from "Victory Lab: The New Science of Winning Campaigns"), *Slate*.

Jonas, J. (2007). "Streaming Analytics vs. Perpetual Analytics (Advantages of Windowless Thinking)." **jeffjonas.typepad.com/jeff_jonas/2007/04/streaming_analy.html** (accessed March 2013).

Kelly, L. (2012). "Big Data: Hadoop, Business Analytics and Beyond." **wikibon.org/wiki/v/Big_Data:_Hadoop,_Business_Analytics_and_Beyond** (accessed January 2013).

Kelly, L. (2013). "Big Data Vendor Revenue and Market Forecast 2012–2017." **wikibon.org/wiki/v/Big_Data_Vendor_Revenue_and_Market_Forecast_2012-2017** (accessed March 2013).

Romano, L. (2012, June 9). "Obama's Data Advantage." *Politico*.

Russom, P. (2013). "Busting 10 Myths about Hadoop: The Big Data Explosion." TDWI's *Best of Business Intelligence,* Vol. 10, pp. 45–46.

Samuelson, D. A. (2013, February). "Analytics: Key to Obama's Victory." INFORMS' *ORMS Today,* pp. 20–24.

Scherer, M. (2012, November 7). "Inside the Secret World of the Data Crunchers Who Helped Obama Win." *Time*.

Shen, G. (2013, January–February). "Big Data, Analytics, and Elections." INFORMS' *Analytics Magazine*.

Watson, H. (2012). "The Requirements for Being an Analytics-Based Organization." *Business Intelligence Journal,* Vol. 17, No. 2, pp. 42–44.

Watson, H., R. Sharda, and D. Schrader. (2012). "Big Data and How to Teach It." Workshop at AMCIS. Seattle, WA.

White, C. (2012). "MapReduce and the Data Scientist." Teradata Aster white paper. **teradata.com/white-paper/MapReduce-and-the-Data-Scientist** (accessed February 2013).

Zikopoulos, P., D. DeRoos, K. Parasuraman, T. Deutsch, D. Corrigan, and J. Giles. (2013). *Harness the Power of Big Data*. New York: MacGraw Hill Publishing.

CHAPTER

14

Business Analytics: Emerging Trends and Future Impacts

LEARNING OBJECTIVES

- Explore some of the emerging technologies that may impact analytics, BI, and decision support
- Describe how geospatial and location-based analytics are assisting organizations
- Describe how analytics are powering consumer applications and creating a new opportunity for entrepreneurship for analytics
- Describe the potential of cloud computing in business intelligence
- Understand Web 2.0 and its characteristics as related to analytics
- Describe organizational impacts of analytics applications
- List and describe the major ethical and legal issues of analytics implementation
- Understand the analytics ecosystem to get a sense of the various types of players in the analytics industry and how one can work in a variety of roles

This chapter introduces several emerging technologies that are likely to have major impacts on the development and use of business intelligence applications. Many other interesting technologies are also emerging, but we have focused on some trends that have already been realized and others that are about to impact analytics further. Using a crystal ball is always a risky proposition, but this chapter provides a framework for analysis of emerging trends. We introduce and explain some emerging technologies and explore their current applications. We then discuss the organizational, personal, legal, ethical, and societal impacts of support systems that may affect their implementation. We conclude with a description of the analytics ecosystem. This section should help readers appreciate different career possibilities within the realm of analytics. This chapter contains the following sections:

14.1 OPENING VIGNETTE: Oklahoma Gas and Electric Employs Analytics to Promote Smart Energy Use

Oklahoma Gas and Electric (OG&E) serves over 789,000 customers in Oklahoma and Arkansas. OG&E has a strategic goal to delay building new fossil fuel generation plants until the year 2020. OG&E forecasts a daily system demand of 5,864 megawatts in 2020, a reduction of about 500 megawatts.

One of the ways to optimize this demand is to engage the consumers in managing their energy usage. OG&E has completed installation of smart meters and other devices on the electronic grid at the consumer end that enable it to capture large amounts of data. For example, currently it receives about 52 million meter reads per day. Apart from this, OG&E expects to receive close to 2 million event messages per day from its advanced metering infrastructure, data networks, meter alarms, and outage management systems. OG&E employs a three-layer information architecture involving data warehouse, improved and expanded integration and data management, and new analytics and presentation capabilities to support the Big Data flow.

With this data, OG&E has started working on consumer-oriented efficiency programs to shift the customer's usage out of peak demand cycles. OG&E is targeting customers with its smart hours plan. This plan encourages customers to choose a variety of rate options sent via phone, text, or e-mail. These rate options offer attractive summer rates for all other hours apart from the peak hours of 2 P.M. to 7 P.M. OG&E is making an investment in customers by supplying a communicating thermostat that will respond to the price signals sent by OG&E and help customers in managing their utility consumption. OG&E also educates its customers on their usage habits by providing 5-minute interval data every 15 minutes to the demand-responsive customers.

OG&E has developed consumer analytics and customer segmentation analytics that will enhance their understanding about individuals' responses to the price signals and identify the best customers to be targeted with specific marketing campaigns. It also uses demand-side management analytics for peak load management/load shed. With Teradata's platform, OG&E has combined its smart meter data, outage data, call center data, rate data, asset data, price signals, billing, and collections into one integrated data platform. The platform also incorporates geospatial mapping of the integrated data using the in-database geospatial analytics that add onto the OG&E's dynamic segmentation capabilities.

Using geospatial mapping and visual analytics, OG&E now views a near–real-time version of data about its energy-efficient prospects spread over geographic areas and comes up with marketing initiatives that are most suitable for these customers. OG&E now has an easy way to narrow down to the specific customers in a geographic region based on their meter usage; OG&E can also find noncommunicating smart meters. Furthermore, OG&E can track the outage, with the deployed crew supporting outages as well as the weather overlay of their services. This combination of filed infrastructure, geospatial data, enterprise data warehouse, and analytics has enabled OG&E to manage its customer demand in such a way that it can optimize its long-term investments.

QUESTIONS FOR THE OPENING VIGNETTE

1. Why perform consumer analytics?
2. What is meant by dynamic segmentation?
3. How does geospatial mapping help OG&E?
4. What types of incentives might the consumers respond to in changing their energy use?

WHAT WE CAN LEARN FROM THIS VIGNETTE

Many organizations are now integrating the data from the different internal units and turning toward analytics to convert the integrated data into value. The ability to view the operations/customer-specific data using in-database geospatial analytics gives organizations a broader perspective and aids in decision making.

Sources: **Teradata.com**, "Utilities Analytic Summit 2012 Oklahoma Gas & Electric," **teradata.com/video/Utilities-Analytic-Summit-2012-Oklahoma-Gas-and-Electric** (accessed March 2013); **ogepet.com**, "Smart Hours," **ogepet.com/programs/smarthours.aspx** (accessed March 2013); **IntellidentUtility.com**, "OGE's Three-Tiered Architecture Aids Data Analysis," **intelligentutility.com/article/12/02/oges-three-tiered-architecture-aids-data-analysis&utm_medium=eNL&utm_campaign=IU_DAILY2&utm_term=Original-Magazine** (accessed March 2013).

14.2 LOCATION-BASED ANALYTICS FOR ORGANIZATIONS

This goal of this chapter is to illustrate the potential of new technologies when innovative uses are developed by creative minds. Most of the technologies described in this chapter are nascent and have yet to see widespread adoption. Therein lies the opportunity to create the next "killer" application. For example, use of RFID and sensors is growing, with each company exploring its use in supply chains, retail stores, manufacturing, or service operations. The chapter argues that with the right combination of ideas, networking, and applications, it is possible to develop creative technologies that have the potential to impact a company's operations in multiple ways, or to create entirely new markets and make a major difference to the world. We also study the analytics ecosystem to better understand which companies are the players in this industry.

Thus far, we have seen many examples of organizations employing analytical techniques to gain insights into their existing processes through informative reporting, predictive analytics, forecasting, and optimization techniques. In this section, we learn about a critical emerging trend—incorporation of location data in analytics. Figure 14.1 gives our classification of location-based analytic applications. We first review applications that make use of static location data that is usually called *geospatial data*. We then examine the explosive growth of applications that take advantage of all the location data being generated by today's devices. This section focuses on analytics applications that are being developed by organizations to make better decisions in managing operations (as was illustrated in the opening vignette), targeting customers, promotions, and so forth. In the following section we will explore analytics applications that are being developed to be used directly by a consumer, some of which also take advantage of the location data.

Geospatial Analytics

A consolidated view of the overall performance of an organization is usually represented through the visualization tools that provide actionable information. The information may include current and forecasted values of various business factors and key performance indicators (KPIs). Looking at the key performance indicators as overall numbers via

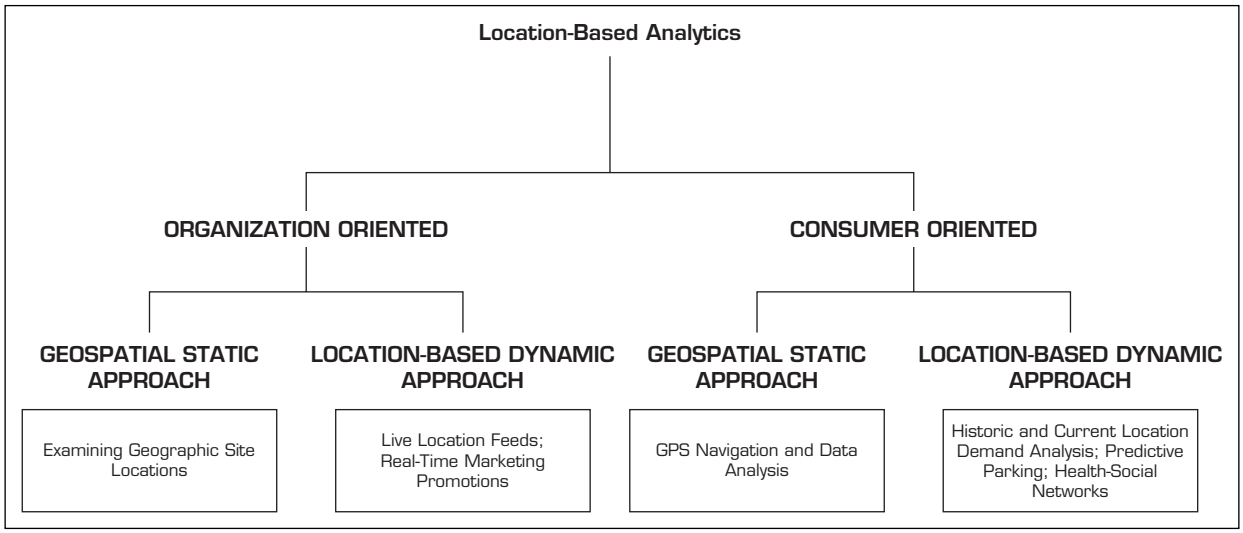

FIGURE 14.1 Classification of Location-Based Analytics Applications.

various graphs and charts can be overwhelming. There is a high risk of missing potential growth opportunities or not identifying the problematic areas. As an alternative to simply viewing reports, organizations employ visual maps that are geographically mapped and based on the traditional location data, usually grouped by the postal codes. These map-based visualizations have been used by organizations to view the aggregated data and get more meaningful location-based insights. Although this approach has advantages, the use of postal codes to represent the data is more of a static approach suitable for achieving a higher level view of things.

The traditional location-based analytic techniques using geocoding of organizational locations and consumers hampers the organizations in understanding "true location-based" impacts. Locations based on postal codes offer an aggregate view of a large geographic area. This poor granularity may not be able to pinpoint the growth opportunities within a region. The location of the target customers can change rapidly. An organization's promotional campaigns might not target the right customers. To address these concerns, organizations are embracing location and spatial extensions to analytics (Gnau, 2010). Addition of location components based on latitudinal and longitudinal attributes to the traditional analytical techniques enables organizations to add a new dimension of "where" to their traditional business analyses, which currently answer questions of "who," "what," "when," and "how much."

Location-based data are now readily available from geographic information systems (GIS). These are used to capture, store, analyze, and manage the data linked to a location using integrated sensor technologies, global positioning systems installed in smartphones, or through radio-frequency identification deployments in retail and healthcare industries.

By integrating information about the location with other critical business data, organizations are now creating location intelligence (LI) (Krivda, 2010). LI is enabling organizations to gain critical insights and make better decisions by optimizing important processes and applications. Organizations now create interactive maps that further drill down to details about any location, offering analysts the ability to investigate new trends and correlate location-specific factors across multiple KPIs. Analysts in the organizations can now pinpoint trends and patterns in revenues, sales, and profitability across geographical areas.

By incorporating demographic details into locations, retailers can determine how sales vary by population level and proximity to other competitors; they can assess the

demand and efficiency of supply chain operations. Consumer product companies can identify the specific needs of the customers and customer complaint locations, and easily trace them back to the products. Sales reps can better target their prospects by analyzing their geography (Krivda, 2010).

Integrating detailed global intelligence, real-time location information, and logistics data in a visual, easy-to-access format, U.S. Transportation Command (USTRANSCOM) could easily track the information about the type of aircraft, maintenance history, complete list of crew, the equipment and supplies on the aircraft, and location of the aircraft. Having this information will enable it to make well-informed decisions and coordinate global operations, as noted in Westholder (2010).

Additionally, with location intelligence, organizations can quickly overlay weather and environmental effects and forecast the level of impact on critical business operations. With technology advancements, geospatial data is now being directly incorporated in the enterprise data warehouses. Location-based in-database analytics enable organizations to perform complex calculations with increased efficiency and get a single view of all the spatially oriented data, revealing the hidden trends and new opportunities. For example, Teradata's data warehouse supports the geospatial data feature based on the SQL/MM standard. The geospatial feature is captured as a new geometric data type called ST_GEOMETRY. It supports a large spectrum of shapes, from simple points, lines, and curves to complex polygons in representing the geographic areas. They are converting the nonspatial data of their operating business locations by incorporating the latitude and longitude coordinates. This process of geocoding is readily supported by service companies like NAVTEQ and Tele Atlas, which maintain worldwide databases of addresses with geospatial features and make use of address-cleansing tools like Informatica and Trillium, which support mapping of spatial coordinates to the addresses as part of extract, transform, and load functions.

Organizations across a variety of business sectors are employing geospatial analytics. We will review some examples next. Sabre Airline Solutions' application, Traveler Security, uses a geospatial-enabled dashboard that alerts the users to assess the current risks across global hotspots displayed in interactive maps. Using this, airline personnel can easily find current travelers and respond quickly in the event of any travel disruption. Application Case 14.1 provides an example of how location-based information was used in making site selection decisions in expanding a company's footprint.

Application Case 14.1

Great Clips Employs Spatial Analytics to Shave Time in Location Decisions

Great Clips, the world's largest and fastest growing salon, has more than 3,000 salons throughout United States and Canada. Great Clips' franchise success depends on a growth strategy that is driven by rapidly opening new stores in the right locations and markets. The company needed to analyze the locations based on the requirements for a potential customer base, demographic trends, and sales impact on existing franchises in the target location. Choosing a good site is of utmost importance. The current processes took a long time to analyze a single site and a great deal of labor requiring intensive analyst resources was needed to manually assess the data from multiple data sources.

With thousands of locations analyzed each year, the delay was risking the loss of prime sites to competitors and was proving expensive: Great Clips employed external contractors to cope with the delay. Great Clips created a site-selection

workflow application to evaluate the new salon site locations by using the geospatial analytical capabilities of Alteryx. A new site location was evaluated by its drive-time proximity and convenience for serving all the existing customers of the Great Clips Salon network in the area. The Alteryx-based solution also enabled evaluation of each new location based on demographics and consumer behavior data, aligning with existing Great Clip's customer profiles and the potential revenue impact of the new site on the existing sites. As a result of using location-based analytic techniques, Great Clips was able to reduce the time to assess new locations by nearly 95 percent. The labor-intensive analysis was automated and developed into a data collection analysis, mapping, and reporting application that could be easily used by the nontechnical real estate managers. Furthermore, it enabled the company to implement proactive predictive analytics for a new franchise location because the whole process now took just a few minutes.

Questions for Discussion

1. How is geospatial analytics employed at Great Clips?
2. What criteria should a company consider in evaluating sites for future locations?
3. Can you think of other applications where such geospatial data might be useful?

Source: **alteryx.com**, "Great Clips," **alteryx.com/sites/default/files/resources/files/case-study-great-chips.pdf** (accessed March 2013).

In addition to the retail transaction analysis applications highlighted here, there are many other applications of combining geographic information with other data being generated by an organization. The opening vignette described a use of such location information in understanding location-based energy usage as well as outage. Similarly, network operations and communication companies often generate massive amounts of data every day. The ability to analyze the data quickly with a high level of location-specific granularity can better identify the customer churn and help in formulating strategies specific to locations for increasing operational efficiency, quality of service, and revenue.

Geospatial analysis can enable communication companies to capture daily transactions from a network to identify the geographic areas experiencing a large number of failed connection attempts of voice, data, text, or Internet. Analytics can help determine the exact causes based on location and drill down to an individual customer to provide better customer service. You can see this in action by completing the following multimedia exercise.

A Multimedia Exercise in Analytics Employing Geospatial Analytics

Teradata University Network includes a BSI video on the case of dropped mobile calls. Please watch the video that appears on YouTube at the following link: **teradatauniversitynetwork.com/teach-and-learn/library-item/?LibraryItemId=893**

A telecommunication company launches a new line of smartphones and faces problems of dropped calls. The new rollout is in trouble, and the northeast region is the worst hit region as they compare effects of dropped calls on the profit for the geographic region. The company hires BSI to analyze the problems arising due to defects in smartphone handsets, tower coverage, and software glitches. The entire northeast region data is divided into geographic clusters, and the problem is solved by identifying the individual customer data. The BSI team employs geospatial analytics to identify the locations where network coverage was leading to the dropped calls and suggests installing

a few additional towers where the unhappy customers are located. They also work with companies on various actions that ensure that the problem is addressed.

After the video is complete, you can see how the analysis was prepared on a slide set at: **slideshare.net/teradata/bsi-teradata-the-case-of-the-dropped-mobile-calls** This multimedia excursion provides an example of a combination of geospatial analytics along with Big Data analytics that assist in better decision making.

Real-Time Location Intelligence

Many devices in use by consumers and professionals are constantly sending out their location information. Cars, buses, taxis, mobile phones, cameras, and personal navigation devices all transmit their locations thanks to network-connected positioning technologies such as GPS, wifi, and cell tower triangulation. Millions of consumers and businesses use location-enabled devices for finding nearby services, locating friends and family, navigating, tracking of assets and pets, dispatching, and engaging in sports, games, and hobbies. This surge in location-enabled services has resulted in a massive database of historical and real-time streaming location information. It is, of course, scattered and by itself not very useful. Indeed, a new name has been given to this type of data mining—**reality mining**. Eagle and Pentland (2006) appear to have been the first to use this term. Reality mining builds on the idea that these location-enabled data sets could provide remarkable real-time insight into aggregate human activity trends. For example, a British company called Path Intelligence (**pathintelligence.com**) has developed a system called Footpath that ascertains how people move within a city or even within a store. All of this is done by automatically tracking movement without any cameras recording the movement visually. Such analysis can help determine the best layout for products or even public transportation options. The automated data collection enabled through capture of cell phone and wifi hotspot access points presents an interesting new dimension in nonintrusive market research data collection and, of course, microanalysis of such massive data sets.

By analyzing and learning from these large-scale patterns of movement, it is possible to identify distinct classes of behaviors in specific contexts. This approach allows a business to better understand its customer patterns and also to make more informed decisions about promotions, pricing, and so on. By applying algorithms that reduce the dimensionality of location data, one can characterize places according to the activity and movement between them. From massive amounts of high-dimensional location data, these algorithms uncover trends, meaning, and relationships to eventually produce human-understandable representations. It then becomes possible to use such data to automatically make intelligent predictions and find important matches and similarities between places and people.

Location-based analytics finds its application in consumer-oriented marketing applications. Quiznos, a quick-service restaurant, used Sense Networks' platform to analyze location trails of mobile users based on the geospatial data obtained from the GPS and target tech-savvy customers with coupons. See Application Case 14.2. This case illustrates the emerging trend in retail space where companies are looking to improve efficiency of marketing campaigns—not just by targeting every customer based on real-time location, but by employing more sophisticated predictive analytics in real time on consumer behavioral profiles and finding the right set of consumers for the advertising campaigns.

Many mobile applications now enable organizations to target the right customer by building the profile of customers' behavior over geographic locations. For example, the Radii app takes the customer experience to a whole new level. The Radii app collects

Application Case 14.2

Quiznos Targets Customers for its Sandwiches

Quiznos, a franchised, quick-service restaurant, implemented a location-based mobile targeting campaign that targeted the tech-savvy and busy consumers of Portland, Oregon. It made use of Sense Networks' platform, which analyzed the location trails of mobile users over detailed time periods and built anonymous profiles based on the behavioral attributes of shopping habits.

With the application of predictive analytics on the user profiles, Quiznos employed location-based behavioral targeting to narrow the characteristics of users who are most likely to eat at a quick-service restaurant. Its advertising campaign ran for 2 months—November and December, 2012—and targeted only potential customers who had been to quick-service restaurants over the past 30 days, within a 3-mile radius of Quiznos, and between the ages of 18 and 34. It used relevant mobile advertisements of local coupons based on the customer's location. The campaign resulted in over 3.7 million new customers and had a 20 percent increase in coupon redemptions within the Portland area.

Questions for Discussion

1. How can location-based analytics help retailers in targeting customers?
2. Research similar applications of location-based analytics in the retail domain.

Source: **Mobilemarketer.com**, "Quiznos Sees 20pc Boost in Coupon Redemption via Location-Based Mobile Ad Campaign," **mobilemarketer.com/cms/news/advertising/14738.html** (accessed February 2013).

information about the user's habits, interests, spending patterns, and favorite locations to understand their personality. Radii uses the Gimbal Context Awareness SDK to gather location and geospatial information. Gimbal SDK's Geofencing functionality enables Radii to pick up the user's interests and habits based on the time they spend at a location and how often they visit it. Depending on the number of users who visit a particular location, and based on their preferences, Radii assigns a personality to that location, which changes based on which type of user visits the location, and their preferences. New users are given recommendations that are closer to their personality, making this process highly dynamic.

Users who sign up for Radii receive 10 "Radii," which is their currency. Users can use this currency at select locations to get discounts and special offers. They can also get more Radii by inviting their friends to use the app. Businesses who offer these discounts pay Radii for bringing customers to their location, as this in turn translates into more business. For every Radii exchanged between users, Radii is paid a certain amount. Radii thus creates a new direct marketing platform for business and enhances the customer experience by providing recommendations, discounts, and coupons.

Yet another extension of location-based analytics is to use augmented reality. Cachetown has introduced a location-sensing augmented reality-based game to encourage users to claim offers from select geographic locations. The user can start anywhere in a city and follow markers on the Cachetown app to reach a coupon, discount, or offer from a business. Virtual items are visible through the Cachetown app when the user points a phone's camera toward the virtual item. The user can then claim this item by clicking on it through the Cachetown app. On claiming the item, the user is given a certain free good/discount/offer from a nearby business, which he can use just by walking into their store.

Cachetown's business-facing app allows businesses to place these virtual items on a map using Google Maps. The placement of this item can be fine-tuned by using Google's Street View. Once all virtual items have been configured with information and

location, the business can submit items, after which the items are visible to the user in real time. Cachetown also provides usage analytics to the business to enable better targeting of virtual items. The virtual reality aspect of this app improves the experience of users, providing them with a "gaming"-type environment in real life. At the same time, it provides a powerful marketing platform for businesses to reach their customers better. More information on Cachetown is at **candylab.com/augmented-reality/**.

As is evident from this section, location-based analytics and ensuing applications are perhaps the most important front in the near future for organizations. A common theme in this section was the use of operational or marketing data by organizations. We will next explore analytics applications that are directly targeted at the users and sometimes take advantage of location information.

SECTION 14.2 REVIEW QUESTIONS

1. How does traditional analytics make use of location-based data?
2. How can geocoded locations assist in better decision making?
3. What is the value provided by geospatial analytics?
4. Explore the use of geospatial analytics further by investigating its use across various sectors like government census tracking, consumer marketing, and so forth.

14.3 ANALYTICS APPLICATIONS FOR CONSUMERS

The explosive growth of the apps industry for smartphone platforms (iOS, Android, Windows, Blackberry, Amazon, and so forth) and the use of analytics are also creating tremendous opportunities for developing apps that the consumers can use directly. These apps differ from the previous category in that these are meant for direct use by a consumer rather than an organization that is trying to mine a consumer's usage/purchase data to create a profile for marketing specific products or services to them. Predictably, these apps are meant for enabling consumers to do their job better. We highlight two of these in the following examples.

Sense Networks has built a mobile application called CabSense that analyzes large amounts of data from the New York City Taxi and Limousine Commission and helps New Yorkers and visitors in finding the best corners for hailing a taxi based on the person's location, day of the week, and time. CabSense rates the street corners on a 5-point scale by making use of machine-learning algorithms applied to the vast amounts of historical location points obtained from the pickups and drop-offs of all New York City cabs. Although the app does not give the exact location of cabs in real time, its data-crunching predictions enable people to get to a street corner that has the highest probability of finding a cab.

CabSense provides an interactive map based on current user location obtained from the mobile phone's GPS locator to find the best street corners for finding an open cab. It also provides a radar view that automatically points the right direction toward the best street corner. The application also allows users to plan in advance, set up date and time of travel, and view the best corners for finding a taxi. Furthermore, CabSense distinguishes New York's Yellow Cab services from the for-hire vehicles and readily prompts the users with relevant details of private service providers that can be used in case no Yellow Cabs are available.

Another transportation-related app that uses predictive analytics has been deployed in Pittsburgh, Pennsylvania. Developed in collaboration with Carnegie Mellon University, this app includes predictive capabilities to estimate parking availability. ParkPGH directs drivers to parking lots in the area where parking is available. It calculates the number of parking spaces available in 10 lots—over 5,300 spaces, and 25 percent of the garage parking in downtown Pittsburgh. Available spaces are updated every 30 seconds, keeping the driver as close to the current availability as possible. The app is also capable of predicting parking availability by the time the driver reaches the destination. Depending on historical demand and current events, the app is able to provide information on which lots will have free space by the time the driver gets to the destination. The app's underlying algorithm uses data on current events around the area—for example, a basketball game—to predict an increase in demand for parking spaces later that day, thus saving commuters valuable time searching for parking spaces in the busy city. Both of these examples show consumer-oriented examples of location-based analytics in transportation. Application Case 14.3 illustrates another consumer-oriented application, but in the health domain. There are many more health-related apps.

Application Case 14.3

A Life Coach in Your Pocket

Most people today are finding ways to stay active and healthy. Although everyone knows it's best to follow a healthy lifestyle, people often lack the motivation needed to keep them on track. 100Plus, a start-up company, has developed a personalized, mobile prediction platform called Outside that keeps users active. The application is based on the quantified self-approach, which makes use of technology to self-track the data on a person's habits, analyze it, and make personalized recommendations.

100 Plus posited that people are most likely to succeed in changing their lifestyles when they are given small, micro goals that are easier to achieve. They built Outside as a personalized product that engages people in these activities and enables them to understand the long-term impacts of short-term activities.

After the user enters basic data such as gender, age, weight, height, and the location where he or she lives, a behavior profile is built and compared with data from Practice Fusion and CDC records. A life score is calculated using predictive analytics. This score gives the estimated life expectancy of the user. Once registered, users can begin discovering health opportunities, which are categorized as "missions" on the mobile interface. These missions are specific to the places based on the user's location. Users can track activities, complete them, and get a score that is credited back to a life score. Outside also enables its users to create diverse, personalized suggestions by keeping track of photographs of them doing each activity. These can be used for suggestions to others, based on their location and preferences. A leader board allows a particular user to find how other people with similar characteristics are completing their missions and inspires the current user to resort to healthier living. In that sense it also combines social media with predictive analytics.

Today, most smartphones are equipped with accelerometers and gyroscopes to measure jerk, orientation, and sense motion. Many applications use this data to make the user's experience on the smartphone better. Data on accelerometer and gyroscope readings is publicly available and can be used to classify various activities like walking, running, lying down, and climbing. Kaggle (**kaggle.com**), a platform that hosts competitions and research for predictive modeling and analytics, recently hosted a competition aimed at identifying muscle motions that may be used to predict the progression of Parkinson's disease. Parkinson's disease is caused by a failure in the central nervous system, which leads to tremors, rigidity, slowness of movement, and postural instability. The objective of the competition is to best identify markers that can lead

(*Continued*)

Application Case 14.3 (Continued)

to predicting the progression of the disease. This particular application of advanced technology and analytics is an example of how these two can come together to generate extremely useful and relevant information.

Questions for Discussion

1. Search online for other applications of consumer-oriented analytical applications.
2. How can location-based analytics help individual consumers?
3. How can smartphone data be used to predict medical conditions?
4. How is ParkPGH different from a "parking space–reporting" app?

Source: Institute of Medicine of the National Academies, "Health Data Initiative Forum III: The Health Datapalooza," **iom.edu/Activities/PublicHealth/HealthData/2012-JUN-05/Afternoon-Apps-Demos/outside-100plus.aspx** (accessed March 2013).

Analytics-based applications are emerging not just for fun and health, but also to enhance one's productivity. For example, Cloze is an app that manages in-boxes from multiple e-mail accounts in one place. It integrates social networks with e-mail contacts to learn which contacts are important and assigns a score—a higher score for important contacts. E-mails with a higher score are shown first, thus filtering less important and irrelevant e-mails out of the way. Cloze stores the context of each conversation to save time when catching up with a pending conversation. Contacts are organized into groups based on how frequently they interact, helping users keep in touch with people with whom they may be losing contact. Users are able to set a Cloze score for people they want to get in touch with and work on improving that score. Cloze marks up the score whenever an attempt at connecting is made.

On opening an e-mail, Cloze provides several options, such as now, today, tomorrow, and next week, which automatically reminds the user to initiate contact at the scheduled time. This serves as a reminder for getting back to e-mails at a later point without just forgetting about them or marking them as "unread," which often leads to a cluttered in-box.

As is evident from these examples of consumer-centric apps, predictive analytics is beginning to enable development of software that is directly used by a consumer. *The Wall Street Journal* (**wsj.com/apps**) estimates that the app industry has already become a $25 billion industry with more growth expected. We believe that the growth of consumer-oriented analytic applications will grow and create many entrepreneurial opportunities for the readers of this book.

One key concern in employing these technologies is the loss of privacy. If someone can track the movement of a cell phone, the privacy of that customer is a big issue. Some of the app developers claim that they only need to gather aggregate flow information, not individually identifiable information. But many stories appear in the media that highlight violations of this general principle. Both users and developers of such apps have to be very aware of the deleterious effect of giving out private information as well as collecting such information. We discuss this issue a little bit further in Section 14.8.

SECTION 14.3 REVIEW QUESTIONS

1. What are the various options that CabSense provides to users?
2. Explore more transportation applications that may employ location-based analytics.
3. Briefly describe how the data are used to create profiles of users.
4. What other applications can you imagine if you were able to access cell phone location data? Do a search on location-enabled services.

14.4 RECOMMENDATION ENGINES

In most decision situations, people rely on recommendations gathered either directly from other people or indirectly through the aggregated recommendations made by others in the form of reviews and ratings posted either in newspapers, product guides, or online. Such information sharing is considered one of the major reasons for the success of online retailers such as **Amazon.com**. In this section we briefly review the common terms and technologies of such systems as these are becoming key components of any analytic application.

The term *recommender systems* refers to a Web-based information filtering system that takes the inputs from users and then aggregates the inputs to provide recommendations for other users in their product or service selection choices. Some recommender systems now even try to predict the rating or preference that a user would give for a particular product or service.

The data necessary to build a recommendation system are collected by Web-based systems where each user is specifically asked to rate an item on a rating scale, rank the items from most favorite to least favorite, and/or ask the user to list the attributes of the items that the user likes. Other information such as the user's textual comments, feedback reviews, amount of time that the user spends on viewing an item, and tracking the details of the user's social networking activity provides behavioral information about the product choices made by the user.

Two basic approaches that are employed in the development of recommendation systems are collaborative filtering and content filtering. In collaborative filtering, the recommendation system is built based on the individual user's past behavior by keeping track of the previous history of all purchased items. This includes products, items that are viewed most often, and ratings that are given by the users to the items they purchased. These individual profile histories with item preferences are grouped with other similar user-item profile histories to build a comprehensive set of relations between users and items, which are then used to predict what the user will like and recommend items accordingly.

Collaborative filtering involves aggregating the user-item profiles. It is usually done by building a user-item ratings matrix where each row represents a unique user and each column gives the individual item rating made by the user. The resultant matrix is a dynamic, sparse matrix with a huge dimensionality; it gets updated every time the existing user purchases a new item or a new user makes item purchases. Then the recommendation task is to predict what rating a user would give to a previously unranked item. The predictions that result in higher item rankings are then presented as recommendations to the users. The user-item based approach employs techniques like matrix factorization and low-rank matrix approximation to reduce the dimensionality of the sparse matrix in generating the recommendations.

Collaborative filtering can also take a user-based approach in which the users take the main role. Similar users sharing the same preferences are combined into a group, and recommendations of items to a particular user are based on the evaluation of items by other users in the same group. If a particular item is ranked high by the entire community, then it is recommended to the user. Another collaborative filtering approach is based on the item-set similarity, which groups items based on the user ratings provided by various users. Both of these collaborative filtering approaches employ many algorithms, such as KNN (*K*-Nearest Neighborhood) and Pearson Correlation, in measuring user and behavior similarity of ratings among the items.

The collaborative filtering approaches often require huge amounts of existing data on user-item preferences to make appropriate recommendations; this problem is most often referred to as *cold start* in the process of making recommendations. Also, in the

typical Web-based environment, tapping each individual's ratings and purchase behavior generates large amounts of data, and applying collaborative filtering algorithms requires separate high-end computation power to make the recommendations.

Collaborative filtering is widely employed in e-commerce. Customers can rate books, songs, or movies and then get recommendations regarding those issues in future. It is also being utilized in browsing documents, articles, and other scientific papers and magazines. Some of the companies using this type of recommender system are **Amazon.com** and social networking Web sites like Facebook and LinkedIn.

Content-based recommender systems overcome one of the disadvantages of collaborative filtering recommender systems, which completely rely on the user ratings matrix, by considering specifications and characteristics of items. In the content-based filtering approach, the characteristics of an item are profiled first and then content-based individual user profiles are built to store the information about the characteristics of specific items that the user has rated in the past. In the recommendation process, a comparison is made by filtering the item information from the user profile for which the user has rated positively and compares these characteristics with any new products that the user has not rated yet. Recommendations are made if there are similarities found in the item characteristics.

Content-based filtering involves using information tags or keywords in fetching detailed information about item characteristics and restricts this process to a single user, unlike collaborative filtering, which looks for similarities between various user profiles. This approach makes use of machine-learning and classification techniques like Bayesian classifiers, cluster analysis, decision trees, and artificial neural networks in order to estimate the probability of recommending similar items to the users that match the user's existing ratings for an item.

Content-based filtering approaches are widely used in recommending textual content such as news items and related Web pages. It is also used in recommending similar movies and music based on the existing individual profile. One of the companies employing this technique is Pandora, which builds a user profile based on the musicians/stations that a particular user likes and makes recommendations of other musicians following the similar genres an individual profile contains. Another example is an app called Patients Like Me, which builds individual patient profiles and recommends patients registered with Patients Like Me to contact other patients suffering from similar diseases.

SECTION 14.4 REVIEW QUESTIONS

1. List the types of approaches used in recommendation engines.
2. How do the two approaches differ?
3. Can you identify specific sites that may use one or the other type of recommendation system?

14.5 WEB 2.0 AND ONLINE SOCIAL NETWORKING

Web 2.0 is the popular term for describing advanced Web technologies and applications, including blogs, wikis, RSS, mashups, user-generated content, and social networks. A major objective of Web 2.0 is to enhance creativity, information sharing, and collaboration.

One of the most significant differences between Web 2.0 and the traditional Web is the greater collaboration among Internet users and other users, content providers, and enterprises. As an umbrella term for an emerging core of technologies, trends, and principles, Web 2.0 is not only changing what is on the Web, but also how it works. Web 2.0 concepts have led to the evolution of Web-based virtual communities and their hosting services, such as social networking sites, video-sharing sites, and more. Many believe

that companies that understand these new applications and technologies—and apply the capabilities early on—stand to greatly improve internal business processes and marketing. Among the biggest advantages is better collaboration with customers, partners, and suppliers, as well as among internal users.

Representative Characteristics of Web 2.0

The following are representative characteristics of the Web 2.0 environment:

- Web 2.0 has the ability to tap into the collective intelligence of users. The more users contribute, the more popular and valuable a Web 2.0 site becomes.
- Data is made available in new or never-intended ways. Web 2.0 data can be remixed or "mashed up," often through Web service interfaces, much the way a dance-club DJ mixes music.
- Web 2.0 relies on user-generated and user-controlled content and data.
- Lightweight programming techniques and tools let nearly anyone act as a Web site developer.
- The virtual elimination of software-upgrade cycles makes everything a *perpetual beta* or work-in-progress and allows rapid prototyping, using the Web as an application development platform.
- Users can access applications entirely through a browser.
- An architecture of participation and *digital democracy* encourages users to add value to the application as they use it.
- A major emphasis is on social networks and computing.
- There is strong support for information sharing and collaboration.
- Web 2.0 fosters rapid and continuous creation of new business models.

Other important features of Web 2.0 are its dynamic content, rich user experience, metadata, scalability, open source basis, and freedom (net neutrality).

Most Web 2.0 applications have a rich, interactive, user-friendly interface based on Ajax or a similar framework. Ajax (Asynchronous JavaScript and XML) is an effective and efficient Web development technique for creating interactive Web applications. The intent is to make Web pages feel more responsive by exchanging small amounts of data with the server behind the scenes so that the entire Web page does not have to be reloaded each time the user makes a change. This is meant to increase the Web page's interactivity, loading speed, and usability.

A major characteristic of Web 2.0 is the global spread of innovative Web sites and start-up companies. As soon as a successful idea is deployed as a Web site in one country, other sites appear around the globe. This section presents some of these sites. For example, approximately 120 companies specialize in providing Twitter-like services in dozens of countries. An excellent source for material on Web 2.0 is Search CIO's *Executive Guide: Web 2.0* (see **searchcio.techtarget.com/general/0,295582,sid19_gci1244339,00.html#glossary**).

Social Networking

Social networking is built on the idea that there is structure to how people know each other and interact. The basic premise is that social networking gives people the power to share, making the world more open and connected. Although social networking is usually practiced in social networks such as LinkedIn, Facebook, or Google+, aspects of it are also found in Wikipedia and YouTube.

We first briefly define *social networks* and then look at some of the services they provide and their capabilities.

A Definition and Basic Information

A *social network* is a place where people create their own space, or homepage, on which they write blogs (Web logs); post pictures, videos, or music; share ideas; and link to other Web locations they find interesting. In addition, members of social networks can tag the content they create and post it with keywords they choose themselves, which makes the content searchable. The mass adoption of social networking Web sites points to an evolution in human social interaction.

Mobile social networking refers to social networking where members converse and connect with one another using cell phones or other mobile devices. Virtually all major social networking sites offer mobile services or apps on smartphones to access their services. The explosion of mobile Web 2.0 services and companies means that many social networks can be based from cell phones and other portable devices, extending the reach of such networks to the millions of people who lack regular or easy access to computers.

Facebook (**facebook.com**), which was launched in 2004 by former Harvard student Mark Zuckerberg, is the largest social network service in the world, with almost 1 billion users worldwide as of February 2013. A primary reason why Facebook has expanded so rapidly is the network effect—more users means more value. As more users become involved in the social space, more people are available to connect with. Initially, Facebook was an online social space for college and high school students that automatically connected students to other students at the same school. Expanding to a global audience has enabled Facebook to become the dominant social network.

Today, Facebook has a number of applications that support photos, groups, events, marketplaces, posted items, games, and notes. A special feature on Facebook is the News Feed, which enables users to track the activities of friends in their social circles. For example, when a user changes his or her profile, the updates are broadcast to others who subscribe to the feed. Users can also develop their own applications or use any of the millions of Facebook applications that have been developed by other users.

Orkut (**orkut.com**) was the brainchild of a Turkish Google programmer of the same name. Orkut was to be Google's homegrown answer to Facebook. Orkut follows a format similar to that of other major social networking sites: a homepage where users can display every facet of their personal life they desire using various multimedia applications. It is more popular in countries such as Brazil than in the United States. Google has introduced another social network called Google+ that takes advantage of the popular e-mail service from Google, Gmail, but it is still a much smaller competitor of Facebook.

Implications of Business and Enterprise Social Networks

Although advertising and sales are the major EC activities in public social networks, there are emerging possibilities for commercial activities in business-oriented networks such as LinkedIn and in enterprise social networks.

USING TWITTER TO GET A PULSE OF THE MARKET Twitter is a popular social networking site that enables friends to keep in touch and follow what others are saying. An analysis of "tweets" can be used to determine how well a product/service is doing in the market. Previous chapters on Web analytics included a significant coverage of social media analytics. This continues to grow in popularity and business use. Analysis of posts on social media sites such as Facebook and Twitter has become a major business. Many companies provide services to monitor and manage such posts on behalf of companies and individuals. One good example is **reputation.com**.

SECTION 14.5 REVIEW QUESTIONS

1. Define *Web 2.0.*
2. List the major characteristics of Web 2.0.
3. What new business model has emerged from Web 2.0?
4. Define *social network.*
5. List some major social network sites.

14.6 CLOUD COMPUTING AND BI

Another emerging technology trend that business intelligence users should be aware of is cloud computing. Wikipedia (**en.wikipedia.org/wiki/cloud_computing**) defines **cloud computing** as "a style of computing in which dynamically scalable and often virtualized resources are provided over the Internet. Users need not have knowledge of, experience in, or control over the technology infrastructures in the cloud that supports them." This definition is broad and comprehensive. In some ways, cloud computing is a new name for many previous, related trends: utility computing, application service provider, grid computing, on-demand computing, *software as a service* (SaaS), and even older, centralized computing with dumb terminals. But the term *cloud computing* originates from a reference to the Internet as a "cloud" and represents an evolution of all of the previously shared/centralized computing trends. The Wikipedia entry also recognizes that cloud computing is a combination of several information technology components as services. For example, *infrastructure as a service* (IaaS) refers to providing computing *platforms as a service* (PaaS), as well as all of the basic platform provisioning, such as management administration, security, and so on. It also includes SaaS, which includes applications to be delivered through a Web browser while the data and the application programs are on some other server.

Although we do not typically look at Web-based e-mail as an example of cloud computing, it can be considered a basic cloud application. Typically, the e-mail application stores the data (e-mail messages) and the software (e-mail programs that let us process and manage e-mails). The e-mail provider also supplies the hardware/software and all of the basic infrastructure. As long as the Internet is available, one can access the e-mail application from anywhere in the Internet cloud. When the application is updated by the e-mail provider (e.g., when Gmail updates its e-mail application), it becomes available to all the customers without them having to download any new programs. Thus, any Web-based general application is in a way an example of a cloud application. Another example of a general cloud application is Google Docs and Spreadsheets. This application allows a user to create text documents or spreadsheets that are stored on Google's servers and are available to the users anywhere they have access to the Internet. Again, no programs need to be installed, "the application is in the cloud." The storage space is also "in the cloud."

A very good general business example of cloud computing is Amazon.com's Web services. Amazon.com has developed an impressive technology infrastructure for e-commerce as well as for business intelligence, customer relationship management, and supply chain management. It has built major data centers to manage its own operations. However, through Amazon.com's cloud services, many other companies can employ these very same facilities to gain advantages of these technologies without having to make a similar investment. Like other cloud-computing services, a user can subscribe to any of the facilities on a pay-as-you-go basis. This model of letting someone else own the hardware and software but making use of the facilities on a pay-per-use basis is the cornerstone of cloud computing. A number of companies offer cloud-computing services, including Salesforce.com, IBM, Sun Microsystems, Microsoft (Azure), Google, and Yahoo!

Cloud computing, like many other IT trends, has resulted in new offerings in business intelligence. White (2008) and Trajman (2009) provided examples of BI offerings related to cloud computing. Trajman identified several companies offering cloud-based data warehouse options. These options permit an organization to scale up its data warehouse and pay only for what it uses. Companies offering such services include 1010data, LogiXML, and Lucid Era. These companies offer feature extract, transform, and load capabilities as well as advanced data analysis tools. These are examples of SaaS as well as *data as a service* (DaaS) offerings. Other companies, such as Elastra and Rightscale, offer dashboard and data management tools that follow the SaaS and DaaS models, but they also employ IaaS from other providers, such as Amazon.com or Go Grid. Thus, the end user of a cloud-based BI service may use one organization for analysis applications that, in turn, uses another firm for the platform or infrastructure.

The next several paragraphs summarize the latest trends in the interface of cloud computing and business intelligence/decision support systems. These are excerpted from a paper written by Haluk Demirkan and one of the co-authors of this book (Demirkan and Delen, 2013).

Service-oriented thinking is one of the fastest growing paradigms in today's economy. Most of the organizations have already built (or are in a process of building) decision support systems that support agile data, information, and analytics capabilities as services. Let's look at the implications of service-orientation on DSS. One of the main premises of service orientation is that service-oriented decision support systems will be developed with a component-based approach that is characterized by reusability (services can be reused in many workflows), substitutability (alternative services can be used), extensibility and scalability (ability to extend services and scale them, increase capabilities of individual services), customizability (ability to customize generic features, and composability—easy construction of more complex functional solutions using basic services), reliability, low cost of ownership, economy of scale, and so on.

In a service-oriented DSS environment, most of the services are provided with distributed collaborations. Various DSS services are produced by many partners, and consumed by end users for decision making. In the meantime, partners play the role of producer and consumer in a given time.

Service-Oriented DSS

In a SODSS environment, there are four major components: information technology as enabler, process as beneficiary, people as user, and organization as facilitator. Figure 14.2 illustrates a conceptual architecture of service-oriented DSS.

In service-oriented DSS solutions, operational systems (1), data warehouses (2), online analytic processing (3), and end-user components (4) can be individually or bundled provided to the users as service. Some of these components and their brief descriptions are listed in Table 14-1.

In the following subsections we provide brief descriptions of the three service models (i.e., data-as-a-service, information-as-a-service, and analytics-as-a-service) that underlie (as its foundational enablers) the service-oriented DSS.

Data-as-a-Service (DaaS)

In the service-oriented DSS environment (such as cloud environment), the concept of data-as-services basically advocates the view that—with the emergence of service-oriented business processes, architecture, and infrastructure, which includes standardized processes for accessing data "where it lives"—the actual platform on which the data resides doesn't matter (Dyche, 2011). Data can reside in a local computer or in a server at a server farm inside a cloud-computing environment. With data-as-a-service, any business

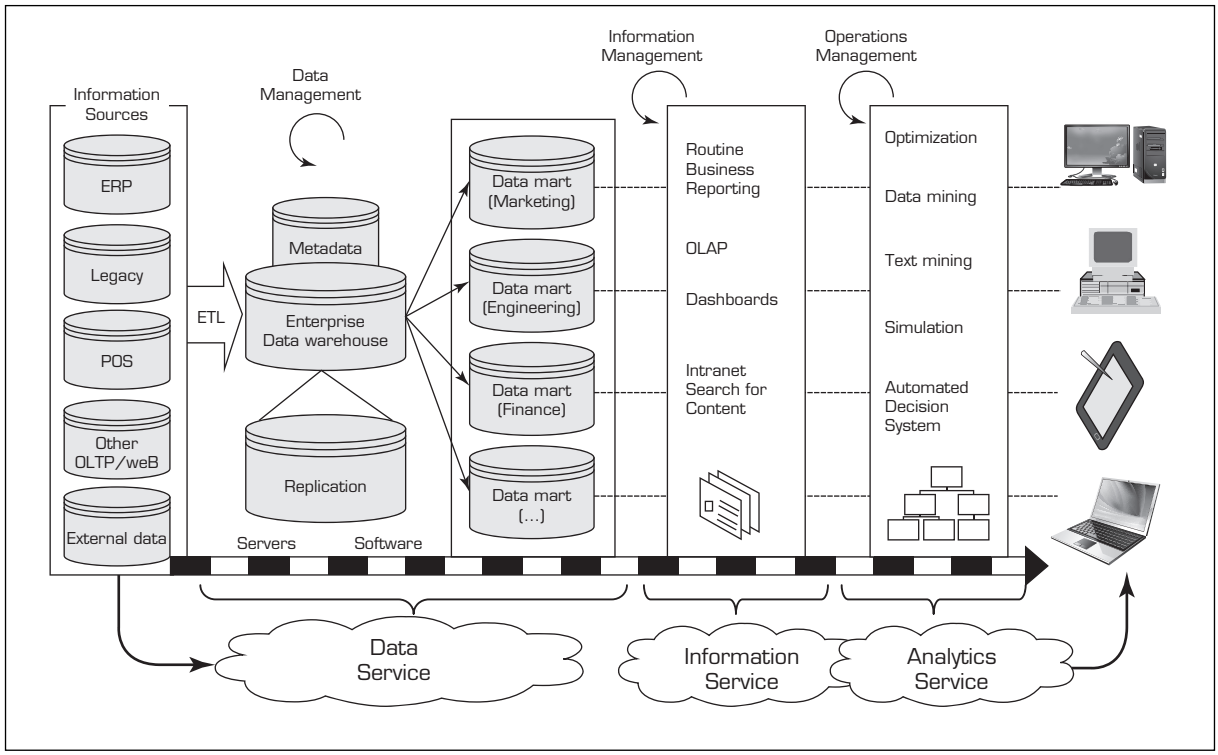

FIGURE 14.2 Conceptual Architecture of Service-Oriented DSS. *Source:* Haluk Demirkan and Dursun Delen, "Leveraging the Capabilities of Service-Oriented Decision Support Systems: Putting Analytics and Big Data in Cloud," *Decision Support Systems,* Vol. 55, No. 1, April 2013, pp. 412–421.

process can access data wherever it resides. Data-as-a-service began with the notion that data quality could happen in a centralized place, cleansing and enriching data and offering it to different systems, applications, or users, irrespective of where they were in the organization, computers, or on the network. This has now been replaced with master data management (MDM) and customer data integration (CDI) solutions, where the record of the customer (or product, or asset, etc.) may reside anywhere and is available as a service to any application that has the services allowing access to it. By applying a standard set of transformations to the various sources of data (for example, ensuring that gender fields containing different notation styles [e.g., M/F, Mr./Ms.] are all translated into male/female) and then enabling applications to access the data via open standards such as SQL, XQuery, and XML, service requestors can access the data regardless of vendor or system.

With DaaS, customers can move quickly thanks to the simplicity of the data access and the fact that they don't need extensive knowledge of the underlying data. If customers require a slightly different data structure or have location-specific requirements, the implementation is easy because the changes are minimal (agility). Second, providers can build the base with the data experts and outsource the presentation layer (which allows for very cost-effective user interfaces and makes change requests at the presentation layer much more feasible—cost-effectiveness), and access to the data is controlled through the data services, which tends to improve data quality because there is a single point for updates. Once those services are tested thoroughly, they only need to be regression tested if they remain unchanged for the next deployment (better data quality). Another important point is that DaaS platforms use NoSQL (sometimes expanded to "not only SQL"), which is a broad class of database management system that differs from classic relational database management systems (RDBMSs) in some significant

TABLE 14-1 Major Components of Service-Oriented DSS

	Component	Brief Description
Data sources	Application programming interface	Mechanism to populate source systems with raw data and to pull operational reports.
Data sources	Operational transaction systems	Systems that run day-to-day business operations and provide source data for the data warehouse and DSS environment.
Data sources	Enterprise application integration/staging area	Provides an integrated common data interface and interchange mechanism for real-time and source systems.
Data management	Extract, transform, load (ETL)	The processes to extract, transform, cleanse, reengineer, and load source data into the data warehouse, and move data from one location to another.
Data services	Metadata management	Data that describes the meaning and structure of business data, as well as how it is created, accessed, and used.
Data services	Data warehouse	Subject-oriented, integrated, time-variant, and nonvolatile collection of summary and detailed data used to support the strategic decision-making process for the organization. This is also used for ad hoc and exploratory processing of very large data sets.
Data services	Data marts	Subset of data warehouse to support specific decision and analytical needs and provide business units more flexibility, control, and responsibility.
Information services	Information	Such as ad hoc query, reporting, OLAP, dashboards, intra- and Internet search for content, data, and information mashups.
Analytics services	Analytics	Such as optimization, data mining, text mining, simulation, automated decision system.
Information delivery to end users	Information delivery portals	Such as desktop, Web browser, portal, mobile devices, e-mail.
Information management	Information services with library and administrator	Optimizes the DSS environment use by organizing its capabilities and knowledge, and assimilating them into the business processes. Also includes search engines, index crawlers, content servers, categorization servers, application/content integration servers, application servers, etc.
Data management	Ongoing data management	Ongoing management of data within and across the environment (such as backup, aggregate, retrieve data from near-line and off-line storage).
Operations management	Operations and administration	Activities to ensure daily operations, and optimize to allow manageable growth (systems management, data acquisition management, service management, change management, scheduling, monitor, security, etc.).
Information sources	Internal and external databases	Databases and files.
Servers	Operations	Database, application, Web, network, security, etc.
Software	Operations	Applications, integration, analytics, portals, ETL, etc.

ways. These data stores may not require fixed table schemas, usually avoid join operations, and typically scale horizontally (Stonebraker, 2010). Amazon offers such a service, called SimpleDB (**http://aws.amazon.com/simpledb**). Google's AppEngine (**http://code.google.com/appengine**) provides its DataStore API around BigTable. But apart from these two proprietary offerings, the current landscape is still open for prospective service providers.

Information-as-a-Service (Information on Demand) (IaaS)

The overall idea of IaaS is making information available quickly to people, processes, and applications across the business (agility). Such a system promises to eliminate silos of data that exist in systems and infrastructure today, to enable sharing real-time information for emerging apps, to hide complexity, and to increase availability with virtualization. The main idea is to bring together diverse sources, provide a "single version of the truth," make it available 24/7, and by doing so, reduce proliferating redundant data and the time it takes to build and deploy new information services. The IaaS paradigm aims to implement and sustain predictable qualities of service around information delivery at runtime and leverage and extend legacy information resources and infrastructure immediately through data and runtime virtualization, and thereby reduce ongoing development efforts. IaaS is a comprehensive strategy for the delivery of information obtained from information services, following a consistent approach using SOA infrastructure and/or Internet standards. Unlike enterprise information integration (EII), enterprise application integration (EAI), and extract, transform, and load (ETL) technologies, IaaS offers a flexible data integration platform based on a newer generation of service-oriented standards that enables ubiquitous access to any type of data, on any platform, using a wide range of interface and data access standards (Yuhanna, Gilpin, and Knoll, The Forrester Wave: Information-as-a-Service, Q1 2010, Forrester Research, 2010). Forrester Research names IaaS as Information Fabric and proposes a new, logical view to better characterize it. Two examples of such products are IBM's Web Sphere Information Integration and BEAs AquaLogic Data Services. These products can take the messy underlying data and present them as elemental services—for example, a service that presents a single view of a customer from the underlying data. These products can be used to enable real-time, integrated access to business information regardless of location or format by means of semantic integration. They also provide models-as-services (MaaS) to provide a collection of industry-specific business processes, reports, dashboards, and other service models for key industries (e.g., banking, insurance, and financial markets) to accelerate enterprise business initiatives for business process optimization and multi-channel transformation. They also provide master data management services (MDM) to enable the creation and management of multiform master data, provided as a service, for customer information across heterogeneous environments, content management services, and business intelligence services to perform powerful analysis from integrated data.

Analytics-as-a-Service (AaaS)

Analytics and data-based managerial solutions—the applications that query data for use in business planning, problem solving, and decision support—are evolving rapidly and being used by almost every organization. Gartner predicts that by 2013, 33 percent of BI functionality will be consumed via handheld devices; by 2014, 30 percent of analytic applications will use in-memory functions to add scale and computational speed, and will use proactive, predictive, and forecasting capabilities; and by 2014, 40 percent of spending on business analytics will go to system integrators, not software vendors (Tudor and Pettey, 2011).

The concept of analytics-as-a-service (AaaS)—by some referred to as Agile Analytics—is turning utility computing into a service model for analytics. AaaS is not limited to a single database or software; rather, it has the ability to turn a general-purpose analytical platform into a shared utility for an enterprise with the focus on virtualization of analytical services (Ratzesberger, 2011). With the needs of Enterprise Analytics growing rapidly, it is imperative that traditional hub-and-spoke architectures are not able to satisfy the demands driven by increasingly complex business analysis and analytics. New and improved architectures are needed to be able to process very large amounts of structured

and unstructured data in a very short time to produce accurate and actionable results. The "analytics-as-a-service" model is already being facilitated by Amazon, MapReduce, Hadoop, Microsoft's Dryad/SCOPE, Opera Solutions, eBay, and others. For example, eBay employees access a virtual slice of the main data warehouse server where they can store and analyze their own data sets. eBay's virtual private data marts have been quite successful—hundreds have been created, with 50 to 100 in operation at any one time. They have eliminated the company's need for new physical data marts that cost an estimated $1 million apiece and require the full-time attention of several skilled employees to provision (Winter, 2008).

AaaS in the cloud has economies of scale and scope by providing many virtual analytical applications with better scalability and higher cost savings. With growing data volumes and dozens of virtual analytical applications, chances are that more of them leverage processing at different times, usage patterns, and frequencies (Kalakota, 2011). A number of database companies such as Teradata, Netezza, Greenplum, Oracle, IBM DB2, DATAllegro, Vertica, and AsterData that provide shared-nothing (scalable) database management applications are well-suited for AaaS in cloud deployment.

Data and text mining is another very promising application of AaaS. The capabilities that a service orientation (along with cloud computing, pooled resources, and parallel processing) brings to the analytic world are not limited to data/text mining. It can also be used for large-scale optimization, highly-complex multi-criteria decision problems, and distributed simulation models. These prescriptive analytics require highly capable systems that can only be realized using service-based collaborative systems that can utilize large-scale computational resources.

We also expect that there will be significant interest in conducting service science research on cloud computing in Big Data analysis. With Web 2.0, more than enough data has been collected by organizations. We are entering the "petabyte age," and traditional data and analytics approaches are beginning to show their limits. Cloud analytics is an emerging alternative solution for large-scale data analysis. Data-oriented cloud systems include storage and computing in a distributed and virtualized environment. These solutions also come with many challenges, such as security, service level, and data governance. Research is still limited in this area. As a result, there is ample opportunity to bring analytical, computational, and conceptual modeling into the context of service science, service orientation, and cloud intelligence.

These types of cloud-based offerings are continuing to grow in popularity. A major advantage of these offerings is the rapid diffusion of advanced analysis tools among the users, without significant investment in technology acquisition. However, a number of concerns have been raised about cloud computing, including loss of control and privacy, legal liabilities, cross-border political issues, and so on. Nonetheless, cloud computing is an important initiative for a BI professional to watch.

SECTION 14.6 REVIEW QUESTIONS

1. Define *cloud computing*. How does it relate to PaaS, SaaS, and IaaS?
2. Give examples of companies offering cloud services.
3. How does cloud computing affect business intelligence?
4. What are the three service models that provide the foundation to service-oriented DSS?
5. How does DaaS change the way data is handled?
6. What is MaaS? What does it offer to businesses?
7. Why is AaaS cost-effective?
8. Why is MapReduce mentioned in the context of Aaas?

14.7 IMPACTS OF ANALYTICS IN ORGANIZATIONS: AN OVERVIEW

Analytic systems are important factors in the information, Web, and knowledge revolution. This is a cultural transformation with which most people are only now coming to terms. Unlike the slower revolutions of the past, such as the Industrial Revolution, this revolution is taking place very quickly and affecting every facet of our lives. Inherent in this rapid transformation are a host of managerial, economic, and social issues.

Separating the impact of analytics from that of other computerized systems is a difficult task, especially because of the trend toward integrating, or even embedding, analytics with other computer-based information systems. Analytics can have both micro and macro implications. Such systems can affect particular individuals and jobs, and they can also affect the work structures of departments and units within an organization. They can also have significant long-term effects on total organizational structures, entire industries, communities, and society as a whole (i.e., a macro impact).

The impact of computers and analytics can be divided into three general categories: organizational, individual, and societal. In each of these, computers have had many impacts. We cannot possibly consider all of them in this section, so in the next paragraphs we touch upon topics we feel are most relevant to analytics.

New Organizational Units

One change in organizational structure is the possibility of creating an analytics department, a BI department, or a knowledge management department in which analytics play a major role. This special unit can be combined with or replace a quantitative analysis unit, or it can be a completely new entity. Some large corporations have separate decision support units or departments. For example, many major banks have such departments in their financial services divisions. Many companies have small decision support or BI/data warehouse units. These types of departments are usually involved in training in addition to consulting and application development activities. Others have empowered a chief technology officer over BI, intelligent systems, and e-commerce applications. Companies such as Target and Walmart have major investments in such units, which are constantly analyzing their data to determine the efficiency of marketing and supply chain management by understanding their customer and supplier interactions.

Growth of the BI industry has resulted in the formation of new units within IT provider companies as well. For example, a few years back IBM formed a new business unit focused on analytics. This group includes units in business intelligence, optimization models, data mining, and business performance. As noted in Sections 14.2 and 14.3, the enormous growth of the app industry has created many opportunities for new companies that can employ analytics and deliver innovative applications in any specific domain.

There is also consolidation through acquisition of specialized software companies by major IT providers. For example, IBM acquired Demandtec, a revenue and promotion optimization software company, to build their offerings after having acquired SPSS for predictive analytics and ILOG to build their prescriptive analytics capabilities. Oracle acquired Hyperion some time back. Finally, there are also collaborations to enable companies to work cooperatively in some cases while also competing elsewhere. For example, SAS and Teradata announced a collaboration to let Teradata users develop BI applications using SAS analytical modeling capabilities. Teradata acquired Aster to enhance their Big Data offerings and Aprimo to add to their customer campaign management capabilities.

Section 14.9 describes the ecosystem of the analytics industry and recognizes the career paths available to analytics practitioners. It introduces many of the industry clusters, including those in user organizations.

Restructuring Business Processes and Virtual Teams

In many cases, it is necessary to restructure business processes before introducing new information technologies. For example, before IBM introduced e-procurement, it restructured all related business processes, including decision making, searching inventories, reordering, and shipping. When a company introduces a data warehouse and BI, the information flows and related business processes (e.g., order fulfillment) are likely to change. Such changes are often necessary for profitability, or even survival. Restructuring is especially necessary when major IT projects such as ERP or BI are undertaken. Sometimes an organization-wide, major restructuring is needed; then it is referred to as *reengineering*. Reengineering involves changes in structure, organizational culture, and processes. In a case in which an entire (or most of an) organization is involved, the process is referred to as **business process reengineering (BPR)**.

The Impacts of ADS Systems

As indicated in Chapter 1 and other chapters, ADS systems, such as those for pricing, scheduling, and inventory management, are spreading rapidly, especially in industries such as airlines, retailing, transportation, and banking. These systems will probably have the following impacts:

- Reduction of middle management
- Empowerment of customers and business partners
- Improved customer service (e.g., faster reply to requests)
- Increased productivity of help desks and call centers

The impact goes beyond one company or one supply chain, however. Entire industries are affected. The use of profitability models and optimization are reshaping retailing, real estate, banking, transportation, airlines, and car rental agencies, among other industries.

Job Satisfaction

Although many jobs may be substantially enriched by analytics, other jobs may become more routine and less satisfying. For example, more than 40 years ago, Argyris (1971) predicted that computer-based information systems would reduce managerial discretion in decision making and lead to managers being dissatisfied. In their study about ADS, Davenport and Harris (2005) found that employees using ADS systems, especially those who are empowered by the systems, were more satisfied with their jobs. If the routine and mundane work can be done using an analytic system, then it should free up the managers and knowledge workers to do more challenging tasks.

Job Stress and Anxiety

An increase in workload and/or responsibilities can trigger job stress. Although computerization has benefited organizations by increasing productivity, it has also created an ever-increasing and changing workload on some employees—many times brought on by downsizing and redistributing entire workloads of one employee to another. Some workers feel overwhelmed and begin to feel anxious about their jobs and their performance. These feelings of anxiety can adversely affect their productivity. Management must alleviate these feelings by redistributing the workload among workers or conducting appropriate training.

One of the negative impacts of the information age is information anxiety. This disquiet can take several forms, such as frustration with the inability to keep up with the amount of data present in our lives. Constant connectivity afforded through mobile

devices, e-mail, and instant messaging creates its own challenges and stress. Research on e-mail response strategies (**iris.okstate.edu/REMS**) includes many examples of studies conducted to recognize such stress. Constant alerts about incoming e-mails lead to interruptions, which eventually result in loss of productivity (and then an increase in stress). Systems have been developed to provide decision support to determine how often a person should check his or her e-mail (see Gupta and Sharda, 2009).

Analytics' Impact on Managers' Activities and Their Performance

The most important task of managers is making decisions. Analytics can change the manner in which many decisions are made and can consequently change managers' jobs. Some of the most common areas are discussed next.

According to Perez-Cascante et al. (2002), an ES/DSS was found to improve the performance of both existing and new managers as well as other employees. It helped managers gain more knowledge, experience, and expertise, and it consequently enhanced the quality of their decision making. Many managers report that computers have finally given them time to get out of the office and into the field. (BI can save an hour a day for every user.) They have also found that they can spend more time planning activities instead of putting out fires because they can be alerted to potential problems well in advance, thanks to intelligent agents, ES, and other analytical tools.

Another aspect of the managerial challenge lies in the ability of analytics to support the decision-making process in general and strategic planning and control decisions in particular. Analytics could change the decision-making process and even decision-making styles. For example, information gathering for decision making is completed much more quickly when analytics are in use. Enterprise information systems are extremely useful in supporting strategic management (see Liu et al., 2002). Data, text, and Web mining technologies are now used to improve external environmental scanning of information. As a result, managers can change their approach to problem solving and improve on their decisions quickly. It is reported that Starbucks recently introduced a new coffee beverage and made the decision on pricing by trying several different prices and monitoring the social media feedback throughout the day. This implies that data collection methods for a manager could be drastically different now than in the past.

Research indicates that most managers tend to work on a large number of problems simultaneously, moving from one to another as they wait for more information on their current problem (see Mintzberg et al., 2002). Analytics technologies tend to reduce the time required to complete tasks in the decision-making process and eliminate some of the nonproductive waiting time by providing knowledge and information. Therefore, managers work on fewer tasks during each day but complete more of them. The reduction in start-up time associated with moving from task to task could be the most important source of increased managerial productivity.

Another possible impact of analytics on the manager's job could be a change in leadership requirements. What are now generally considered good leadership qualities may be significantly altered by the use of analytics. For example, face-to-face communication is frequently replaced by e-mail, wikis, and computerized conferencing; thus, leadership qualities attributed to physical appearance could become less important.

The following are some potential impacts of analytics on managers' jobs:

- Less expertise (experience) is required for making many decisions.
- Faster decision making is possible because of the availability of information and the automation of some phases in the decision-making process.
- Less reliance on experts and analysts is required to provide support to top executives; managers can do it by themselves with the help of intelligent systems.

- Power is being redistributed among managers. (The more information and analysis capability they possess, the more power they have.)
- Support for complex decisions makes them faster to make and be of better quality.
- Information needed for high-level decision making is expedited or even self-generated.
- Automation of routine decisions or phases in the decision-making process (e.g., for frontline decision making and using ADS) may eliminate some managers.

In general, it has been found that the job of middle managers is the most likely job to be automated. Midlevel managers make fairly routine decisions, which can be fully automated. Managers at lower levels do not spend much time on decision making. Instead, they supervise, train, and motivate nonmanagers. Some of their routine decisions, such as scheduling, can be automated; other decisions that involve behavioral aspects cannot. However, even if we completely automate their decisional role, we could not automate their jobs. The Web provides an opportunity to automate certain tasks done by frontline employees; this empowers them, thus reducing the workload of approving managers. The job of top managers is the least routine and therefore the most difficult to automate.

SECTION 14.7 REVIEW QUESTIONS

1. List the impacts of analytics on decision making.
2. List the impacts of analytics on other managerial tasks.
3. Describe new organizational units that are created because of analytics.
4. How can analytics affect restructuring of business processes?
5. Describe the impacts of ADS systems.
6. How can analytics affect job satisfaction?

14.8 ISSUES OF LEGALITY, PRIVACY, AND ETHICS

Several important legal, privacy, and ethical issues are related to analytics. Here we provide only representative examples and sources.

Legal Issues

The introduction of analytics may compound a host of legal issues already relevant to computer systems. For example, questions concerning liability for the actions of advice provided by intelligent machines are just beginning to be considered.

In addition to resolving disputes about the unexpected and possibly damaging results of some analytics, other complex issues may surface. For example, who is liable if an enterprise finds itself bankrupt as a result of using the advice of an analytic application? Will the enterprise itself be held responsible for not testing the system adequately before entrusting it with sensitive issues? Will auditing and accounting firms share the liability for failing to apply adequate auditing tests? Will the software developers of intelligent systems be jointly liable? Consider the following specific legal issues:

- What is the value of an expert opinion in court when the expertise is encoded in a computer?
- Who is liable for wrong advice (or information) provided by an intelligent application? For example, what happens if a physician accepts an incorrect diagnosis made by a computer and performs an act that results in the death of a patient?
- What happens if a manager enters an incorrect judgment value into an analytic application and the result is damage or a disaster?
- Who owns the knowledge in a knowledge base?
- Can management force experts to contribute their expertise?

Privacy

Privacy means different things to different people. In general, **privacy** is the right to be left alone and the right to be free from unreasonable personal intrusions. Privacy has long been a legal, ethical, and social issue in many countries. The right to privacy is recognized today in every state of the United States and by the federal government, either by statute or by common law. The definition of *privacy* can be interpreted quite broadly. However, the following two rules have been followed fairly closely in past court decisions: (1) The right of privacy is not absolute. Privacy must be balanced against the needs of society. (2) The public's right to know is superior to the individual's right to privacy. These two rules show why it is difficult, in some cases, to determine and enforce privacy regulations (see Peslak, 2005). Privacy issues online have specific characteristics and policies. One area where privacy may be jeopardized is discussed next. For privacy and security issues in the data warehouse environment, see Elson and LeClerc (2005).

COLLECTING INFORMATION ABOUT INDIVIDUALS The complexity of collecting, sorting, filing, and accessing information manually from numerous government agencies was, in many cases, a built-in protection against misuse of private information. It was simply too expensive, cumbersome, and complex to invade a person's privacy. The Internet, in combination with large-scale databases, has created an entirely new dimension of accessing and using data. The inherent power in systems that can access vast amounts of data can be used for the good of society. For example, by matching records with the aid of a computer, it is possible to eliminate or reduce fraud, crime, government mismanagement, tax evasion, welfare cheating, family-support filching, employment of illegal workers, and so on. However, what price must the individual pay in terms of loss of privacy so that the government can better apprehend criminals? The same is true on the corporate level. Private information about employees may aid in better decision making, but the employees' privacy may be affected. Similar issues are related to information about customers.

The implications for online privacy are significant. The USA PATRIOT Act also broadens the government's ability to access student information and personal financial information without any suspicion of wrongdoing by attesting that the information likely to be found is pertinent to an ongoing criminal investigation (see Electronic Privacy Information Center, 2005). Location information from devices has been used to locate victims as well as perpetrators in some cases, but at what point is the information not the property of the individual?

Two effective tools for collecting information about individuals are cookies and spyware. Single-sign-on facilities that let a user access various services from a provider are beginning to raise some of the same concerns as cookies. Such services (Google, Yahoo!, MSN) let consumers permanently enter a profile of information along with a password and use this information and password repeatedly to access services at multiple sites. Critics say that such services create the same opportunities as cookies to invade an individual's privacy.

The use of artificial intelligence technologies in the administration and enforcement of laws and regulations may increase public concern regarding privacy of information. These fears, generated by the perceived abilities of artificial intelligence, will have to be addressed at the outset of almost any artificial intelligence development effort.

MOBILE USER PRIVACY Many users are unaware of the private information being tracked through mobile PDA or cell phone use. For example, Sense Networks' models are built using data from cell phone companies that track each phone as it moves from one cell tower to another, from GPS-enabled devices that transmit users' locations, and from PDAs transmitting information at wifi hotspots. Sense Networks claims that it is extremely careful and protective of users' privacy, but it is interesting to note how much information is available through just the use of a single device.

HOMELAND SECURITY AND INDIVIDUAL PRIVACY Using analytics technologies such as mining and interpreting the content of telephone calls, taking photos of people in certain places and identifying them, and using scanners to view your personal belongings are considered by many to be an invasion of privacy. However, many people recognize that analytic tools are effective and efficient means to increase security, even though the privacy of many innocent people is compromised.

The U.S. government applies analytical technologies on a global scale in the war on terrorism. In the first year and a half after September 11, 2001, supermarket chains, home improvement stores, and other retailers voluntarily handed over massive amounts of customer records to federal law enforcement agencies, almost always in violation of their stated privacy policies. Many others responded to court orders for information, as required by law. The U.S. government has a right to gather corporate data under legislation passed after September 11, 2001. The FBI now mines enormous amounts of data, looking for activity that could indicate a terrorist plot or crime.

Privacy issues abound. Because the government is acquiring personal data to detect suspicious patterns of activity, there is the prospect of improper or illegal use of the data. Many see such gathering of data as a violation of citizens' freedoms and rights. They see the need for an oversight organization to "watch the watchers," to ensure that the Department of Homeland Security does not mindlessly acquire data. Instead, it should acquire only pertinent data and information that can be mined to identify patterns that potentially could lead to stopping terrorists' activities. This is not an easy task.

Recent Technology Issues in Privacy and Analytics

Most providers of Internet services such as Google, Facebook, Twitter, and others depend upon monetizing their users' actions. They do so in many different ways, but all of these approaches in the end amount to understanding a user's profile or preferences on the basis of their usage. With the growth of Internet users in general and mobile device users in particular, many companies have been founded to employ advanced analytics to develop profiles of users on the basis of their device usage, movement, and the contacts of the users. *The Wall Street Journal* has an excellent collection of articles titled "What They Know" (**wsj.com/wtk**). These articles are constantly updated to highlight the latest technology and privacy/ethical issues. Some of the companies that have been mentioned in this series include companies such as Rapleaf (**rapleaf.com**). Rapleaf claims to be able to provide a profile of a user by just knowing their e-mail address. Clearly, their technology enables them to gather significant information. Similar technology is also marketed by X+1 (**xplusone.com**). Another company that aims to identify devices on the basis of their usage is Bluecava (**bluecava.com**). All of these companies employ technologies such as clustering and association mining to develop profiles of users. Such analytics applications definitely raise thorny questions of privacy violation for the users. Of course, many of the analytics start-ups in this space claim to honor user privacy, but violations are often reported. For example, a recent story reported that Rapleaf was collecting unauthorized user information from Facebook users and was banned from Facebook. A column in *Time Magazine* by Joel Stein (2011) reports that an hour after he gave his e-mail address to a company that specializes in user information monitoring (**reputation.com**), they had already been able to discover his Social Security number. This number is a key to accessing much private information about a user and could lead to identity theft. So, violations of privacy create fears of criminal conduct based on user information. This area is a big concern overall and needs careful study. The book's Web site will constantly update new developments. *The Wall Street Journal* site "What They Know" is a resource that ought to be consulted periodically.

Another application area that combines organizational IT impact, Big Data, sensors, and privacy concerns is analyzing employee behaviors on the basis of data collected from sensors that the employees wear in a badge. One company **Sociometric Solutions** (**sociometricsolutions.com**) has reported several such applications of their sensor-embedded badges that the employees wear. These sensors track all movement of an employee. **Sociometric Solutions** has reportedly been able to assist companies in predicting which types of employees are likely to stay with the company or leave on the basis of these employees' interactions with other employees. For example, those employees who stay in their own cubicles are less likely to progress up the corporate ladder than those who move about and interact with other employees extensively. Similar data collection and analysis have helped other companies determine the size of conference rooms needed or even the office layout to maximize efficiency. This area is growing really fast and has resulted in another term—people analytics. Of course, this creates major privacy issues. Should the companies be able to monitor their employees this intrusively? Sociometric has reported that its analytics are only reported on an aggregate basis to their clients. No individual user data is shared. They have noted that some employers want to get individual employee data, but their contract explicitly prohibits this type of sharing. In any case, sensors are leading to another level of surveillance and analytics, which poses interesting privacy, legal, and ethical questions.

Ethics in Decision Making and Support

Several ethical issues are related to analytics. Representative ethical issues that could be of interest in analytics implementations include the following:

- Electronic surveillance
- Ethics in DSS design (see Chae et al., 2005)
- Software piracy
- Invasion of individuals' privacy
- Use of proprietary databases
- Use of intellectual property such as knowledge and expertise
- Exposure of employees to unsafe environments related to computers
- Computer accessibility for workers with disabilities
- Accuracy of data, information, and knowledge
- Protection of the rights of users
- Accessibility to information
- Use of corporate computers for non–work-related purposes
- How much decision making to delegate to computers

Personal values constitute a major factor in the issue of ethical decision making. The study of ethical issues is complex because of its multi-dimensionality. Therefore, it makes sense to develop frameworks to describe ethics processes and systems. Mason et al. (1995) explained how technology and innovation expand the size of the domain of ethics and discuss a model for ethical reasoning that involves four fundamental focusing questions: Who is the agent? What action was actually taken or is being contemplated? What are the results or consequences of the act? Is the result fair or just for all stakeholders? They also described a hierarchy of ethical reasoning in which each ethical judgment or action is based on rules and codes of ethics, which are based on principles, which in turn are grounded in ethical theory. For more on ethics in decision making, see Murali (2004).

SECTION 14.8 REVIEW QUESTIONS

1. List some legal issues of analytics.
2. Describe privacy concerns in analytics.

3. Explain privacy concerns on the Web.
4. List ethical issues in analytics.

14.9 AN OVERVIEW OF THE ANALYTICS ECOSYSTEM

So, you are excited about the potential of analytics, and want to join this growing industry. Who are the current players, and what do they do? Where might you fit in? The objective of this section is to identify various sectors of the analytics industry, provide a classification of different types of industry participants, and illustrate the types of opportunities that exist for analytics professionals. The section (indeed the book) concludes with some observations about the opportunities for professionals to move across these clusters.

First, we want to remind the reader about the three types of analytics introduced in Chapter 1 and described in detail in the intervening chapters: descriptive or reporting analytics, predictive analytics, and prescriptive or decision analytics. In the following sections we will assume that you already know these three categories of analytics.

Analytics Industry Clusters

This section is aimed at identifying various analytics industry players by grouping them into sectors. We note that the list of company names included is not exhaustive. These merely reflect our own awareness and mapping of companies' offerings in this space. Additionally, the mention of a company's name or its capability in one specific group does not mean that is the only activity/offering of that organization. We use these names simply to illustrate our descriptions of sectors. Many other organizations exist in this industry. Our goal is not to create a directory of players or their capabilities in each space, but to illustrate to the students that many different options exist for playing in the analytics industry. One can start in one sector and move to another role altogether. We will also see that many companies play in multiple sectors within the analytics industry and, thus, offer opportunities for movement within the field both horizontally and vertically.

Figure 14.3 illustrates our view of the analytics ecosystem. It includes nine key sectors or clusters in the analytics space. The first five clusters can be broadly termed technology providers. Their primary revenue comes from developing technology, solutions, and training to enable the user organizations employ these technologies in the most effective and efficient manner. The accelerators include academics and industry organizations whose goal is to assist both technology providers and users. We describe each of these next, briefly, and give some examples of players in each sector.

Data Infrastructure Providers

This group includes all of the major players in the data hardware and software industry. These organizations provide hardware and software targeted at providing the basic foundation for all data management solutions. Obvious examples of these would include all major hardware players that provide the infrastructure for database computing—IBM, Dell, HP, Oracle, and so forth. We would also include storage solution providers such as EMC and NetApp in this sector. Many companies provide both hardware and software platforms of their own (e.g., IBM, Oracle, and Teradata). On the other hand, many data solution providers offer database management systems that are hardware independent and can run on many platforms. Perhaps Microsoft's SQL Server family is the most common example of this. Specialized integrated software providers such as SAP also are in this family of companies. Because this group of companies is well known and represents

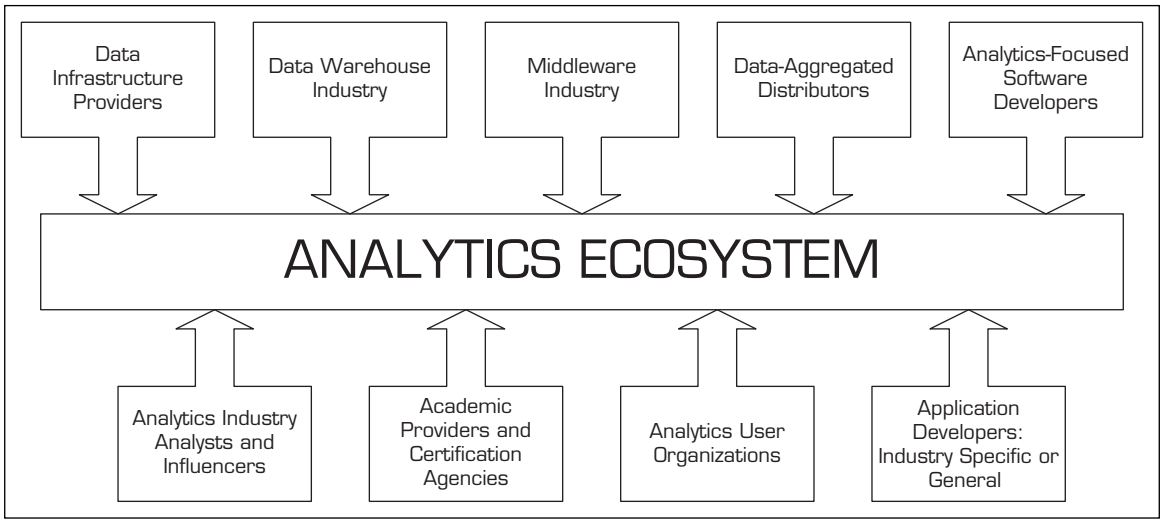

FIGURE 14.3 Analytic Industry Clusters.

a massive overall economic activity, we believe it is sufficient to recognize the key roles all these companies play. By inference, we also include all the other organizations that support each of these companies' ecosystems. These would include database appliance providers, service providers, integrators, and developers.

Several other companies are emerging as major players in a related space, thanks to the network infrastructure enabling cloud computing. Companies such as Amazon and **Salesforce.com** pioneered to offer full data storage (and more) solutions through the cloud. This has now been adopted by several of the players already identified.

Another group of companies that can be included here are the recent crop of companies in the Big Data space. Companies such as Cloudera, Hortonworks, and many others do not necessarily offer their own hardware but provide infrastructure services and training to create the Big Data platform. This would include Hadoop clusters, MapReduce, NoSQL, and other related technologies for analytics. Thus, they could also be grouped under industry consultants or trainers. We include them here because their role is aimed at enabling the basic infrastructure.

Bottom line, this group of companies provides the basic data and computing infrastructure that we take for granted in the practice of any analytics.

Data Warehouse Industry

We distinguish between this group and the preceding group mainly due to differences in their focus. Companies with data warehousing capabilities focus on providing integrated data from multiple sources so an organization can derive and deliver value from its data assets. Many companies in this space include their own hardware to provide efficient data storage, retrieval, and processing. Recent developments in this space include performing analytics on the data directly in memory. Companies such as IBM, Oracle, and Teradata are major players in this arena. Because this book includes links to Teradata University Network (TUN), we note that their platform software is available to TUN participants to explore data warehousing concepts (Chapter 3). In addition, all major players (EMC, IBM, Microsoft, Oracle, SAP, Teradata) have their own academic alliance programs through which much data warehousing software can be obtained so that students can develop familiarity and experience with the software. These companies clearly work with all the other sector players to provide data warehouse solutions and services within their

ecosystem. Because players in this industry are covered extensively by technology media as well as textbooks and have their own ecosystems in many cases, we will just recognize them as a backbone of the analytics industry and move to other clusters.

Middleware Industry

Data warehousing began with the focus on bringing all the data stores into an enterprise-wide platform. By making sense of this data, it becomes an industry in itself. The general goal of this industry is to provide easy-to-use tools for reporting and analytics. Examples of companies in this space include MicroStrategy, Plum, and many others. A few of the major players that were independent middleware players have been acquired by companies in the first two groups. For example, Hyperion became a part of Oracle. SAP acquired Business Objects. IBM acquired Cognos. This segment is thus merging with other players or at least partnering with many others. In many ways, the focus of these companies has been to provide descriptive analytics and reports, identified as a core part of BI or analytics.

Data Aggregators/Distributors

Several companies realized the opportunity to develop specialized data collection, aggregation, and distribution mechanisms. These companies typically focus on a specific industry sector and build upon their existing relationships. For example, Nielsen provides data sources to their clients on retail purchase behavior. Another example is Experian, which includes data on each household in the United States. (Similar companies exist outside the United States, as well.) Omniture has developed technology to collect Web clicks and share such data with their clients. Comscore is another major company in this space. Google compiles data for individual Web sites and makes a summary available through Google Analytics services. There are hundreds of other companies that are developing niche platforms and services to collect, aggregate, and share such data with their clients.

Analytics-Focused Software Developers

Companies in this category have developed analytics software for general use with data that has been collected in a data warehouse or is available through one of the platforms identified earlier (including Big Data). It can also include inventors and researchers in universities and other organizations that have developed algorithms for specific types of analytics applications. We can identify major industry players in this space along the same lines as the three types of analytics outlined in Chapter 1.

Reporting/Analytics

As seen in Chapters 1 and 4, the focus of reporting analytics is on developing various types of reports, queries, and visualizations. These include general visualizations of data or dashboards presenting multiple performance reports in an easy-to-follow style. These are made possible by the tools available from the middleware industry players or unique capabilities offered by focused providers. For example, Microsoft's SQL Server BI toolkit includes reporting as well as predictive analytics capabilities. On the other hand, specialized software is available from companies such as Tableau for visualization. SAS also offers a visual snalytics tool for similar capacity. Both are linked through TUN. There are many open source visualization tools as well. Literally hundreds of data visualization tools have been developed around the world. Many such tools focus on visualization of data from a specific industry or domain. A Google search will show the latest list of such software providers and tools.

Predictive Analytics

Perhaps the biggest recent growth in analytics has been in this category. Many statistical software companies such as SAS and SPSS embraced predictive analytics early on and developed the software capabilities as well as industry practices to employ data mining techniques, as well as classical statistical techniques, for analytics. SPSS was purchased by IBM and now sells IBM SPSS Modeler. SAS sells its software called Enterprise Miner. Other players in this space include KXEN, Statsoft, Salford Systems, and scores of other companies that may sell their software broadly or use it for their own consulting practices (next group of companies).

Two open source platforms (R and RapidMiner) have also emerged as popular industrial-strength software tools for predictive analytics and have companies that support training and implementation of these open sources tools. A company called Alteryx uses R extensions for reporting and predictive analytics, but its strength is in delivery of analytics solutions processes to customers and other users. By sharing the analytics process stream in a gallery where other users can see what data processing and analytic steps were used to arrive at a result from multiple data sources, other users can understand the logic of the analysis, even change it, and share the updated process with other users if they so choose.

In addition, many companies have developed specialized software around a specific technique of data mining. A good example includes a company called Rulequest, which sells proprietary variants of decision tree software. Many neural network software companies such as NeuroDimensions would also fall under this category. It is important to note that such specific software implementations may also be part of the capability offered by general predictive analytics tools identified earlier. The number of companies focused on predictive analytics is so large that it would take several pages to identify even a partial set.

Prescriptive Analytics

Software providers in this category offer modeling tools and algorithms for optimization of operations. Such software is typically available as management science/operations research (MS/OR) software. The best source of information for such providers is through *OR/MS Today*, a publication of INFORMS. Online directories of software in various categories are available on their Web site at **orms-today.org**. This field has had its own set of major software providers. IBM, for example, has classic linear and mixed-integer programming software. IBM also acquired a company (ILOG) that provides prescriptive analysis software and services to complement their other offerings. Analytics providers such as SAS have their own OR/MS tools—SAS/OR. FICO acquired another company, XPRESS, that offers optimization software. Other major players in this domain include companies such as AIIMS, AMPL, Frontline, GAMS, Gurobi, Lindo Systems, Maximal, and many others. A detailed delineation and description of these companies' offerings is beyond the scope of our goals here. Suffice it to note that this industry sector has seen much growth recently.

Of course, many techniques fall under the category of prescriptive analytics. Each group has its own set of providers. For example, simulation software is a category in its own right. Major companies in this space include Rockwell (ARENA) and Simio, among others. Palisade provides tools that include many software categories. Similarly, Frontline offers tools for optimization with Excel spreadsheets as well as predictive analytics. Decision analysis in multiobjective settings can be performed using tools such as Expert Choice. There are also tools from companies such as Exsys, XpertRule, and others for generating rules directly from data or expert inputs.

Some new companies are evolving to combine multiple analytics models in the Big Data space. For example, Teradata Aster includes its own predictive and prescriptive analytics capabilities in processing Big Data streams. We believe there will be more opportunities for companies to develop specific applications that combine Big Data and optimization techniques.

As noted earlier, all three categories of analytics have a rich set of providers, offering the user a wide set of choices and capabilities. It is worthwhile to note again that these groups are not mutually exclusive. In most cases a provider can play in multiple components of analytics.

Application Developers or System Integrators: Industry Specific or General

The organizations in this group focus on using solutions available from the data infrastructure, data warehouse, middleware, data aggregators, and analytics software providers to develop custom solutions for a specific industry. They also use their analytics expertise to develop specific applications for a user. Thus, this industry group makes it possible for the analytics technology to be truly useful. Of course, such groups may also exist in specific user organizations. We discuss those next, but distinguish between the two because the latter group is responsible for analytics within an organization whereas these application developers work with a larger client base. This sector presents excellent opportunities for someone interested in broadening their analytics implementation experience across industries. Predictably, it also represents a large group, too numerous to identify. Most major analytics technology providers clearly recognize the opportunity to connect to a specific industry or client. Virtually every provider in any of the groups identified earlier includes a consulting practice to help their clients employ their tools. In many cases, revenue from such engagements may far exceed the technology license revenue. Companies such as IBM, SAS, Teradata, and most others identified earlier have significant consulting practices. They hire graduates of analytics programs to work on different client projects. In many cases the larger technology providers also run their own certification programs to ensure that the graduates and consultants are able to claim a certain amount of expertise in using their specific tools.

Companies that have traditionally provided application/data solutions to specific sectors have recognized the potential for the use of analytics and are developing industry-specific analytics offerings. For example, Cerner provides electronic medical records (EMR) solutions to medical providers. Their offerings now include many analytics reports and visualizations. This has now extended to providing athletic injury reports and management services to sports programs in college and professional sports. Similarly, IBM offers a fraud detection engine for the health insurance industry and is working with an insurance company to employ their famous Watson analytics platform (which is known to have won against humans in the popular TV game show *Jeopardy!*) in assisting medical providers and insurance companies with diagnosis and disease management. Another example of a vertical application provider is Sabre Technologies, which provides analytical solutions to the travel industry including fare pricing for revenue optimization, dispatch planning, and so forth.

This group also includes companies that have developed their own domain-specific analytics solutions and market them broadly to a client base. For example, Axiom has developed clusters for virtually all households in the United States based upon all the data they collect about households from many different sources. These cluster labels allow a client organization to target a marketing campaign more precisely. Several companies provide this type of service. Credit score and classification reporting companies (such as FICO and Experian) also belong in this group. Demandtec (a company now owned

by IBM) provides pricing optimization solutions in the retail industry. They employ predictive analytics to forecast price-demand sensitivity and then recommend prices for thousands of products for retailers. Such analytics consultants and application providers are emerging to meet the needs of specific industries and represent an entrepreneurial opportunity to develop industry-specific applications. One area with many emerging start-ups is Web/social media/location analytics. By analyzing data available from Web clicks/smartphones/app uses, companies are trying to profile users and their interests to be better able to target promotional campaigns in real time. Examples of such companies and their activities include Sense Networks, which employs location data for developing user/group profiles; X+1 and Rapleaf, which profile users on the basis of e-mail usage; Bluecava, which aims to identify users through all device usage; and Simulmedia, which targets advertisements on TV on the basis of analysis of a user's TV-watching habits.

Another group of analytics application start-ups focuses on very specific analytics applications. For example, a popular smartphone app called Shazam is able to identify a song on the basis of the first few notes and then let the user select it from their song base to play/download/purchase. Voice-recognition tools such as Siri on iPhone and Google Now on Android are likely to create many more specialized analytics applications for very specific purposes in analytics applied to images, videos, audio, and other data that can be captured through smartphones and/or connected sensors.

This start-up activity and space is growing and in major transition due to technology/venture funding and security/privacy issues. Nevertheless, the application developer sector is perhaps the biggest growth industry within analytics at this point.

Analytics User Organizations

Clearly, this is the economic engine of the whole analytics industry. If there were no users, there would be no analytics industry. Organizations in every other industry, size, shape, and location are using analytics or exploring use of analytics in their operations. These include the private sector, government, education, and the military. It includes organizations around the world. Examples of uses of analytics in different industries abound. Others are exploring similar opportunities to try and gain/retain a competitive advantage. We will not identify specific companies in this section. Rather, the goal here is to see what types of roles analytics professionals can play within a user organization.

Of course, the top leadership of an organization is critically important in applying analytics to its operations. Reportedly, Forrest Mars of the Mars Chocolate Empire said that all management boiled down to applying mathematics to a company's operations and economics. Although not enough senior managers seem to subscribe to this view, the awareness of applying analytics within an organization is growing everywhere. Certainly the top leadership in information technology groups within a company (such as chief information officer) need to see this potential. For example, a health insurance company executive once told me that his boss (the CEO) viewed the company as an IT-enabled organization that collected money from insured members and distributed it to the providers. Thus, efficiency in this process was the premium they could earn over a competitor. This led the company to develop several analytics applications to reduce fraud and overpayment to providers and promote wellness among those insured so they would use the providers less often. Virtually all major organizations in every industry we are aware of are considering hiring analytical professionals. Titles of these professionals vary across industries. Table 14-2 includes selected titles of the MS graduates in our MIS program as well as graduates of our SAS Data Mining Certificate program (courtesy of Dr. G. Chakbraborty). This list indicates that most titles are indeed related to analytics. A "word cloud" of all of the titles of our analytics graduates, included in Figure 14-4, confirms the general results of these titles. It shows that analytics is already a popular title in the organizations hiring graduates of such programs.

TABLE 14-2 Selected Titles of Analytics Program Graduates

Advanced Analytics Math Modeler	Media Performance Analyst
Analytics Software Tester	Operation Research Analyst
Application Developer/Analyst	Operations Analyst
Associate Director, Strategy and Analytics	Predictive Modeler
Associate Innovation Leader	Principal Business Analyst
Bio Statistical Research Analyst	Principal Statistical Programmer
Business Analysis Manager	Procurement Analyst
Business Analyst	Project Analyst
Business Analytics Consultant	Project Manager
Business Data Analyst	Quantitative Analyst
Business Intelligence Analyst	Research Analyst
Business Intelligence Developer	Retail Analytics
Consultant Business Analytics	Risk Analyst—Client Risk and Collections
Credit Policy and Risk Analyst	SAS Business Analyst
Customer Analyst	SAS Data Analyst
Data Analyst	SAS Marketing Analyst
Data Mining Analyst	SAS Predictive Modeler
Data Mining Consultant	Senior Business Intelligence Analyst
Data Scientist	Senior Customer Intelligence Analyst
Decision Science Analyst	Senior Data Analyst
Decision Support Consultant	Senior Director of Analytics and Data Quality
ERP Business Analyst	Senior Manager of Data Warehouse, BI, and Analytics
Financial/Business Analyst	Senior Quantitative Marketing Analyst
Healthcare Analyst	Senior Strategic Marketing Analyst
Inventory Analyst	Senior Strategic Project Marketing Analyst
IT Business Analyst	Senior Marketing Database Analyst
Lead Analyst—Management Consulting Services	Senior Data Mining Analyst
Manager of Business Analytics	Senior Operations Analyst
Manager Risk Management	Senior Pricing Analyst
Manager, Client Analytics	Senior Strategic Marketing Analyst
Manager, Decision Support Analysis	Senior Strategy and Analytics Analyst
Manager, Global Customer Strategy and Analytics	Statistical Analyst
Manager, Modeling and Analytics	Strategic Business Analyst
Manager, Process Improvement, Global Operations	Strategic Database Analyst
Manager, Reporting and Analysis	Supply Chain Analyst
Managing Consultant	Supply Chain Planning Analyst
Marketing Analyst	Technical Analyst
Marketing Analytics Specialist	

FIGURE 14.4 Word Cloud of Titles of Analytics Program Graduates.

Of course, user organizations include career paths for analytics professionals moving into management positions. These titles include project managers, senior managers, directors … all the way up to chief information officer or chief executive officer. Our goal is here is to recognize that user organizations exist as a key cluster in the analytics ecosystem.

Analytics Industry Analysts and Influencers

The next cluster includes three types of organizations or professionals. The first group is the set of professional organizations that provides advice to analytics industry providers and users. Their services include marketing analyses, coverage of new developments, evaluation of specific technologies, and development of training/white papers, and so forth. Examples of such players include organizations such as the Gartner Group, The Data Warehousing Institute, and many of the general and technical publications and Web sites that cover the analytics industry. The second group includes professional societies or organizations that also provide some of the same services but are membership based and organized. For example, INFORMS, a professional organization, has now focused on promoting analytics. The Special Interest Group on Decision Support Systems (SIGDSS), a subgroup of the Association for Information Systems, also focuses on analytics. Most of the major vendors (e.g., Teradata and SAS) also have their own membership-based user groups. These entities promote the use of analytics and enable sharing of the lessons learned through their publications and conferences. They may also provide placement services.

A third group of analytics industry analysts is what we call analytics ambassadors, influencers, or evangelists. These folks have presented their enthusiasm for analytics through their seminars, books, and other publications. Illustrative examples include Steve Baker, Tom Davenport, Charles Duhigg, Wayne Eckerson, Bill Franks, Malcolm

Gladwell, Claudia Imhoff, Bill Inman, and many others. Again, the list is not inclusive. All of these ambassadors have written books (some of them bestsellers!) and/or given presentations to promote the analytics applications. Perhaps another group of evangelists to include here is the authors of textbooks on business intelligence/analytics (such as us, humbly) who aim to assist the next cluster to produce professionals for the analytics industry.

Academic Providers and Certification Agencies

In any knowledge-intensive industry such as analytics, the fundamental strength comes from having students who are interested in the technology and choose that industry as their profession. Universities play a key role in making this possible. This cluster, then, represents the academic programs that prepare professionals for the industry. It includes various components of business schools such as information systems, marketing, and management sciences. It also extends far beyond business schools to include computer science, statistics, mathematics, and industrial engineering departments across the world. The cluster also includes graphics developers who design new ways of visualizing information. Universities are offering undergraduate and graduate programs in analytics in all of these disciplines, though they may be labeled differently. A major growth frontier has been certificate programs in analytics to enable current professionals to retrain and retool themselves for analytics careers. Certificate programs enable practicing analysts to gain basic proficiency in specific software by taking a few critical courses. Power (2012) published a partial list of the graduate programs in analytics, but there are likely many more such programs, with new ones being added daily.

Another group of players assists with developing competency in analytics. These are certification programs to award a certificate of expertise in specific software. Virtually every major technology provider (IBM, Microsoft, MicroStrategy, Oracle, SAS, Teradata) has its own certification programs. These certificates ensure that potential new hires have a certain level of tool skills. On the other hand, INFORMS has just introduced a Certified Analytics Professional (CAP) certificate program that is aimed at testing an individual's general analytics competency. Any of these certifications give a college student additional marketable skills.

The growth of academic programs in analytics is staggering. Only time will tell if this cluster is overbuilding the capacity that can be consumed by the other eight clusters, but at this point the demand appears to outstrip the supply of qualified analytics graduates.

The purpose of this section has been to create a map of the landscape of the analytics industry. We identified nine different groups that play a key role in building and fostering this industry. It is possible for professionals to move from one industry cluster to another to take advantage of their skills. For example, expert professionals from providers can sometimes move to consulting positions, or directly to user organizations. Academics have provided consulting or have moved to industry. Overall, there is much to be excited about the analytics industry at this point.

SECTION 14.9 REVIEW QUESTIONS

1. Identify the nine clusters in the analytics ecosystem.
2. Which clusters represent technology developers?
3. Which clusters represent technology users?
4. Give examples of an analytics professional moving from one cluster to another.

Chapter Highlights

- Geospatial data can enhance analytics applications by incorporating location information.
- Real-time location information of users can be mined to develop promotion campaigns that are targeted at a specific user in real time.
- Location information from mobile phones and PDAs can be used to create profiles of user behavior and movement. Such location information can enable users to find other people with similar interests and advertisers to customize their promotions.
- Location-based analytics can also benefit consumers directly rather than just businesses. Mobile apps are being developed to enable such innovative analytics applications.
- Web 2.0 is about the innovative application of existing technologies. Web 2.0 has brought together the contributions of millions of people and has made their work, opinions, and identity matter.
- User-created content is a major characteristic of Web 2.0, as is the emergence of social networking.
- Large Internet communities enable the sharing of content, including text, videos, and photos, and promote online socialization and interaction.
- Business-oriented social networks concentrate on business issues both in one country and around the world (e.g., recruiting, finding business partners). Business-oriented social networks include LinkedIn and Xing.
- Cloud computing offers the possibility of using software, hardware, platform, and infrastructure, all on a service-subscription basis. Cloud computing enables a more scalable investment on the part of a user.
- Cloud-computing–based BI services offer organizations the latest technologies without significant upfront investment.
- Analytics can affect organizations in many ways, as stand-alone systems or integrated among themselves, or with other computer-based information systems.
- The impact of analytics on individuals varies; it can be positive, neutral, or negative.
- Serious legal issues may develop with the introduction of intelligent systems; liability and privacy are the dominant problem areas.
- Many positive social implications can be expected from analytics. These range from providing opportunities to disabled people to leading the fight against terrorism. Quality of life, both at work and at home, is likely to improve as a result of analytics. Of course, there are also negative issues to be concerned about.
- Analytics industry consists of many different types of stakeholders.

Key Terms

business process reengineering (BPR)
cloud computing
mobile social networking
privacy
reality mining
Web 2.0

Questions for Discussion

1. With regards to location-based analytics, discuss the decision Google took to develop Android OS and provide free regular updates.
2. Discuss the types of services your organization/university could offer with location-based analytics.
3. Select an app and examine the type of data it needs from the user to support its features.
4. Evaluate the right balance between making collaborative filtering more efficient and the need to preserve users' data integrity. In other words, how far should companies go in collecting data on their users?
5. Search the Web for the number of proposals Mark Zuckerberg has received until now to buy out Facebook.
6. Illustrate why the emergence of cloud computing will dramatically increase the quantity of IT terminals.
7. Explain why Amazon – originally being a click and mortar company – now offers cloud computing solutions to businesses and consumers.

8. Appraise IBM transformation from a hardware company to a service only provider for business and discuss the failed attempt of HP to sell its PC division.
9. Illustrate the symptoms of information anxiety and devise solutions to overcome that issue in the workplace.
10. Debate the necessity of the USA Patriot Act with regards to the NSA gathering data and voice conversations from world leaders in October 2013.
11. Discuss the concept of Big Data and how governments and companies are trying to solve governance and business issues.
12. Analyze the benefits that Massive Open Online Courses (MOOC) providers such as Coursera can offer to universities.

Exercises

TERADATA UNIVERSITY NETWORK (TUN) AND OTHER HANDS-ON EXERCISES

1. Go to **teradatauniversitynetwork.com** and search for case studies. Read the Continental Airlines cases written by Hugh Watson and his colleagues. What new applications can you imagine with the level of detailed data an airline can capture today.
2. Also review the Mycin case at **teradatauniversitynetwork.com**. What other similar applications can you envision?
3. At **teradatauniversitynetwork.com**, go to the podcasts library. Find podcasts of pervasive BI submitted by Hugh Watson. Summarize the points made by the speaker.
4. Go to **teradatauniversitynetwork.com** and search for BSI videos. Review these BSI videos and answer case questions related to them.
5. Discuss the pros and cons of location-tracking–based services, and determine a balance between personalized services and challenges for privacy.
6. Write a short essay around the topics of "ethics," "data," "business needs," and "citizen rights."
7. Discuss how pervasive computing with cloud-based services can enhance personalised services on the web.
8. Evaluate how legal and societal implications will evolve within the next five years with the continuous expansion of data collection in consumers' daily lives.
9. Search the Web for big data and healthcare. How relevant are the two terms in current global debates?
10. Discuss the difference between BI and Business Analytics.
11. Open and use Spotify or Deezer and identify how data is transformed into useful information to personalise its services to consumers.
12. List the different cloud storage providers and identify their strengths and weaknesses.
13. News search engines are now able to compile information and create their own content. List the names of a few news search engines.
14. Web 1.0 was mainly institutions producing content, and Web 2.0 is users producing contents. What do you think Web 3.0 will be?
15. Smartphones have revolutionized the way we consume and produce data. Which type of hardware do you think will bring about the next revolution?
16. Search the Internet for the amount of data being produced every year globally and discuss the evolution of data storage capacity.

    It contains accelerometer and gyroscope readings on 30 subjects who had the smartphone on their waist. The data is available in a raw format, and involves some data preparation efforts. Your objective is to identify and classify these readings into activities like walking, running, climbing, and such. More information on the data set is available on the download page. You may use clustering for initial exploration and gain an understanding on the data. You may use tools like R to prepare and analyze this data.

End-of-Chapter Application Case

Southern States Cooperative Optimizes Its Catalog Campaign

Southern States Cooperative is one of the largest farmer-owned cooperatives in the United States, with over 300,000 farmer members being served at over 1,200 retail locations across 23 states. It manufactures and purchases farm supplies like feed, seed, and fertilizer and distributes the products to farmers and other rural American customers.

Southern States Cooperative wanted to maintain and extend their success by better targeting the right customers in its direct-marketing campaigns. It realized the need to continually optimize marketing activities by gaining insights into its customers. Southern States employed Alteryx modeling tools, which enabled the company to solve the main business challenges of determining the right set of customers to be targeted for mailing the catalogs, choosing the right combination of storage keeping units (SKUs) to be included in the catalog, cutting down mailing costs, and increasing customer response, resulting in increased revenue generation, ultimately enabling it to provide better services to its customers.

SSC first built a predictive model to determine which catalogs the customer was most likely to prefer. The data for the analysis included Southern States' historical customer transaction data; the catalog data including the SKU information; farm-level data corresponding to the customers; and geocoded customer locations—as well as Southern States outlets. In performing the analysis, data from one year was analyzed on the basis of recency, frequency, and monetary value of customer transactions. In marketing, this type of analysis is commonly known as RFM analysis. The number of unique combinations of catalog SKUs and the customer purchase history of particular items in SKUs were used to predict the customers who were most likely to use the catalogs and the SKUs that ought to be included for the customers to respond to the catalogs. Preliminary exploratory analysis revealed that all the RFM measures and the measure of previous catalog SKU purchases had a diminishing marginal effect. As a result, these variables were natural-log transformed for logistic regression models. In addition to the logistic regression models, both a decision tree (based on a recursive partitioning algorithm) and a random forest model were also estimated using an estimation sample. The four different models (a "full" logistic regression model, a reduced version of the "full" logistic regression model based on the application of both forward and backward stepwise variable selection, the decision tree model, and the random forest model) were then compared using a validation sample via a gains (cumulative captured) response chart. A model using logistic regression was selected in which the most significant predictive factor was customers' past purchase of items contained in the catalog.

Based on the predictive modeling results, an incremental revenue model was built to estimate the effect of a customer's catalog use and the percentage revenues generated from the customer who used a particular catalog in a particular catalog period. Linear regression was the main technique applied in estimating the revenue per customer responding to the catalog. The model indicated that there was an additional 30 percent revenue per individual who used the catalog as compared to the non-catalog customers.

Furthermore, based on the results of the predictive model and the incremental revenue model, an optimization model was developed to maximize the total income from mailing the catalogs to customers. The optimization problem jointly maximizes the selection of catalog SKUs and customers to be sent the catalog, taking into account the expected response rate from mailing the catalog to specific customers and the expected profit margin in percentage from the purchases by that customer. It also considers the mailing cost. This formulation represents a constrained non-linear programming problem. This model was solved using genetic algorithms, aiming to maximize the combined selection of the catalog SKUs and the customers to whom the catalog should be sent to result in increased response, at the same time increasing the revenues and cutting down the mailing costs.

The Alteryx-based solution involved application of predictive analytics as well as prescriptive analytics techniques. The predictive model aimed to determine the customer's catalog use in purchasing selected items and then prescriptive analytics was applied to the results generated by predictive models to help the marketing department prepare the customized catalogs containing the SKUs that suited the targeted customer needs, resulting in better revenue generation.

From the model-based counterfactual analysis of the 2010 catalogs, the models quantified that the people who responded to the catalogs spent more in purchasing goods than those who had not used a catalog. The models indicated that in the year 2010, targeting the right customers with catalogs containing customized SKUs, Southern States Cooperative would have been able to reduce the number of catalogs sent by 63 percent, while improving the response rate by 34 percent, for an estimated incremental gross margin, less mailing cost, of $193,604—a 24 percent increase. The models were also applied toward the analysis of 2011 catalogs, and they estimated that with right combination and targeting of the 2011 catalogs, the total incremental gross margin would have been $206,812. With the insights derived from results of the historical data analysis, Southern States Cooperative is now planning to make use of these models in their future direct-mail marketing campaigns to target the right customers.

Questions for the End-of-Chapter Application Case

1. What is main business problem faced by Southern States Cooperative?
2. How was predictive analytics applied in the application case?
3. What problems were solved by the optimization techniques employed by Southern States Cooperative?

What We Can Learn from This End-of-Chapter Application Case

Predictive models built on historical data can be used to help quantify the effects of new techniques employed, as part of a retrospective assessment that otherwise cannot be quantified. The quantified values are estimates, not hard numbers, but obtaining hard numbers simply isn't possible. Often in a real-world scenario, many business problems require application of more than one type of analytics solution. There is often a chain of actions associated in solving problems where each stage relies on the outputs of the previous stages. Valuable insights can be derived by application of each type of analytic technique, which can be further applied to reach the optimal solution. This application case illustrates a combination of predictive and prescriptive analytics where geospatial data also played a role in developing the initial model.

Sources: **Alteryx.com**, "Southern States Cooperative Case Study," and direct communication with Dr. Dan Putler, **alteryx.com/sites/default/files/resources/files/case-study-southern-states.pdf** (accessed February 2013).

References

Alteryx.com. "Great Clips." **alteryx.com/sites/default/files/resources/files/case-study-great-chips.pdf** (accessed March 2013).

Alteryx.com. "Southern States Cooperative Case Study." Direct communication with Dr. Dan Putler. **alteryx.com/sites/default/files/resources/files/case-study-southern-states.pdf** (accessed February 2013).

Anandarajan, M. (2002). "Internet Abuse in the Workplace." *Communications of the ACM*, Vol. 45, No. 1, pp. 53–54.

Argyris, C. (1971). "Management Information Systems: The Challenge to Rationality and Emotionality." *Management Science*, Vol. 17, No. 6, pp. B-275.

Chae, B., D. B. Paradice, J. F. Courtney, and C. J. Cagle. (2005). "Incorporating an Ethical Perspective into Problem Formulation." *Decision Support Systems*, Vol. 40, No. 2, pp. 197–212.

Davenport, T. H., and J. G. Harris. (2005). "Automated Decision Making Comes of Age." *MIT Sloan Management Review*, Vol. 46, No. 4, p. 83.

Delen, D., B. Hardgrave, and R. Sharda. (2007). "RFID for Better Supply-Chain Management Through Enhanced Information Visibility." *Production and Operations Management,* Vol. 16, No. 5, pp. 613–624.

Dyche, J. (2011). "Data-as-a-Service, Explained and Defined." **searchdatamanagement.techtarget.com/answer/Data-as-a-service-explained-and-defined** (accessed March 2013).

Eagle, N., and A. Pentland. (2006). "Reality Mining: Sensing Complex Social Systems." *Personal and Ubiquitous Computing*, Vol. 10, No. 4, pp. 255–268.

Electronic Privacy Information Center. (2005). "USA PATRIOT Act." **epic.org/privacy/terrorism/usapatriot** (accessed March 2013).

Elson, R. J., and R. LeClerc. (2005). "Security and Privacy Concerns in the Data Warehouse Environment." *Business Intelligence Journal*, Vol. 10., No. 3, p. 51.

Emc.com. "Data Science Revealed: A Data-Driven Glimpse into the Burgeoning New Field." **emc.com/collateral/about/news/emc-data-science-study-wp.pdf** (accessed February 2013).

Fritzsche, D. (1995, November). "Personal Values: Potential Keys to Ethical Decision Making." *Journal of Business Ethics*, Vol. 14, No. 11.

Gnau, S. (2010). "Find Your Edge." *Teradata Magazine Special Edition Location Intelligence.* **teradata.com/articles/Teradata-Magazine-Special-Edition-Location-Intelligence-AR6270/?type=ART** (accessed March 2013).

Gupta, A., and R. Sharda. (2009). "SIMONE: A Simulator for Interruptions and Message Overload in Network Environments." *International Journal of Simulation and Process Modeling*, Vol. 4, Nos. 3/4, pp. 237–247.

Institute of Medicine of the National Academies. "Health Data Initiative Forum III: The Health Datapalooza." **iom.edu/Activities/PublicHealth/HealthData/2012-JUN-05/Afternoon-Apps-Demos/outside-100plus.aspx** (accessed February 2013).

IntellidentUtility.com. "OGE's Three-Tiered Architecture Aids Data Analysis." **intelligentutility.com/article/12/02/oges-three-tiered-architecture-aids-data-analysis&utm_medium=eNL&utm_campaign=IU_DAILY2&utm_term=Original-Magazine** (accessed March 2013).

Kalakota, R. (2011). "Analytics-as-a-Service: Understanding How **Amazon.com** Is Changing the Rules." **practicalanalytics.wordpress.com/2011/08/13/analytics-as-a-service-understanding-how-amazon-com-is-changing-the-rules** (accessed March 2013).

Krivda, C. D. (2010). "Pinpoint Opportunity." *Teradata Magazine Special Edition Location Intelligence.* **teradata.com/articles/Teradata-Magazine-Special-Edition-Location-Intelligence-AR6270/?type=ART** (accessed March 2013).

Liu, S., J. Carlsson, and S. Nummila. (2002, July). "Mobile E-Services: Creating Added Value for Working Mothers." *Proceedings DSI AGE 2002*, Cork, Ireland.

Mason, R. O., F. M. Mason, and M. J. Culnan. (1995). *Ethics of Information Management*. Thousand Oaks, CA: Sage.

Mintzberg, H., et al. (2002). *The Strategy Process,* 4th ed. Upper Saddle River, NJ: Prentice Hall.

Mobilemarketer.com. "Quiznos Sees 20pc Boost in Coupon Redemption via Location-Based Mobile Ad Campaign." **mobilemarketer.com/cms/news/advertising/14738.html** (accessed February 2013).

Murali, D. (2004). "Ethical Dilemmas in Decision Making." *BusinessLine*.

Ogepet.com. "Smart Hours." **ogepet.com/programs/smarthours.aspx** (accessed March 2013).

Perez-Cascante, L. P., M. Plaisent, L. Maguiraga, and P. Bernard. (2002). "The Impact of Expert Decision Support Systems on the Performance of New Employees." *Information Resources Management Journal*.

Peslak, A. P. (2005). "Internet Privacy Policies." *Information Resources Management Journal*.

Power, D. P. (2012). "What Universities Offer Master's Degrees in Analytics and Data Science?" **dssresources.com/faq/index.php?action=artikel&id=250** (accessed February 2013).

Ratzesberger, O. (2011). "Analytics as a Service." **xlmpp.com/articles/16-articles/39-analytics-as-a-service** (accessed September 2011).

Sensenetworks.com. "CabSense New York: The Smartest Way to Find a Cab." **sensenetworks.com/products/macrosense-technology-platform/cabsense** (accessed February 2013).

Stein, J. "Data Mining: How Companies Now Know Everything About You." *Time Magazine.* **time.com/time/magazine/article/0,9171,2058205,00.html** (accessed March 2013).

Stonebraker, M. (2010). "SQL Databases V. NoSQL Databases." *Communications of the ACM*, Vol. 53, No. 4, pp. 10–11.

Teradata.com. "Sabre Airline Solutions." **teradata.com/t/case-studies/Sabre-Airline-Solutions-EB6281** (accessed March 2013).

Teradata.com. "Utilities Analytic Summit 2012 Oklahoma Gas & Electric." **teradata.com/video/Utilities-Analytic-Summit-2012-Oklahoma-Gas-and-Electric** (accessed March 2013).

Trajman, O. (2009, March). "Business Intelligence in the Clouds." *InfoManagement Direct.* **information-management.com/infodirect/2009_111/10015046-1.html** (accessed July 2009).

Tudor, B., and C. Pettey. (2011, January 6). "Gartner Says New Relationships Will Change Business Intelligence and Analytics." *Gartner Research*.

Tynan, D. (2002, June). "How to Take Back Your Privacy (34 Steps)." *PC World*.

WallStreetJournal.com. (2010). "What They Know." **online.wsj.com/public/page/what-they-know-2010.html** (accessed March 2013).

Westholder, M. (2010). "Pinpoint Opportunity." *Teradata Magazine Special Edition Location Intelligence.* **teradata.com/articles/Teradata-Magazine-Special-Edition-Location-Intelligence-AR6270/?type=ART** (accessed March 2013).

White, C. (2008, July 30). "Business Intelligence in the Cloud: Sorting Out the Terminology." **BeyeNetwork.b-eye-network.com/channels/1138/view/8122** (accessed March 2013).

Winter, R. (2008). "E-Bay Turns to Analytics as a Service." **informationweek.com/news/software/info_management/210800736** (accessed March 2013).

Yuhanna, N., M. Gilpin, and A. Knoll. (2010). "The Forrester Wave: Information-as-a-Service, Q1 2010." **forrester.com/rb/Research/wave%26trade%3B_information-as-a-service%2C_q1_2010/q/id/55204/t/2** (accessed March 2013).

GLOSSARY

active data warehousing *See* real-time data warehousing.

ad hoc DSS A DSS that deals with specific problems that are usually neither anticipated nor recurring.

ad hoc query A query that cannot be determined prior to the moment the query is issued.

agency The degree of autonomy vested in a software agent.

agent-based models A simulation modeling technique to support complex decision systems where a system or network is modeled as a set of autonomous decision-making units called agents that individually evaluate their situation and make decisions on the basis of a set of predefined behavior and interaction rules.

algorithm A step-by-step search in which improvement is made at every step until the best solution is found.

analog model An abstract, symbolic model of a system that behaves like the system but looks different.

analogical reasoning The process of determining the outcome of a problem by using analogies. It is a procedure for drawing conclusions about a problem by using past experience.

analytic hierarchy process (AHP) A modeling structure for representing *multi-criteria* (multiple goals, multiple objectives) *problems*—with sets of criteria and alternatives (choices)—commonly found in business environments.

analytical models Mathematical models into which data are loaded for analysis.

analytical techniques Methods that use mathematical formulas to derive an optimal solution directly or to predict a certain result, mainly in solving structured problems.

analytics The science of analysis—to use data for decision making.

application service provider (ASP) A software vendor that offers leased software applications to organizations.

Apriori algorithm The most commonly used algorithm to discover association rules by recursively identifying frequent itemsets.

area under the ROC curve A graphical assessment technique for binary classification models where the true positive rate is plotted on the *Y*-axis and the false positive rate is plotted on the *X*-axis.

artificial intelligence (AI) The subfield of computer science concerned with symbolic reasoning and problem solving.

artificial neural network (ANN) Computer technology that attempts to build computers that operate like a human brain. The machines possess simultaneous memory storage and work with ambiguous information. Sometimes called, simply, a *neural network*. *See* neural computing.

association A category of data mining algorithm that establishes relationships about items that occur together in a given record.

asynchronous Occurring at different times.

authoritative pages Web pages that are identified as particularly popular based on links by other Web pages and directories.

automated decision support (ADS) A rule-based system that provides a solution to a repetitive managerial problem. Also known as *enterprise decision management (EDM)*.

automated decision system (ADS) A business rule–based system that uses intelligence to recommend solutions to repetitive decisions (such as pricing).

autonomy The capability of a software agent acting on its own or being empowered.

axon An outgoing connection (i.e., terminal) from a biological neuron.

backpropagation The best-known learning algorithm in neural computing where the learning is done by comparing computed outputs to desired outputs of training cases.

backward chaining A search technique (based on if-then rules) used in production systems that begins with the action clause of a rule and works backward through a chain of rules in an attempt to find a verifiable set of condition clauses.

balanced scorecard (BSC) A performance measurement and management methodology that helps translate an organization's financial, customer, internal process, and learning and growth objectives and targets into a set of actionable initiatives.

best practices In an organization, the best methods for solving problems. These are often stored in the knowledge repository of a knowledge management system.

Big Data Data that exceeds the reach of commonly used hardware environments and/or capabilities of software tools to capture, manage, and process it within a tolerable time span.

blackboard An area of working memory set aside for the description of a current problem and for recording intermediate results in an expert system.

black-box testing Testing that involves comparing test results to actual results.

bootstrapping A sampling technique where a fixed number of instances from the original data is sampled (with replacement) for training and the rest of the data set is used for testing.

bot An intelligent software agent. Bot is an abbreviation of robot and is usually used as part of another term, such as knowbot, softbot, or shopbot.