CHAPTER

Introduction to Big Data

1

INTRODUCTION

The biggest phenomenon that has captured the attention of the modern computing industry today since the "Internet" is "Big Data". These two words combined together was first popularized in the paper on this subject by McKinsey & Co., and the foundation definition was first popularized by Doug Laney from Gartner.

The fundamental reason why "Big Data" is popular today is because the technology platforms that have emerged along with it, provide the capability to process data of multiple formats and structures without worrying about the constraints associated with traditional systems and database platforms.

Big Data

Data represents the lowest raw format of information or knowledge. In the computing world, we refer to data commonly in terms of rows and columns of organized values that represent one or more entities and their attributes. Long before the age of computing or information management with electronic processing aids, data was invented with the advent of counting and trade, preceding the Greeks. Simply put, it is the assignment of values to numerals and then using those numerals to mark the monetary value, population, calendars, taxes, and many historical instances to provide ample evidence to the fascination of the human mind with data and knowledge acquisition and management.

Information or data management according to a series of studies by Carnegie Mellon University entails the process of organizing, acquiring, storing, retrieving, and managing data. Data collected from different processes is used to make decisions feasible to the understanding and requirements of those executing and consuming the results of the process. This administrative behavior was the underlying theme for Herbert Simon's view of bounded rationality[1], or the limited field of vision in human minds when applied to data management. The argument presented in the decision-making behaviors and administrative behaviors makes complete sense, as we limit the data in the process of modeling, applying algorithmic applications, and have always been seeking discrete relationships within the data as opposed to the whole picture.

In reality, however, decision making has always transcended beyond the traditional systems used to aid the process. For example, patient treatment and management is not confined to computers and programs. But the data generated by doctors, nurses, lab technicians, emergency personnel, and medical devices within a hospital for each patient can now, through the use of unstructured data integration techniques and algorithms, be collected and processed electronically to gain mathematical or

[1] March, J. G., & Simon, H. A. (1958) Organizations (*http://www.amazon.com/Organizations-James-G-March/dp/063118631X*).

statistical insights. These insights provide visible patterns that can be useful in improving quality of care for a given set of diseases.

Data warehousing evolved to support the decision-making process of being able to collect, store, and manage data, applying traditional and statistical methods of measurement to create a reporting and analysis platform. The data collected within a data warehouse was highly structured in nature, with minimal flexibility to change with the needs of data evolution. The underlying premise for this comes from the transactional databases that were the sources of data for a data warehouse. This concept applies very well when we talk of transactional models based on activity generated by consumers in retail, financial, or other industries. For example, movie ticket sales is a simple transaction, and the success of a movie is based on revenues it can generate in the opening and following weeks, and in a later stage followed by sales from audio (vinyl to cassette tapes, CDs', and various digital formats), video ('DVDs and other digital formats), and merchandise across multiple channels. When reporting sales revenue, population demographics, sentiments, reviews, and feedback were not often reported or at least were not considered as a visible part of decision making in a traditional computing environment. The reasons for this included rigidity of traditional computing architectures and associated models to integrate unstructured, semi-structured, or other forms of data, while these artifacts were used in analysis and internal organizational reporting for revenue activities from a movie.

Looking at these examples in medicine and entertainment business management, we realize that decision support has always been an aid to the decision-making process and not the end state itself, as is often confused.

If one were to consider all the data, the associated processes, and the metrics used in any decision-making situation within any organization, we realize that we have used information (volumes of data) in a variety of formats and varying degrees of complexity and derived decisions with the data in nontraditional software processes. Before we get to Big Data, let us look at a few important events in computing history.

In the late 1980s, we were introduced to the concept of decision support and data warehousing. This wave of being able to create trends, perform historical analysis, and provide predictive analytics and highly scalable metrics created a series of solutions, companies, and an industry in itself.

In 1995, with the clearance to create a commercial Internet, we saw the advent of the "dot-com" world and got the first taste of being able to communicate peer to peer in a consumer world. With the advent of this capability, we also saw a significant increase in the volume and variety of data.

In the following five to seven years, we saw a number of advancements driven by web commerce or e-commerce, which rapidly changed the business landscape for an organization. New models emerged and became rapidly adopted standards, including the business-to-consumer direct buying/selling (website), consumer-to-consumer marketplace trading (eBay and Amazon), and business-to-business-to-consumer selling (Amazon). This entire flurry of activity drove up data volumes more than ever before. Along with the volume, we began to see the emergence of additional data, such as consumer review, feedback on experience, peer surveys, and the emergence of word-of-mouth marketing. This newer and additional data brings in subtle layers of complexity in data processing and integration.

Along the way between 1997 and 2002, we saw the definition and redefinition of mobility solutions. Cellular phones became ubiquitous and the use of voice and text to share sentiments, opinions, and trends among people became a vibrant trend. This increased the ability to communicate and create a crowd-based affinity to products and services, which has significantly driven the last decade of technology innovation, leading to even more disruptions in business landscape and data management in terms of data volumes, velocity, variety, complexity, and usage.

The years 2000 to 2010 have been a defining moment in the history of data, emergence of search engines (Google, Yahoo), personalization of music (iPod), tablet computing (iPad), bigger mobile solutions (smartphones, 3G networks, mobile broadband, Wi-Fi), and emergence of social media (driven by Facebook, MySpace, Twitter, and Blogger). All these entities have contributed to the consumerization of data, from data creation, acquisition, and consumption perspectives.

The business models and opportunities that came with the large-scale growth of data drove the need to create powerful metrics to tap from the knowledge of the crowd that was driving them, and in return offer personalized services to address the need of the moment. This challenge was not limited to technology companies; large multinational organizations like P&G and Unilever wanted solutions that could address data processing, and additionally wanted to implement the output from large-scale data processing into their existing analytics platform.

Google, Yahoo, Facebook, and several other companies invested in technology solutions for data management, allowing us to consume large volumes of data in a short amount of time across many formats with varying degrees of complexity to create a powerful decision support platform. These technologies and their implementation are discussed in detail in later chapters in this book.

Defining Big Data

Big Data can be defined as volumes of data available in varying degrees of complexity, generated at different velocities and varying degrees of ambiguity, that cannot be processed using traditional technologies, processing methods, algorithms, or any commercial off-the-shelf solutions.

Data defined as Big Data includes machine-generated data from sensor networks, nuclear plants, X-ray and scanning devices, and airplane engines, and consumer-driven data from social media. Big Data producers that exist within organizations include legal, sales, marketing, procurement, finance, and human resources departments.

Why Big Data and why now?

These are the two most popular questions that are crossing the minds of any computing professional: Why Big Data? Why now? The promise of Big Data is the ability to access large volumes of data that can be useful in gaining critical insights from processing repeated or unique patterns of data or behaviors. This learning process can be executed as a machine-managed process with minimal human intervention, making the analysis simpler and error-free. The answer to the second question—Why now?—is the availability of commodity infrastructure combined with new data processing frameworks and platforms like Hadoop and NoSQL, resulting in significantly lower costs and higher scalability than traditional data management platforms. The scalability and processing architecture of the new platforms were limitations of traditional data processing technologies, though the algorithms and methods existed.

The key thing to understand here is the data part of Big Data was always present and used in a manual fashion, with a lot of human processing and analytic refinement, eventually being used in a decision-making process. What has changed and created the buzz with Big Data is the automated data processing capability that is extremely fast, scalable, and has flexible processing.

While each organization will have its own set of data requirements for Big Data processing, here are some examples:

- *Weather data*—there is a lot of weather data reported by governmental agencies around the world, scientific organizations, and consumers like farmers. What we hear on television or radio is an analytic key performance indicator (KPI) of temperature and forecasted conditions based on several factors.
- *Contract data*—there are many types of contracts that an organization executes every year, and there are multiple liabilities associated with each of them.
- *Labor data*—elastic labor brings a set of problems that organizations need to solve.
- *Maintenance data*—records from maintenance of facilities, machines, non-computer-related systems, and more.
- *Financial reporting data*—corporate performance reports and annual filing to Wall Street.
- *Compliance data*—financial, healthcare, life sciences, hospitals, and many other agencies that file compliance data for their corporations.
- *Clinical trials data*—pharmaceutical companies have wanted to minimize the life cycle of processing for clinical trials data and manage the same with rules-based processing; this is an opportunity for Big Data.
- *Processing doctors' notes on diagnosis and treatments*—another key area of hidden insights and value for disease state management and proactive diagnosis; a key machine learning opportunity.
- *Contracts*—every organization writes many types of contracts every year, and must process and mine the content in the contracts along with metrics to measure the risks and penalties.

Big Data example

In order to understand the complexities of the different types of data and their associated content including integration challenges, let us examine a sample of text and analytics mixed within content from different sources. In this example of a large and popular restaurant chain organization that wants to know the correlation between its sales and consumer traffic based on weather conditions. There are both historic patterns and current patterns that need to be integrated and analyzed, and added to this complexity is the social media sharing of consumer perspectives about shopping experiences and where weather is mentioned as a key factor. All this data needs to be added to the processing.

Upon deeper examination of the requirements, we find we need the data sets described in Table 1.1.

The complexities that exist in the data sets across the different sources in terms of data quality, volume, and variety make it difficult to integrate the same seamlessly. Let us examine the sample data to gain a better perspective. The next section discusses the example of Social Media posts and has several example websites that provide insight into the importance of content and context.

Social Media posts

Drive-Thru Windows Still Put the Fast in Fast Food Restaurants, May 30, 2012, *www.npd.com/wps/portal/npd/us/news/pressreleases/pr_120530a*.

> *Last year 12.4 billion visits were made through fast-food drive-thrus, a two percent. … In the total quick service/fast food restaurant segment, carry-out … about their drive-thru experience it's because they somehow lost time.*

Table 1.1 Sample Data Sets Needed for Restaurant Service Performance Analytics

Data	Features	Source	Complexity
Weather	Structured and semi-structured Available for all latitude and longitude codes Has 'weatherperson's editorial commentary and user shared data	Governmental agencies Public news channels Social media	Metadata Geo-coding Language Image and video formats
Customer sentiment	Voice Text Images Videos Blogs, forums, and other Internet channels	Call center Social media Campaign Databases Customer resource management (CRM)	Metadata Context Language Formats
Product	Corporate product data	In-house	HierarchiesMenu packaging
Competition	Available in structured formats from data federators Available unstructured from social media, forums, and Internet	Third party Social media Internal research	Metadata Data quality Context
Location	Structured Unstructured	Internal MDM Social media Third party Surveys	Metadata Data quality Formats: images and videos
Campaign	Structured	Internal	Multiple campaigns at any given time across geographies

Quick Fast Food Service Crucial to Success, *QSR Magazine*, *www.qsrmagazine.com/ordering/fast-food-fast*.

> *Fast Food Fast—The fast food industry is based on the principles of quality food … for customers who join the drive-thru or in-store queue: to get quality food fast … and it takes 10 minutes, you'll be annoyed by the time you get service.*

Drive Thru Study for Quick Serve Restaurants, *QSR Magazine*, March 16, 2011, *www2.qsrmagazine.com/reports/drive-thru-experience*.

> *Drive Thru Experience Study—You're fast, you're accurate, but today's. … In that same time frame, domestic and international same-store sales … when it comes to fast food or fast casual let alone any restaurant experience.*

Evolution Fast Food—Banker's Hill—San Diego, CA, *www.yelp.com› Restaurants › Vegetarian*.

> *I never realized how douchey using a drive-thru could make you look as a … much to say other than decent food, if a bit overpriced, friendly service, if a bit slow.*

Survey data analysis

The data shown in Figures 1.1, 1.2, and 1.3 represents a sample of consumer sentiment across three categories as measured. The trend from this data set shows the customers' expectations and the reality of service at a fast-food drive-thru.

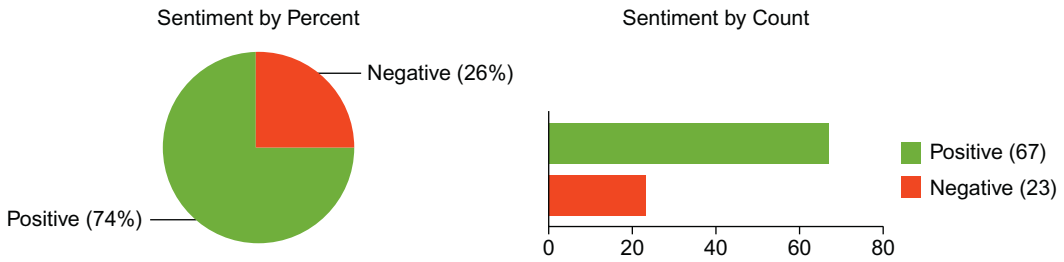

FIGURE 1.1

User sentiment analysis for burgers across the United States.

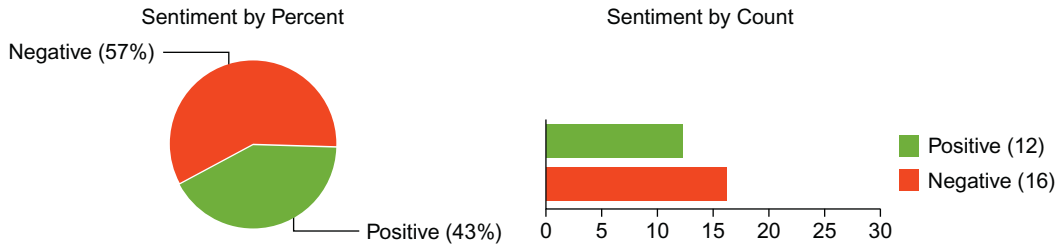

FIGURE 1.2

User sentiment analysis of drive-thru performance of U.S.-based fast-food restaurants.

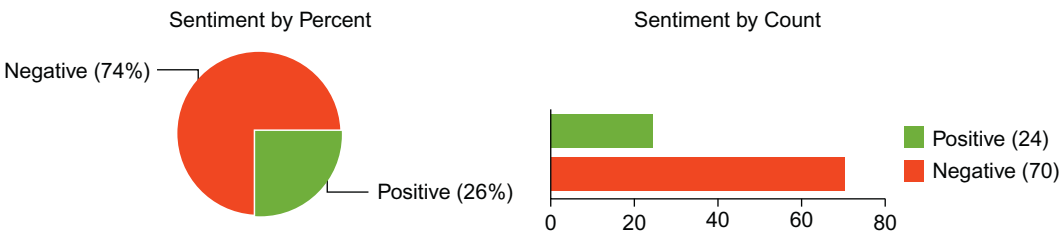

FIGURE 1.3

User sentiment analysis of fast-food quality in the United States.

Survey data

The data sample in Figure 1.4 is from a survey of fast-food restaurants from FoodSource (*www.foodsource.com*). A survey typically will consist of hundreds of such responses and needs to be processed for creating a cluster from a population and their associated responses. Often the data becomes cumbersome to manage due to lack of demographics of the survey targets.

How do you make your fast food order a little healthier?
Don't order large size 18%
I don't 16%
Chicken instead of beef 13%
No mayo/spread 12%
No cheese 11%
Skip the side item 9%
Low or no-calorie drink 9%
No bread (protein style) 4%
Substitute veggies for fries 4%
Eat half now, half later 3%
Saled instead of burger 3%

Other tricks to ordering more "healthy" fast food
Ask for knife and fork and ditch the bun after eating contents,
Double garden; double lettuce, tomato, pickles on sandwich; replacing mayo with juicy Vegetables,
If you order fries, get them with no salt.
less fried
Moderations is the #1 tip
places that use fresh items
Whole wheat or multigrain bun

How often do you feed your family fast food rather than cooking at home?
about once a week 40%
1-2 times per month 29%
more than once a week 18%
Never 13%

How often do you eat fast food?
more than once a week 52%
about once a week 28%
1-2 times per month 20%

Where do you live?
U.S. Southwest 35%
U.S. Midwest 24%
U.S. Southeast 13%
U.S. Northeast 11%
U.S. Northwest 11%
U.S. South 4%

Are you male or female?
Male 55%
Female 45%

(a)

Who makes your favorite burger?
McDonald's 18%
Burger King 16%
In-N-Out 11%
Carl's Jr, 9%

What's your favorite burger?
Big Mac 13%
Whopper 8%
Whopper Jr, 8%

Who has the best french fries?
McDonald's 21%
Rallys/Checkers 9%
Wendy's 8%
Burger King 7%
Del Taco 7%
Steak N Shake 6%

What's the best place for fast food? other than burgers?
Taco Bell 20%
Chick-fil-a 10%
KFC 10%

Best chain for order accuracy
Arby's 10%
Carl's Jr. 10%

Worst chain for order accuracy?
Burger King 23%
McDonald's 19%
El Pollo Loco 11%
Taco Bell 11%

Best chain for drive-thru speed
McDonald's 38%
Taco Bell 10%

Worst drive-thru speed
El Pollo Loco 11%

Friendlest chains
Subway 9%
Taco Bell 9%
In-n-out 8%
Carl's Jr, 7%

Least friendly chains
McDonald's 20%
Burger King 12%

Best value
McDonald's 20%
Taco Bell 15%

Best quality
McDonald's 11%
Chipotle 9%
Chick-fil-a 9%

Best chain for kids
McDonald's 61%
Dairy queen 7%
Note: This was the largest margin in the entire survey

How old are you?
36-45 30%
26-35 20%
18-25 15%
Under 18 15%
Over 55 11%
46-55 9%

(b)

FIGURE 1.4

(a) Survey sample and (b) the data.

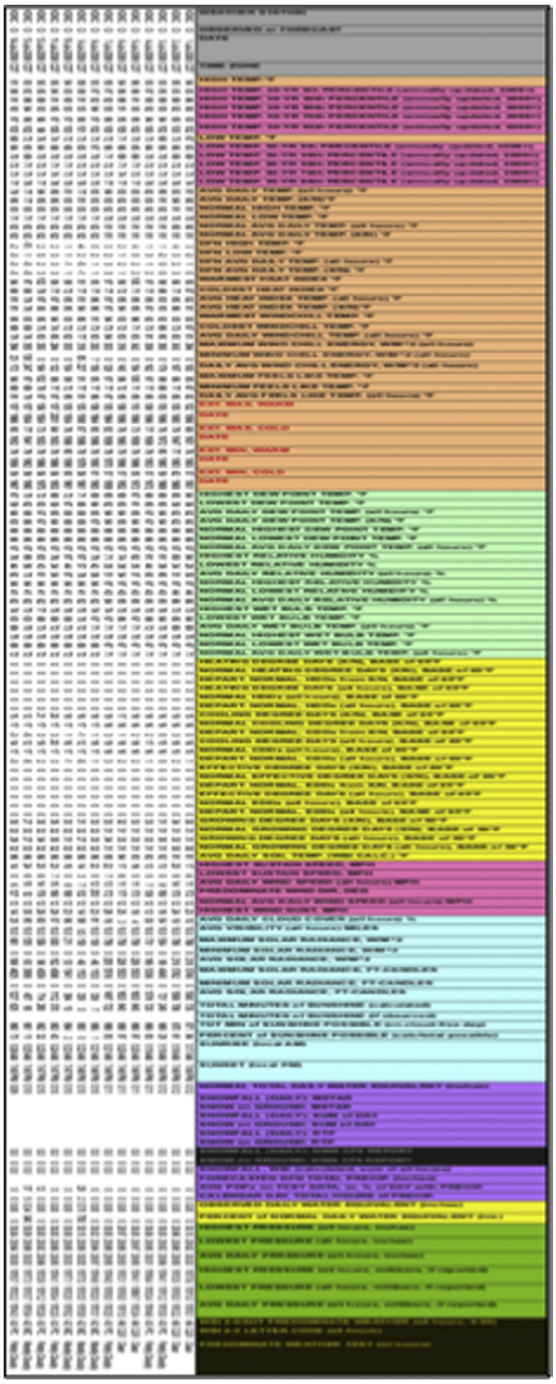

FIGURE 1.5

Sample weather data.

(Source: www.weatherbank.com.)

FIGURE 1.6

Sample weather map.

Weather data

Weather data is another common data type that companies today want to integrate in their data analysis and discovery platform. Presented in Figures 1.5 and 1.6 are common formats of weather data that is available for integration and consumption.

Twitter data

Another popular and most requested data for sentiments, trends, and other relevant topics is the social media channel Twitter. Twitter is a microblog and often contains less than 140 characters per tweet. The complexity is to understand the context of the flow and the associated content before proceeding with analysis (Figure 1.7). Twitter data has more complexity than any other web data, the reason being the cryptic notational format used by different consumers of the platform.

Integration and analysis

Based on the data examples discussed here, you can see that a lot of information is available in different sources and formats that can be harnessed into powerful analytics to create a disruptive differentiator for an organization. The fast-food company in the example that is observing the correlation between weather and food sales can answer the following types of questions more effectively:

- What sales occurred across the entire United States for a given day/week/month/quarter/year, and under what weather conditions?
- Did people prefer drive-thru in extreme weather conditions irrespective of the geography?

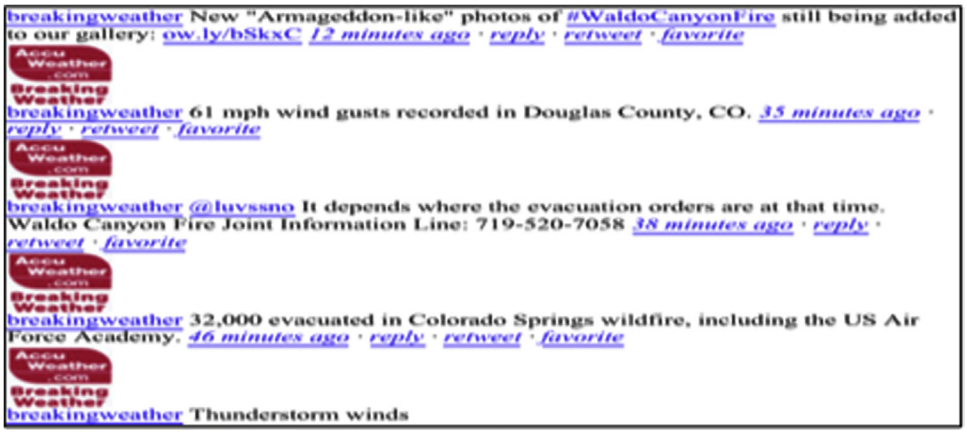

FIGURE 1.7

Tweets about weather.

- Did restaurants along the highway get more traffic in drive-thru during regular versus abnormal weather?
- Did service interruptions occur due to weather?
- What is the propensity for business impact in abnormal weather?
- Did customers wait at one zip code more than another zip code of the same population or demographics? More importantly, what factor did weather play?
- Does coffee sell more than burgers in the winter in Boston, and during the same time, do more cold beverages sell in Orlando?
- Do restaurants need to staff more in different weather conditions? What is the budget impact in such situations?
- What is the customer sentiment, especially during different weather patterns? What drives customers to the store in these circumstances? How can quality of service be improved and sustained?
- What are the customer expectations of pricing? Do they provide feedback by phone, email, or social media?
- What is the competition comparison by customers in social media?
- Do customers measure quality of service alone as the yardstick or are there other mentions in social media?
- Do customers differ in their purchase behaviors across geographies?

You might ask, can we not answer these questions without all this additional data? Or, do we not answer these questions today? The answer to both these questions is countered with another question: What is the effectiveness of the decisions you make today in your business? How aligned are these decisions to the market and, most importantly, your customer? Considering the agile requirements in decision making based on today's fast-paced market and changing economic conditions, every business needs to have a 360° view of data across their organization. A 360° view of your customer or your market or your organization is combined as inside-out (i.e., your view of the market

Table 1.2 Data Output

Source	Data	Metric
Weather	Latitude/longitude Temperature Forecast Time zone Date and hour	Average daily temperature Average snowfall/day Maximum temperature Minimum temperature
Customer sentiment	Sentiment: happy, disappointed, frustrated Tone Channel Influence Followers Posts	Total number of posts Average number of posts Average repost Total positive Total negative Total followers Amplification
Competition	Competitor name Product/service Channel Posts Authors	Total number of posts Total number of authors Average post/channel Average post/author Average compare/product Average compare/author
Contracts	Type Date range Liabilities	Total number of contracts Total type of contracts Contracts/date range—expiry Contract/type of liability
Location	Address Date and time Staff friendliness Cleanliness Quality of service	Number of visits Service time Wait time Quality of service Cleanliness

and customer) and outside-in (i.e., your customer or market's view of you) viewpoints in the form of data and its associated analytics and visualizations. This extends to including data such as contracts, compliance reporting, Excel spreadsheets, safety reports, surveys and feedback, and other data sets. The next section discusses examples of additional data and metrics associated with the data that form portions of the Big Data needed by an organization.

Additional data types

Let us proceed further and assume that all the data has been extracted and transformed from various sources. The output for each of them will look as outlined in Table 1.2.

When all the data is integrated with the data existing in the current business intelligence platforms, the fast-food company can get better insights into the following subject areas:

- Customers
- Markets
- Products

- Vendors/suppliers
- Contracts
- Labor management
- Campaign
- Location management

The analytics and trends that can be created with these additional metrics will provide analysts within the fast-food organization better insights into what drives business and how weather can form a powerful disruption to the business and, more importantly, the consumer. Additionally, in a business-to-business scenario, the data from contracts and liabilities provides context-related information that can aid in negotiations and renewal situations.

The promise of Big Data as seen from the fast-food example in this chapter proves one of the basic reasons why the entire industry is abuzz with wanting to adopt Big Data within their organization. Based on several examples we have discussed in this chapter, Big Data is complex, and this complexity is driven by three characteristics: volume, velocity, and variety. At this juncture, consider your organization and write down a list of missing information that is due to volume, velocity, variety, or complexity of processing issues.

SUMMARY

In this chapter, we discussed an example-driven approach to understanding Big Data. The issue of finding value is dwarfed when compared to the complexity and ambiguity associated with Big Data. In the next chapter, we will discuss the complexities associated with Big Data, and how to derive value from the complexity.

Further reading

Hedberg, B. (1981). How organizations learn and unlearn. In: Nyström, P. C., & Starbuck, W. H. (Eds.), *Handbook of Organizational Design*. Oxford University Press, USA.

Mullins, L. J. (1993). *Management and Organizational Behaviors*, (3rd ed.).

www.mckinsey.com

www.forrester.com

www.gartner.com

www.tdwi.org