



Chapter 1

WHAT ARE BIG DATA AND ANALYTICS?

LEARNING OBJECTIVES

After completing this chapter, you should be able to do the following:

- Identify the three different types of data.
- Recall what type of data volume Big Data represents.
- Recognize Big Data terminology.

INTRODUCTION

In the early 20th century, businesses kept track of financial and operational results using paper and ink. It was difficult enough just to record the date of the transactions, let alone summarize information with financial statements. The main form of automation that helped improve the efficiency of accounting clerks was limited to innovations in carbon copy paper, mimeograph machines, copy machines, and the like. When computers finally were available for operational and financial use, the systems were based on a batch recording of transactions. Again the focus was on capturing internal data to help an organization understand its financial and operational results. As computers advanced and became more powerful, the focus increased in obtaining more internally generated operational and financial information as well as analyzing the myriad information as a result of increased computing power, increased data, and more user-friendly tools.

Prior to the advent of the Internet, an organization worked mainly with its internal data. With the subsequent advances in Internet use in the latter half of the 20th century and the beginning of the 21st

century, external information became accessible that could be integrated with internal data. Companies moved from producing batch information to employees generating information (on both the corporate and personal level), to sensors producing data about all aspects of our lives. This last point can be frightening because appliances, sensors, and different apparatuses are generating more data in shorter periods of time than ever before. This has resulted in a flood of information, the concept of *Big Data*, and predictive analytics.

DEFINITION—WHAT IS BIG DATA?

What is *Big Data*? Big Data is a set of high-volume, high-velocity, and high-variety information that demands cost-effective, innovative forms of information processing for enhanced insight and decision making.¹

The end goal of Big Data should be to leverage the information resulting in increased value to the customer and an organization.

HOW BIG IS BIG? VOLUME LEVELS IN BIG DATA

In addition to transactional data and user-created data, the advent of the Internet opened the floodgates to new databases, new forms of data, and data that no longer needed to be created by human intervention.

DOMO.com created an analysis of the amount of data that is processed or created every minute over the Internet.² Consider the following by-the-minute volumes:

- YouTube users upload 400 hours of new video.
- Snapchat users watch 6,944,444 videos.
- Facebook messenger shares 216,302 photos.
- Amazon makes \$222,283 in sales.
- Instagram users like 2,430,555 posts.
- Siri answers 99,206 requests.
- Dropbox users upload 833,333 new files.

From Internet Stats Live³

- Email users send 160,000,000 messages.⁴
- Twitter users tweet 450,793 times.
- Internet users in the world: 3,588,643,537
- Facebook active users: 1,867,648,665

¹ IT glossary, Gartner, www.gartner.com/it-glossary/big-data

² “Data Never Sleeps 4.0,” <https://www.domo.com/learn/data-never-sleeps-4-0> Accessed 3/28/17

³ www.internetlivestats.com/one-second/#google-band, accessed 3/28/17

⁴ Timed live at 7:35 p.m. on 3/15/2017, www.internetlivestats.com/one-second/#google-band

- Pinterest active users: 203,925,085
- Websites hacked: 63,345⁵

Just a couple of years ago, the number of global Internet users was estimated to be in excess of 2.4 billion people. Internet Stats Live estimates that there are now 3.6 billion global users.

The amount of data continues to grow exponentially. There's nothing on the horizon that suggests this increase of information will not continue. The challenge for the accountant is managing the expansion of information in terms of collecting, archiving, accessing, and interpreting. The growth in structured data, unstructured data, streaming data, and the like will only continue.

Why has everyone become so interested in the explosion of data that has become known as Big Data? The McKinsey Global Institute published research in 2011 in which it estimated that “retailers exploiting data analytics at scale across their organizations could increase their operating margins by more than 60 percent and that the U.S. healthcare sector could reduce costs by 8 percent through data-analytics efficiency and quality improvements.”⁶

KNOWLEDGE CHECK

1. How can Big Data best be described?
 - a. Large systems in multi-national companies.
 - b. Structured data, unstructured data, and streaming data.
 - c. Enterprise resource planning (ERP) systems with all software applications in the organization.
 - d. Data processed with serial processing.
2. It is estimated that Snapchat users watch how many videos every minute of the day?
 - a. Nearly 1 million.
 - b. Nearly 4 million.
 - c. Nearly 7 million
 - d. Nearly 10 million.

⁵ This figure is from 3/ 14/ 2017.

⁶ David McCourt, “Getting Big Impact with Big Data,” *McKinsey Quarterly*, January 2015, www.mckinsey.com/insights/business_technology/getting_big_impact_from_big_data

EXAMPLES OF VOLUME

What type of data volumes does Big Data involve?

Table 1-1		
Acronym	Description	Size
(B)	Byte	= 8 bits
(KB)	Kilobyte	= 1,000 bits
(MB)	Megabyte	= 1,000,000 bits
(GB)	Gigabyte	= 1,000,000,000 bits
(TB)	Terabyte	= 1,000,000,000,000 bits
(PB)	Petabyte	= 1,000,000,000,000,000 bits
(EB)	Exabyte	= 1,000,000,000,000,000,000 bits
(ZB)	Zettabyte	= 1,000,000,000,000,000,000,000 bits
(YB)	Yottabyte	= 1,000,000,000,000,000,000,000,000 bits

MEGABYTES, GIGABYTES, TERABYTES ... WHAT ARE THEY?

How much data could be contained in the preceding measurements? We turned to WhatsAByte.com to find out.⁷

Byte: 100 bytes equates to an average sentence like this one.

Kilobyte: 100 kilobytes equals a page of words like the one you're reading now.

Megabyte: 100 megabytes equals a couple volumes of encyclopedias. 600 megabytes is about the amount of data that will fit on a CD-ROM disk.

Gigabyte: 100 gigabytes could contain an entire library floor of academic journals.

Terabyte: A terabyte could hold 1,000 copies of the Encyclopedia Britannica. Ten terabytes could hold the printed collection of the Library of Congress.

⁷ "Megabytes, Gigabytes, Terabytes...What Are They?" What's a Byte?, accessed 3/ 28/ 17, www.whatsabyte.com/

Petabyte: A petabyte could hold approximately 20 million four-door filing cabinets full of text. It could contain 500 billion pages of standard printed text.

Exabyte: It's estimated that five exabytes would be equal to all of the words ever spoken by mankind.

Zettabyte: 1 ZB is equivalent to approximately 152 million years of high-definition video.⁸

KNOWLEDGE CHECK

3. A petabyte could contain how many billion pages of standard text?
- a. 100.
 - b. 500.
 - c. 900.
 - d. 750.

THE ACCOUNTANT AND BIG DATA

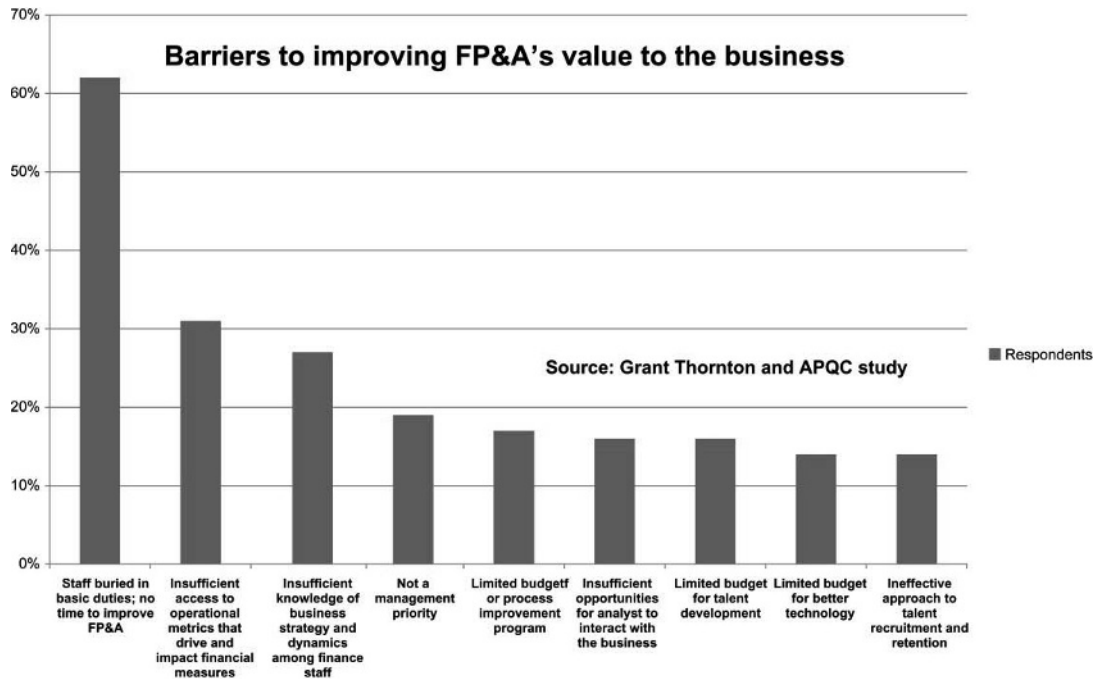
Although many organizations have sought to leverage Big Data applications and resources, they have not had the time or resources to pursue the dream fully. The American Productivity and Quality Center (APQC) conducted a study sponsored by Grant Thornton entitled "Financial Planning and Analysis: Influencing Corporate Performance with Stellar Processes, People, and Technology."⁹ One of the study's conclusions was that the finance staff has not been at the forefront of the battle of Big Data. Two-thirds of survey respondents indicated that they spent too much time on basic financial management duties to improve financial planning and analysis. When asked what the most significant barriers to improving financial planning and analysis value to the business were, they responded as follows (see table 1-2):¹⁰

⁸ www.economist.com/news/leaders/21707538-internet-not-american-whatever-ted-cruz-thinks-road-surfdom, accessed 3/ 28/ 17

⁹ "Financial Planning and Analysis: Influencing Corporate Performance with Stellar Processes, People, and Technology," Grant Thornton website, March 23, 2015, www.grantthornton.com/-/media/content-page-files/advisory/pdfs/2015/BAS-APQC-financial-planning-and-analysis-final.ashx, accessed 3/ 28/ 17

¹⁰ Ibid. 11.

Table 1-2



According to the study, “financial planning and analysis departments are consumed by the basics: data management, process administration, managing the machinery of periodic forecasting and variance analysis and working with the accounting staff to correct posting errors.”

Having limited time to focus on data analysis, what were the major areas the financial planning and analysis group could focus on?

- Simple aggregation of exposures and losses 60 percent
- Basic cause-and-effect analysis 57 percent
- Scenarios and “what-if” analyses to identify possible outcomes 36 percent
- Predictive analysis techniques to project probable outcomes 24 percent

ACCOUNTING'S BIG DATA PROBLEM¹¹

According to CFO.com, unless accountants and finance executives work for companies in businesses that provide or deliver data products and services, they may not be participants in the Big Data trend because most of them have been trained almost exclusively on structured data (data that fits into tables, Excel spreadsheets, databases, and the like) rather than unstructured data.

Keep in mind that unstructured data represents the most significant segment of existing data and will probably yield the largest benefit.

¹¹ Katz, David M., “Accounting's Big Data Problem,” *CFO*, March 4, 2014, ww2.cfo.com/management-accounting/2014/03/accountings-big-data-problem/, accessed 3/28/17

One such example of the unstructured data comes from Trax—a Singapore-based firm that provides an image-recognition app to gather data from photos taken of shelves at retail stores. The photos allow an organization to better manage inventories.

Another example of unstructured data can be found in corporations' published text in the following sources:

- 10-Ks and 10-Qs Management's Discussion and Analysis
- Press releases
- Interviews with corporate executives

BIG DATA TERMINOLOGY

As in any new field, Big Data has some terms that must be mastered. The following list is not meant to be all-inclusive, but it identifies many of the terms related to Big Data, analytics, and business intelligence.

Business intelligence (BI). The integration of data, technology, analytics, and human knowledge to optimize business decisions and ultimately drive an enterprise's success. BI programs usually combine an enterprise data warehouse and a BI platform or toolset to transform data into usable, actionable business information.¹²

Data analytics (DA). The science of examining raw data with the purpose of drawing conclusions from that information. Data analytics is used in many industries to allow companies and organizations to make better business decisions, and in the sciences to verify or disprove existing models or theories.¹³

Cloud computing. A model for delivering information technology services in which resources are retrieved from the Internet through web-based tools and applications rather than a direct connection to a server. Data and software packages are stored in servers. However, cloud computing allows access to information as long as an electronic device has access to the web. This type of system allows employees to work remotely.¹⁴

Dashboards. A business intelligence dashboard (BI dashboard) is a BI software interface that provides preconfigured or customer-defined metrics, statistics, insights, and visualization into current data. It allows the end and power users of BI software to view instant results into the live performance state of business or data analytics.¹⁵

Data mining. The practice of searching through large amounts of computerized data to find useful patterns or trends.¹⁶

¹² "Featured Education, Research and Resources," TDWI, <https://tdwi.org> Accessed 3/28/17

¹³ "Data analytics," *TechTarget*, Accessed 3/28/17, searchdatamanagement.techtarget.com/definition/data-analytics

¹⁴ "Cloud computing," *Investopedia*, Accessed 3/28/17, www.investopedia.com/terms/c/cloud-computing.asp

¹⁵ "Dashboards," *Techopedia*, Accessed 3/28/17, www.techopedia.com/definition/13773/business-intelligence-dashboard-bi-dashboard

¹⁶ "Data mining," *Merriam-Webster's Collegiate Dictionary*, 11th edition, Springfield, MA, Merriam Webster, 2003. Also available at www.merriamwebster.com.

Data scientist. An employee or BI consultant who excels at analyzing data, particularly large amounts of data, to help a business gain a competitive edge.¹⁷

Data visualization. The presentation of data in a pictorial or graphic format.

Hadoop. A free, java-based programming framework that supports the processing of large data sets in a distributed computing environment. It is part of the Apache project, sponsored by the Apache Software Foundation.¹⁸

OLAP. (OnLine Analytical Processing). OLAP is a powerful technology for data discovery, including capabilities for limitless report viewing, complex analytical calculations, and predictive “what-if” scenario (budget, forecast) planning.¹⁹

Predictive analytics. The practice of extracting information from existing data sets to determine patterns and predict future outcomes and trends. Predictive analytics does not tell you what will happen in the future. It forecasts what might happen in the future with an acceptable level of reliability, and includes what-if scenarios and risk assessment.²⁰

Prescriptive analytics. A type of business analytics that focuses on finding the best course of action for a given situation and belongs to a portfolio of analytic capabilities that include descriptive and predictive analytics.²¹

Semi-structured data. Data that has not been organized into a specialized repository, such as a database, but that nevertheless has associated information, such as metadata, that makes it more amenable to processing than raw or unstructured data. For example, a Word document contains metadata or tagging that allows for keyword searches, but it does not have as much relational structure or utility as the information in a database.²²

Structured data. Data that resides in a fixed field within a record or file. This includes data contained in relational databases and spreadsheets.²³

Unstructured data. Information that doesn’t reside in a traditional row-column database. It often includes text and multimedia. Examples include email messages, word processing documents, videos, photos, audio files, presentations, web pages, and many other kinds of business documents. Although these files may have an internal structure, they are considered “unstructured” because the data is not contained in a database. Experts estimate that 80 percent to 90 percent of the data in any organization is unstructured.²⁴

¹⁷ “Data scientist,” *TechTarget*, Accessed 3/ 28/ 17, [searchbusinessanalytics.techtarget.com/ definition/ Data-scientist](http://searchbusinessanalytics.techtarget.com/definition/Data-scientist)

¹⁸ Ibid., “Hadoop,” *TechTarget*, [searchcloudcomputing.techtarget.com/ definition/ Hadoop](http://searchcloudcomputing.techtarget.com/definition/Hadoop)

¹⁹ “OLAP,” *OLAP.com*, Accessed 3/ 28/ 17, [olap.com/ olap-definition/](http://olap.com/olap-definition/)

²⁰ “Predictive analytics,” *Webopedia*, Accessed 3/ 28/ 17, [www.webopedia.com/ TERM/ P/ predictive_analytics.html](http://www.webopedia.com/TERM/P/predictive_analytics.html)

²¹ “Prescriptive analytics,” *TechTarget*, Accessed 3/ 28/ 17, [searchcio.techtarget.com/ definition/ Prescriptive-analytics](http://searchcio.techtarget.com/definition/Prescriptive-analytics)

²² Ibid., “Semi-structured data,” [whatis.techtarget.com/ definition/ semi-structured-data](http://whatis.techtarget.com/definition/semi-structured-data), [whatis.techtarget.com/ definition/ semi-structured-data](http://whatis.techtarget.com/definition/semi-structured-data)

²³ “Structured data,” *Webopedia*, [whatis.techtarget.com/ definition/ semi-structured-data](http://whatis.techtarget.com/definition/semi-structured-data), [www.webopedia.com/ TERM/ S/ structured_data.html](http://www.webopedia.com/TERM/S/structured_data.html)

²⁴ Ibid., “Unstructured data,” [whatis.techtarget.com/ definition/ semi-structured-data](http://whatis.techtarget.com/definition/semi-structured-data), [www.webopedia.com/ TERM/ U/ unstructured_data.html](http://www.webopedia.com/TERM/U/unstructured_data.html)

FOUR TYPES OF DATA ANALYTICS

Adding to the examples and sources of Big Data, let's examine how some of the definitions and related terminology we've just learned fit into processes used to examine data. Different types of analytics can be used to analyze Big Data for different purposes.

Descriptive Analytics

Descriptive analytics is information that has happened in the past. From an accounting perspective, this would represent traditional historical financial information. Consider the following examples:

- An assessment of customer credit risk can be predicted based on that company's past financial performance.
- A prediction of sales results can be created from customers' product preferences and sales cycle.
- Current product reviews can be used to predict future sales.
- Employee evaluation can be used to predict turnover.

Diagnostic Analytics

Diagnostic analysis describes the reason for the historical results. It attempts to answer the question "Why did this happen?" as in the following examples:

- In traditional finance, variance analysis uncovers the underlying reasons for differences in budgeted and actual results.
- Causal analysis can be used to describe why certain results occurred.
- Analytic dashboards can be used to describe why something happened. For example, during the Ebola outbreak in Africa, it was possible to view the daily spread of the virus as it occurred in different geographic regions.
- Tracking the increase in views, posts, fans, followers, and so forth, as a result of purchasing additional views on Facebook to increase the exposure of a particular post, video, or picture.

Discovery Analysis (Insight)

Although not technically one of the four types of data analytics, the step of discovery analysis could be inserted between diagnostic and predictive analytics. During discovery analysis or insight, research and analysis can be undertaken to identify whether there is a relationship between the historical information and another database.

Predictive Analytics

Predictive analytics attempts to determine what will happen by analyzing historical data and trends. Consider the following examples of predictive analytics:

- An accounting department prepares a cash flow projection report.
- Preparing an estimate of inventory levels
- Predicting an outcome based on changed assumptions. The revenue will increase by a specific percentage if an additional 5 percent is spent on the marketing budget.
- The issuance of additional coupons or promotions for a retail organization is projected to result in a 10 percent increase in revenue.
- Based on historical results, ads released the week of Black Friday are predicted to generate greater than normal sales for the Black Friday holiday shopping season.

Here is another well-known example of predictive analytics from the sports world:

During the early 2000s, the New York Yankees were the most acclaimed team in Major League Baseball. But on the other side of the continent, the Oakland A's were racking up success after success, with much less fanfare—and much less money.

While the Yankees paid its star players tens of millions, the A's managed to be successful with a low payroll. How did they do it? When signing players, they didn't just look at basic productivity values such as RBIs, home runs, and earned-run averages. Instead, they analyzed hundreds of detailed statistics from every player and every game, attempting to predict future performance and production. Some statistics were even obtained from videos of games using video recognition techniques. This allowed the team to sign great players who may have been lesser-known, but who were equally productive on the field. The A's started a trend, and predictive analytics began to penetrate the world of sports with a splash, with copycats using similar techniques.²⁵

Perhaps predictive analytics will someday help bring Major League salaries into line.

There are tools that can also be used as part of the predictive analysis. One such example that will be addressed in a later chapter is the Net Promoter Score. The Net Promoter Score provides an indication of how willing a customer is to promote or recommend your products, on a scale of 1–10. Companies want to achieve a 9 or 10 with each customer. At this level, the customer will be “promoting” your product to other potential customers.

Prescriptive Analytics

Prescriptive analytics uses the information from descriptive, diagnostic, and predictive analytics to suggest specific decisions or changes in approach to a business strategy. It could also be described as the best scenario to take to achieve the desired outcome. The following are examples of prescriptive data analytics:

- Airline seat prices and the manner in which the cost per seat regularly increases as the departure date draws near. A related component of this decision is the airlines' decision to overbook flights and offer incentives to placate passengers who are inconvenienced.
- Applications such as Facebook suggest to the user that there are additional friends they may wish to connect with. This “prescription for connecting” is based on the analysis of common friends in both of the individuals' profiles. Hence, the new friends and our potential friends are suggested as contacts.
- The most common prescriptive analytics would be medical drugs that have been known to alleviate certain medical issues (statin drugs, diabetic drugs, blood pressure drugs, and the like). The medications can also have negative predictions due to potential problem interactions.

²⁵ “Predictive Analytics: A Primer,” *Amberoon*, accessed 3/ 28/ 17, www.amberoon.com/CarpeDatumRx/bid/236799/Predictive-Analytics-A-Primer

BENEFITS OF BIG DATA

Now that you know what Big Data is, you may be wondering how it will help you in your practice. What are the benefits an organization can derive from Big Data? A study from IBM²⁶ showed that organizations competing on analytics outperform competitors by:

- 1.6 x revenue growth
- 2.5 x stock price appreciation
- 2.0 x EBITDA (earnings before interest, taxes, depreciation and amortization) growth.

Also, the World Economic Forum in 2012 stated that data gathering is a new class of economic asset, like currency and gold.

What are the benefits of Big Data?

Big Data offers strategic benefits for businesses, including the following:

- Better strategic decisions
- Quicker arrival of new products and services to market
- Increased innovation
- Better insight into the business
- Better insight into the competition
- Real-time change for existing products, services, or offers
- Environmental scans for threats or opportunities

Big Data also aids in decision capability enhancement such as the following:

- Increase retained and analyzed amount of data
- Increase the speed of data analysis
- Produce more accurate results
- Better decision-making processes
- Improved forecasting
- More accurate identification of root cause analysis
- Smarter decisions—leverage new sources of data to improve the quality of decision-making
- Faster decisions—enabled more real-time data capture and analysis to support decision-making at the point of impact, such as when a customer is navigating our website or on the telephone with a customer service representative
- Decisions that make a difference—focus Big Data efforts toward areas that provide true differentiation
- Analysis based on entire data sets as opposed to sample sets
- Enhanced transparency of data

Businesses will also experience efficiency improvements, including the following:

- Reduce or eliminate manual processes
- Cost savings
- Increased productivity
- Automated routine decisions

²⁶ “Analytics: The Real-World Use of Big Data,” IBM, Accessed 3/28/17, www.ibm.com/services/us/gbs/thoughtleadership/ibv-big-data-at-work.html

- Improved manufacturing productivity and maintenance
- Integration of previously related databases
- Improved scalability

Customer relationships and sales can also benefit from utilizing Big Data. Some of the benefits in these areas include the following:

- Improved customer satisfaction
- Better customer service
- Increased input from customers
- Improved sales results via cross-selling and upselling
- Increased attraction and retention of customers
- Increased targeted marketing via social media

Finally, Big Data can enhance a business's governance and compliance efforts through the following:

- Improved fraud detection
- Improved risk assessment and management
- Tools that can scan and access corporate data to prevent unauthorized release of data

KNOWLEDGE CHECK

4. Describe decision making within a Big Data framework.
 - a. Smarter, faster, more accurate.
 - b. Slower, more detailed, more structured.
 - c. Slower, more accurate, more transparent.
 - d. Same speed, more accurate, more structured.

Practice Questions

1. What are the components of Big Data?
2. What are the four stages of analytics?
3. What is Hadoop?
4. Describe what a data scientist does.

