16 Big Data

Learning Objectives

* Understand what is Big Data
* Know the business benefits of Big Data
* Outline the technologies involved in Big Data
* Summarize how Big Data is managed

### INTRODUCTION

Big Data is an umbrella term for a collection of datasets so large and complex that it becomes difficult to process them using traditional data management tools. There has been increasing democratization of the process of content creation and sharing over the Internet using social media applications. The combination of cloud-based storage, social media applications, and mobile access devices is helping crystallize the big data phenomenon. The leading management consulting firm, McKinsey & Co. created a flutter when it published a report in 2011 showing the huge impact of such big data on business and other organizations. They also reported that there will be millions of new jobs in the next decade related to the use of big data in many industries.

Big data can be used to discover new insights from a 360-degree view of a situation that can allow for a completely new perspective on situations, new models of reality, and potentially new types of solutions. It can help spot business trends and opportunities. For example, Google can predict the spread of dis-ease by tracking the use of search terms related to the symptoms of the disease over the globe in real-time. Big data can help determine the quality of research, prevent diseases, link legal citations, combat crime, and determine real-time roadway traffic conditions. Big data is enabling evidence-based medicine and many other innovations.

Data has become the new natural resource. Organizations have a choice in how to engage with this exponentially growing volume, variety, and velocity of data. They can choose to be buried under the avalanche or they can choose to use it for competitive advantage. Challenges in big data include the entire range of operations from capture, curation, storage, search, sharing, analysis, and visualization. Big data is more valuable when analyzed as a whole. More and more

information is derivable from analysis of a single large set of related data, as compared to separate smaller sets. However, special tools and skills are needed to manage such extremely large datasets.

#### Caselet: Personalized Promotions at Sears

*A couple of years ago, Sears Holdings concluded that it needed to generate greater value from the huge amounts of customer, product, and promotion data is collected from its many brands. Sears required about eight weeks to generate personalized promotions, at which point many of them were no longer optimal for the company. It took so long mainly because the data required for these large scale analyses were both voluminous and highly fragmented—housed in many databases and “data warehouses” maintained by the various brands. Sears turned to the technologies and practices of big data. As one of its first steps, it set up a Hadoop cluster using a group of inexpensive commodity servers.*

*Sears started using the Hadoop cluster to store incoming data from all its brands and existing data warehouses. It then conducted analyses on the cluster directly, avoiding the time-consuming complexities of pulling data from various sources and combining them so that they can be analyzed. Sears’s Hadoop cluster stores and processes several petabytes of data at a fraction of the cost of a comparable standard data warehouse. The time needed to generate a comprehensive set of promotions dropped from eight weeks to one. And these promotions are of higher quality because they’re more timely, more granular, and more personalized.*

*(Source: McAfee & Brynjolfsson HBS, October 2012)*

1. *What are other ways in which Sears can benefit from big data?*
2. *What are the challenges in making use of big data?*

### DEFINING BIG DATA

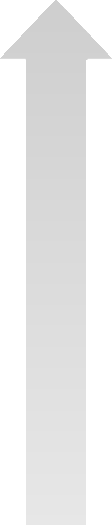
In 2000, there were 800,000 petabytes of data in the world. It is expected to grow to 35 zettabytes by the year 2020. About a million books' worth of data is being created daily on social media alone. Big data is big, fast, unstructured, and of many types. There are several unique features of big data

*Variety* There are many types of data, including structured and unstructured data. Structured data consists of numeric and text fields. Unstructured data includes images, video, audio, and many other types. There are also many sources of data. The traditional sources of structured data include data from ERPs systems and other operational systems. Sources for unstructured data include social media, web, RFID, machine data, and others. Unstructured data comes in a variety of sizes, resolutions, and are subject to different kinds of analysis.

For example, video files can be tagged with labels and they can be played, but video data is typically not computed which is the same with audio data. Graphic data can be analyzed for network distances. Facebook texts and tweets can be analyzed for sentiments but cannot be directly compared.

*Velocity* The Internet greatly increases the speed of movement of data, from e-mails to social media to video files, data can move quickly. Cloud-based storage makes sharing instantaneous and easily accessible from anywhere. Social media applications enable people to share their data instantly. Mobile access to these applications also speeds up the generation and access to data (Figure 16.1).

Big Data = Transactions + Interactions + Observations

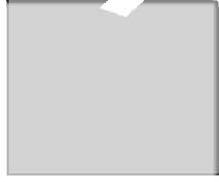
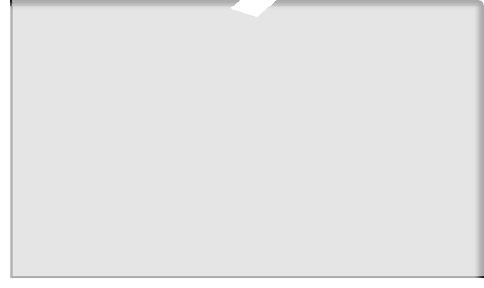
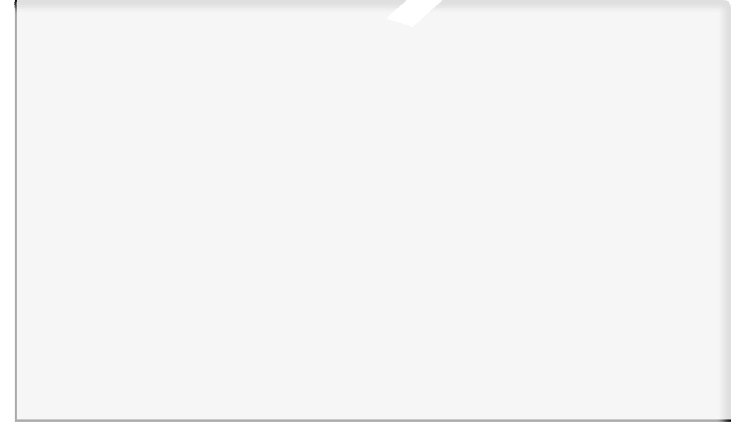


Petabytes

Terabytes

Gigabytes

Megabytes



BIG DATA

Sensors/RFID/Devices

User Generated Content

Mobile Web

Sentiment

Social Interactions & Feeds

User Click Stream

Spatial & GPS Coordinates

External Demographics

Business Data Feeds HD Video, Audio, Images

Search Marketing

Speech to Text

Behavioral Targeting

Customer Touches

Product/Service Logs

Support Contacts

Dynamic Funnels

SMS/MMS

Increasing Data Variety and Complexity

ERP

Purchase detail Purchase record Payment record

Segmentation

Offer details

CRM

Affiliate Networks

Dynamic Pricing

Offer history

A/B testing

WEB

Web logs

*Source:* Contents of the above graphic created in partnership with Teradata, Inc.

FIGURE 16.1 Sources of Big Data (Source: hortonworks.com)

*Volume* Websites have become great sources and repositories for many kinds of data. User clickstreams are recorded and stored for future use. Social media applications such as Facebook, Twitter, Pinterest, and other applications have enabled users to become prosumers (producers and consumers)of data. There is an increase in the number of data shares and also the size of each data element. High-definition videos can increase total shared data. There are autonomous data streams of video, audio, text, data, and so on coming from social media sites, websites, RFID applications, and so on.

*Sources of Data* There are several sources of data including some new ones. Data from outside the organization may be incomplete and of different quality and accuracy.

1. *Social Media* All activities on the web and social media are considered stores and are accessible. The email was the first major source of new data. Google searches, Facebook posts, Tweets, YouTube videos, and blogs enable people to generate data for one another.
2. *Organizations* Business organizations and government are a major source of data. ERP systems, e-commerce systems, user-generated content, web- access logs, and many other sources of data generate valuable data for organizations.
3. *Machines* The Internet of things is evolving. Many machines are connected to the web and autonomously generate data that is untouched by humans. RFID tags and telematics are the two major applications that generate enormous amounts of data. Connected devices such as phones and refrigerators generate data about their location and status.
4. *Metadata* There is enormous data about the data itself. Web crawlers and web-bots scan the web to capture new webpages, their HTML structure, and their metadata. This data is used by many applications including web search engines.

The data also includes a varied quality of data. It depends upon the purpose of collecting the data and how carefully it has been collected and curated. Data from within the organization is likely to be of a higher quality. Publicly available data would include some trustworthy data such as from the government.

### BIG DATA LANDSCAPE

Big data can be understood at many levels (Figure 16.2). At the highest level are business applications to suit particular industries or to suit business intelligence for executives. A unique concept of ‘data as a service’ is also possible for particular industries. At the next level, there are infrastructure elements for broad cross-industry applications, such as analytics and structured databases. This also includes offering this infrastructure as a service with some operational management services built-in. At the core, big data is about technologies and standards to store and manipulate the large fast streams of data and make them available for rapid data-based decision-making.

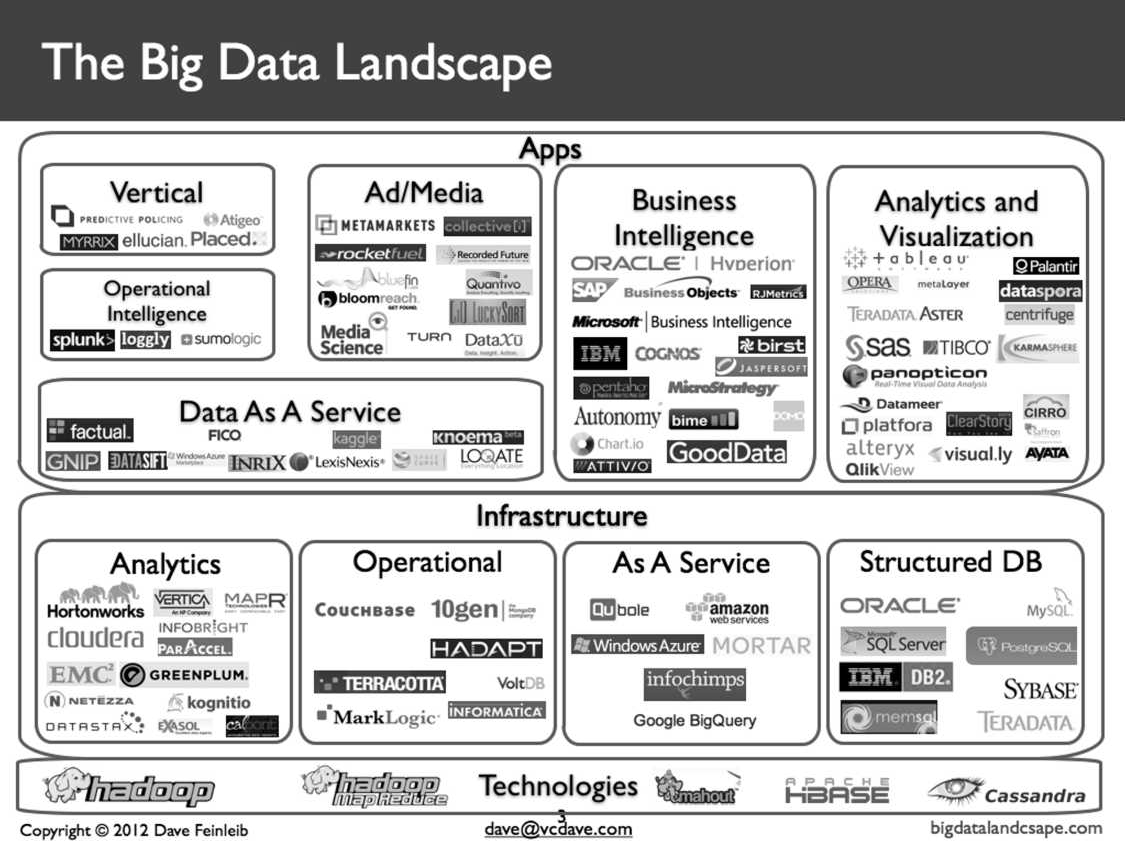


FIGURE 16.2 The Big Data Landscape (*Source: *bigdatalandscape.com)

### BUSINESS IMPLICATIONS OF BIG DATA

*“Big data will disrupt your business. Your actions will determine whether these disruptions are positive or negative.” (Gartner, 2012)*

Any industry that produces information-based products is most likely to be disrupted. Thus, the newspaper industry has taken a hit from digital distribution channels, as well as from published-on-web-only blogs. Entertainment has also been impacted by digital distribution and piracy as well as by user-generated-and- uploaded content on the internet. The education industry is being disrupted by massively online open courses (MOOCs) and user-uploaded content. Healthcare delivery is impacted by electronic health records and digital medicine. The retail industry has been highly disrupted by e-commerce companies. Fashion companies are impacted by quick feedback on their designs on social media. The banking industry has been impacted by the cost-effective online self-serve banking applications and this will impact employment levels in the industry.

There is a rapid change in business models enabled by big data technologies. Steve Jobs, the ex-CEO of Apple, conceded that his company’s products and business models would be disrupted. He preferred his older products to be cannibalized by his new products rather than by those of the competition.

Every other business too will likely be disrupted. The key issue for businesses is how to harness big data to generate growth opportunities and to leapfrog the competition. Organizations need to learn how to organize their businesses so that they do not get buried in high volume, velocity, and variety of data, but instead use it smartly and proactively to obtain a quick but decisive advantage over their competition. Organizations need to figure out how to use big data as a strategic asset in real-time, to identify opportunities, thwart threats, build new capabilities, and enhance operational efficiencies. Organizations can now effectively fuse strategy and digital business, and then strive to design innovative ‘digital business strategy’ around digital assets and capabilities.

### TECHNOLOGY IMPLICATIONS OF BIG DATA

*Big data forces organizations to address the variety of information assets and how fast these new asset types are changing information management demands. (Gartner, 2012)*

The growth of data is made possible in part by the advancement of storage technology. The attached graph shows the growth of disk drive average capacities. The cost of storage is falling, the size of the storage is getting smaller, and the speed of access is going up (Figure 16.3). Flash drives have become cheaper. Random ac- cess memory storage used to be expensive but now is so inexpensive that entire databases can be loaded and processed quickly instead of swapping sections of it into and out of high-speed memory.

New data management and processing technologies have emerged. IT professionals integrate big data structured assets with content and must increase their business requirement identification skills. Big data is going democratic. Business functions will be protective of their data and will begin initiatives around exploit- ing it. IT support teams need to find ways to support end-user-deployed big data solutions. Enterprise data warehouses will need to include big data in some form. The IT platform needs to be strengthened to help provide the enablement of a ‘digital business strategy’ around digital assets and capabilities.

### BIG DATA TECHNOLOGIES

New tools and techniques have arisen in the last 10-15 years to handle this large and still growing data. There are technologies for storing and accessing this data.

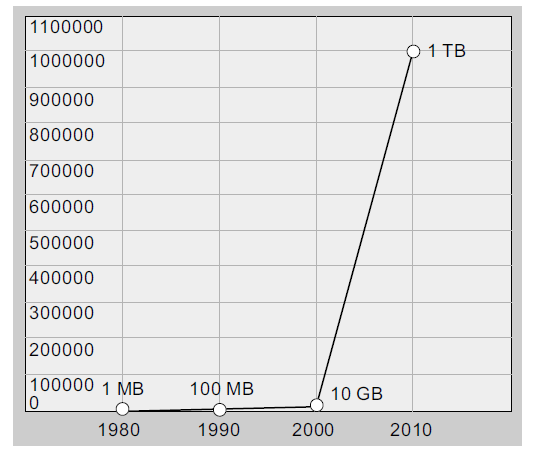


FIGURE 16.3 Data Storage Density Trend

*Non-relational Data Structures* Big data is stored using nontraditional data structures. Large nonrelational databases like *Hadoop* have emerged as a leading data management platform for big data. In Hadoop’s Distributed File System (HDFS), data is stored as ‘key and data-value’ combinations. Google BigFile is another prominent technology. NoSQL is emerging as a popular language to ac- cess and manage non-relational databases. There is a matching Data Warehousing system called Hive along with its PigSQL language. The open-source stack of programming languages (such as Pig) and other tools help make Hadoop a powerful and popular tool.

*Massively Parallel Computing* Given the size of data, it is useful to divide and conquer the problem quickly using multiple processors simultaneously. Parallel processing allows for the data to be processed by multiple machines so that results can be achieved sooner. *The Map-Reduce* algorithm, originally generated at Google for doing searches faster, has emerged as a popular parallel processing mechanism. The original problem is divided into smaller problems, which are then mapped to multiple processors that can operate in parallel. The outputs of these processors are passed to an output processor that reduces the output to a single stream, which is then sent to the end-user. Figure 16.4 shows an example of a Map-Reduce algorithm.

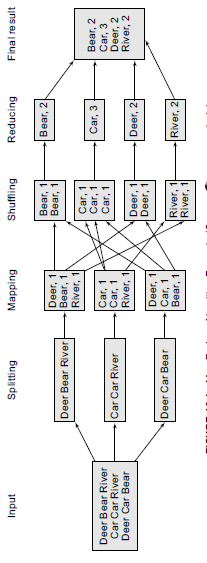




FIGURE 16.4 Map-Reduce Algorithm Example (*Source:* www.cs.uml.edu)

*Unstructured Information Management Architecture (UIMA)* This is one of the elements in the ‘secret sauce’ behind IBM Watson’s system that read massive amounts of data and organizes for just-in-time processing. Watson beat the Jeopardy (quiz program) champion in 2011 and is now used for many business applications, like diagnosis in healthcare situations. Natural language processing is another capability that helps extend the power of big data technologies.

*Spark* Developed by UC Berkeley’s AMP labs, this is a relatively very new way of processing big data that is 10x faster than the Hadoop system. It is a general unified big data computing genuine that covers batch as well as stream processing of data. It integrates the functionality of many of the tools around Hadoop into Spark itself. Besides, Spark does most of its work in memory, thus obtaining the 10x speed.

### MANAGEMENT OF BIG DATA

Many organizations have started initiatives around the use of big data. However, most organizations do not necessarily have a grip on it. Here are some emerging insights into making better use of big data.

* Across all industries, the business case for big data is strongly focused on addressing *customer-centric objectives*. The first focus on deploying big data initiatives is to protect and enhance customer relationships and customer experience.
* *Solve a real pain-point*. Big data should be deployed for specific business objectives to avoid being overwhelmed by the sheer size of it all.
* Organizations are beginning their *pilot* implementations by using existing and newly accessible internal sources of data. It is better, to begin with, data under one’s control and where one has a superior understanding of the data.
* Put *humans and data together* to get the most insight. Combining data-based analysis with human intuition and perspectives is better than going just one way.
* Advanced *analytical capabilities* are required, yet lacking, for organizations to get the most value from big data. There is a growing awareness of building or hiring those skills and capabilities.
* Use more *diverse data*, not just more data. This would provide a broader perspective on reality and better quality insights.
* The *faster* you analyze the data, the more its predictive value. The value of data depreciates with time. If the data is not processed in five minutes, then the immediate advantage is lost.
* *Don’t throw away data* if no immediate use can be seen for it. Data has value beyond what you initially anticipate. Data can add perspective to other data later in a multiplicative manner.
* *Maintain one copy* of your data, not multiple. This would help avoid confusion and increase efficiency.
* Plan for *exponential growth*. Data is expected to continue to grow at exponential rates. Storage costs continue to fall, data generation continues to grow, and data-based applications continue to grow in capability and functionality.
* A *scalable and extensible* information management foundation is a prerequisite for big data advancement. Big data builds upon a resilient, secure, efficient, flexible, and real-time information processing environment.
* Big data is transforming business, just like IT did. Big data is a new phase representing a *digital world*. Business and society are not immune to its strong impacts.

## Conclusion

Big data is a new natural force and natural resource. The exponentially growing volume, variety, and velocity of data are constantly disrupting businesses across all industries, at multiple levels from product to business models. Organizations need to begin initiatives around big data; acquire skills, tools, and technologies; and show the vision to disrupt their industry and come out ahead.

## Questions

1. What are the 3 Vs of big data?
2. How does big data impact the business models?
3. What is Hadoop?
4. How does the Map-Reduce algorithm work?
5. What are the key issues in managing big data?