Big Data

We are accumulating data and information at an increasingly rapid pace from such diverse sources as company documents, e-mails, Web pages, credit card swipes, phone messages, stock trades, memos, address books, and radiology scans. New sources of data and information include blogs, podcasts, videocasts (think of YouTube), digital video surveillance, and RFID tags and other wireless sensors

Organizations and individuals must process an unimaginably vast amount of data that is growing ever more rapidly. According to IDC (a technology research firm), the world generates exabytes of data each year (an exabyte is one trillion terabytes). Furthermore, the amount of data produced worldwide is increasing by 50 percent each year. As we discussed at the beginning of the chapter, we refer to the superabundance of data available today as Big Data. (We capitalize Big Data to distinguish the term from large amounts of traditional data.) We are awash in data that we have to make sense of and manage. To deal with the growth and the diverse nature of digital data, organizations must employ sophisticated techniques for data management

Defining Big Data

It is difficult to define Big Data. Here we present two descriptions of the phenomenon. First, the technology research firm Gartner (www.gartner.com) defines Big Data as diverse, high-volume, high-velocity information assets that require new forms of processing to enable enhanced decision making, insight discovery, and process optimization.

Second, the Big Data Institute (TBDI; www.the-bigdatainstitute.com) defines Big Data as vast data sets that:

- Exhibit variety;
- Include structured, unstructured, and semi-structured data;
- Are generated at high velocity with an uncertain pattern;
- Do not fit neatly into traditional, structured, relational databases ; and
- Can be captured, processed, transformed, and analyzed in a reasonable amount of time only by sophisticated information systems

Big Data generally consists of the following. Keep in mind that this list is not inclusive. It will expand as new sources of data emerge.

- Traditional enterprise data—examples are customer information from customer relationship management systems, transactional enterprise resource planning data, Web store transactions, operations data, and general ledger data.
- Machine-generated/sensor data—examples are smart meters; manufacturing sensors; sensors integrated into smartphones, automobiles, airplane engines, and industrial machines; equipment logs; and trading systems data.

• Social data—examples are customer feedback comments; microblogging sites such as Twitter; and social media sites such as Facebook, YouTube, and LinkedIn.
• Images captured by billions of devices located throughout the world, from digital cameras and camera phones to medical scanners and security cameras.

Let's take a look at a few specific examples of Big Data
• When the Sloan Digital Sky Survey in New Mexico was launched in 2000, its telescope collected more data in its first few weeks than had been amassed in the entire history of astronomy. By 2013, the survey's archive contained hundreds of terabytes of data. However, the Large Synoptic Survey Telescope in Chile, due to come online in 2016, will collect that quantity of data every five days.
• In 2013 Google was processing more than 24 petabytes of data every day. • Facebook members upload more than 10 million new photos every hour. In addition, they click a "like" button or leave a comment nearly 3 billion times every day.
• The 800 million monthly users of Google's YouTube service upload more than an hour of video every second.
• The number of messages on Twitter grows at 200 percent every year. By mid-2013 the volume exceeded 450 million tweets per day.

Characteristics of Big Data Big Data has three distinct characteristics: volume, velocity, and variety.
These characteristics distinguish Big Data from traditional data.
• Volume: We have noted the incredible volume of Big Data in this chapter. Although the sheer volume of Big Data presents data management problems, this volume also makes Big Data incredibly valuable. Irrespective of their source, structure, format, and frequency, data are always valuable. If certain types of data appear to have no value today, it is because we have not yet been able to analyze them effectively. For example, several years ago when Google began harnessing satellite imagery, capturing street views, and then sharing these geographical data for free, few people understood its value. Today, we recognize that such data are incredibly useful (e.g., consider the myriad of uses for Google Maps). Consider machine-generated data, which are generated in much larger quantities than nontraditional data. For instance, sensors in a single jet engine can generate 10 terabytes of data in 30 minutes. With more than 25,000 airline flights per day, the daily volume of data from just this single source is incredible. Smart electrical meters, sensors in heavy industrial equipment, and telemetry from automobiles increase the volume of Big Data.
• Velocity: The rate at which data flow into an organization is rapidly increasing. Velocity is critical because it increases the speed of the feedback loop between a

company and its customers. For example, the Internet and mobile technology enable online retailers to compile histories not only on final sales, but on their customers' every click and interaction. Companies that can quickly utilize that information—for example, by recommending additional purchases—gain competitive advantage.

• Variety: Traditional data formats tend to be structured, relatively well described, and they change slowly. Traditional data include financial market data, point-of-sale transactions, and much more. In contrast, Big Data formats change rapidly. They include satellite imagery, broadcast audio streams, digital music files, Web page content, scans of government documents, and comments posted on social networks.

Managing Big Data Big

Data makes it possible to do many things that were previously impossible; for example, spot business trends more rapidly and accurately, prevent disease, track crime, and so on. When properly analyzed, Big Data can reveal valuable patterns and information that were previously hidden because of the amount of work required to discover them. Leading corporations, such as Walmart and Google, have been able to process Big Data for years, but only at great expense. Today's hardware, cloud computing , and open- source software make processing Big Data affordable for most organizations.

The first step for many organizations toward managing Big Data was to integrate information silos into a database environment and then to develop data warehouses for decision making. After completing this step, many organizations turned their attention to the business of information management—making sense of their proliferating data. In recent years, Oracle, IBM, Microsoft, and SAP have spent billions of dollars purchasing software fi rms that specialize in data management and business intelligence.

In addition, many organizations are turning to NoSQL databases (think of them as "not only SQL" databases) to process Big Data. These databases provide an alternative for fi rms that have more and different kinds of data (Big Data) in addition to the traditional, structured data that fit neatly into the rows and columns of relational databases.

Traditional relational databases such as Oracle and MySQL store data in tables organized into rows and columns. Each row is associated with a unique record, for instance a customer account, and each column is associated with a fi eld that defines an attribute of that account (e.g., customer name, customer identification number, customer address, etc.). In contrast, NoSQL databases can manipulate structured as well as unstructured data and inconsistent or missing data. For this reason, NoSQL databases are particularly useful when working with Big Data. Many products utilize

NoSQL databases, including Cassandra (http:// cassandra.apache.org), CouchDB (http://couchdb.apache.org), MongoDB (www.mongodb .org), and Hadoop (http://hadoop.apache.org). The following example focuses on MongoDB, a leading NoSQL database vendor