chapter 25

Big Data Technologies Based on MapReduce and Hadoop[1]

The amount of data worldwide has been growing ever since the advent of the World Wide Web around 1994. The early search engines—namely, AltaVista (which was acquired by Yahoo in 2003 and which later became the Yahoo! search engine) and Lycos (which was also a search engine and a Web portal—were established soon after the Web came along. They were later overshadowed by the likes of Google and Bing. Then came an array of social networks such as Facebook, launched in 2004, and Twitter, founded in 2006. LinkedIn, a professional network launched in 2003, boasts over 250 million users worldwide. Facebook has over 1.3 billion users worldwide today; of these, about 800 million are active on Facebook daily. Twitter had an estimated 980 million users in early 2014 and it was reported to have reached the rate of 1 billion tweets per day in October 2012. These statistics are updated continually and are easily available on the Web.

One major implication of the establishment and exponential growth of the Web, which brought computing to laypeople worldwide, is that ordinary people started creating all types of transactions and content that generate new data. These users and consumers of multimedia data require systems to deliver user-specific data instantaneously from mammoth stores of data at the same time that they create huge amounts of data themselves. The result is an explosive growth in the amount of data generated and communicated over networks worldwide; in addition, businesses and governmental institutions electronically record every transaction of each customer, vendor, and supplier and thus have been accumulating data in so-called data warehouses (to be discussed in Chapter 29). Added to this mountain of data is the data

[1]We acknowledge the significant contribution of Harish Butani, member of the Hive Program Management Committee, and Balaji Palanisamy, University of Pittsburgh, to this chapter.

generated by sensors embedded in devices such as smartphones, energy smart meters, automobiles, and all kinds of gadgets and machinery that sense, create, and communicate data in the internet of things. And, of course, we must consider the data generated daily from satellite imagery and communication networks.

This phenomenal growth of data generation means that the amount of data in a single repository can be numbered in petabytes (10**15 bytes, which approximates to 2**50 bytes) or terabytes (e.g., 1,000 terabytes). The term *big data* has entered our common parlance and refers to such massive amounts of data. The McKinsey report[2] defines the term *big data* as datasets whose size exceeds the typical reach of a DBMS to capture, store, manage, and analyze that data. The meaning and implications of this data onslaught are reflected in some of the facts mentioned in the McKinsey report:

- A $600 disk can store all of the world's music today.
- Every month, 30 billion of items of content are stored on Facebook.
- More data is stored in 15 of the 17 sectors of the U.S. economy than is stored in the Library of Congress, which, as of 2011, stored 235 terabytes of data.
- There is currently a need for over 140,000 deep-data-analysis positions and over 1.5 million data-savvy managers in the United States. Deep data analysis involves more knowledge discovery type analyses.

Big data is everywhere, so every sector of the economy stands to benefit by harnessing it appropriately with technologies that will help data users and managers make better decisions based on historical evidence. According to the Mckinsey report,

> If the U.S. healthcare [system] could use the big data creatively and effectively to drive efficiency and quality, we estimate that the potential value from data in the sector could be more than $300 billion in value every year.

Big data has created countless opportunities to give consumers information in a timely manner—information that will prove useful in making decisions, discovering needs and improving performance, customizing products and services, giving decision makers more effective algorithmic tools, and creating value by innovations in terms of new products, services, and business models. IBM has corroborated this statement in a recent book,[3] which outlines why IBM has embarked on a worldwide mission of enterprise-wide big data analytics. The IBM book describes various types of analytics applications:

- **Descriptive and predictive analytics**: Descriptive analytics relates to reporting what has happened, analyzing the data that contributed to it to figure out why it happened, and monitoring new data to find out what is happening now. Predictive analytics uses statistical and data mining techniques (see Chapter 28) to make predictions about what will happen in the future.

---

[2]The introduction is largely based on the McKinsey (2012) report on big data from the McKinsey Global Institute.

[3]See IBM (2014): *Analytics Across the Enterprise: How IBM Realizes Business Value from Big Data and Analytics.*

- **Prescriptive analytics:** Refers to analytics that recommends actions.
- **Social media analytics:** Refers to doing a sentiment analysis to assess public opinion on topics or events. It also allows users to discover the behavior patterns and tastes of individuals, which can help industry target goods and services in a customized way.
- **Entity analytics:** This is a somewhat new area that groups data about entities of interest and learns more about them.
- **Cognitive computing:** Refers to an area of developing computing systems that will interact with people to give them better insight and advice.

In another book, Bill Franks of Teradata[4] voices a similar theme; he states that tapping big data for better analytics is essential for a competitive advantage in any industry today, and he shows how to develop a "big data advanced analytics ecosystem" in any organization to uncover new opportunities in business.

As we can see from all these industry-based publications by experts, big data is entering a new frontier in which big data will be harnessed to provide analytics-oriented applications that will lead to increased productivity, higher quality, and growth in all businesses. This chapter discusses the technology that has been created over the last decade to harness big data. We focus on those technologies that can be attributed to the MapReduce/Hadoop ecosystem, which covers most of the ground of open source projects for big data applications. We will not be able to get into the applications of the big data technology for analytics. That is a vast area by itself. Some of the basic data mining concepts are mentioned in Chapter 28; however, today's analytics offerings go way beyond the basic concepts we have outlined there.

In Section 25.1, we introduce the essential features of big data. In Section 25.2, we will give the historical background behind the MapReduce/Hadoop technology and comment on the various releases of Hadoop. Section 25.3 discusses the underlying file system called Hadoop Distributed File System for Hadoop. We discuss its architecture, the I/O operations it supports, and its scalability. Section 25.4 provides further details on MapReduce (MR), including its runtime environment and high-level interfaces called Pig and Hive. We also show the power of MapReduce in terms of the relational join implemented in various ways. Section 25.5 is devoted to the later development called Hadoop v2 or MRv2 or YARN, which separates resource management from job management. Its rationale is explained first, and then its architecture and other frameworks being developed on YARN are explained. In Section 25.6 we discuss some general issues related to the MapReduce/Hadoop technology. First we discuss this technology vis-à-vis the parallel DBMS technology. Then we discuss it in the context of cloud computing, and we mention the data locality issues for improving performance. YARN as a data service platform is discussed next, followed by the challenges for big data technology in general. We end this chapter in Section 25.7 by mentioning some ongoing projects and summarizing the chapter.

---

[4]See Franks (2013) : *Taming The Big Data Tidal Wave.*

25.1 What Is Big Data?

Big data is becoming a popular and even a fashionable term. People use this term whenever a large amount of data is involved with some analysis; they think that using this term will make the analysis look like an advanced application. However, the term *big data* legitimately refers to datasets whose size is beyond the ability of typical database software tools to capture, store, manage, and analyze. In today's environment, the size of datasets that may be considered as big data ranges from terabytes (10**12 bytes), or petabytes (10**15 bytes), to exabytes (10**18 bytes). The notion of what is Big data will depend on the industry, how data is used, how much historical data is involved and many other characteristics. The Gartner Group, a popular enterprise-level organization that industry looks up to for learning about trends, characterized big data in 2011 by the three V's: volume, velocity, and variety. Other characteristics, such as veracity and value, have been added to the definition by other researchers. Let us briefly see what these stand for.

Volume. The volume of data obviously refers to the size of data managed by the system. Data that is somewhat automatically generated tends to be voluminous. Examples include sensor data, such as the data in manufacturing or processing plants generated by sensors; data from scanning equipment, such as smart card and credit card readers; and data from measurement devices, such as smart meters or environmental recording devices.

The **industrial internet of things** (IIOT or IOT) is expected to bring about a revolution that will improve the operational efficiency of enterprises and open up new frontiers for harnessing intelligent technologies. The IOT will cause billions of devices to be connected to the Internet because these devices generate data continuously. For example, in gene sequencing, next generation sequencing (NGS) technology means that the volume of gene sequence data will be increased exponentially.

Many additional applications are being developed and are slowly becoming a reality. These applications include using remote sensing to detect underground sources of energy, environmental monitoring, traffic monitoring and regulation by automatic sensors mounted on vehicles and roads, remote monitoring of patients using special scanners and equipment, and tighter control and replenishment of inventories using radio-frequency identification (RFID) and other technologies. All these developments will have associated with them a large volume of data. Social networks such as Twitter and Facebook have hundreds of millions of subscribers worldwide who generate new data with every message they send or post they make. Twitter hit a half billion tweets daily in October 2012.[5] The amount of data required to store one second of high-definition video may equal 2,000 pages of text data. Thus, the multimedia data being uploaded on YouTube and similar video hosting platforms is significantly more voluminous than simple numeric or text data. In 2010, enterprises stored over 13 exabytes (10**18 bytes) of data, which amounts to over 50,000 times the amount of data stored by the Library of Congress.[6]

---

[5]See Terdiman (2012): http://www.cnet.com/news/report-twitter-hits-half-a-billion-tweets-a-day/

[6]From Jagadish et al. (2014).

Velocity. The definition of *big data* goes beyond the dimension of volume; it includes the types and frequency of data that are disruptive to traditional database management tools. The Mckinsey report on big data[7] described velocity as the speed at which data is created, accumulated, ingested, and processed. High velocity is attributed to data when we consider the typical speed of transactions on stock exchanges; this speed reaches billions of transactions per day on certain days. If we must process these transactions to detect potential fraud or we must process billions of call records on cell phones daily to detect malicious activity, we face the velocity dimension. Real-time data and streaming data are accumulated by the likes of Twitter and Facebook at a very high velocity. Velocity is helpful in detecting trends among people that are tweeting a million tweets every three minutes. Processing of streaming data for analysis also involves the velocity dimension.

Variety. Sources of data in traditional applications were mainly transactions involving financial, insurance, travel, healthcare, retail industries, and governmental and judicial processing. The types of sources have expanded dramatically and include Internet data (e.g., clickstream and social media), research data (e.g., surveys and industry reports), location data (e.g., mobile device data and geospatial data), images (e.g., surveillance, satellites and medical scanning), e-mails, supply chain data (e.g., EDI—electronic data interchange, vendor catalogs), signal data (e.g., sensors and RFID devices), and videos (YouTube enters hundreds of minutes of video every minute). Big data includes structured, semistructured, and unstructured data (see discussion in Chapter 26) in different proportions based on context.

Structured data feature a formally structured data model, such as the relational model, in which data are in the form of tables containing rows and columns, and a hierarchical database in IMS, which features record types as segments and fields within a record.

Unstructured data have no identifiable formal structure. We discussed systems like MongoDB (in Chapter 24), which stores unstructured document-oriented data, and Neo4j, which stores data in the form of a graph. Other forms of unstructured data include e-mails and blogs, PDF files, audio, video, images, clickstreams, and Web contents. The advent of the World Wide Web in 1993–1994 led to tremendous growth in unstructured data. Some forms of unstructured data may fit into a format that allows well-defined tags that separate semantic elements; this format may include the capability to enforce hierarchies within the data. XML is hierarchical in its descriptive mechanism, and various forms of XML have come about in many domains; for example, biology (bioML—biopolymer markup language), GIS (gML—geography markup language), and brewing (BeerXML—language for exchange of brewing data), to name a few. Unstructured data constitutes the major challenge in today's big data systems.

Veracity. The veracity dimension of big data is a more recent addition than the advent of the Internet. Veracity has two built-in features: the credibility of the source, and the suitability of data for its target audience. It is closely related to trust;

[7]See Mckinsey (2013).

listing veracity as one of the dimensions of big data amounts to saying that data coming into the so-called big data applications have a variety of trustworthiness, and therefore before we accept the data for analytical or other applications, it must go through some degree of quality testing and credibility analysis. Many sources of data generate data that is uncertain, incomplete, and inaccurate, therefore making its veracity questionable.

We now turn our attention to the technologies that are considered the pillars of big data technologies. It is anticipated that by 2016, more than half of the data in the world may be processed by Hadoop-related technologies. It is therefore important for us to trace the MapReduce/Hadoop revolution and understand how this technology is positioned today. The historical development starts with the programming paradigm called MapReduce programming.

25.2 Introduction to MapReduce and Hadoop

In this section, we will introduce the technology for big data analytics and data processing known as Hadoop, an open source implementation of the MapReduce programming model. The two core components of Hadoop are the MapReduce programming paradigm and HDFS, the Hadoop Distributed File System. We will briefly explain the background behind Hadoop and then MapReduce. Then we will make some brief remarks about the Hadoop ecosystem and the Hadoop releases.

25.2.1 Historical Background

Hadoop has originated from the quest for an open source search engine. The first attempt was made by the then Internet archive director Doug Cutting and University of Washington graduate student Mike Carafella. Cutting and Carafella developed a system called Nutch that could crawl and index hundreds of millions of Web pages. It is an open source Apache project.[8] After Google released the Google File System[9] paper in October 2003 and the MapReduce programming paradigm paper[10] in December 2004, Cutting and Carafella realized that a number of things they were doing could be improved based on the ideas in these two papers. They built an underlying file system and a processing framework that came to be known as Hadoop (which used Java as opposed to the C++ used in MapReduce) and ported Nutch on top of it. In 2006, Cutting joined Yahoo, where there was an effort under way to build open source technologies using ideas from the Google File System and the MapReduce programming paradigm. Yahoo wanted to enhance its search processing and build an open source infrastructure based on the Google File System and MapReduce. Yahoo spun off the storage engine and the processing parts of Nutch as **Hadoop** (named after the stuffed elephant toy of Cutting's son). The

[8]For documentation on Nutch, see http:nutch.apache.org

[9]Ghemawat, Gbioff, and Leung (2003).

[10]Dean and Ghemawat (2004).