Culminating Projects in Information Assurance

Department of Information Systems

12-2017

# Analyzing Big Data Using Hadoop

Sudip Pariyar

pasu1201@stcloudstate.edu

Follow this and additional works at: https://repository.stcloudstate.edu/msia_etds

**Analyzing Big Data Using Hadoop**


by


Sudip Pariyar



A Starred Paper

Submitted to the Graduate Faculty of

St. Cloud State University

in Partial Fulfillment of the Requirements

for the Degree

Master of Science

in Information Assurance



December, 2017



Starred Paper Committee
Changsoo Sohn, Chairperson
Mark B. Schmidt
Lynn A. Collen

**Abstract**

Due to growing development of advanced technology, data is produced in an increasing rate and dumped without analyzing it. Data sets are coming in large quantities through many mediums like, Networking sites, Stock exchanges, Airplane's black boxes etc. People who used to have 44 kb small floppy disk in the past are not happy with 1 TB external hard-drives nowadays. Big companies who are forced to add more servers in order to maintain the velocity of the incoming large data sets, are still looking for an easy way to control, handle big data. Traditional methods of handling big data are causing a variety of issues such as slow system performance, and lack of scalability. This research paper explores through the alternative method of handling big data which can address issues of the traditional methods. The goal of this research paper is to highlight an importance of a new method that can replace the traditional method of handling big data. This paper mainly consists of analyzed past work done by several fellow researchers. The outcomes of this paper will be useful for students and researchers alike who would like to work in the field of big data.

## Acknowledgements

I am very much honored to appreciate people who have helped me to complete this research work. I would like to thank Professor Changsoo Sohn, who is also Committee Chair, for his guidance. He was available anytime I needed to work on this research study. I have learned so many things from his class and academic meetings. I would also sincerely like to thank professor Mark B. Schmidt and professor Lynn A. Collen for their guidance to complete this research study. Their time and advise have helped me to accomplish the goal of this research study.

# Table of Contents

Chapter                                                                                                                              Page

**List of Tables**

**List of Figures**

## Chapter 1: Introduction

**Introduction**

Storing and analyzing large amounts of data has been very challenging these days since data is being generated at an exponential rate. Big data contains very important informational data, and when not securely disposed may lead to huge financial and reputational loss. So, the generated data should be either properly managed or properly destroyed. Traditional methods of handling data are becoming more and more challenging because of the size and variety of the data being generated. Hadoop is one of the software that easily address the issues of traditional method of handling big data. It is HDFS and MapReduce components that help find the solution to the big companies who were facing issues to store and process substantial amounts of data. The features of Hadoop software are already able to attract big companies like Yahoo and Amazon to implement the Hadoop application. The wordcount example from this research paper elaborates how easy it is to implement the new software application that can help understand the method of storing and processing of the data in the Hadoop environment. This paper analyzes how data is being treated in the past and highlights the advantages of using the Hadoop over the traditional methods of handling big data. The software needed to start a project with Hadoop software, which is Java and IntelliJ IDEA, is discussed to give an idea for starting a new project. In short, this paper discusses about big data and talks how Hadoop can be a good solution to address the current issues of handling big data using the traditional methods.

**Problem Statement**

Everyday data generated are increased at exponential rate. As of 2012, 90 % of the data generated in the world was generated only in the last two years (Friedman, 2012). This increasing rate of generated data is challenging the traditional method of handling big data in many ways. Storing and processing of large data is one of the biggest challenges of big data that uses the traditional method of handling data. The purpose of this study is to find the software platform as a solution that can address the challenges of big data.

**Nature and Significance of the Problem**

Big data has become the basis of marketing strategy for many small and big companies. They turn raw data to informational data using varieties of data analyzing tools. Companies using traditional methods of handling big data are facing many challenges regarding storing and handling big data. Those companies may lose many opportunities to improve the declined progress of their business performance. The data generated contains the valuable information. Without using the suitable techniques, the data generated are useless and valueless. It will just occupy a large space in the system and triggers the performance of the system. So, this study is very necessary to find a solution that can address the challenges of handling big data. Academically, this research study will be very useful for students who are pursuing master's degree in the computer related field specially Information Assurance degree. Working in the security field and analyzing data is one of the areas where Information Assurance students can utilize their degree. The literature review will help students and researchers to understand the concept of big data and how it is handled in the past. The required hardware and software details will allow them to install the application required for

the new technology, which stores and processes the big data. Also, the wordcount example will help students and researchers to start a project on Hadoop at their own. Overall, students and researcher can take advantages of this research study to build their knowledge on big data field and start a project on Hadoop. Companies also can take advantages of this research study. They can review this research study to find out the benefits of using the Hadoop software, so they can make a decision when to replace the traditional method of handling big data. Following are some contributions of this research study:

1) Academically student can learn the concept of opportunity and challenges of big data, Hadoop Components, and programming languages like Java, R, Python.

2) Commercially business can take advantages of the concepts of big data and Hadoop to improve business strategy and decision making.

3) Contributes to the society by creating opportunities to the people who aim to be good data analytics, and those business owners who want to expand their businesses.

4) Can be an academic reference for future researcher in big data field.

**Objective of the Study**

One of the key objectives of this research work is to define how HDFS and MapReduce help address the challenges of big data. The HDFS component will describe how big data are stored and MapReduce will define how big data are processed. With this main objective, the research study will also have other objectives which are as follows:

1) To address the challenges of big data using Hadoop.

2) Using Hadoop can easily store and process big data.

3) HDFS and MapReduce help reduce the challenges of storing and processing of big data.

**Study Proposition**

The outcome of the project will answer some statements that are listed below:

**Statement 1**: Big data can be easily stored and retrieved using Hadoop components HDFS and MapReduce.

**Statement 2**: Hadoop helps resolving the challenges of big data.

**Statement 3**: Analyzing big data using Hadoop will help companies improve operational performance.

**Limitations of the Study**

Working on this research project supports the basis of discussion for how effective the new software is over the traditional software. Finding real-time large volume of data for learning purpose is the biggest challenges during this research study. If we are able to get large volume of data, then adding more spaces in the local machine is very difficult. For adding more space in the local machine, it may cost 3,000 to 5,000 thousand dollars. This research paper is planned to be conducted at a free of cost or minimal cost. At this moment, only one server is used for conducting this research study. So, the scalability characteristics of big data using Hadoop is limited to the single node Cluster for this research study.

**Definition of Terms**

There are some technical terms that are used in this research study. They are as follows:

**Data**: In this research paper data are referred to the collection of raw unit of information generated by the users.

**Big data**: Big data in this research paper is referring as the large volume, variety, and velocity of data generated.

**HDFS**:  Also known as Hadoop Distributed File System, is a Hadoop component which is used for storing large amount of data and contains the lists of files stored in the Hadoop database.

**Cluster:** It is referring to one of the components of HDFS which consist of Master node and Slave Node.

**MapReduce:** It is referring to one of the Hadoop modules which handles the processing of big data.

**Byte**: It is term used for measuring the unit of data volume. The smallest unit of measuring data is bit and 8 bits is equal to 1 byte.

**Megabyte**: A unit of measuring data equal to 1024 bytes.

**Gigabytes**: A unit of measuring data equal to 1024 megabytes

**Terabytes**: A unit of measuring data equal to 1024 gigabytes.

**RDBMS**: Is also known as relational database management system. RDBMS is traditional methods of handling data based on the relational model.

**Summary**

This Chapter covers the basic idea about how Hadoop is important software to replace the old method of storing and processing the big data. Also, the objectives of this research paper to analyze the advantages of using Hadoop software that can easily store and process

the big data. As mentioned above, this section directs the reader to the specific purpose of the research project which is to analyze the big data using Hadoop. This research study will also address the proposition mentioned in the Study proposition. In the next chapter, the basic concept of big data will be discussed using the academic article from the past researchers.

**Chapter 2: Background and Review of Literature**

**Introduction**

The main objective of this chapter is to discuss the past academic work related with the big data and Hadoop. Various articles that discusses about the term big data, its characteristics, and how they are being stored and processed will be analyzed to draw some challenges of traditional methods that handles the big data. Also, this chapter will highlight why the traditional method is not sufficient to handle the big data. The analysis included in this chapter does not modify any past work in the related field but tries to consolidate same idea taken from different perspectives about the big data and its challenges.

**Background Related to the Problem**

Data has become very essential source of improvement for many companies in this competitive market era. Almost every company gathers information from their customers and use them to learn the shopping habits of customers. Data is not only used for improving marketing strategy, but it is also used by different sectors such as airlines, hospitals for better quality service. The data generated is increasing every day and becoming more and more challenges to the traditional methods of handling data. Without proper data handling techniques, only 5% of the data generated are analyzed and used. Users who used traditional methods of handling data are facing many challenges such as storing and processing big data, privacy, timeliness, etc. In one hand there are many opportunities of big data but also have challenges as well. So, for storing and processing of big data without any challenges, there is a need of new software that can easily addresses the issues of big data.

**Literature Related to the Problem**

Academic articles are the reliable source to learn about the big data and how they are challenges to the traditional methods of handling data. For this research study, the St. Cloud State University Library database and IEEE articles are used as the main sources of references to the problem. Past researchers' academic articles are used and analyze to learn the basic concepts of big data and its challenges. The analysis process of the research paper from the past will be focused on the problem and solution of the problem. As a focus on problem, literature review related to the big data and its challenges will be analyzed using appropriate academic articles from the past researchers. As a part of problem focus, other aspects of big data will also be discussed. As a solution to the problem, articles related to Hadoop software will be analyzed and examine whether the software is suitable solution to the problem. Some of the discoveries from the literature review of the problem will be described in next chapters.

**What is big data?** Traditionally, big data is defined as a massive volume of the data. But nowadays it is very hard to define a term big data. There are no such measurements that makes the data as a big data, so defining a big in a single term is very hard and does not have any particular definition (Kale & Jones, 2016). Big data is referred to as an enormous amount of structured and unstructured data collected from various means of sources. Data are being generated, stored and transferred from one medium to another. Every year the amount of data generated is bigger than the data generated in the previous years. So, with the time passing, data generated are becoming bigger and bigger. Some of the sources of big data are Web logs, GPS tracking systems, social networks, Internet-based text documents, detail call records, astronomy, atmospheric science, biology, genomics, nuclear physics, biochemical experiments, medical records, scientific research, military surveillance, multimedia archives, airlines, meteorology, etc. mobile device has become one of the useful sources of

data for analyzing people behavior. Some of the areas are shown in the figure 1. As seen in the figure 1, the location recorded in the mobile phone can be used for tracking people: where they are and where they traveled. Also, people uses social media sites on their mobile device. Data generated from people using social media can be easily used for various purposes such as monitoring them and finding peoples' social ties. Some of the other data generated from the mobile sources and the fields that can use mobile phone data are shown in the figure below:
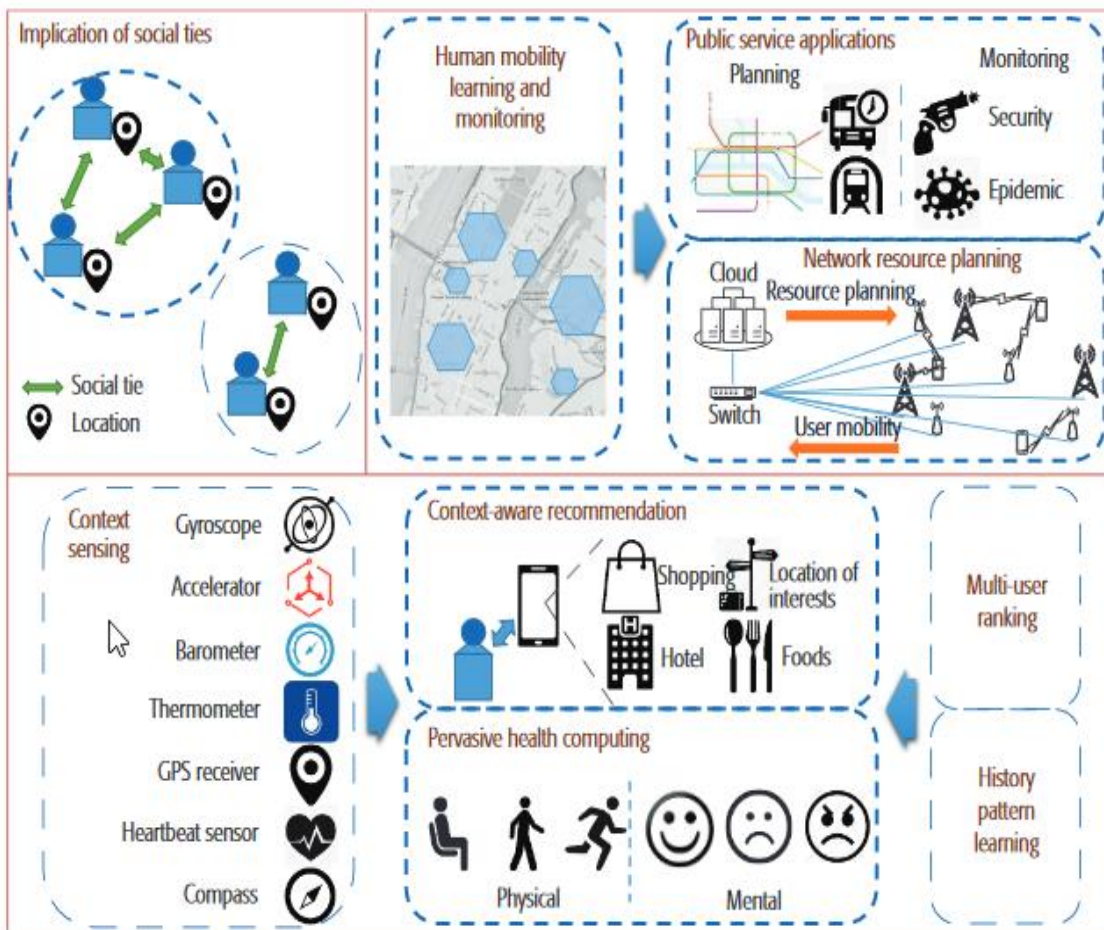


*Figure 1*. Data generated from mobile device (Cheng, Fang, Yang, & Hong, 2017).

According to the article, "Cisco Visual Networking Index," "the annual global IP traffic will surpass the zettabyte threshold in 2016 and will reach 2.3 ZB by 2020." And, the individual will have to spend 5,000,000 years to watch the complete video generated in each month of year 2020 (Cisco VNI Forecast and Methodology, 2016). Further it can be defined with V's: Volume, Volatility, Velocity, Variety, Variability, Veracity, and Value which are defined in more details in the characteristics of big data section (Heller et al., 2016).

**History of big data**. With the invention of computer, the data generated from the computers has increased at a rapid rate (Yaqoob et al., 2016). Initially data were generated by the people who work in big companies. Their daily tasks such as entering client's data and generating office reports were the initial source of data. When the social media like Facebook, Yahoo, and Hotmail were introduced, then users were able to create their own data. When users were able to create their own data by producing videos and uploading photos, the data begin to generate in increasing way. Currently, machines are generating enormous amount of data as more and more businesses start conducting online business and move away from the typical brick and mortar business model. Activities such as monitoring the world in every second due to security reasons or for the purpose of using the data for better marketing strategies, generating weather predictions, taking pictures and videos surveillance etc. are creating more volume of data.

According to the article written by Uri Friedman (2012), 90% of the total data generated in the world are generated only in past 2 years. The history of data generated is dated back in 1887-1890 when Herman Hollerith invented an electric machine that can read holes punched into paper cards to tabulate 1890 census data. This modern device of that time

reduced the amount of time taken to complete census from 8 years to 1 year. During 1935-1937, President Franklin D. Roosevelt's launched Social Security Act project that helped to track records of 26 million working Americans and 3 million employers. In 1943, a device named "Colossus" was developed in British facility for breaking the Nazi codes during World War II. This device is considered as the first programmable electronic computer. The time taken to decode the message were reduced from weeks to hours through this machine.  In 1961, more than 12,000 cryptologists confronted that the information was over flooded during cold war. Seventeen thousand reels of tape were generated just in the month of July 1961. Due to public criticism in 1965-1966, the U.S. Government failed to transfer all its documents which includes 742 tax returns and 175 million sets of fingerprints to magnetic computer tape. This incident later brought out the 1974 Privacy Act limiting federal agencies from sharing public records. The concept of sharing file in the internet was introduced by U.S. Government in 1960s. Using those concepts from the past used by U.S. Government, the British computer scientist Tim Berners-Lee proposed internet system to transfer files throughout the globe using hypertext system called World Wide Web. In 1996, President Bill Clinton announced developing a supercomputer that can-do calculation in a second that might take human to calculate in 30,000 years. In 1997, for the first time the term Big Data was introduced or used by the NASA researchers Michael Cox and David Ellsworth. According to them, "[D]ata sets are generally quite large, taxing the capacities of main memory, local disk, and even remote disk," "We call this the problem of big data" (Friedman, 2012). After 9/11 attack, the U.S. government started data mining techniques from large volume of data to fight against the terrorism. The data collected from various sources were fused in a grand database to find the

suspicious individuals. By 2010 the NSA's 30,000 employees have already stored 1.7 billion emails, and phone calls. In 2007-2008 with in the popularity of social media, the massive amounts of data and applied mathematics started replacing every device that were used in the past. In January 2009, the Indian government generated the largest biometric database by taking 1.2 billion peoples' finger print, photographs, and assigning every individual with 12-digit identification numbers. In May 2009 President Barack Obama introduced data.gov to track down everything from flight to product recalls. In August 2010, Google CEO Eric Schmidt mentioned that 5 exabytes of data were generated between human civilizations to 2003. And currently the same amount of data is generated in just 2 days. In February 2011, IBM's Watson Computer system scanned 200 million pages of information which is equal to 4 terabytes in seconds. With the growing technology used, every year data generated is increased as shown in the below figure. The trend of data production is growing every year by hundreds of thousands of exabytes. In figure 2, it is clearly seen that in year 2009 few hundreds exabytes of data is produced and by the end of the year 2020 the data produce will exceed to 40,000 exabytes (Kale & Jones, 2016). The history of big data has shown whenever the data generated are grown, the new technique has been improved to handle it. When we look back into the punch card era, it was the new technique that reduced the work load by revolutionizing the time taken for analyzing the data. So, the rate of data generated is increasing as well as the new techniques for handling the enormous amount of data are being improved.
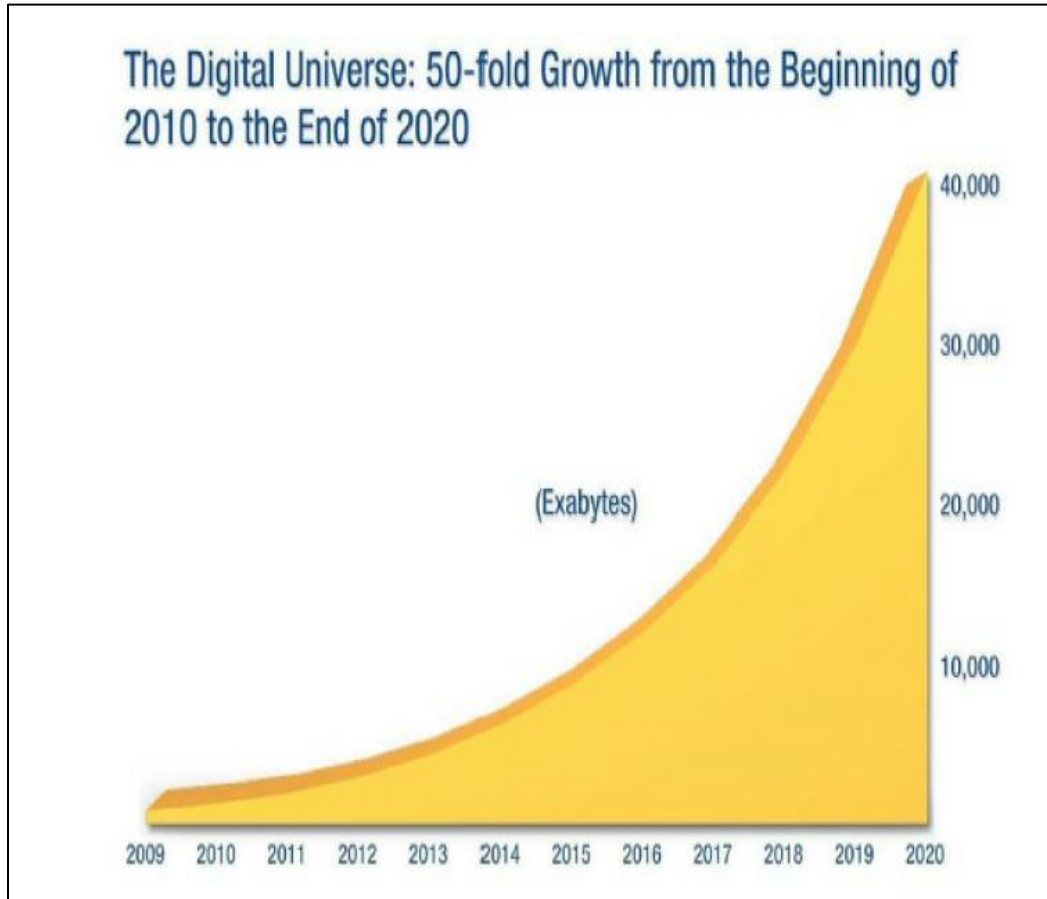
*Figure 2.* Yearly data generated (Kale & Jones, 2016).

In the article by Han Hu and team, the history of big data is defined into four major

milestones. During those four major milestones, the increased data generated has been

recorded and explained how big data were handled at those milestones. They are as follows

(Hu, Wen, Chua, & Li, 2014):

A) ***Megabyte to Gigabyte***: This milestone of big data was achieved during 1970s and

1980s. The data were increased from megabyte to gigabyte mostly generated from

businesses. With the increase data size, businesses were concerned about system

performance and focused on hardware improvement. Also improved the database system to cope with increased volume of data.

B) *Gigabyte to Terabyte*: According to paper by Hu and team, in late 1980s the data generated were increased to Terabyte with the popularization of digital technology. This increase volume of data brought challenges to the single operated large computer. Then the distributed system of file was proposed to get more storage and better performance. As a result, parallel database such as shared memory database, shared disk- database were built to solve issues of big data.

C) *Terabyte to Petabyte*: By 1990s the parallel database was widely used and as a result, the data generated grow from terabyte to petabyte. The development of World Wide Web brought internet era. The internet was widely used in 1990s which generated the structured and unstructured data increasing data from terabyte to petabyte. The petabyte of data brought storing and processing challenges to the traditional methods of handling data. In 2003, to address big data challenges, Google proposed the concept of Google File System and MapReduce. During mid 2000s data mining techniques and user generated content brought the massive volume of data.

D) *Petabyte to Exabyte*: Previous few year, the data generated has shifted from Petabyte to Exabyte. Big companies like Google processed 25 petabytes of data each day (Khan et al., 2014). The data generated is increasing at an exponential rate and soon the data will grow more than Exabytes.

**Characteristics of big data**. Defining big data in a single term is very difficult. Data have certain characteristics that makes it easier to define the term big data. Following are some characteristic that makes data a big data:

A) *Volume***:** Volume is one of the features of Big data that defines big data in terms of size they are generated. Depending upon the work load of the company, data are generated low or high. In terms of handling volume of big data, it can be simply handled by adding more database center. One airplane generates 10 TB bytes of data in every 30 mins. There are more than 25,000 of airlines flights that increase this number daily to Petabytes (Dijcks, 2013). The data generated are growing at an exponential rate. The increased volume of data offers a lot of opportunities to the companies as it contains valuable information. In terms of volume data are measured in bytes (Yaqoob et al., 2016) as shown in table 1. The size of the data starts from 1 bit and goes higher as shown in the table 1.

B) *Variety***:** Another feature of big data is variety which is referred to the various forms and sources of the data. Traditionally data were generated from business sources. But the advancement in technology has enabled users to create their own data which might varies in the form. Users' day-to-day activities, social media, medical industry, geo location device are some other sources that generate the various forms of data. These data generated from various sources are classified as structured, unstructured, and semi-structured data. Structured data are those data that are stored in the databases and can be easily retrieved. Numbers and alphabets are example of structured data. Unstructured data are those data which are stored in the forms of files which can be further structured in order use or analyze it. Videos, photos, text, emails are some example of unstructured data.

Table 1

*Data in Terms of Byte* (Hu, Wen, Chua, & Li, 2014)

| Name | Equal to | Size in Byte |
|---|---|---|
| Bit | 1 bit | 1/8 |
| Nibble | 4 bits | 1/2 (rare) |
| Byte | 8 bits | 1 |
| Kilobyte | 1,024 bytes | 1,024 |
| Megabyte | 1,024 kilobytes | 1,048,576 |
| Gigabyte | 1,024 megabytes | 1,073,741,824 |
| Terabyte | 1,024 gigabytes | 1,099,511,627,776 |
| Petabyte | 1,024 terabytes | 1,125,899,906,842,624 |
| Exabyte | 1,024 petabytes | 1,152,921,504,606,846,976 |
| Zettabyte | 1,024 Exabyte | 1,180,591,620,717,411,303,424 |
| Yottabyte | 1,024 zettabytes | 1,208,925,819,614,629,174,706,176 |

C) *Velocity***:** Another feature of big data is velocity at which data are generated. The velocity of data refers to the speed of data which are generated as well as the speed of data taken to process it. Data are generated from the sources and stored in the database at very fast speed. But it is very challenging to analyze which data will be important and which one does not (Dijcks, 2013). Almost every person has some kind of electronic devices that generates the data. Most of the people are connected to each other through socials media like Facebook,

twitter, yahoo etc. They like each other comments, send message, and uploads video and photos which generates the enormous amount of data. Social media is one of the sources that are generating data in an exponential rate.

D) *Value***:** Every information has certain value on it unless it is not properly extracted. Depending upon the nature of the data generated, it can be very useful in the future. These days many big companies are utilizing scratch data into meaningful data by analyzing it. Companies can use data to restructure the marketing strategy. Human behavior itself is a challenge for big data.  In order to find out the value of big data, it is important to know how to utilize this information or what is the point of collecting the data, and plan for using it in the future. It is important to know the pattern of data to make some useful assumptions about it (Heller et al., 2016).

It was estimated that in 2010, the big data was valued at $3.2 billion and will be increased to $16.9 billion in near future (Khan et al., 2014).

**Classification of big data.** In the past, the sources of data used to be very limited and the data generated used to be single data type. With the improvement of technologies, types of data generated are also getting more and more complex. Sometimes, the data generated by the single source has different data types. For example, mobile phone is very popular among people and yearly the people friendly app and feature are making it more popular for using it. People take and upload videos on social media, take photos, and write messages generating their own data. One mobile phone alone generates lots of data type. So, based on the complexity of data types, big data are classified into three groups. The classification of big data are as follows:

A) *Structured:* Structured data are those kinds of data which are easily formatted so that computer can easily read and use the data. In this kind of data, the data are organized in the form of number, string and date format within the column and rows so that computer can easily identify the data. The example of a structured data can be a Student's data found in the university database. The student data may contain valuable information such as name, age, sex, contact address, year enrolled etc. in the form of table's column and row. For decades, Structured Query language has been used for maintaining the structured data (Asthana & Chari, 2015).

B) *Unstructured***:** An opposed to structured data, unstructured data are those data which are not in this general format and can be of various length and format. This kind of data is hard to read and understand by computers. Example of unstructured data are pictures, audio recordings and video recordings, etc. Traditional databases can only process the single format data. So, to process the unstructured data, the data should be first converted to the structured data and only they can be further processed to analyze. But software like Hadoop can easily process and analyze both structured and unstructured data.

C) *Semi structured***:** Semi structured data are those kinds of data which may consists the structure and unstructured data. For example, the data generated from the history of people visiting the website is an example of semi-structured data.

**Big data vs traditional data.** If we look back to the history about data, we can see how the data are generated in increased rate, and how new techniques are developed to handle it. Data in the past used to be small and can be store and process easily. But nowadays data generated are growing in an exponential rate making those traditional data to be called as big

data. Here the traditional data are referred to those data which were generated before the use

parallel computing. Following are some differences that distinguish the traditional data from

big data (Hu, Wen, Chua, & Li, 2014):

A) *Data size*: The data size of the traditional data was low in comparison to the size of

big data. Traditionally, the size of the data varies from 1 bit to Gigabytes. But big data are

measured in petabytes or exabytes and the size is increasing every year.

B) *Data generated rate*: Traditional data were generated at a very slow rate. They

usually were generated hourly or daily and measure up to gigabytes. But big data are growing

at a very fast rate.

C) *Data structure*: The data in the past were generated in structured format and were

easy to store and process. Currently, big data are generated in various forms such as semi

structured and even un-structured. Hadoop is a software that can store and process data in any

form.

D) *Data source*: Traditional data were centralized and must process through single

server. But big data or modern data are processed in distributed servers computing parallelly.

E) *Data storage*: Traditional data are stored in the RDBMS format where as big data

are stored in HDFS or NoSQl. For storing and processing traditional data, SQL programming

language is used. If the data are unstructured or semi-structure, the format need to be changed

in single format and store. For storing and processing big data, various programming language

are used such as Java, R, Python, Hive Query language, NoSQL, etc.

**Sources of big data**. There are various sources that are generating the Big data. Some

of the sources of Big data are Internet of Things (IoT), self-quantified, multi-media, and social

media (Hasem et al., 2016). IoT data are generated by GPS devices, mobile, Personal Device

Assistants, alarms, window sensor, refrigerators etc. At this point, IoT data is not considered

as big data but by the time 2020, this kind of data will be a big part of big data. Self-

quantified data are those kinds of data which are generated by individual to track their

personal behavior like, monitoring calories burned, steps walked, exercising habits, etc.

Internet is one of the prime sources of big data. according to the article by Khan and team

(2014), 2 billion of people are connected to the internet. And 5 billion people are using some

kind of electronic device which generate large amount of data. Multi-media data are the data

generated by the individuals who are connected to the internet. Text, images, audio send over

the internet contribute this kind of data.  Social media data are generated from various social

media, like google, twitter, Facebook, YouTube, Instagram, etc. Some of lists of social media

that generate large amount of data are shown in the figure 3. As seen in the figure 3, as of

2014, users upload 100 hours of videos every minute in YouTube. YouTube is opened by 1

billion new users. About 6 billion hours are spent watching YouTube every month which is

equivalent to watching an hour by every user (Khan et al., 2014). Figure 3 shows other top

social media that generates the large amount of data such as twitter, Facebook, Google,

LinkedIn etc. Traditional enterprise data are another source of big data (Dijcks, 2013). This

enterprise data includes customers information from CRM system, transactional data etc.

Nowadays, companies generate and store customers data and analyze them to improve the

marketing strategy. Machine generate data are another source of big data. According to the

article by Dijcks (2013), airplanes generate 10 terabytes of data in every 30 minutes. More

than 25,000 commercial flights are operated daily, and these flights generate 1 petabyte of

data every day. So, big data are generated from various sources which may include social

media, enterprise, and automated machines.

| Social media data | Data production scenario |
| --- | --- |
| YouTube (Youtube, 2014) | • Users upload 100 h of new videos every minute<br>• More than 1 billion unique users open YouTube each month<br>• Over 6 billion hours are spent watching videos each month; that is, almost an hour for every person on Earth and 50% more than last year |
| Facebook (Facebook, 2014) | • Receives 34,722 "likes" every minute<br>• 100 terabytes of data uploaded daily<br>• Currently 1.4 billion users<br>• Employs 70 languages |
| Twitter (Twitter, 2014) | • Over 645 million users<br>• 175 million tweets per day |
| Google+ (plus, 2014) | • 1 billion accounts |
| Google (Google, 2014a) | • Receives over 2 million search queries per minute<br>• Processes 25 petabytes of data each day |
| Pinterest (Pinterest, 2014) | • 70 million users by October 2013 |
| Apple (Apple, 2014) | • Receives around 47,000 application downloads per minute |
| Tumblr (Tumblr, 2014) | • Blog owners publish 27,000 new posts per minute |
| Instagram (Instagram, 2014) | • Users share 40 million photos daily |
| Flickr (Flickr, 2014) | • Snappers upload 3125 new photos per minute |
| LinkedIn (LinkedIn, 2014) | • 2.1 million groups |
| Foursquare (Foursquare, 2014) | • Over 2000 check-ins<br>• 571 new websites launched per minute |
| WordPress (Wordpress, 2014) | • Bloggers publish nearly 350 new blogs per minute |

*Figure 3*. Data generated by social media (Khan et al., 2014)

**Importance of big data**. Proper processing and data analytics can assist businesses in

making smart decisions. It has helped in determining timely alerts for defects, generating sale

coupons based on client's buying habits, detecting fraudulent activities or recalculating risks

in minutes, etc. Here are some advantages that various fields can have with big data:

Big data consists of valuable information that can be very useful for big companies.

Upon analyzing raw big data into meaningful data, business can learn insightful

understanding of their business which help them to improve productivity, a stronger

competition position, and greater innovation. For example, Wal-Mart can analyze their daily,

monthly, or yearly sales transaction and find out the combination of products that goes

together. This can be obtained by finding the total number of customers who buy two or three

combinations of product together. Suppose, 100 customers bought milk and bread together in

same transaction, then the company can put milk and bread in same cabinet in order to boost

sales of Milk or Bread. Further, businesses can find out which product is selling better with

big profit and which product is not doing well. This way business would be able to plan future

marketing strategy using the results obtained from analyzing the big data.

Manufacturing companies are using devices to track the success and failure of its

product by installing monitoring device in the product. Most motor vehicle manufacturing

companies gather big data. These kinds of device usually capture customer usage pattern,

failure and success of product which can help manufacturing company to reduce the cost of

development and assemble of the product.

Many businesses using big data to target new customers. Many tracking devices such

as mobile apps, GPS, internet web browsing help businesses to find the buying pattern of

customers. And customers are automatically suggested new product or their favorite products

through their tracking devices. So, big data can be very helpful to find out the potential

customer.

Big data has been used for investigating the criminal activities and suspicious activity

by government and other institutions. After 9/11 attack, US government has used big data to

fight against the terrorism (Friedman, 2012). By 2010, National Security Agency has already

stored 1.7 billion emails and phone calls to monitor the suspicious activity. Cell phone are

becoming the reliable source for monitoring the people activity (Cheng, Fang, Yang, & Hong,

2017). Many banks are using customer shopping habits to find the fraudulent transactions.

Big data has also been used by politicians for their election campaign. According to

the textbook by Kerzner & Maniyam (2013), big data analysis helped Barack Obama to win

the presidential election. The article states that the characteristics of voters were analyzed, and

the election campaign directly addresses the voters. So, analyzing big data is becoming more

and more popular method among election candidates to interact with voters.

Big data has been very useful to the health sector. Health services are improved by

analyzing large patterns of health data. Even the large health care data has allowed

pharmaceutical companies to improve personalized medicine ensuring the better and fast

recovery (Yaqoob et al., 2016).

So, big data has become a very important source of improvement for various sectors.

Big data has been used in many ways and the results are always positive. Commercial

businesses are using big data for marketing purpose, health sectors are using big data for

improving health care services and improving medicine, governments are using big data for

data mining to monitor criminal activities. So, proper analysis of big data may be beneficial to

many sectors.

**Challenges of big data**. Data has various characteristics that makes it big data. Due to its

various characteristics, big data are becoming more and more challenging to the traditional methods of

handling data. Some the challenges of handling big data are explained below:

A) ***Heterogeneity and incompleteness***: Usually machine analysis algorithm aspects homogenous data. Various formats data are generated and should be changed to single format in order to get the accurate results. Suppose, the health records of the patient are generated from a hospital and some of the attributes like employment status, educational back ground are missing then the analysis would not completely satisfactory. The machine analysis system will consider the empty attribute as null value in the job section, and the result may be affected as people hide their employment status who actually have jobs. When the attributes are null, the analysis process will end as incomplete and error is likely to remain in the data. Such incompleteness and error should be addressed during data analysis process which is very challenging.

B) ***Scalability***: One of biggest challenges while handling big data is its size. The uncontrolled amount of data generated every day is challenging the companies to store and process it. In the past, companies used to get rid of this kind of issue by making the processor faster following Moore's law (Agrawal et al., 2012). But these days, the volume of data are increased in a speed more than the computing resources. As the volume of data increases, the capacity of servers needs to be increased. Only increasing servers to address the storage deficiency is not a solution. The added server should be able to perform the job as faster as it was before adding the server. Modern technology allows adding more servers without any issues but when the data volume gets very large then the companies may find scalability as a biggest challenge of the big data.

C) ***Timeliness***: Timeliness is another one of the challenges of big data. As volume of the data increases, the time taken to analyze the data will also increases. Some cases like

analyzing fraudulent activity needs very fast processing. But it is not possible to get a full analysis of the fraudulent activity before the fraudulent transactions can occur. So, more the volume and variety of data is, the more time is needed for a complete analysis of big data. So, companies like bank and security companies faces timeliness challenges of analyzing the big data.

D) *Privacy***:** Another challenge of big data is maintaining the privacy of the data generated. These days people log into various social media. Every user is asked to give their information accurately and have to agree terms and conditions of the applications. Many social networking sites, retails store, and health sector transfer people's data electronically for internal or external purpose. So, there is always a fear of misuse of personal data as the data generated are increasing exponentially.

E) *Energy management***:** Managing the energy required to operate large computing system is another challenge of big data (Hu, Wen, Chua, & Li, 2014). Maintaining big data using traditional method is not economic and environmental feasible. The energy consumption by large machines is increasing with the increasing demand of analyzing the big data. The more analysis of data means the more amount of energy needed. So, maintaining energy is another concern when handling big data with traditional methods of handling big data.

**Security challenges of big data**. The exponential growth of the data has opened many opportunities for researchers, students, and industries as well as cyber criminals (Schwieger & Ladwig, 2016). Cyber criminals can destroy both the industry and its customers with a single opportunity. The effect of weak securities can lead to the destruction of industry's reputation

and may be subject to millions of dollars loss as a data breach settlement. So, the industries collecting large volume of data should be aware of security challenges while storing and processing the big data. Some of the security challenges are explained below:

**A)** *Privacy*: Privacy is considered as one of the primary security challenges of big data. These days lots of companies trades customer sensitive information based on the user's location and preferences (Agrawal et al., 2012). Hackers can easily track the identity of the users by analyzing the location and pattern of user's activity. Once they are able to track the personal information, they can use that information to create duplicate debit and credit cards to be sold online.  When the data are transferred from one company to another, there should be some guarantee that the company that receives the data will fairly use the data. Sometimes the fair use of data results personal harm too.

One of the retailers was tracking the shopping habits of customers and concluded the teenage customer was pregnant. The retailer started sending deals and coupons related with the pregnancy products in her mailing address thinking it might be useful to that customer. This deals and coupons unintentionally disclosed her father about the pregnancy. Even though the retailer used an accurate result from the analysis, it violated the privacy of the teenage customer. Future data scientists should be aware of consequences of using information from the analysis of big data.

**B)** *Quantity of loss affected from the security breaches*: Other potential security challenges of big data are the amount of loss affected from the security breach. The more data is generated, the more devastating consequences will result from the data compromised than that we have from the normal data (Lafuente, 2015).

Another security challenge of big data is maintaining the granular access of the big data. There will be lots of users trying to access to the big data. Being large in size, big data needs more work to classify the importance of the data and decide whom to give access of it.

Most companies collect data from various sources and store at the central storing site. Data contains very sensitive information since it may contain the information of people, employee, financial information, and trade secret. This sensitive information might be potentially vulnerable from attacker since they can easily attack to the central database.

Additional security challenges of big data are maintaining the compliance with the law. Some law restricts where the data should be stored and processed. Since company has to store and transfer data over the internet, it is potentially vulnerable for companies to maintain the law. They might face security issues, or they might be violating the privacy of individuals without noticing it.

**Securing big data**. Securing big data is crucial and need extra efforts. Since big data involves large sets of data, the effect is also in large scale when it is compromised. Recently half of Americans were affected by the Equifax data breach. According to the article by Bremmer (2012), the numbers of user hacked in LinkedIn, Heartland, and Dropbox were 167, 130, and 68 million respectively. Companies also suffered huge financial loss as a settlement of this data breach. In the article by Schwieger and Ladwig (2016), it states Robert Jandoli who is also Global Chief Information Security Officer suggest some consideration that can be made while handling the big data to protect privacy:

A) *Collection*: While gathering the data, the people involve in gathering the information should have knowledge what they have collected. They should able to answer

whether the data collected is reliable, and also whether data collected is secure. Sensitive

information may require extra effort to handle such as social security number, date of birth,

family health record, etc.

B) *Storage*: Storing a data plays a crucial role on the security of the data. One mistake

can open a door for hacker. The user should be aware of vulnerability and should make sure

that security precaution has been applied on the data.

C) *Uses/User*: When the data are used, it is necessary to make sure that the users are

authorized to access the data information and should be aware how the data are being used.

D) *Transfer*: Lots of the companies these days exchange data from one place to

another. The privacy can be maintained when the purpose of sending the data is clear. Also,

the medium of transfer should be carefully monitored while transferring the data.

E) *Destruction*: Destroying of the data is very important aspect of the security

challenges of big data. When data occupies certain spaces, it can be possible that the data can

be retrieved and misused by third party. So, while destroying the data it is very important that

the data is properly destroyed (Schwieger & Ladwig, 2016).

Big data can be secured by using other techniques. Some of the major techniques

while securing the big data are follows (Lafuente, 2015):

A) *Data anonymization and generalization*: Data anonymization is the process of

sanitizing the data for the purpose of hiding the sensitive information from the data. In the

data anonymization process, some the information which can lead to the disclosure of the

identity of the data holder are replaced or deleted from the original data. For example, in

medical records data, the name of the patients, their age, and address are removed so that the

patient's privacy is maintained. In this case, the records may only have few general records

which protects the patient's details. Sometimes Generalization technique is used along with

the data anonymization which helps to maintain the privacy. In generalization of data, the

patients with age 25, is generalized as age 20-30. So, the privacy of big data can be protected

by simply applying the technique of data anonymization and generalization.

B) *Data encryption*: Another way to secure big data is by applying encryption

technique. Encryption can be applied at the source or at the big data platform (Tankard,

2017). In encryption techniques, the data information is changed from plain text to cipher text.

Cipher text are symbolic format of the original text, which needs special key to read the text.

The key can be secured more by applying salted password. Salted password makes hacker

hard to guess the key required to read the information. There are some pre-computed

passwords that makes hackers easy to hack passwords that secure the big data. Rainbow tables

are the precomputed list of possible hash values used for reversing cryptographic hash

function. The simply generated hashed passwords are easy to crack since the possible output

is already computed in a table. When a salt value is attached along with the password, the new

password's length of the password is increased and also gives a unique value for each

individual user. In order to attack the salted password, hacker needs large amount of storage

which might not be feasible for their computer systems (Vishwakarma & Madhavan, 2014).

Since the salted password are randomly selected values, hackers need to crack through every

individual password. Suppose an organization system is hacked which contains the data

information of the user. Hacker may be successful to crack user's username and password one

by one but take long enough time to do so. Which will increase the time and storage to crack

the password. So, salted password system helps prevent attacks from precomputed tables like Rainbow tables which makes big data more secure from hackers.

C) ***Access control***: Applying access control on big data is another way we can secure the big data from unauthorized users.  In case of big data, only basic forms of access control can be applied (Julio, Manuel, & Eduardo, 2016). According to the article written by Julio, some of researcher focus access control in the MapReduce and suggests a framework to enforce access control at the key value level (Ulusoy et al., 2015).

D) ***User Authentication***: User Authentication is another technique that can be used for securing the big data. Applying mechanism for asking user credential to access the big data can be applied to prevent the unauthorized user from accessing the sensitive data. Username and password can be applied as a credential. The password can be protected applying hash algorithm or applying salted password to prevent the password attacks such as rainbow table. In case of mobile device, the sensitive data can be protected using biometric credential or facial detection techniques as a user authentication (Cheng, Fang, Yang, & Hong, 2017).

So, there are several ways that can be used for securing the big data. Big data contains valuable information that needs extra effort to secure from unauthorized user. Using one of the techniques discuss above may help companies fight against the data breach, which may save them a lot of money and consequences from the data breach. We cannot stop hackers from attacking the database but learning the potential threats on time may keep hackers away from the secure database.

**How data has been handled in the past.** According to the article by Friedman (2012), electric machine was invented in 1890 by Herman Hollenrith to read the holes punched into the paper cards to calculate the 1890 census data. This machine enables to count the population in one year that was instead taken eight years in the past.

Then in 1943, device named "Colossus" was invented by the British official to break the Nazi codes. This machine was able to read 5,000 characters per second reducing the time that might take up to weeks.

In 1961, U.S National Security Agency (NSA) collected and processed signals automatically with computers storing in magnetic tape (17, 000 reels of tape in 1961). And the internet era was proposed in 1989 by British computer scientist Tim Berners-Lee to share information globally through the system called World Wide Web.

In August 1996, U.S. President Bill Clinton announced developing a supercomputer that will compute a math in a second which might take 30,000 years to a human with a hand-held calculator. In 2002, U.S. Government started data mining techniques to start fighting against the terrorism. This was leading towards the era of setting a huge database. People can store large amounts of data in database using various applications. In 2004, sharing of large amount of data started where retail stores like Walmart started recording the shopping behaviors of the customers.

In February 2011, IBM's Watson computer system scanned the 4 terabytes of data in a matter of seconds (Friedman, 2012).

One of the popular techniques used in the past is database technology. Database technology has been popular for decades for handling the data. Among various users such as

researchers and entrepreneur of 1970s and 1980s, database technology was very popular (Kevin, 2012). In an academic paper the term "database" was mentioned in 1960s. In 1970, IBM's San Jose research team worked on developing the collection of database technologies dubbed system R based on Edgar F. Codd's relational model paper.

Looking at the history of how data are handled in the past, computer scientists are always looking for new instruments to address the volume of data being generated. From punched hole era to this big database era, there has been a discovery of a machine that makes computing a lot easier than it was before. Tasks taking more than weeks are reduced to the matter of seconds. The history has witnessed the demand of the machines that not only process the tasks quickly but also store large amount of data.

**Literature Related to the Methodology**

For decades, RDBMS has been popular for storing and processing the data. But with the increasing volume, variety, and velocity of data generated, these traditional methods of handling data are becoming more and more challenges. RDBMS use a single data types and data are stored in the form of row and column. But there is a need of new software that can handle big data without any challenges. To find a solution to the traditional methods of handling data, the academic articles from the past will be analyzed. And the software will also be installed to examine the small project related with the storing and processing big data. Academic articles will be used for getting the basic concepts of the software that handles big data. And the software will be installed in the local machine to support the concept obtained from the academic articles.

**Modern Techniques of Handling Big Data**

There is various application which can be used for the analyzing big data. In order to address the challenges of big data, Hadoop software can be used as an alternative solution. Hadoop is an open source software which can be downloaded from apache.org for free of cost. To get started using Hadoop software there are some recommended hardware and software details provided in the hadoop.apache.org website. According to the hadoop.apache.org, the hardware and software pre-requisites for using single node cluster Hadoop is listed below:

**Hardware requirement.**

A) The Computer with Linux or windows operating system.

Hadoop project can be performed in the Linux or Windows operating system. For this research study, Windows 10 version of operating system has been used.

B) RAM size 4gb or higher:

For this research study, 8 GB RAM has been used but the recommended RAM size is 4gb or higher.

C) 2 or more core processor:

The 2 core processors were allocated for this research study. And for installing Hadoop in the system, 2 or more core processor is needed.

**Software requirement.**

A) Java 7 or higher

B) Hadoop 2.7 or higher

C) Eclipse or IntelliJ Community Version

D) Virtual Machine and Cloudera (optional)

Hadoop is owned by Apache Foundation, and is available for free for downloading. Java, Eclipse, and IntelliJ are also available for downloading for free. Students and researchers can use these software for free. The software is also available from commercial vendors which provide support when needed. Cloudera and Horton works are the companies which provides Hadoop supports, but they charge for service. Their free version can be used but the software support is not available when needed.

Once the requirements are meet, the Hadoop software can be installed for free of cost to get started with the simple project. Later the software and hardware can be increased to work on more complex project with bigger volumes and variety of big data. The details for installing the Hadoop software will be discussed in the Tools and Technique section of Chapter 4.

**Summary**

Big has become an essential information for most of the small and big companies. Almost every transaction made these days are electronically done. Electronic transactions are generating large volume of data which are making more challenges to the traditional methods of handling data. To improve the company's performance and customer satisfaction, many companies have started to collect the data from their customer. But the volume and variety of data have challenged the traditional methods of handling big data. The characteristics and opportunities of big data were analyzed using literature in the field of big data in this chapter. Various work from the scholar concludes that although big data have significant opportunities, it is becoming more challenges to the traditional methods. And there is a need

of new software that can handle the big data challenges. This chapter analyzes the article from

past to discover the challenges of handling big data, and recommends the feasible way to

handle the big data using Hadoop. The next chapter will briefly discuss on the concept of

Hadoop software and how data can be handled using the Hadoop software.

**Chapter 3: Methodology**

**Introduction**

This chapter will explore Hadoop in more details. Using the past work related with Hadoop, the characteristics and architecture of Hadoop components will be discussed with the example. Also, the chapter will discuss about the tools and techniques that are used as an alternative software for handling the big data. The main goal of this chapter is to establish Hadoop as an alternative source for analyzing big data.

**What is Hadoop?**

Hadoop is an open-source framework maintained by Apache Software foundation which can stores and process big data. One of the creators, Doug Cutting's, who coined the name Hadoop in honor of his son's little yellow toy "elephant." Hadoop is the implementation of MapReduce based on the model proposed by google in October 2003. Every year the data generated are growing at an exponential rate, and the generated data are in different format. These increasing rates of data generated, and different format are challenging the traditional method of handling data. To deal with the challenges of traditional method of handling data, the concept of Hadoop software was introduced (DeRoos, 2014). Hadoop can easily handle large data set at any scale without any issues. Hadoop ecosystem consists of two widely used components, one is used for storing and another is used for processing the big data. The storage part is known as Hadoop Distributed File System (HDFS) and the processing part is MapReduce model. Hadoop consists of Cluster which are made of various kinds of nodes. Typically, the nodes found in the clusters are master nodes, and slave nodes as shown in the

Figure 4. Hadoop software is composed of Hadoop module and Hadoop project. User can use these modules and projects to store and process the big data.

**Hadoop modules**. According to the website hadoop.aoache.org, Hadoop framework consists of four major modules which are as follows:

a) **Hadoop Common**: Hadoop common is defined as utilities that can be used by other components of the Hadoop. Hadoop common have Java archive files which are required for starting the Hadoop Software.

b) **Hadoop Distributed file system**: Also known as HDFS which is a data storage module of Hadoop. HDFS contains a metadata which has details about the file stored in the HDFS. More about the HDFS will be discussed later in the data analysis section of chapter 4.

c) **Hadoop Yarn**: Yarn is an acronym used for Yet Another Resource Management, which maintains the job scheduling and clusters in the Hadoop world (White, 2009). Basically, yarn was developed to improve the MapReduce implementation but also support other parallel distribute computing system. Yarn maintains the cluster in two ways: one through resource manager and another through node manager. Resource manager can maintain the use of resource over the cluster whereas the node manager can launch and monitor the container. When the user requests to run the application in the Yarn, the resource manager is contacted, and the resource manager finds node manager, who can run the application in the container (White, 2009). Yarn is designed to overcome the limitations of MapReduce such as scalability and availability.

d) **Hadoop MapReduce**: Hadoop MapReduce is one of the Hadoop modules that processes the data very fast. MapReduce has two jobs: one is Mapper and the another one is Reducer. In this module, the Master nodes assign the task to the data nodes, i.e., Mapper and Reducer. Mapper will find the data from the blocks and Reducer will sort the results. More about the MapReduce will be discussed later in the data analysis section.

**Hadoop ecosystem**. Hadoop is an open source framework maintained by the Apache Foundation for reliable, scalable, and distributed computing (Welcome to Apache™ Hadoop®!, 2014). According to the website hadoop.apache.org, the components of Hadoop are defined as projects which function different to each other's. Some of the widely used Hadoop components are as follows:

a) Ambari: According to the website ambary.apache.org, Ambari is an application which is aimed for provisioning, managing, and monitoring the Hadoop clusters. Ambari also have tools that are used for adding and removing the slave nodes.

b) Avro: Avro is used for data serialization which provides a container file for storing persistent data. Avro was created by Doug Cutting for making Hadoop to be writable in many programming languages such as C, C++, C#, Java, JavaScript, Python, Ruby. Avro provides a schema resolution capability which allow old and new user to perform task without any difficulties. The schema can be declared using optional field for reading and writing the data. Avro has metadata where files are stored.

c) Cassandra: Hadoop Cassandra provides database that can be easily scalable and highly available without interruption in the job performance.

d) Chukwa: Chukwa is a data collections system which is mainly used for displaying, monitoring, and analyzing the outcomes of the collected data.

e) HBase: HBase is defined as an application which enables a scalable, distributed database for large tables in a structured format. It is a column oriented database added on the top of HDFS. HBase contains table which are made of row and column, but the data are stored and determined as a column family. Usually, HBase applications are used for reading and writing the real-time data. HBase address the big data scalability issue by just adding the nodes on the HDFS. HBase is not similar to RDBMS but can perform a job that cannot be done by RDBMS such as populating tables in the cluster and managing the large sets of data. In October 2007, first HBase project was introduced along with Hadoop (White, 2009).

f) Hive: Hive is a data warehouse system which uses the Hive query language commands to manage the databases. This Hive query language is a mixture of MySQL, SQL-92, and Oracle's SQL. Hive supports datatypes such as numeric, Boolean, string, timestamp, arrays, maps, structs etc. and stores the data in row or file format. Hive was initially developed by Facebook to learn and manage the large volume of data generated by the Facebook users. Hive was designed for programmer who has a good knowledge of SQL. Hive uses schema on read which makes it faster than the traditional database. Hive architecture is different from

traditional database system, but is similar with traditional database in terms of using SQL interface. Every year the architecture of Hive is improved, which makes it looks similar as a traditional database (White, 2009).

g) Mahout: Mahout is a data mining software that can be easily scalable. Mahout offers java libraries or scalable machine learning algorithm which can be used for analyzing the data. These machine learning algorithms allow user to perform a task such as classification, clustering, association rule analysis, and predictive analysis (Deroos et al., 2014). For supporting the statistical analysis, mahout algorithms are classified into three categories. They are collaborative filtering, clustering, and classification. Basically, mahout is designed as a recommendation system. For example, amazon.com recommends its customers new products based on their reviews and shopping habits from the past.

h) Pig: Pig is a Hadoop component used for parallel computing. Pig can be very useful while using multiple data structures. They use joins to connect the multiple data structures which is not available in the MapReduce project. MapReduce project may contain a lot of steps completing the job such as writing mapper and reducer code, compiling codes, submitting jobs and writing the output. Pig is run through very few lines of code. Pig consist of two parts: Pig Latin as a language for data flows and execution environment for running the Pig Latin language White, 2009).

i) Spark: Spark is a computing system which is used for configuring the Hadoop cluster for fast procession of Hadoop data. Spark does not use MapReduce job of

execution engine to run the job. It uses its own distributed runtime to complete the job. The interface supports three kinds of languages: Java, Python, and Scala. Scala is similar to Java programming language but use few less lines of code than Java programming. Analytic tools such as MLlib, GraphX, Spark Streaming, and Spark SQL can be added to spark for analyzing the data. Spark works in the concept of job which are equivalent to Map or Reduce job. Spark application can run multiple jobs in a series or parallel.

j) Tez: "Tez is a data-flow programming language build in the Hadoop Yarn to execute an arbitrary DAG of tasks to process data for both batch and interactive use-case" (Welcome to Apache™ Hadoop®!, 2014).

k) ZooKeeper: ZooKeeper is a high-performance co-ordination system used for distributed applications. It helps maintaining the configuration information, and provides distributed synchronization. Zookeeper allows user to build a distributed application that can handle the partial failure of the jobs. Zookeeper is highly available as it runs multiple machines. Zookeeper has a common library where programmer can write or use a program from a library.

**Architecture of Hadoop**. Based on the MapReduce model, Hadoop consists of three layers and they are as follows:

Application Layer/end user access layer, MapReduce workload layer, and distributed parallel file systems/data layer (DeRoos et al., 2014).

A) *Application Layer/end user access layer*: This layer serves as a medium of communication between the end user and the Hadoop. End users interact with Hadoop

through application software which employ the programming interface such as java. Eclipse is one of the applications through which end user can easily perform work on java based Hadoop projects.

B) *MapReduce workload management layer*: The performance and reliability of Hadoop depends on this layer. In this layer, all aspects of Hadoop environment including scheduling and launching jobs, balancing resources work load. This layer also deals with the failure and issues of the Hadoop.

C) *Distributed parallel file system or data layer system*: This layer consists of the storage of information. Most of the information in the Hadoop is stored in the form of Hadoop distributed file system.

**HDFS**. HDFS also known as Hadoop Distributed File System is one the Hadoop components which handles the storage of big data. When users need to add more storage in the system, then they can easily increase the storage capacity by adding servers. HDFS consist of number of clusters depending upon the user configurations. The cluster consists of Master and Slave nodes. The data in the Hadoop cluster are broken into many small blocks which are 128 MB sizes by default. These blocks are stored in the different slaves' nodes in the Hadoop clusters. These blocks are highly scalable and can be increased when needed.

In HDFS, users can create a new file, append content to the end of file, delete or rename the file, and modify file attributes (DeRoos et al., 2014). In comparison to traditional method of handling data, Hadoop's storage can be scalable at a very low cost because Hadoop uses commodity hardware.

Hadoop is composed of clusters. And cluster have Master node and Slave node as shown in the figure 4. Master node is also known as name node which assigns jobs to the slave nodes. Beside assigning jobs to the slave nodes, master node manages the file system namespace. All the details are store in the form of namespace image and edit log (White, 2009). Cluster have only one master node, where as it may have multiple slave nodes. The function of slave nodes is to store data in the form of blocks and performed a job assigned by the master node.
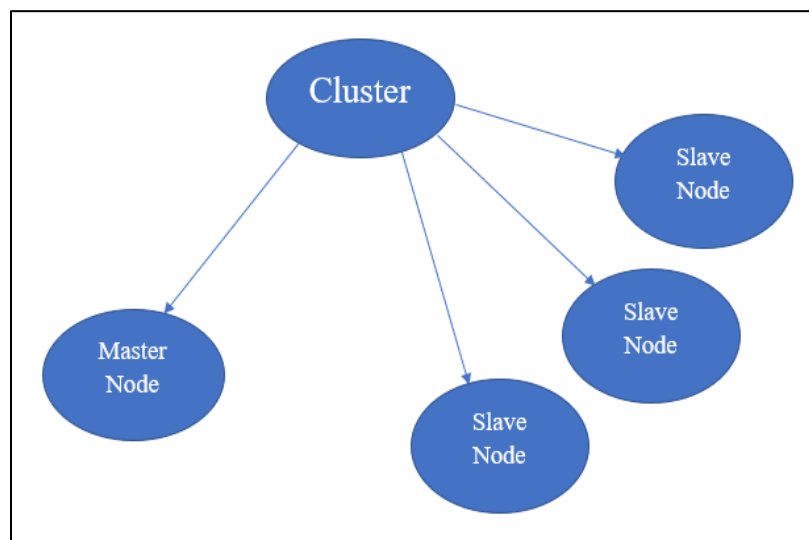


*Figure 4*. Cluster architecture.

**MapReduce**. MapReduce is a programming model that processes and generates the big data. In the past, Google has implemented hundreds of computational processing that handles large data. The computational at that time was straight forward but were having lots of issues like, parallel computation, distribution of the data over various machine, and handling errors while processing large amount of data with the simple computational design. So, in 2005 Google came up with a solution to resolve these kinds of issues which is based on the Map and Reduce operation. The Map operation applies computation of key/value pairs in

an input and Reduce operation combines all the result value that is computed from the result of Map operation. As shown in the figure 5 below, the users divide the input files in different blocks of 128MB size and these blocks generate the number of copies program in the clusters. Every cluster has different programs with one master nodes and several data node. Data nodes are also known as a worker node and may be assigned Map work or Reduce work by the master node.
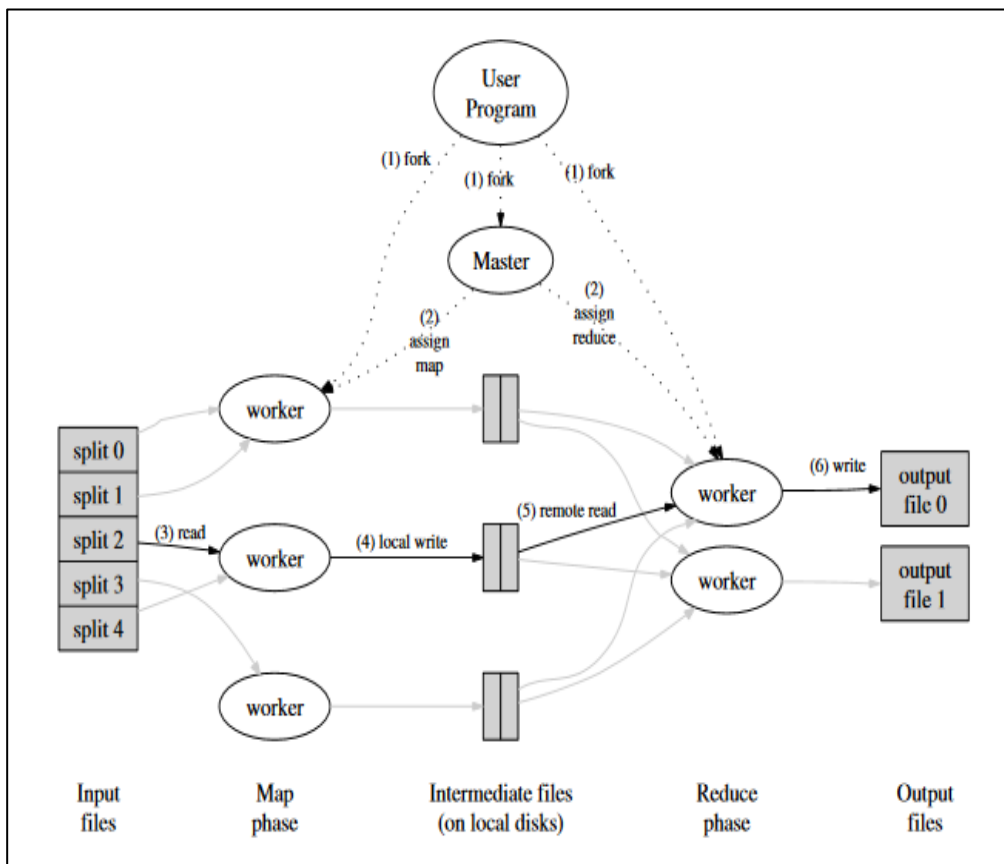


*Figure 5*, Map and Reduce operation (Dean & Ghemawat, n.d.).

Once the user defines the input files, the master nodes assigns the worker node for Map function. Those worker nodes who are assigned for Map work reads files from different input files and writes the file in local disk. Once the Map worker nodes finished their work by

writing the result in local disk, another sets of worker nodes are assigned for Reduce function. The assigned worker nodes read the files from local disk and write it to the output files. In this way, the retrieved process is completed in the Hadoop MapReduce (Dean & Ghemawat, n.d.).

Wordcount is a Hello World program of the Hadoop world. In this program, the word is counted from the file in such a way that it will count how many times the word is repeated in the file. In wordcount program, the words are sorted in various steps which are as follows (Guo, Rao, Cheng, & Zhou, 2017):

A) *Input:* In this step, the data are copied in the HDFS as an input. These inputs contain the words in any format.

B) *Splitting*: Once the data are copied as input in the HDFS, the data are split into different blocks depending on the sizes of the file and block. If the size of the file is 500 MB and the size of the block is 128 MB, the file will be split into four blocks. In the blocks, the file is arranged in such a way that every word is separated by space, comma, etc.

C) *Mapping*: The counting of words occurs in this phase. The words are counted and given a 1 value for every word. In this step, the results are arranged in the form of key and value pair. Each pair will have one key word and value 1. Suppose if the file has "This is a wordcount example," then the result in this phase will be following:

This,1

Is,1

A, 1

Wordcount,1

Example,1

D) *Shuffle*: The result obtained from the Mapping steps are further sorted and arranged in such a way that repeated words are put together as a group. But still the key and value appear to be the same, and the repeated word are listed as many time as they appear in the file. For word a "Apple" appearing 3 times and "Mango" 2 times in the file, it will be arranged as:

Apple, 1

Apple,1

Apple, 1

Mango, 1

Mango, 1

E) *Reduce*: In this step, the sorted group of word from the shuffle step is further sorted in such a way that the repeated words appear only one time and the number of time repeated will be added to the value. In this phase, the example from the shuffle phase after sorted will looks like following:

Apple, 3

Mango, 2

If there are more words in the file, then in the same way rest of the repeated words are sorted and the value are added to the number of times they are repeated.

F) *Results:* Finally, the sorted words from the Reduce phase are put together into the output where users can read the results.

This graphical representation of the entire process is shown in figure 6. As we can see

the wordcount process is separated into three phases in the figure 6: Input Split, Map Task,

and Reduce Task.
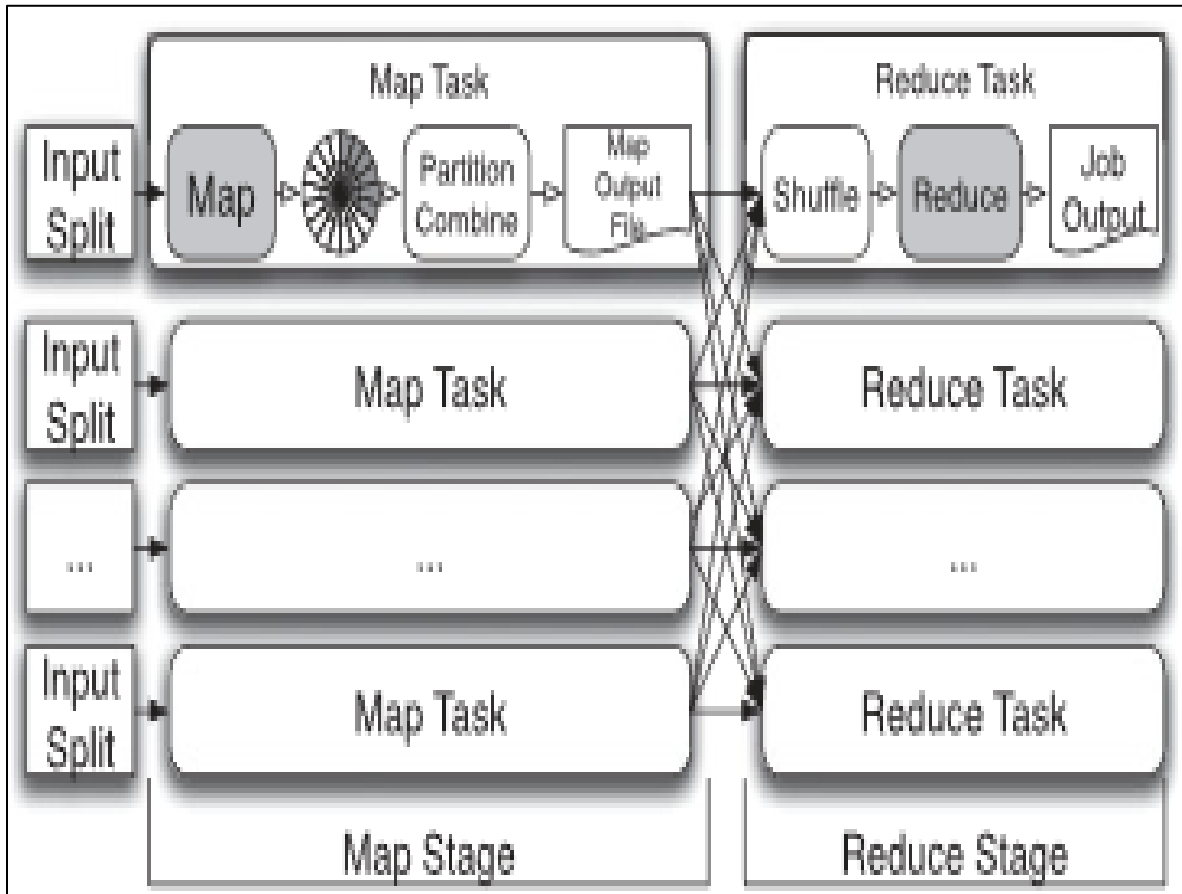


*Figure 6*. Wordcount program (Guo et al., 2017).

The example code that runs the MapReduce program explained above can be written

in Scala programming language which is similar to Java: all of the steps involved in the above

description is as follows:

**package** com.ims

**import** org.apache.spark.{SparkContext, SparkConf}

**object** WordCount {

```
  def main(args: Array[String]) {

    val conf = new SparkConf().setAppName("word count").setMaster("local[2]")

    val sc = new SparkContext(conf)

    System.setProperty("hadoop.home.dir", "C:\\winutil\\")

    val input = sc.textFile("data/inputs/wordcount.txt")

    val mapOutput =  input.flatMap(line => line.split(" ")).map(x => (x, 1))

    mapOutput.reduceByKey(_ + _).collect.foreach(println)

     mapOutput.reduceByKey(_ + _,1).saveAsTextFile("data\\output")

  }

}
```

**Hadoop VS traditional database**. Although the function of both Hadoop and traditional database are same, they differ in many ways. Some of the key differences are as follows:

A) *Data types***:** Most of the traditional database expects the data in the form of structured format such as XML or tables. If the data are in the form of unstructured, then the data should be first turn into the structured format so that the system can support for further processing. When data generated started to evolve in the mixed structured form then it becomes more challenging for traditional methods to handle the data type.  Whereas Hadoop MapReduce can handle both kinds of data format i.e. structured and unstructured (Natarajan, 2012). During data processing, usually traditional databases care about the redundancy of the data. The more redundancy there will more performance issues but in case of Hadoop the data

are replicated into different blocks for a better performance. So, data type is one of the major

differences between the Hadoop and traditional database.

B) *Servers***:** Both systems may have several servers but in traditional database system

if one server goes down then it will affect the whole database. But in Hadoop, if one system

fails then it does not affect the other system, since the files are divided and replicated in small

blocks of 128 MB. Adding more server to address stability issues can be done in traditional

database but the performance will be affected. The system will run slow after adding the

server in traditional database but in Hadoop several more servers can be added without any

speed issues (Ozcan, 2013).

C) *Programming Language*: Hadoop projects can be performed in various

programming languages. It uses programming language such as Java, Python, Hive Query

language, and R. But RDBMS uses only one programming language, Structured Query

Language.

Traditional database system is widely used by many companies for decades whereas

Hadoop is becoming popular in recent years to handle the big data. Hadoop was introduced in

2005 by Doug Cuttings and Mike Cafarella. Figure 7 shows the comparison of some of the

database techniques. In the figure 7, the applications such as Hive, Giraph, HBase and

RDBMS are compared on the basis of changeable data, data layout, data types, hardware,

high availability, indexes, and query language.

| Criteria | Hive | Giraph | HBase | RDBMS |
|---|---|---|---|---|
| Changeable data | No | | Yes | Yes |
| Data layout | Raw files stored in HDFS; Hive supports proprietary row-oriented or column-oriented formats. | | A sparse, distributed, persistent multidimensional sorted map | Row-oriented or column-oriented |
| Data types | Bytes; data types are interpreted on query. | | | Rich data type support |
| Hardware | Hadoop-clustered commodity x86 servers; five or more is typical because the underlying storage technology is HDFS, which by default requires three replicas. | | | Typically large, scalable multiprocessor systems |
| High availability | Yes; built into the Hadoop architecture | | | Yes, if the hardware and RDBMS are configured correctly |
| Indexes | Yes | No | Row-key only or special table required | Yes |
| Query language | HiveQL | Giraph API | HBase API commands (get, put, scan, delete, increment, check), HiveQL | SQL |

*Figure 7*. Comparison of Hadoop-bases storage and RDBMS (DeRoos et al., 2014)

**Scope of Hadoop**. With the emergence of big data concept and its opportunities, the demand of a Hadoop developer has also been rise. According to the site indeed.com, the demand of jobs in Hadoop field has been increased since 2009 as shown in the below figure 8. Hadoop can be seen at no. 7 in the job trend. And the job trend for Hadoop is growing every year. Having a Hadoop skill is to have wide variety of skills such as Linux, Java, R, SQL, and python. There are many job role opportunities in Hadoop such as Hadoop Developer, Data Scientists, Hadoop Admin, Business Analyst, etc. (Kerzner & Maniyam, 2013). There are

many big companies who have already implemented Hadoop software. The complete lists of companies using Hadoop software can be found in the hadoop.apache.org website. Not only companies are implementing the Hadoop software, but Government are also pouring money on big data. In March 2012, Obama administration announce to spend $200 million on big data research (Hu, Wen, Chua, & Li, 2014). Based on this job trend and investment on big data, the future of Hadoop field is very bright. There are lots of opportunities in Hadoop field for individuals and companies who want to implement Hadoop software.



*Figure 8*. Hadoop job trend (Kerzner & Maniyam, 2013).

**Implementation of Hadoop**. The components of Hadoop have been implemented by several big companies among which Amazon.com, google, Facebook, E-bay are currently using Hadoop software. Some of the examples of big companies using Hadoop project are seen in figure 9. Yahoo, IBM, Rackspace, Facebook, E-bay, Google, LinkedIn are some examples that are implementing Hadoop software to store and manage the large amount of data. There are some vendors who has started 24/7 support service for its customer. Companies can pay very little amount as a service fee and take advantage of commercial Hadoop software. Cloudera and Horton Works are such vendors who provide Hadoop Software support services. The complete lists of the company implementing the Hadoop software can be found at hadoop.apache.org. E-bay, one of the Hadoop users, has 532 nodes. Each node has 8 core processors, and the storage capacity is 5.3 petabytes (Powered by Apache Hadoop, n.d.). Every year the investment on the big data is increased by 30% and it is estimated to reach $114 billion by 2018 (Asthana & Chari, 2015). Most of the companies are benefitted by using Hadoop software in many ways, some of them are as follows:

- Improving operational performance
- Advertisement, marketing, and sales
- Security and system management
- Education and training
- Database management
- Data analytics
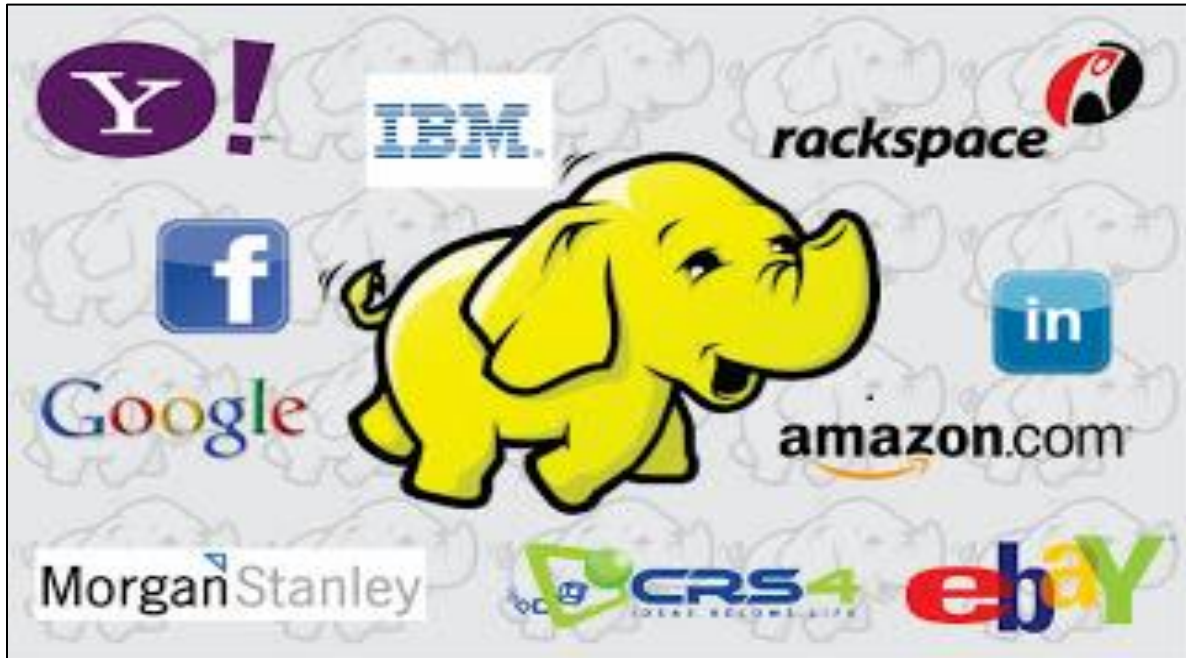
*Figure 9*. Companies using Hadoop (Marathe, 2017).

**Design of the Study**

The results of this research study solely depend on the academic research works done

in the past. Rather than taking a survey, this study analyzes the academic articles and

conclude the similar ideas presented by different authors. So, Qualitative Study is the study

design appropriate for this research study. The documents on big data from the past will be

selected and analyzed to learn the characteristics of big data. Also, the academic articles are

used for determining the basic concepts of big data and concluding the challenges when

traditional methods are used for handling big data. With the help of simple project and articles

related with the software that handle big data, will be analyzed and the results of the research

work will be noted. Using Hadoop components, an analytical research method will be applied

to get the informational meaning from the data generated in the large quantity. The result

obtained from the analysis of previously done research will be used to create the problem and solutions for this research study.

**Data Collection**

Being the qualitative nature of the study, the data for this research study will be collected from various academic journal article. Based on the research done in the past, the concluded results from various article will be analyzed to create a research problem and results. Also, the wordcount example using the solution recommended from the previous research works will be included in the research study. This includes a simple project that defines the wordcount project defining the MapReduce function to explore the techniques used for addressing the issues with traditional method of handling big data.

**Tools and Techniques**

In this research study, various tools will be used for supporting the conclusion of the study. This research study mainly uses open source software and simple, easy programming language to perform a basic project in Hadoop software.

The tools used during this research work are:

A) Hadoop Software,

B) Java Programming or Scala Programming,

C) Eclipse, and

D) Virtual Machine with Cloudera (optional)

**Hardware and software environment.** To conduct this research project, Hadoop software will be used for analyzing big data. To run the basic project on Hadoop, the required hardware and software are explained below:

A) *Hardware***:** The computer with 4 GB RAM, 2 core processors, and 1 terabytes of hard-disk is recommended. The Hadoop project can be run in the Linux or windows operating system.

B) *Software***:**

1) *Hadoop*: Latest version of Hadoop can be downloaded from hadoop.apache. org. website for free. The document related to download Hadoop software can be found in the website too. The program to run this software is written in Java or Scala programming language. Java programming language is a core part of the Hadoop project since the configuration settings and program are written in Java programming language.

2) *Java*: Java is another software that is needed for running Hadoop project. Java is an open source which can be downloaded for free from www.java.com. The benefit of using Java are as follows (Urquhart, 1997):

A) Open source: Java is an open source which can be downloaded from the internet for free.

B) Freely available resources: Java is very easy to learn. There are lots of resources found in the internet for learning purpose. W3school provides free trainings related to java which can found in the following website: http://www.w3schools.in/java-tutorial/intro/ . There are so many

community forums available in the internet that can help in the specific

issues related with the java programming.

C) Easy to learn: Java uses a simple model of network computing that is easy

to understand and program.

D) Auto error detection features: Some common errors like memory leaks,

dangling pointers, and incompatible object assignment are detected

automatically by automatic garbage collection and type safe reference.

E) Multitasking feature: The multithreading feature found in Java helps user

to work on different project at the same time.

3) *IntelliJ IDEA or Eclipse:* For this research study, Intellij IDEA software is

used for running the example of Hadoop project. Intellij is an open source

which can be downloaded for free from www.jetbrains.com. The Community

version of this software is available for free which students can take advantage

of for learning purpose. IntelliJ Idea provides Integrated development

environment (IDE) for Java. As an alternative for Intellij IDEA, users also can

use eclipse which is also an open software. This software can be downloaded

from https://www.eclipse.org/downloads/.

**How to Install Hadoop Software**. To install single cluster Hadoop

in the local machine, three open source software needs to be installed:Java JDK, IDE for Java,

and Hadoop software. According to the article "Installing Hadoop-2.6.X on Windows 10" by

Shantanu Sharma, Hadoop software can be install through several steps and need latest

version of Java, Eclipse, and latest version of Hadoop software. To get started with installing

the Hadoop, the Java needs to be installed first. The latest version of Java can be downloaded

from the following website for free:

http://www.oracle.com/technetwork/java/javase/downloads/index.html. In this website the

latest version of JDK can be downloaded by clicking the tab as shown in the figure 10.
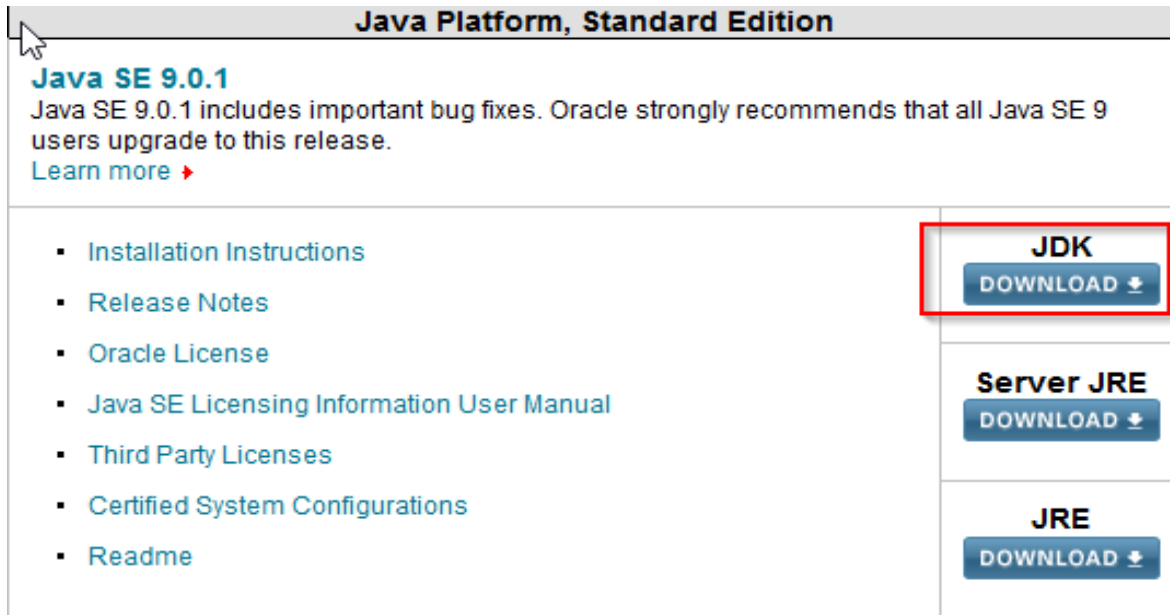


*Figure 10*. Java JDK downloading window.

In a download window, select JDK download and new window will pop-up. In new

window, select accept to the term and condition of the Oracle corporation and start

downloading according to your operating system as shown in the figure 11. By clicking

windows option, the new window will pop up asking user to save the file as shown in figure

12. This file can be saved in the C drive. This path will be later needed to configure the

environment variable setting.
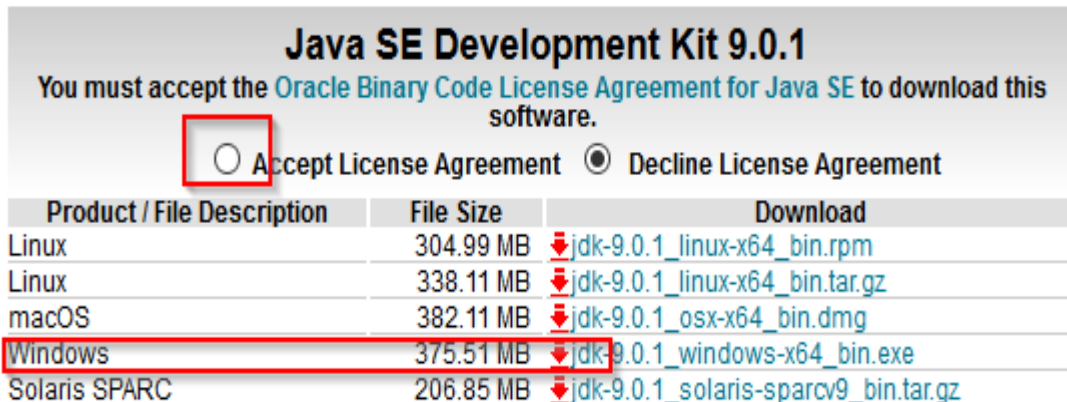
*Figure 11*. Java JDK for Windows 64

The Java JDK needs to be saved in the local machine and installed by double clicking

the saved file.
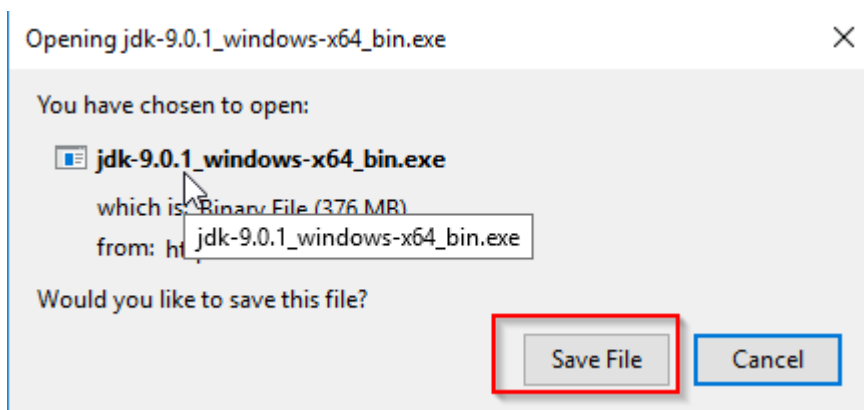


*Figure 12*. Windows prompt for saving file for Java JDK

Once the Java is installed successfully, it can be verified by using the "java -version"

command in the command prompt window. Upon installing java, the environment variable for

java should set up. The variable will be JAVA_HOME and the value for this variable will the

path from the local system where Java JDK is stored. And the system variable should be

changed. To do so click on Path and select edit and add %JAVA_HOME%bin to set up

system variable for Java.
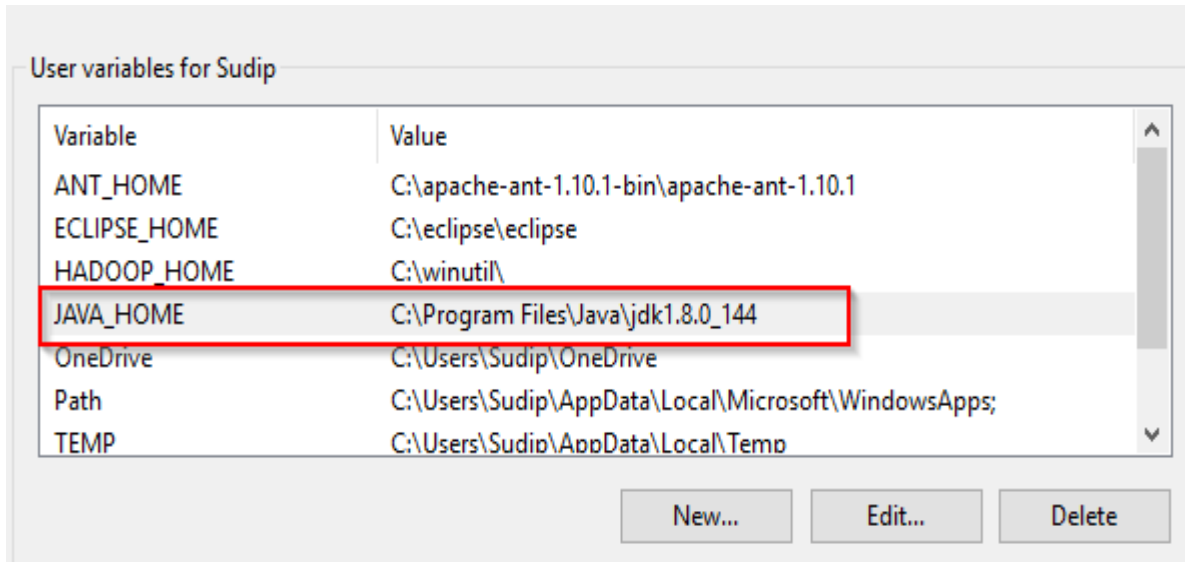
*Figure 13*. Environment variable setting window.

After setting up the environmental variables for Java, the integrated development

environment also known as IDE should be downloaded. Either Intellij IDEA or Eclipse can be

downloaded as IDE. Eclipse can be downloaded from https://eclipse.org/downloads/



*Figure 14*. Window for downloading Eclipse Oxygen version.

The downloaded file can be installed by double clicking the downloaded file from the download folder. Now, installed software files should be save in the C folder. The environment variable for eclipse should be set up. This can be done by going to system properties window and selecting environment variable. In environment variable window, the user variable should be ECLIPSE_HOME and value for this should be the path where the eclipse is saved in the local machine. The system variables also need to be edited. To do so select the Path from system variable and select edit. After selecting edit, add %ECLIPSE_HOME%\bin to the list. Now the plugging master for eclipse should be downloaded and it can be downloaded from [https://github.com/winghc/hadoop2x-eclipse-plugin/tree/master/release](https://github.com/winghc/hadoop2x-eclipse-plugin/tree/master/release) . Download three jar files from the website and save it to eclipse\dropins folder. If the eclipse folder is save in C folder of local machine, then the path will be C:\eclipse\dropins or C:\eclipse\eclipse\dropins. Now slf4j-1.7.21 needs to be downloaded and copy all jar files to eclipse\plugins.

After installing latest Eclipse in the local system, Hadoop software can be downloaded from hadoop.apache.org website [http://www.apache.org/dyn/closer.cgi/hadoop/common/hadoop-2.7.4/hadoop-2.7.4-src.tar.gz](http://www.apache.org/dyn/closer.cgi/hadoop/common/hadoop-2.7.4/hadoop-2.7.4-src.tar.gz). The selection of the version can be done by doing google research or the latest release version can be selected. Once in the download window, it will suggest link to download and can be followed to download latest Hadoop as shown in figure 15.

*Figure 15*. Download page for latest version of Hadoop.

The stable release can be downloaded which will be in the zipped tar file. The zipped tar file should be unzipped and save in the local machine. The Hadoop common files should be downloaded from following website https://github.com/amihalik/hadoop-common-2.6.0-bin/tree/master/bin . These 11 files as shown in figure 16, need to be downloaded and copied in the hadoop\bin folder.



*Figure 16*. List of Hadoop common file

After installing Hadoop software some of the files from the Hadoop folder need to be configured. The Hadoop XML files need to be configured. The common files that need to be configure are: core-site.xml, hdfs-site.xml, map-red.xml, and yarn-site.xml. All of the files that need changed is found in hadoop/etc/hadoop folder. The hadoop-env.cmd file should be configured as shown in figure 17.



*Figure 17*. hadoop-env.cmd configuration file.

Now, the environment variable for Hadoop also needs to be configured. The variable for Hadoop should HADOOP_HOME and the value should be the path where the unzipped Hadoop file is stored. The system variable also needs to be set up. To so, on system variable select PATH and select edit. After selecting edit, following path should be added in the system variable (Sharma, n.d.):

%HADOOP_HOME%\bin

%HADOOP_HOME%\sbin

%HADOOP_HOME%\share\hadoop\common\*

%HADOOP_HOME%\share\hadoop\hdfs

%HADOOP_HOME%\share\hadoop\hdfs\lib\*

%HADOOP_HOME%\share\hadoop\hdfs\*

%HADOOP_HOME%\share\hadoop\yarn\lib\*

%HADOOP_HOME%\share\hadoop\yarn\*

%HADOOP_HOME%\share\hadoop\mapreduce\lib\*

%HADOOP_HOME%\share\hadoop\mapreduce\*

%HADOOP_HOME%\share\hadoop\common\lib\*

Once the environment variables are configured properly, Hadoop software is ready for use. Users can write program in Java language and compile it through eclipse.

**Advantages of Hadoop**

Hadoop is an open source software available for downloading for free of cost. Besides being an open source, it has lot of benefits which are discussed below. Following are some advantages of using Hadoop software to handle big data.

A) **Cost effective:** Traditional methods of handling big data involve lots of additional servers for data storage and extra computational devices. But could save lots of money in both professional and individual level. Traditional methods may cost $25,000-$50,000 for maintaining 1 terabytes of data for a year. But using Hadoop may cost few thousand dollars for maintaining 1 terabytes of data (Asthana & Chari, 2015). For learning and individual purposes, Hadoop is available as an open source. The only concern with the personal computer while using Hadoop is processing RAM. But can be added at a very low price. The suggested RAM capacity for the using Hadoop software is 4 GB. In the professional level, big companies can buy a storing and processing space in the cloud at a minimum service fee charge which can save lots of money in the infrastructure required for handling the big data.

Cloudera and Horton works are among some companies who provide 24/7 support with a nominal fee charge.

B) **Open-source:** Hadoop is an open source software which can be downloaded for free from Apache Foundation website. It is operated by Apache foundation and several computer programmers contribute the maintenance and progression of the software. It is widely available for download through Apache Hadoop website http://hadoop.apache.org/ . This website also has well documented article for how to install Hadoop in the system and also walk through some basic Hadoop projects. All of the software and materials related to Hadoop are available for free.

C) **Scalability**: Hadoop can store large number of data sets over the several servers and can process it in parallel. Hadoop allows user to add more servers as needed without any complexity. Traditional database management system can't handle big data due to scalability issues, since the computing processing is not parallel. In Hadoop, HDFS allows easy scalable storage for big data. The Hadoop storage can be increased by adding more nodes. HDFS consist of blocks which are 128 MB by default. In comparison to the traditional database, the size of the block in HDFS is very large. The default size for traditional disk are 512 bytes.

D) **Fast processing:** The architecture of Hadoop enables user to process data very fast. Since it has a distributed file system, the data are distributed over several clusters. During the program execution, the Map finds the appropriate clusters and combines the result as a reduce operation. Hadoop can process unstructured terabytes of data in minutes and petabytes of data in hours. According to the website hadoop.apache.org, in July 2008, one of Yahoo's

Hadoop clusters sorted 1 terabyte of data in 209 seconds, which was record at that time. This

achievement is called as Hadoop wins terabyte sort benchmark.

**Hadoop Data Security**

Hadoop has simple user authentication procedure while installing the software. Other

than that Hadoop does not provide the authentication services. But the data in Hadoop can be

secured by using Kerberos. Kerberos allows a secure communication between the client and

server in Hadoop for a user authentication without any exchange of password in the network

(Parmar, Bhattacharyya, Bandyopadhyay, & Kim, 2017). As shown in the figure 18, client

requests Kerberos Key Distribution Center (KDC) for authentication with requesting Ticket

Granting Ticketing (TGT). The client's request is granted by KDC along with the secure

randomly generated key which is also known as session key. Once the Client receives TGT

along with session key, requests the Ticket Granting Server for service request by sending the

TGT along with the request. Then the Ticket Granting Server grants the service request and

sends the session key and Ticket granted. Now the client and server establish communication

with the help of authentication service ticket. In this way the communication between the

client and server is secured in the network without transferring the authentication password
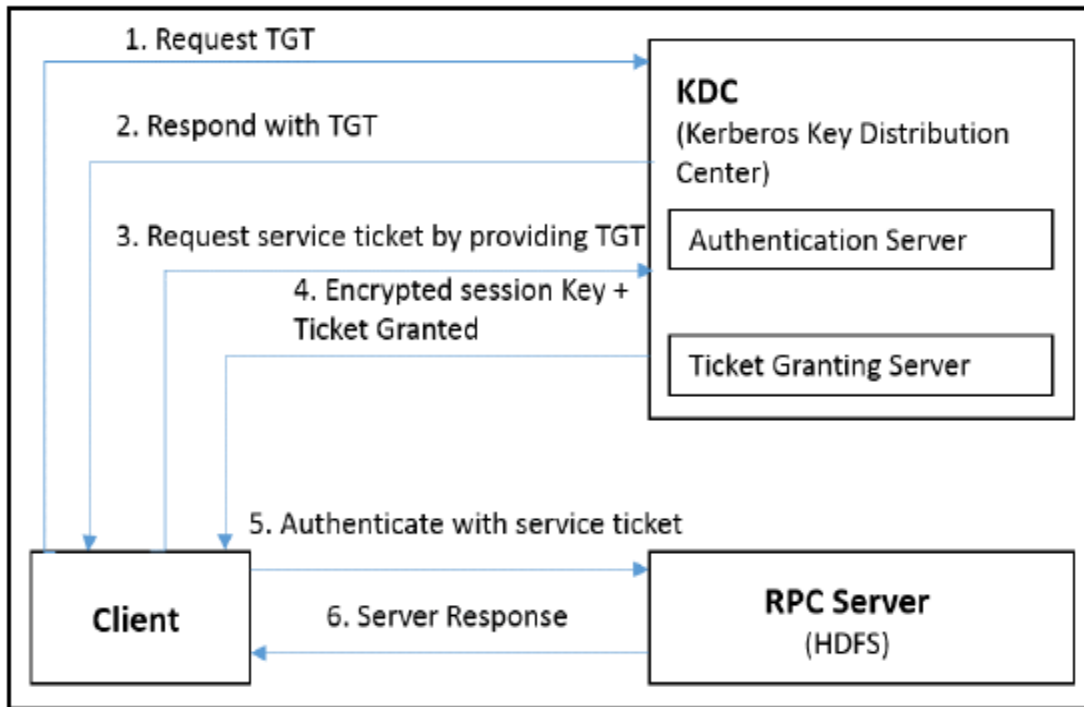
over the network.

*Figure 18.* Applying Kerberos in Hadoop (Parmar et al., 2017).

According to the paper by Sharma and Joshi (2015), Hadoop can be more secured by

the incremental security model. This security model suggests following considerations while

handling the big data:

A) While accessing data for access, modify, and controlling jobs, the role based

access control and attribute based access control can be applied. Applying such

access control, the system will decide which users can access, control or modify

the data in the Hadoop.

B) Applying the encryption technique Hadoop data.

C) Auditing the account of event of data accessed and modified by each user.

D) Making sure that all the security compliance for storing data are followed without

any duplication.

E) Cleaning the data, sanitizing the data, and proper destructing of the unnecessary

data.

**Summary**

Big data has various opportunities that can help companies improve their business

strategy and sales performance. Without right tools, the opportunities from big data may turns

into the challenges. Various big data challenges like storing, processing, scalability, etc. are

easily solved by using the Hadoop software. The components of Hadoop software HDFS

allows to store large volume of data and the MapReduce allows to process the large amount of

data in a fraction of seconds. This chapter discussed about the new software that can easily

handle the issues of big data. The goal of this chapter was to get the basic concepts of Hadoop

software. Next chapter will discuss about the analysis of Hadoop in more detail and how it

helps to address the challenges of big data.

## Chapter 4: Data Presentation and Analysis

**Introduction**

To learn about which software handles big data properly, it is necessary to learn the characteristics of the big data and challenges while handling through the traditional method. Various works from the past has been done on the nature of big data and how the volume of data is being generated and affecting the computing world. Academic articles are the basis for finding the research problem and solution. This section covers how data are collected and analyzed.

**Data Presentation**

For this research study, the academic journal articles were retrieved from Saint Cloud State University library database and IEEE articles. The background of the problem was searched in various paper and the findings is discussed in Chapter 2. In Chapter 2, the definition of term big data and how it is challenges to traditional data handling methods was learned. The various challenges of big data were discussed in the literature related to the methodology section and possible methods for addressing the big data issues was discussed in Chapter 3. Chapter 4 contains details about how the research problem is analyzed and suggest the possible solution to the problem. The data presented in this research study are solely based on the academic articles written by the past researchers. Also, the example code and projects support the result drawn by the literature review.

**Data Analysis**

Big data has become a basis of improving the organizational performance through analyzing the data collected from the various sources. Learning customer's pattern of buying

habits helps organization makes a future marketing strategy for maximum sales. Also, they can analyze the customer feedback for improving the sales goals. But big data has become the opportunity as well as challenges for companies by struggling to store and process the incoming volume of the data. System Scalability, timely analyze of data, handling variety of data and fast processing are some challenges while using traditional methods of handling big data. To address the challenges of big data, Hadoop is an alternative software that can easily solve the issues that are caused in the traditional methods of handling big data.  Here are some reasons why Hadoop software can be helpful tools to analyze the big data without any issues (Kerzner & Maniyam, 2013):

A) **Storage**: Storing data is one of the biggest challenges for traditional methods of handling large sets of incoming data. Storing large data is associated with many other challenges. It directly affects the processing performance, cost related to add more server. HDFS is one of the components of Hadoop that helps maintaining the storage of big data. HDFS consist of single cluster or multiple clusters. Every cluster consists of blocks which is 128 MB by default. When a user defines input, the contents of the input are equally divided into the blocks. And the data are replicated into the data nodes. Usually the traditional computing machine does have a disk system as a block which is 512 bytes in size. Typically, the size of the block in HDFS are very large. One cluster can store petabytes of data. The Cluster can be added easily at low cost whenever the scalability is required (White, 2009). So, HDFS allows user to store the lots of dataset and when needed, more servers can be added at a very low cost.

B) **Processing**: When the size of the dataset is larger, the time taken to process it is also longer while using the traditional methods of handling data. More servers are added to store the large quantity of data, but server does not support the parallel computing. In case of Hadoop the processing of data is parallel computing which saves the time to process. In traditional methods, 100 MB of data can process in 1 second when the data processing rate is 100mb/s. When another server is added for 100MB of data, it will take 2 seconds to process the whole data. Because the data is processed from one server and starts processing another one, once the first server is done. In case of Hadoop for the same dataset it will take 1 second because the server is processed parallelly at the same time. There many Hadoop components that handles the processing of very large datasets. Ambari, Spark, ZooKeeper are Hadoop components that can help maintaining the large amount of dataset. According to the hadoop.apache.org site, one of the Yahoo's Hadoop cluster sorted one terabytes of data in 209 seconds. So, Hadoop can sort large dataset very fast with high performance (Parmar et al., 2017). Wordcount is a simple example of Hadoop project that explains the fast storing and processing of the data which is explained later in this chapter.

C) **Cost efficiency**: Maintaining the database at a minimum cost is the one of the most important challenges of the Big data. Companies using traditional method of handling big data are spending $25,000 to $50,000 per year for 1 terabyte of data. As a result, companies using traditional methods might think twice before storing more data. But using Hadoop software can reduce this cost into few thousand dollars per terabyte per year (Kerzner & Maniyam, 2013). Commercial companies like Cloudera and Horton Works provide 24/7 support service at $3,382 per node per year for 2 terabytes of data. When using open source

Hadoop, then companies can save even more money. For 18 TB of data, Hadoop can be

installed and maintained at $1266 for a year (Asthana & Chari, 2015). The details of cost for

installing and maintaining Hadoop is shown in the table 2 below.

Table 2

*Yearly Cost of having Hadoop for 18TB Data* (Asthana & Chari, 2015)

| Investment Title | Cost in US Dollar |
|---|---|
| Software Cost/year | $41 |
| Maintenance Cost/year | $91 |
| Deployment Cost/one time | $320 |
| Administration Cost/year | $795 |
| Facilities Cost/year | $7 |
| Provisioning Cost/year | $12 |
| **Total Cost** | **$1266** |

  D)  **Allows more data to capture**: Due to cost related issues many companies do not

capture the large volume of data. In Traditional methods of handling data, the cost of

maintaining the 1 terabyte of data per year is $25000 to $50000. But when Hadoop software is

used, companies are saving lots of costs of maintaining the data. So, extra data can be stored

at the same price if we use Hadoop instead of using traditional method of handling big data.

This allows companies to capture more and more data at a low cost.

  E)  **Provides scalable analytics**: The HDFS and MapReduce components of Hadoop

allows a parallel storing and processing of data. With the increase of volume of the data the

analytics can be scalable in parallel distributed way. The MapReduce works for data that are

of petabytes in size.

F) **Provides rich analytics**: Hadoop has a unique quality of handling big data in

different programming languages. The project in Hadoop can be done using one of these

coding languages Java, Python, SQL, R, and Ruby.

There are other reasons why Hadoop can be used as alternative solution for big data

challenges, they are open source, easy to learn the programming languages. So, due to these

above unique features of Hadoop, the traditional methods of analyzing big data can be

replaced.

Following is wordcount example run in the Cloudera. The complete phases of

MapReduce project are explained with the example code. Following are the steps involved in

the word count job.

**Step 1:** To get started with the wordcount job which is also a hello world program of

the Hadoop world, the terminal in the Cloudera system is being used. The first step in this

terminal would be creating a directory on Hadoop. To do so following command will create a
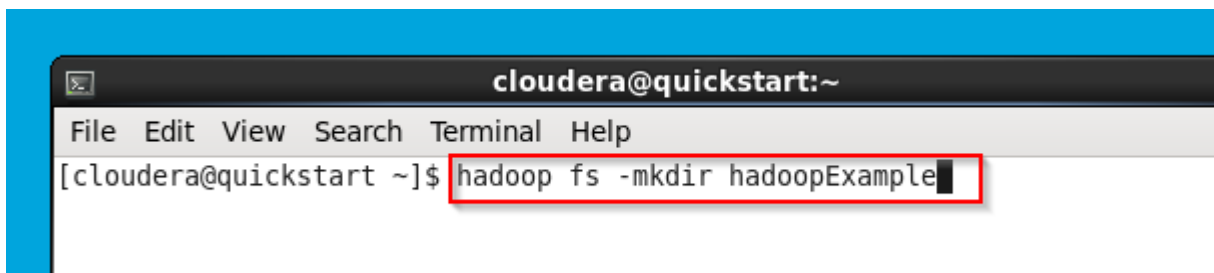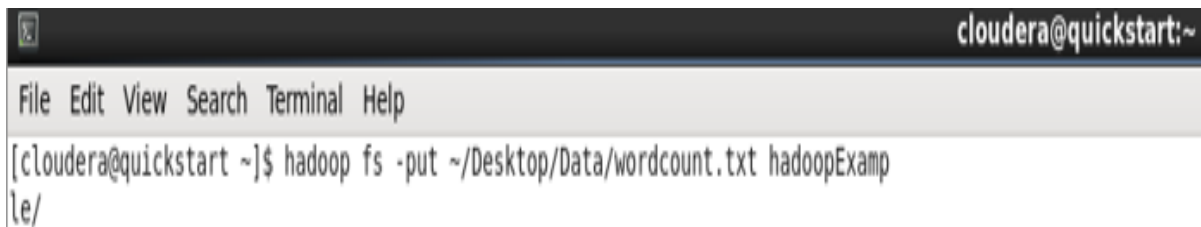
directory in the Hadoop system:



*Figure 19*. Screenshot of code of making directory in HDFS.

**Step 2:** Once the directory is created, the file need to be created in the local system. Since, Cloudera is being used, the folder Data was created, and inside the Data folder wordcount.txt file was created.

**Step 3:** Now, the file is copied in the HDFS from the local machine. Following code is being used for copying the data from local file to the directory in the HDFS.
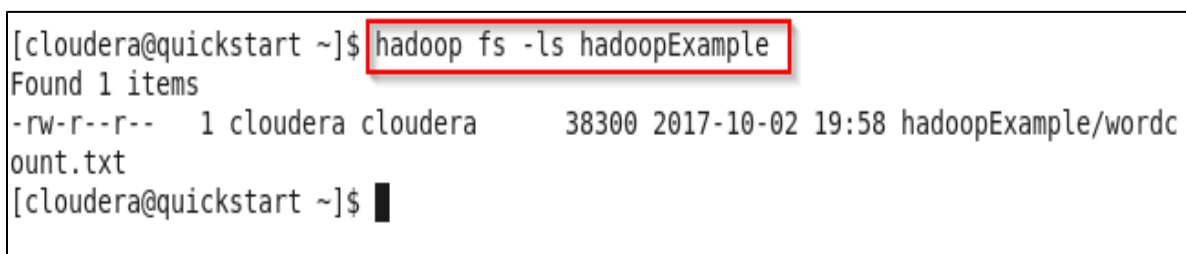


*Figure 20*. Screenshot of code adding file in directory.

In the above code, ~/Desktop/Data/wordcount.txt is the local file which is being copied and pasted to the destination folder as hadoopExample/

**Step 4:** Once the command is completed we can check the content of the HDFS directory hadoopExample by running this command as shown in figure 21:



*Figure 21*. Screenshot of code for checking file on directory.

In the above result it is confirmed that the hadoopExample directory contains the wordcount.txt file.

Also, we can see the content of the wordcount.txt by running the following command as shown in figure 22:

```
ount.txt
[cloudera@quickstart ~]$ hadoop fs -cat hadoopExample/wordcount.txt▮
```

*Figure 22*. Screenshot of code for checking content of file.

**Step 5:** Now, to write a code that run the wordcount MapReduce job, we need to run it through the Scala. After login to the Scala first we take wordcount.txt file from HDFS as an input as shown in figure 23.

```
scala> val input = sc.textFile("hadoopExample/wordcount.txt")
input: org.apache.spark.rdd.RDD[String] = hadoopExample/wordcount.txt MapPartiti
onsRDD[1] at textFile at <console>:27
```

*Figure 23*. Screenshot of code for copying file on HDFS.

We can also see the backend process of the code that we just executed as shown in figure 24. Here by executing input.toDebugString we can see what is going on during the execution of input code.

```
scala> input.toDebugString
res1: String =
(2) hadoopExample/wordcount.txt MapPartitionsRDD[1] at textFile at <console>:27
[]
 |  hadoopExample/wordcount.txt HadoopRDD[0] at textFile at <console>:27 []
```

*Figure 24*. Screenshot of code showing RDD path.

**Step 6:** Once we put txt file as an input, we split the words in a line so that the line is made in every space.

```
scala> input.flatMap(line => line.split(" "))
res2: org.apache.spark.rdd.RDD[String] = MapPartitionsRDD[2] at flatMap at <cons
ole>:30
```

Figure 25. Screenshot of code for splitting phase.

**Step 7:** The result obtained from the above code is put as input for the second input as shown in figure 26:

```
scala> val inputSecond = input.flatMap(line => line.split(" "))
inputSecond: org.apache.spark.rdd.RDD[String] = MapPartitionsRDD[3] at flatMap at <cons
ole>:29
```

Figure 26. Screenshot of code for making split result as input.

The second input is mapped in such a way that every line is taken as a word and the output will give a count as numbers . Here the word is considered as key and the numbers are considered as value:

```
scala> inputSecond.map(word => (word,1))
```

*Figure 27*. Screenshot of code for Map Job.

**Step 8:** Now result from the above code is considered as a Third input as shown in figure 28.

```
scala> val inputThird = inputSecond.map(word => (word,1))
```

*Figure 28.* Screenshot of code making Map result as input.

**Step 9:** Now the reduce function is applied on the Third input where the input are sorted by key as shown in figure 29. The repeated key will be reduced to one key and value will be added for every repeated key.



```
scala> inputThird.reduceByKey( _ + _ )
```

*Figure 29.* Screenshot of code for Reduce Job.

The result from the above code can be print using the following code.



```
scala> inputThird.reduceByKey( _ + _ ).collect.foreach(println)
```

*Figure 30.* Screenshot of code for printing the results.

In this way, a simple wordcount program can be executed from the Spark Shell program in Cloudera which is a basic program that shows the process of Map and Reduce job in Hadoop.

**Summary**

Big data are by nature is challenging to the traditional method of handling data. Their volume, variety, velocity properties are making traditional analyzing method act slow and unable to process it. Hadoop has all those properties that can address the issues of big data.

Hadoop HDFS project allows user to store lots of data and other Hadoop projects such as

MapReduce, Hive allows to process the data very fast. This chapter analyzes the method that

were used in the past and explains how data can be stored and process using HDFS and

MapReduce components of Hadoop. In the next chapter, the results of the research paper will

be briefly discussed.

**Chapter 5: Results, Conclusion, and Recommendations**

**Introduction**

The outcome of this study is very specific and addresses the problem statement. Using literature review the issues of the bid data were noted, and through methodology the probable solution for the big data challenges were discuss. Using the past work done by the scholars, it is noticed that Hadoop can be one of the alternative solutions for the big data challenges.

**Results**

When the study was started, there was only known fact that big data has become the challenges to store and process while using the traditional methods of handling data. But the real-time example that includes wordcount project during this study has helped how easily Hadoop framework can solve the challenges of big data. Some the important results obtained from this research study are as follows:

A) Handling big data is challenging using the traditional method of handling data: Due to various nature of data, it becomes more difficult to store and process data for companies who rely on data analytics. Traditional methods like RDBMS was primarily used for decades to store and process the data until the data started changing to big data. The volume, variety, and velocity characteristics are data are becoming harder and harder to maintain. Performance wise, cost feasible wise companies are not able to store and process the large amount of data using traditional methods of handling data.

B) Hadoop has characteristics of handling big data, that is challenges for traditional methods of handling big data. Recently, big companies are using Hadoop project for storing and processing large amount of dataset. Using Hadoop software, users can easily scalable the

storage capacity just by adding slave nodes to the server. The hardware required for adding

storage capacity is very low in cost which enables to store lots of extra data. Its large block

size enables users to store large amount of data. Also, the parallel computing properties runs

Hadoop project vary fast. So, most of the issues of traditional method of handling data are

addressed by Hadoop software.

**Conclusion**

In conclusion, big data are very challenging to handle by nature. Using proper

technique can help small and big companies maintain the data and it can used for future goals

like marketing strategy and sales performances. Mishandling big data may suffer company's

huge financial and reputational loss. Lots of data breach are costing companies in the data

breach settlement. Using proper data handling techniques, companies can save data from

hackers and utilize them for companies benefits. Using traditional method of handling data,

one can be able to store and process the data. But the time taken to store, and process may

take longer than ever. Nowadays taking a longer time to analyze data may cost the company.

In banking sector, the data analyze needs to be very fast because the longer time taken to

analyze data may not save the fraudulent transaction. Using traditional method has many

challenges while handling big data. With the speed and volume of data generated, it's almost

impossible for small companies to handle the big data with traditional methods because of the

time involved to store and process data, and the cost related with maintaining the database.

Hadoop can be one of the good choices to solve the issues that traditional is unable to handle.

Hadoop being open source, easy to maintain, cost efficiency makes likeable among the data

scientists, small companies, and big companies. So, Hadoop is one the big data handling

techniques that can replace the traditional methods handling big data.

**Future Work**

   Hadoop is an open source which can be easily installed and used. But being an open

source there is no 24 hours software support systems for companies that uses Hadoop for free.

Currently, some companies have started software service at their own, charging a small

amount who wish to use the software support services. The scope of commercial use of

Hadoop needs to be explored in detail so that people or companies who wish to use Hadoop

software can learn more about using Hadoop commercially. The future works should focus on

how Hadoop is commercially blooming, which companies provides full software supports,

and what are the standards to become a Hadoop software service provider. The results

obtained from the future work can be very useful for researchers, students, and data scientists

who wish to enter to the Hadoop world.

## References

Agrawal, D., Bernstein, P., Bertino, E., Davidson, S., Dayal, U., Franklin, M., . . . . Widom, J. (2012). *Challenges and opportunities with big data*. A white paper prepared for the Computing Community Consortium committee of the Computing Research Association. Retrieved from http://cra.org/ccc/resources/ccc-led-whitepapers/.

Asthana, A., & Chari, S. (2015). *Cost-benefit analysis: comparing the IBM PureData System with Hadoop Implementations for Structured Analytics*. Retrieved October 31, 2017, from http://www.cabotpartners.com/Downloads/TCO-Study-Pure-Data-versus-Hadoop-May-2015.pdf.

Bremmer, I. (2015). What does America stand for? *Time*, *185*(20), 26-31.

Cheng, X., Fang, L., Yang, L., & Hong, X. (2017). Exploiting mobile big data: Sources, features, and applications. *IEEE Network*, *31*(1), 72-79. doi:10.1109/ MNET.2017.1500295NM

Dean, J., & Ghemawat, S. (n.d.). *MapReduce: Simplified data processing on large clusters*. Retrieved April 12, 2017, from https://static.googleusercontent.com/media/ research.google.com/en//archive/mapreduce-osdi04.pdf.

DeRoos, D., Zikopoulos, P. C., Melnyk, R. B., Brown. B, & Coss, R. (2014). *Hadoop for Dummies*. John Wiley & Sons.

Dijcks, J. (2013). Oracle: Big data for the enterprises. Oracle Corporation. Retrieved March 11, 2017, from http://www.oracle.com/us/products/database/big-data-for-enterprise-519135.pdf.

Friedman, U. (2012). Anthropology of an idea: Big data. *Foreign Policy*, (196), 30.

Guo, Y., Rao, J., Cheng, D., & Zhou, X. (2017). IShuffle: Improving Hadoop performance with shuffle-on-write. *IEEE Transactions on Parallel And Distributed Systems*, *28*(6), 1649-1662. doi:10.1109/TPDS.2016.2587645

Hashem, I. A. T., Chang, V., Anuar, N. B., Adewole, K., Yaqoob, I., Gani, A., et al. (2016). The role of big data in smart city. *International Journal of Information Management, 36*(5), 748-758.

Heller, P., Piziak, D., Stackowiak, R., Licht, A., Luckenbach, T., Cuathen, B., … Knudsen, J. (2016). An enterprise architect's guide to big data. Oracle Corporation. Retrieved March 11, 2017, from http://www.oracle.com/technetwork/topics/entarch/articles/oea-big-data-guide-1522052.pdf.

Hu, H., Wen Y., Chua, T., & Li, X. (2014). Toward scalable systems for big data analytics: A technology tutorial. *IEEE Access, Access, IEEE*, p. 652. doi:10.1109/ACCESS.2014.2332453

Julio, M., Manuel A. S., & Eduardo, F. (2016). Main issues in big data security. *Future Internet, 8*(3), 44. doi:10.3390/fi8030044

Kale, N., & Jones, N. (2016). Practical analytics (1st ed.). Epistemy Press.

Kerzner, M., & Maniyam, S. (2013). *Hadoop illuminated*. Hadoop Illuminated, LLC. Retrieved on November 12, 2017 from http://hadoopilluminated.com/hadoop_illuminated/hadoop-illuminated.pdf.

Kevin, D. (2012). From punched cards to "big data": A social history of database populism. *Communication +1, 1,* 1-33. doi:10.7275/R5B8562P

Khan, N., et al. (2014). Big data: Survey, technologies, opportunities, and challenges. *The Scientific World Journal,* 18.

Lafuente, G. (2015). The big data security challenge. *Network Security*, *2015*(1), 12-14. doi:10.1016/S1353-4858(15)70009-7

Marathe, S. (2017). *An introduction to Hadoop*. Retrieved November 23, 2017, from https://www.mindtory.com/an-introduction-to-hadoop/.

Natarajan, R. (2012). *Apache Hadoop fundamental–HDFS and MapReduce explained with a diagram*. Retrieved on April 24, 2017 from http://www.thegeekstuff.com/2012/01/ hadoop-hdfs-mapreduce-intro/.

Ozcan, S. (2013). *Difference between Hadoop and RDBMS*. Retrieved April 24, 2017 from http://oraclesys.com/2013/04/03/difference-between-hadoop-and-rdbms/.

Parmar, R., Roy, S., Bhattacharyya, D., Bandyopadhyay, S., & Kim, T. (2017). Large-scale encryption in the Hadoop environment: Challenges and solutions. *IEEE Access*, *5,* 7156-7163. doi:10.1109/ACCESS.2017.2700228

*Powered by Apache Hadoop.* (n.d.). Retrieved April 30, 2017, Retrieved from https://wiki.apache.org/hadoop/PoweredBy.

Schwieger, D., & Ladwig, C. (2016). Protecting privacy in big data: A layered approach for curriculum integration. *Information Systems Education Journal*, *14*(3), 45-54.

Sharma, S. (n.d.). *Hadoop-2.6.X on Windows 10*. Department of Computer Science, Ben-Gurion University. Retrieved November 27, 2017, from http://www.ics.uci.edu/~shantas/Install_Hadoop-2.6.0_on_Windows10.pdf

Sharma, V., & Joshi, N. K. (2015). The evolution of big data security through Hadoop

    incremental security model. *International Journal of Innovative Research in Science,*

    *Engineering, and Technology*, 3297:2007. Retrieved November 11, 2017, from

    https://www.ijirset.com/upload/2015/may/97_30_The.pdf.

Tankard, C. (2017). Encryption as the cornerstone of big data security. *Network Security*, (3),

    5-7. doi:10.1016/S1353-4858(17)30025-9

Ulusoy, H., Colombo, P., Ferrari, E. Kantarcioglu, M., E. & Guard, M. R. (2015). Fine-

    grained security policy enforcement for MapReduce systems. In *Proceedings of the*

    *10th ACM Symposium on Information, computer and Communication Security*,

    Singapore, April 14-17, 2015, pp. 285-296.

Urquhart, K. (1997). Java's open future. *IEEE Micro, Micro, IEEE*, (3), 10.

    doi:10.1109/MM.1997.591649

Vishwakarma, D., & Madhavan, C. (2014). Efficient dictionary for salted password analysis.

    In *IEEE CONECCT 2014-2014 IEEE International Conference on Electronics,*

    *Computing and Communication Technologies.* doi:10.1109/CONECCT.2014.6740293

Welcome to Apache™ Hadoop®! (2014). Retrieved October 31, 2017, from

    http://hadoop.apache.org/.

Cisco VNI Forecast and Methodology, 2015-2020. (2016, August 11). Retrieved March 11,

    2017, from http://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-

    networking-index-vni/complete-white-paper-c11-481360.html.

White, T. (2009). *Hadoop: The definitive guide*. Beijing; Sebastopol, CA: O'Reilly.

Yaqoob, I., Hashem, I. T., Gani, A., Mokhtar, S., Ahmed, E., Anuar, N. B., & Vasilakos, A.

    V. (2016). Big data: From beginning to future. *International Journal of Information*

    *Management*, *36,* 1231-1247. doi:10.1016/j.ijinfomgt.2016.07.009