# BIG DATA: THE ARCHITECTURE, STORAGE AND ANALYSIS

BY

# MAKGOBA BOITUMELO (2020)

A RESEARCH REPORT SUBMITTED FOR THE HONOURS OF SCIENCE DEGREE IN THE DEPARTMENT OF COMPUTER SCIENCE, SCHOOL OF MATHEMATICAL AND COMPUTER SCIENCES, FACULTY OF SCIENCE AND AGRICULTURE, UNIVERSTIY OF LIMPOPO, SOUTH AFRICA

# SUPERVISOR:PROFESSOR S MOKWENA

# SEPTEMBER 2020

# DEDICATION

I would like to dedicate this dissertation to my family, my twin brother Tumelo Makgoba who has been nothing but supportive and my mother Rebeccah Makgoba who has been my pillar of strength.

# DECLARATION

I Boitumelo Makgoba hereby declare that the research project submitted for Honours Science (Computer Science) in the Faculty of Science and Agriculture at the University of Limpopo is my own work and has not been submitted to any institution of higher education. All the sources used for this research have been duly cited and referenced.

Full Names Date

Student number

# ACKNOWLEGDEMENTS

This work is funded by the National Research Counsil (NRF) at the University of Limpopo, School of Mathematical and Computer Science, Department of Computer Science under the supervision of Prof S Mokwena

# ABSTRACT

The evolution of big data technology has certainly raised a lot of eyebrows particularly in the Information Technology (IT) industry and academia. The big data technology has grown so tremendously that it is at the forefront of the technology industry. Huge amounts of data are generated every second through platforms such as social media (i.e. Facebook, twitter, Instagram, LinkedIn etc.) and the Internet. According to Statista (the statistics portal that allows access to market and consumer data as well as from research institutions for better decision making), the number of worldwide active social media users is expected to reach an estimate of 3.02 billion monthly active social media users by 2021. This figure represents around a third of earth’s entire population. The amount of data generated daily is so huge that personal computers and workstations cannot handle it.

This research work studies and investigates how the large amounts of data can be efficiently stored, analysed and what tools are most appropriate to manage such data. The Apache Hadoop framework is a powerful tool which will be used to investigate effective handling of the large datasets. The Apache Hadoop framework is a an open source tool that we adopted and used to virtualize how the large datasets are managed. Within the Apache Hadoop ecosystem, the storage and the analytics components are investigated. We explore how the Apache Hadoop supports various forms of data (structured, semi-structured and unstructured). We further investigate big data analytics methodologies and techniques with the aim of identifying the most optimal and efficient tools given different input data types. The Cloudera platform also has been used to benchmark against the virtualized local Hadoop performance. The experimental results conducted in this research demonstrate the great capability and potential that Hadoop has when dealing with datasets over a gigabyte in size.

*Keywords: Big Data, Hadoop, MapReduce, HDFS, Big Data Analytics*

# TABLE OF CONTENT

[DEDICATION i](#_bookmark0)

[DECLARATION ii](#_bookmark1)

[ACKNOWLEGDEMENTS iii](#_bookmark2)

[ABSTRACT iv](#_bookmark3)

[LIST OF TABLES vii](#_bookmark4)

[LIST OF FIGURES viii](#_bookmark5)

[LIST OF ABBREVIATIONS x](#_bookmark6)

* 1. [Background 1](#_bookmark7)
     1. [Volume 1](#_bookmark8)
     2. [Velocity 2](#_bookmark9)
     3. [Variety 2](#_bookmark10)
     4. [Veracity 2](#_bookmark11)
  2. [Problem Statement 3](#_bookmark13)
  3. [Aim and Objectives 3](#_bookmark14)
  4. [Research Questions 3](#_bookmark15)
  5. [Purpose of study 4](#_bookmark16)

[1.7 Conclusion 4](#_bookmark17)

[CHAPTER TWO: LITERATURE REVIEW 5](#_bookmark18)

* 1. [Hadoop Architecture 5](#_bookmark19)
  2. [MapReduce 6](#_bookmark21)
  3. [File System (HDFS) 7](#_bookmark23)

[2.7 Conclusion 14](#_bookmark26)

[CHAPTER THREE: METHODOLOGY 15](#_bookmark27)

* 1. [Tools used 15](#_bookmark28)
  2. [The Advantages of MySQL 18](#_bookmark34)
  3. [Supplementary Tools 21](#_bookmark36)
  4. [The datasets 22](#_bookmark38)
  5. [Conclusion 22](#_bookmark39)

[CHAPTER FOUR: RESULTS AND ANALYSIS 23](#_bookmark40)

* 1. [Experiments 23](#_bookmark41)
     1. [MySQL data loading performance 23](#_bookmark42)
     2. [Hadoop cluster data loading performance 24](#_bookmark45)
     3. [Cloudera performance 27](#_bookmark51)
     4. [R studio data load performance 28](#_bookmark54)
  2. [Analysis 29](#_bookmark55)

[CHAPTER FIVE: CONCLUSION 34](#_bookmark61)

[REFERENCES 35](#_bookmark62)

# LIST OF TABLES

[**Table 3.1 The Hadoop configuration and setup 20**](#_bookmark35)

[**Table 3.2 Cloudera setup 22**](#_bookmark37)

[**Table 4. 1: Summarized performance of MySQL, Hadoop and R studio 29**](#_bookmark56)

# LIST OF FIGURES

# [Figure 1. 1: The 4Vs of big data 2](#_bookmark12)

# [Figure 2. 1: The main components of the Hadoop Architecture 6](#_bookmark20)

# [Figure 2. 2: The architecture of MapReduce 7](#_bookmark22)

# [Figure 2. 3: The HDFS Architecture 8](#_bookmark24)

# [Figure 2. 4: Hadoop ecosystem 9](#_bookmark25)

# [Figure 3. 1: A successfully installed VMware Workstation 14 Player installed on](#_bookmark29) [Windows 10 16](#_bookmark29)

# [Figure 3. 2: A successfully installed Ubuntu 18.04 inside the VMware](#_bookmark30) [Workstation 14 Player 16](#_bookmark30)

# [Figure 3. 3: The Hadoop-2.6.7 successfully installed 17](#_bookmark31)

# [Figure 3. 4: Hadoop cluster applications status 17](#_bookmark32)

# [Figure 3. 5: Sqoop-1.4.7 successfully installed 18](#_bookmark33)

# [Figure 4. 1: Importing dataset into MySQL database 23](#_bookmark43)

# [Figure 4. 2: Dataset successfully loaded in the MySQL database 24](#_bookmark44)

# [Figure 4. 3 : Commands to launch Hadoop 24](#_bookmark46)

# [Figure 4. 4: The running processes in the Hadoop cluster 25](#_bookmark47)

# [Figure 4. 5 : Command to Import data into HDFS using Sqoop 25](#_bookmark48)

# [Figure 4. 6: Results of data loading into Hadoop through sqoop 26](#_bookmark49)

# [Figure 4. 7: Errors resulted by loading data into Hadoop through sqoop 26](#_bookmark50)

# [Figure 4. 8: The graph showing the performance of the Impala 27](#_bookmark52)

# [Figure 4. 9: A graph showing the performance of the HDFS 28](#_bookmark53)

# [Figure 4. 10: Summarized dataset 30](#_bookmark57)

# [Figure 4. 11: Summarized dataset 30](#_bookmark58)

# [Figure 4. 12: A graph showing the relationship between the customer service](#_bookmark59) [and how they are rated. 32](#_bookmark59)

# [Figure 4. 13: The boxplot of average stars 33](#_bookmark60)

# LIST OF ABBREVIATIONS

AI : Artificial Intelligence 1

CPU: Central Processing Unit 21

CSV : Comma Separated Values 2

DBMS : Database Management System 19

EDW : Enterprise-Data-Warehouse 12

HDFS :Highly Distributed File System 6

HTML : Hypertext Markup Language 2

IT : Information Technology iv

RDBMS : Relational Database Management System 1

UDF : User Defined Function 10

XML : Extensible Markup Language 2

# CHAPTER ONE: INTRODUCTION

# Background

The transformation of big data technology most certainly came a long way with a bang, from the earliest examples of humans storing and analysing data called tally sticks to the rise of artificial intelligence (AI) where people are using mobile devices to access digital data rather than the use of an office or home computers. Big data is mostly referred to as a new word since it was only discovered a few years ago. It is often seldom to hear one talking of vast amount of data, but currently this is the challenge faced. The growth of data is constantly increasing every second i.e. over Zettabytes of data is being generated from various sources, which is approximately equivalent to 1 trillion of gigabytes. With such enormous amount of data being generated every single day from various sources, improved analysis is of importance in order to extract the right information which is appropriate for the goal that needs to be achieved.

Data size has increased exponentially especially given today’s technology in multiple sectors such as manufacturing, business science and web application. Data can be in different formats such as structured, semi-structured while others are structured, and combine them with different data types such as documents, records, pictures and videos. The size and structure of this particular data have become incredibly dynamic (Aminu, 2014). Big data organizes and extracts the most important information from a very fast growing, huge volumes, and different forms. The changes are often data sets collected from a lot of and autonomous sources in the minimal possible time by the use of multiple statistical and machine learning techniques (Kune et.al, 2011).

Big data is large data sets which are unable to be processed by the traditional relational database management systems (RDBMS) or any other traditional systems such as laptops and computers (Venkateshalet.al, 2015). With that being said big data can be further defined by its four characteristics which are well known as the 4Vs. These are known as Volume, Velocity, Variety and Veracity.

# Volume

Volume is defined as how large data are in big data we speak of data larger than terabytes and petabytes. The large size of data makes it difficult to store and analyse using traditional tools (Jaseena et.al, 2014).

# Velocity

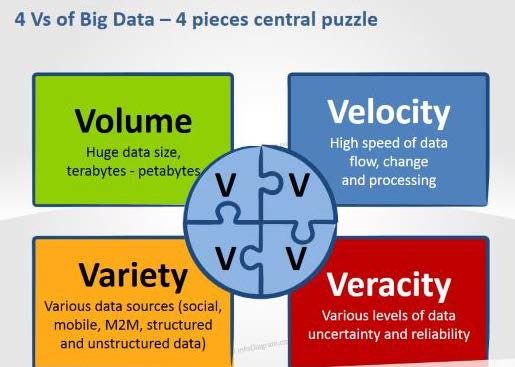
This involves the speed of the data coming from various sources. This means how fast the data can travel using big data technologies since with traditional tools the data will be at a slow pace (Lawal, 2014).

# Variety

Data being generated is not of the same form. All this data is entirely dissimilar consisting of raw, structured (CSV files), semi-structured (html and xml files) and even unstructured (text files) data which is hard to be handled by the easily acquired traditional analytic systems (Jaseena et.al, 2014).

# Veracity

Veracity involves the trustworthiness and reliability of the data set. Accuracy or truthfulness is required to produce analysis that allows taking informed decisions. Due to the high value of the 4Vs it makes it difficult to store, analyse and process the data. The Figure 1.1 below depicts the 4Vs of big data.



# Figure 1. 1: The 4Vs of big data

It is believed that big data will play an imperative role in the future in all walks of life and the societies. For instance, the government has started extracting contents of social media and other sources in order to track down suspicious organizations as well as to predict events which may be harmful to the society and be able to take informed

decisions (Che et.al, 2013). Big data may be regarded as a very powerful tool which requires other powerful tools such as the Hadoop framework to be able to store and analyse data in order to come up with better solutions as well as make much informed decisions.

Technology has played an imperative role in the increase of the growth of digital data. It has also made communication very easy. The penetration of technology in different sectors, there has been a powerful forward of data available in digital form (Tulasi et.al, 2015).

# Problem Statement

Data has grown tremendously over the years so much that it comes with complexities. Complexities such as where can we efficiently store the data and what appropriate tools can be used to even further analyse it.

# Aim and Objectives

The aim of this research work is to study and investigate how large amounts of data from sources can be efficiently stored, processed and analysed. We evaluate the big data tools for efficiency in terms of performance. The objectives of the study are broken down into the following:

* Study and understand the Hadoop framework and how it works.
* Measure the performance of the Hadoop framework when loading and processing the data.
* Perform the analytics from a given dataset to discover the meaning of data as well as other insights into the dataset.

# Research Questions

* How can we comprehend the Hadoop framework architecture and all of its capabilities?
* How can the performance of Hadoop be measured as compared to the other tools?
* How can the insights of the big datasets be discovered?

# Purpose of study

Big data is generated every second from various sources such as the social media. With this form of data, it can be used to combat unfortunate events such as crime through analysis. The purpose of this study is to figure out the tools of big data in order to perform analysis for better solutions.

# Significance of the study

The significance of this study is to thoroughly explain the methods, processes and procedures of how we can store big data and finally analyse it using the Hadoop framework with the hope of achieving good results.

# Conclusion

Big data has been introduced as a large dataset with complexities which require specific tools to deal with such a challenge. The big data technological tools have been made freely available in order to be able to modify the algorithms for improving the tools which results in better performance, better predictions and better solutions. Data analytics is seen as the driving force in changing the working of organization in different sectors. In this chapter, we outlined the inherent problems and challenges that big data pose. The aim and objectives of the research investigation, research questions as well as the significance of the study.

# CHAPTER TWO: LITERATURE REVIEW

The growth of big data is taking place so rapidly that challenges are encountered. Big data is processed from various sources such as social media or any other organization which are then stored. The stored data are spent on much cost while they are neglected or deleted due to inadequate space (Acharjya, 2016). With such peculiar characteristics, big data is known as data that does not conform to the traditional relational database (Zakir, 2015), therefore storage becomes a challenge. Furthermore, storage is not the only problem we encounter, given the fact that with such vast amount of data we need to do analysis to make informed decisions.

These huge amounts of data are generated within seconds, analysis becomes a problem without the appropriate tools. The technology of big data is presently taking a huge turn with the known databases. The change of efficient data arrangement, massive computations and the process of data workloads with advanced analysis tools. This includes the techniques of statistics and machine learning (Kune, 2015).

# Hadoop Architecture

According to (Jaseena, 2014), one of the upcoming technological tools i.e. Hadoop framework and MapReduce provides fresh and innovative methods to process and change big data. Big data are large datasets that are complicated and are usually unstructured. The Hadoop then comes into play to change it into information that is clear and understandable.

Hadoop is known as an open source that is scalable, and fault tolerant Virtual Grid operating system architecture, which stores data and process it (Acharjya, 2016). The Apache which manages the Hadoop defines it as a software library framework which performs distribution processing of large amounts of data by the use of clusters of computers that use models of programming. It is made up in such a way that it is able to measure from single servers to thousands of machines. Each provides local computation and storage. Instead of depending on hardware to provide high availability, the library is made in such a way that it tracks and deals with failures at the application layer, so provision of highly-available service at the top of the clusters

are likely to encounter failures. It consists of two main components namely MapReduce and Highly Distributed File System (HDFS) as shown in Figure 2.1.

HDFS

MapReduce

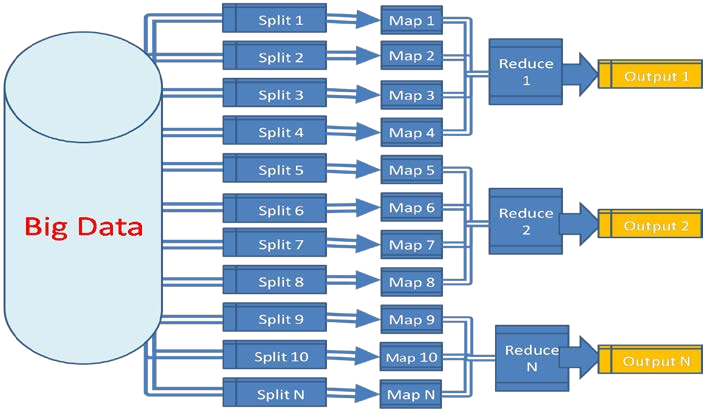
Hadoop

# Figure 2. 1: The main components of the Hadoop Architecture

# MapReduce

MapReduce is a programming model which was initiated by Google to execute the process of parallelism on vast amounts of datasets in a reliable and fault tolerant way (Venkatesh, 2015). Firstly the application is broken down into single blocks to process the individual map jobs in a parallel way (Venkatesh, 2015). Secondly, the master node breaks down the input into small sub problems and further distributes them to worker nodes in the next map step. Then the master node mix the outputs for all the sub problems in reduce step (Acharjya, 2016)

The MapReduce consist of two functions, i.e. map() and reduce(). The mapper is responsible for filtering and sorting or organizing the data and the reducer simply summarizes the results (Jaseena et.al, 2014). Users are able to achieve their own processing logic by creating their own customized map() and reduce() function. The map() function yields a list of intermediate value pairs by simply taking the input value pairs (Jaseena et.al, 2014). The MapReduce runtime system gathers the intermediate pairs available, put them together according to the intermediate keys and passes them to reduce() function for yielding the end results. It introduces the solution to problems like large scale data processing. Problem of handling a lot of processors that are up to a thousand and handling parallel and distributed environment is even more challenging. The Figure 2.2 below depicts how the MapReduce works.



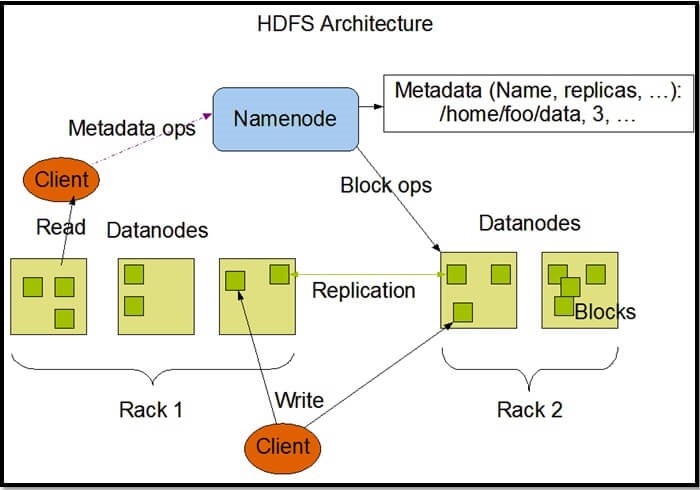
# Figure 2. 2: The architecture of MapReduce

# File System (HDFS)

The HDFS is a primary storage system used by the projects within the hadoop. It makes use of the Namenode and Datanode architecture to perform a distributed file system that offers high performance access to data across highly scalable Hadoop clusters (Hammad et.al, 2015). The HDFS has the following characteristics (Chakravorthy, 2014):

* + - Uses huge files
    - Perform huge block sequential reads for analytics processing
    - Access huge files using sequential I/O in append mode

The Figure 2.3 depicts the architecture of the HDFS together with its main components.



# Figure 2. 3: The HDFS Architecture

The HDFS cluster entails a single Namenode, it is a master server machine which controls the file system and maintains access to the file system with the help of the clients, (Raste, 2014). The secondary Namenode does the work of regulating the validity of the Namenode and keeping the Namenode information up to date from time to time (Hammad, 2015).

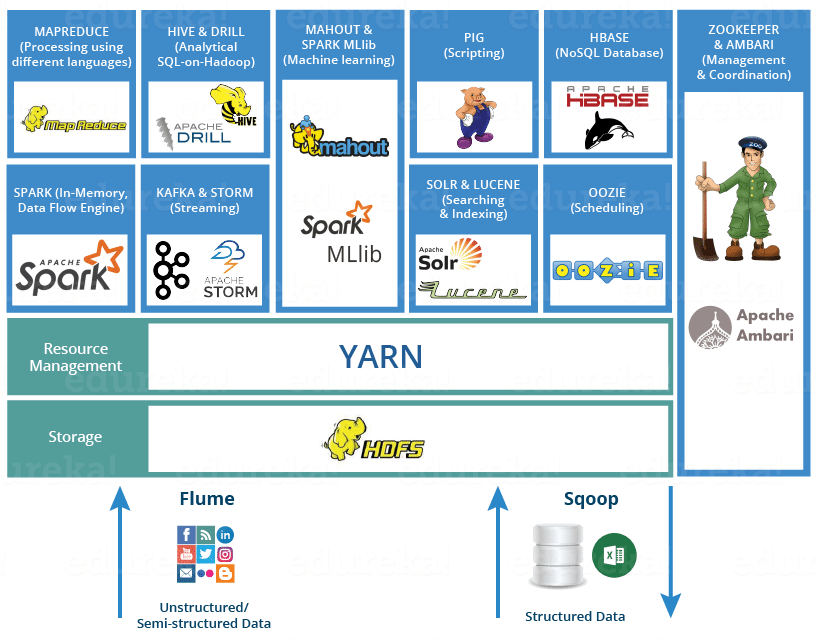
Multiple data nodes are present in each cluster. As depicted in Figure 2.3 the data is partitioned into blocks and finally stored in the data nodes in a distributed manner. The DataNode takes the responsibilities of data read and write operations, through the process of performing data analysis. The Rack awareness concept is followed in Hadoop, basically it means in order to improve the network traffic while reading/writing HDFS file, the Namenode picks the Datanode that is nearby the rack to Read/Write request. Namenode acquires rack information by regulating the rack’s ids of each Datanode. This is called Rack Awareness in Hadoop.

In big data Hadoop rack awareness is required due to the following reasons:

* + - To achieve data high availability and reliability.
    - Achieve the performance of the cluster.
    - To keep bulk data in the rack when possible
    - Avoid data loss in case an entire rack fails, although it is highly unlikely as compared to node failure
    - To keep bulk data in the rack when possible
    - An assumption that in-rack id’s higher bandwidth, lower latency

# Hadoop Ecosystem

The Hadoop consist of Hbase, Hcatalog, Pig, Hive, Oozie, Zookeeper, and Kafka which is known as Hadoop ecosystem as shown in Figure 2.4.However, the most commonly known paradigm are the HDFS and MapReduce for Big Data (Khan,2014).



# Figure 2. 4: Hadoop ecosystem

The Figure 2.4 depicts the Hadoop ecosystem components explained as follows (Khan et.al, 2014)

* + - HDFS: it is a storage medium implemented when the amount of data exceeds what a single machine can handle. It acts as storage for the data that is coming in.
    - Hbase: It is a system that manages the performance of the data. It is an open source that is column base, which allows acceleration of performance of operations over the same values across huge data sets.
    - Zookeeper: it sustains designs and labels huge volume of data. Distributed synchronization is provided as well. This situation allows distributed processes to manage and contribute one another through a name space of data registers that is shared and hierarchical, such as a file system. Separately, Zookeeper is

a distributed service that contains master and slave nodes and stores configuration information.

* + - Hcatalog: It regulates the HDFS. It stores metadata and creates tables for huge volume of data
    - Hive: Hive structures warehouses in HDFS and other input sources, such as Amazon S3. It is a subplatform in the Hadoop ecosystem and produces its own query language (HiveQL). The language is compiled by MapReduce and enables user-defined functions (UDFs).
    - Pig: it produces a high-level scripting language and achieves a run-time platform that allows users to execute MapReduce on Hadoop
    - Mahout: Mahout is a library for machine learning and data mining. It is broken down into four main groups’ i.e. collective filtering, categorization, clustering, and mining of parallel frequent patterns.
    - Oozie: Oozie coordinates, executes, and manages job flow.

Although Hadoop is known as a very powerful tool to process large amounts of data, it comes along with advantages as well as disadvantages. In the study of Big Data Computing and Storage Tools authors in (Prasad, 2016) have identified the advantages of big data as follows:

* + - It is an open source which means that is freely accessible to anyone
    - Cost effective: the hadoop is very much affordable
    - Scalable: it has the ability to handle vast amounts of data.
    - Fault tolerant: in an event of a failure in a system, the system continues to operate properly.
    - High throughput: high throughput is achieved due to batch processing
    - Portability: Hadoop architecture can be effectively ported while working with multiple commodities of operating system and hardware that maybe heterogeneous.
    - On-Demand-Service: it can be set manually on lent computing nodes on cloud.

The authors in (Prasad, 2016) further identified the disadvantages of big data as follows:

* + - Single point failure: Hadoop suffer from a single point failure. This means that if the NameNode fails then Hadoop cluster will become out of the way. In case, if NameNode failure occurs then it requires manual intervention of Hadoop Administrator to recover the NameNode with assistance from the secondary NameNode.
    - Low efficiency than the Database Management System. Hadoop has proven to have low efficiency due to its inability to switch to the next stage prior to completion of the previous stage.
    - Inefficient in dealing with small files: HDFS is designed for high throughput optimization, it is not suitable to read small files.
    - Not suitable for real time access: MapReduce and HDFS applies batch processing architecture hence, it does not fit for real-time accesses.
    - Hadoop does not support iterative behaviour.

Hadoop is the most powerful tool which enables us to process and analyse huge amounts of data. Not only is convenient but it enables accurate and informed decisions. Informed decisions are imperative because they make us avoid hasty decisions and implement accurate predictions.

There are various big data analytics techniques. According to (Kune, 2015) data analytics is defined as the process of scrutiny that data undergoes by the use of statistical models, computing technologies and data mining techniques. Since big data is not only huge, but also different and grows rapidly, multiple technologies and analytical techniques are required in order to try to extract relevant information (Maltby, 2011).

# The areas and techniques of big data

The area techniques of big data can be defined as methods or different ways in which analysis can be performed. Data is generated and extracted from various sources. The different areas and techniques of big data analysis can be defined as follows (Kune et. Al, 2011):

# Text analysis

The procedure of extracting information from the text sources such as semi-structured data which includes social media postings, blogs and web materials.

# In Memory analytics

The process of absorbing vast volume of data from various sources directly into the memory of the system for efficient query and calculation performance.

# Predictive analytics

The process of predicting future events using historical data, statistics, modelling, machine learning and data mining.

# Graph analytics

It is the study of the pattern analysis of different connected components.

# Enterprise Data Warehouse (EDW)

Enterprise data warehouse are databases used in data analysis. It is considered as a core component of business intelligence.

# Visualization products

A few challenges with big data are finding a way to visually represent results. Visualization can assist in the information search. It allows comparing different dataset thoroughly (Maltby, 2011).

# NoSQL databases

These types’ databases are structured in a way that it is capable to handle huge amounts of information that do not utilize a relational model. They are scalable and are frequently used for tracking and analysing real-time list which grow rapidly.

# Association rule learning

It is a rule-based method for unpacking intriguing relations between variables in large databases. It is often used in data mining.

# Data Mining

Data mining is the process of extracting information from sources such as social media by the use of statistical and machine learning model to directly point patterns in large datasets.

# Cluster analysis

Cluster analysis is a data mining technique that breaks down a large group into small groups of similar objects. The characteristics of the datasets are much more similar but it is not known in advance.

# Crowdsourcing

It is the collection of data from a large group of people through an open call.

# Machine learning

Machine learning is a computer science field that uses statistical techniques to provide a system with the ability to “learn” from data without human intervention.

# Levels of data analytics

Analysis of huge dataset requires more computational complexities. The main challenge is to handle inconsistencies and uncertainty available in the dataset. Tools that are available for big data analysis have poor performance in dealing with computational complexities, uncertainty and inconsistencies (Acharjya et.al, 2016). Analytics can be divided into three various levels that is descriptive, predictive and prescriptive.

# Descriptive

Descriptive analytics provides details of data distribution and visualization. It looks at the success and failure performance by looking at the historic performance (Assunção et.al, 2015).

# Predictive

Predictive analytics is focused on predicting trends and possibilities. It uses historical information (Zakir et.al, 2015) such as the mortality rate in the year 2012, and the use of rules and algorithms.

# Prescriptive

Prescriptive analytics is aimed at providing insights from data that could be used in decision making (Assunção, 2015). It looks at the what, when and why it will happen.

# Conclusion

According to (Zakir, 2015) big data is not just large and complicated, it requires advanced technology to analyse and process. Analysis is a powerful technique which can actually turn particular events around by a simple accurate prediction such as being able to assess a business cycle as to whether the business will go through a recession by merely using a graph. However, prior to analysis, the storage and processing of data would have taken place. The authors in (Maltby, 2011) further explain that big data analytics attempts to take advantage of the excess of information to use it productively. There are several benefits that vary and range from higher quality education to medical research and top secret government activities.

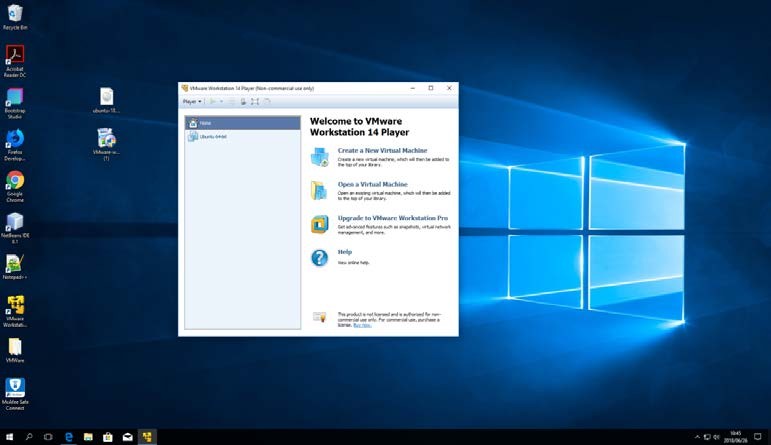
Big data has been introduced as large amounts of datasets which are complex in chapter one. The greatest weapon to big data is the Hadoop framework as described in this chapter. It is the most powerful tool which is able to break down large amounts of data into blocks which can be further analysed by the use of various techniques of analysis.

# CHAPTER THREE: METHODOLOGY

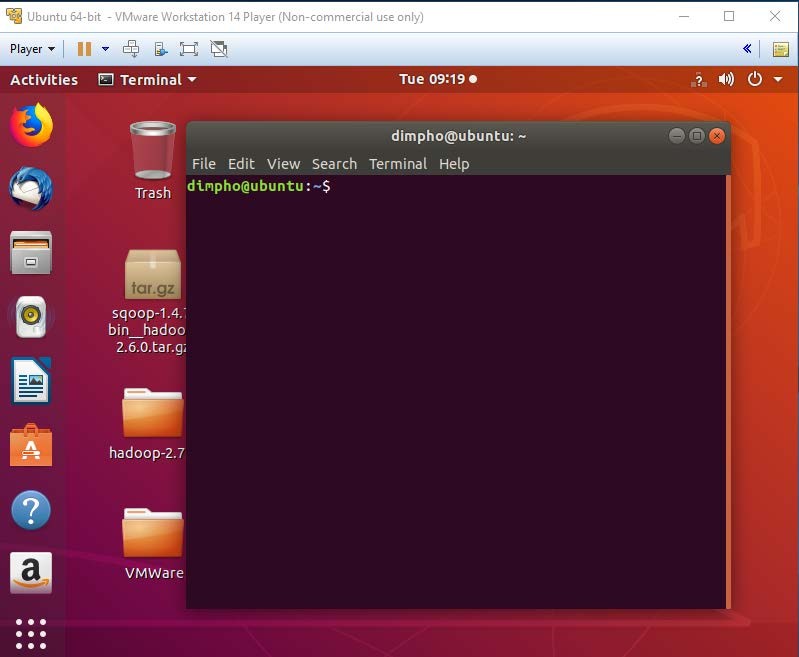
Every big data application is dependent on a set of data and accurate processing of the data is a key for harnessing the potential of such data and its application. This chapter focuses on the implementation of the big data tools and projects. The study seeks to answer such questions as what, how, and why in order to uncover the insights. The research questions clearly state that we are on a quest to unravel how big data can be analysed, how it can be stored efficiently and mostly the importance of the Hadoop framework. The Hadoop framework lies at the heart of our research undertaking although other tools such as the high-performance computing may still be used but they come at a high cost.

# 3.1 Tools used

Before implementing the Hadoop it needs to be layered in an environment where it will run efficiently. In this case VMware virtualization software is employed and is installed in a Windows environment with the Windows 10 as the host operating system. Within the virtual machine (VMware) the Ubuntu 18.04 operating system is installed as a guest operating system, since Hadoop works better on the linux-based operating systems. The VMware Workstation Player allows a second, isolated operating system to run on a single computer ([www.vmware.com](http://www.vmware.com/)) as shown in Figure 3.1. The accessibility of these tools and software are free as they are open source software and no amount of money is required.



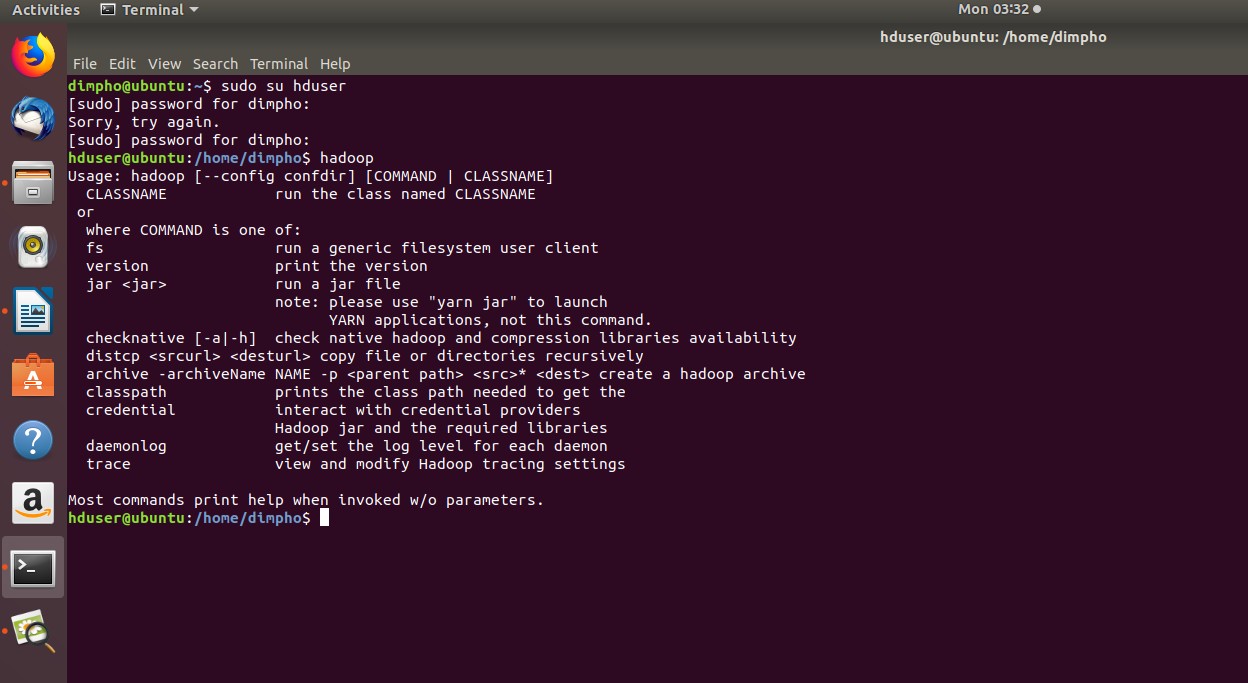
# Figure 3. 1: A successfully installed VMware Workstation 14 Player installed on Windows 10



**Figure 3. 2: A successfully installed Ubuntu 18.04 inside the VMware Workstation 14 Player**

Once the environment has been configured and layered, a java runtime environment needs to be installed as well since Hadoop specifically runs on the java environment. Once the java environment has been configured, the Hadoop may be installed. Figure

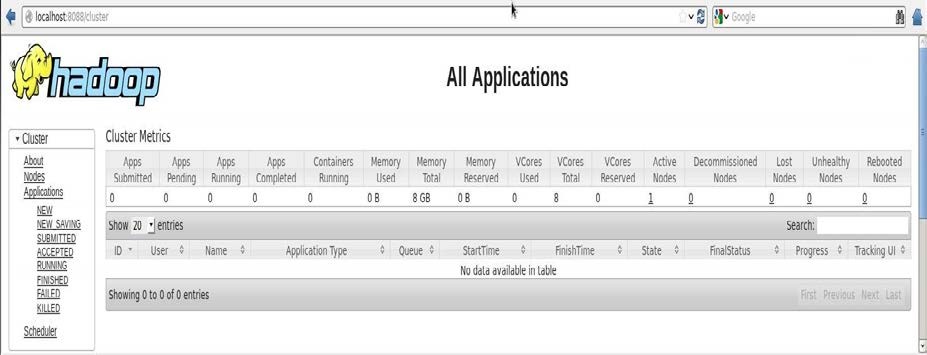
3.3 depicts the commands of a Hadoop once it has been successfully installed. The command “hadoop” is used to check as to whether the Hadoop has been successfully installed.



# Figure 3. 3: The Hadoop-2.6.7 successfully installed

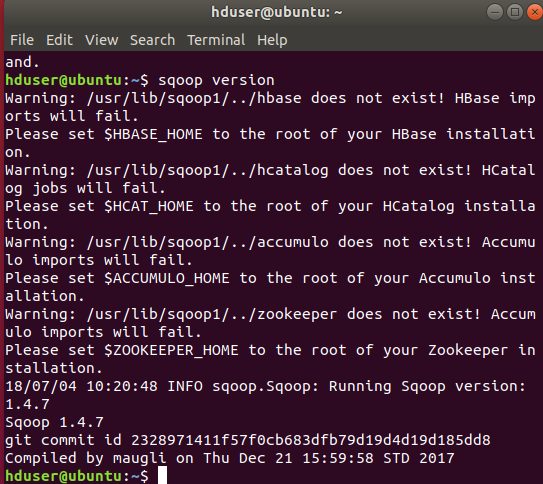
With the Hadoop being installed, several projects (processes) that work inside the Hadoop were also installed. These projects are responsible for the transferring, storage and analysis of data. The project Sqoop which is used to transfer data between relational databases and Hadoop (sqoop.apache.org) has been set up as shown in Figure 3.5. The command “sqoop help” is used to verify the installation and the command “sqoop version” is used to check which version of the sqoop is running as shown in Figure 3.5. Furthermore, MySQL database is installed as well, which is then connected to the sqoop by the utility called JDBC connector to allow transfer of data.

Once the Hadoop is set up you can verify all application clusters from the browser by typing the link https://localhost:8088 which gives the results depicted in Figure 3.4.



# Figure 3. 4: Hadoop cluster applications status

The MySQL connector/j is the official JDBC driver for MySQL which allows connection between the MySQL and the sqoop. This connection allows the dataset in the MySQL to be imported directly into the Hadoop cluster.



# Figure 3. 5: Sqoop-1.4.7 successfully installed

Sqoop imports data from an RBMS such as MySQL or oracle Database into HDFS and then export data back after data has been transformed using MapReduce. In this case MySQL is connected to Sqoop. Once that has been set up the following commands allows a successful load of the data into the Hadoop cluster. Since the experimental dataset size is 1.4GB it is unable to open with the traditional applications such as excel. We use the open source EmEditor application to explore the contents of such big files l. The reason for this is to be able to check to verify the columns and the rows so that they can correspond with those created in the MySQL for a smooth loading.

Once this is done the data is loaded and stored into the HDFS and then analysed using the R language. The MySQL was chosen as the database because it is an open source and freely available database that facilitates effective management of databases. It is a database management system (DBMS) tool that is stable, reliable and provides powerful solutions. The following section provides advantages of MySQL DBMS (https:[//w](http://www.datamation.com/storage/8-major-advantages-of-using-)ww[.datamation.com/storage/8-major-advantages-of-using-](http://www.datamation.com/storage/8-major-advantages-of-using-) mysql.html).

# The Advantages of MySQL

* + 1. **Data Security**

MySQL is known as the most secure and reliable database management system. It is used by popular web applications such as WordPress, Facebook and Twitter.

# On-Demand Scalability

MySQL has the ability to handle huge amounts of data, which makes it reliable. This open source solution allows complete customization to ecommerce business with unique database server requirements.

# High Performance

MySQL is capable of ensuring optimum speed, full text indexes and unique memory caches for enhanced performance.

# Round-the-clock Uptime

MySQL comes with assurance of 24 hours, 7 days uptime and offers a wide range of high availability solutions like specialized cluster and master/slave replication configurations.

# Comprehensive Transactional Support

MySQL is one of the most robust transactional database engines available on the market. It consists of complete atomic, consistent, isolated, durable transaction support, multi-version transaction support, it is the perfect solution for data integrity. MySQL ensures immediate deadlock identification through server-enforced referential integrity.

# Complete Workflow Control

The usability of MySQL is very fast. The average time to download and install is less than 30 minutes regardless of your platform, whether Linux, Microsoft, Macintosh or UNIX.

# Reduced Total Cost of Ownership

MySQL is cost effective. It produces reliable results while allowing you to save as much as you can.

# The flexibility of Open Source

The fact that it is freely available, it makes maintenance, debugging and upgrades fast and easy while improving the end-user experience. The users are to choose solutions that suit their own needs.

Finally, we used the R programming language for analysis and our choice for this analysis tool being:

* + - * It is an open source as well which means it is freely available
      * It has better visualization
      * It works well with big data because you can integrate R with Hadoop
      * It is much easier to understand

The summarized Hadoop experimental processes are as follows:

* + - * Move data with sqoop/terminal into the HDFS
      * Transform HDFS file
      * Extract and load data from the HDFS into the R
      * Write algorithms to perform analytics using R

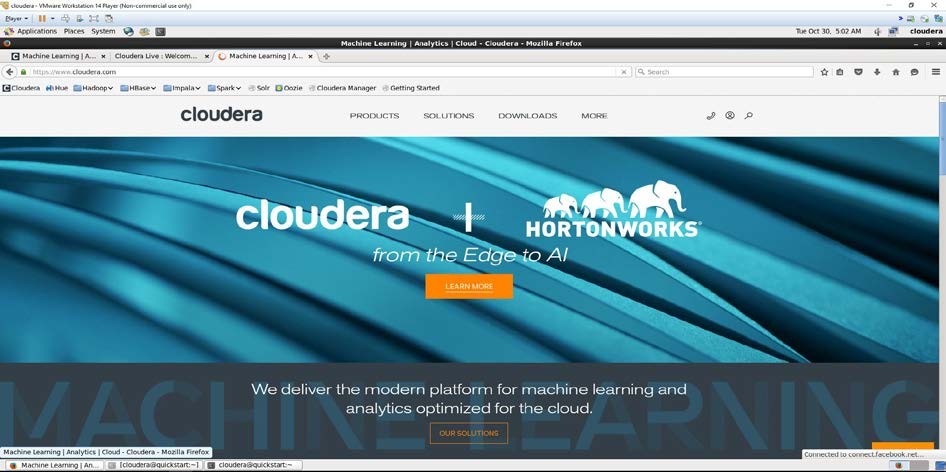
Table 3.1 depicts the summarized Hadoop configuration and setup. The table shows all the tools that were used and their versions.

# Table 3.1 The Hadoop configuration and setup

|  |  |
| --- | --- |
| **Parameter** | **Description** |
| CPU | Intel® Core™ i5-6500 CPU @ 3.20GHz  3.19GHz |
| Virtual Machine  Software | VMware-workstation-full-14.1.2-8497320 |
| OS | Windows 10 |
| Guest OS | Ubuntu 18.04.1 |
| Hadoop | 2.7.6 |
| Sqoop | 1.4.7 |
| R | 3.4.4 |
| MySQL | 5.7.23 |

# Supplementary Tools

Experiments and investigations have a tendency of yielding undesired results. In this study supplementary tools have been introduced should the investigation yield undesired results. The Cloudera has been introduced as a supplemental tool to benchmark our virtualized results with that of the Cloudera cluster for performance comparisons. Cloudera is a software company which provides a platform for big data storage, processing and analysis. The Figure 3.7 depicts the installed Cloudera in the VMware.



# Figure 3. 6: The Cloudera setup and configuration

The Cloudera consist of the Hadoop ecosystem, which means that there is no need for installing Hadoop projects as everything is built in. In this study the Impala was adopted as the database for the dataset which automatically loads the data into the HDFS in the Hadoop.

The Apache Impala project provides high performance, low-latency SQL queries on data stored in popular Hadoop file formats (cloudera.com/document/enterprises/5-8- x/topic/impala). The following are key advantages of Impala:

* Impala produces results within seconds which indicates high performance
* Impala is pioneering the use of the parquet file format, a columnar storage layout that is optimized for large scale queries usually in warehouse scenarios.

The Impala is more convenient than sqoop because it does not require long processes to load the data into the HDFS. The table 3.2 below summarizes the Cloudera setup thoroughly.

# Table 3.2 Cloudera setup

|  |  |
| --- | --- |
| **Main tools** | **Version** |
| Cloudera | 5.13 |
| Impala | 2.10.0 |
| Hadoop | 2.6.0 |

# The datasets

Data that has been used for this specific research has been acquired from the kaggle website ([www.kaggle.com](http://www.kaggle.com/)). Kaggle is a space for models of predictions and analytics competitions in which professions such statisticians and data miners challenge each other to yield the best models for predicting and describing the datasets uploaded by organizations and users (www.kaggle.com). The form of data which was acquired is structured data. Structured data is data that is well organized and in this study a comma-separated values (CSV) is used , it is a file in which data is delimited by commas and it is in double quotation marks.

This form of data is used because it is easy to understand and much simpler to work with. It does not need to be organized or prepared first prior to using the tools on it. Structured data is easily traceable.

# Conclusion

In this chapter, the tools used to achieve the research aim and objectives were explained thoroughly and clearly. The data introduced are structured data which are easy to work with as you can make more clear conclusions and appropriate decisions.

# CHAPTER FOUR: RESULTS AND ANALYSIS

Several experiments were conducted in order measure and compare the performance of the tools. A 1.4GB dataset was used throughout the experiments. The dataset consists of 22 columns and over one million three hundred and twenty-six thousand one hundred observations, 1326100 to be specific.

# Experiments

Conducted experiments are:

* + - MySQL data loading performance
    - Hadoop cluster data loading performance
    - R studio data loading performance and analysis

# MySQL data loading performance

This experiment focuses on how long it took to load the dataset in the database. Firstly a database called yelp2 was created along with the number of columns that corresponds with columns in the CSV file. Then the commands in Figure 4.1 where used to smoothly import the dataset into the table called yelp\_user.

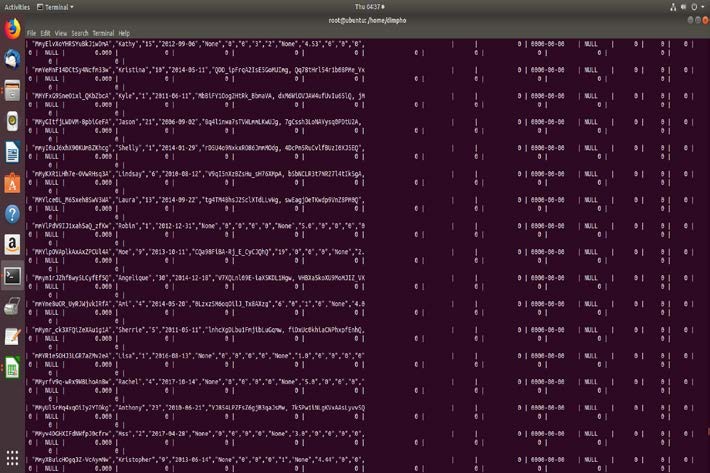
*LOAD DATA LOCAL INFILE ‘home/dimpho/Desktop/yelp\_user.csv’ INTO TABLE yelp\_user*

*FIELDS TERMINATED BY ‘,’ ENCLOSED BY ‘”’ LINES TERMINATED BY ‘\r\n’*

# Figure 4. 1: Importing dataset into MySQL database

According to the commands in Figure 4.1 the LOAD DATA LOCAL INFILE reads the rows from the CSV file which is a comma separated file into a table at a very quickly and rapidly pace. The field of the file is terminated by a comma indicated by FIELD TERMINATED BY ‘,’ and enclosed by double quotation marks specified by ENCLOSED BY ‘”’. Each row of the CSV file is terminated by a new line character indicated by LINES TERMINATED BY ‘/n’.

After the dataset being successfully loaded into the database yelp2, the statement *SELECT \* FROM yelp\_user* was used to verify dataset loaded in the database and the Figure 4.1 confirms successful data loading process.



# Figure 4. 2: Dataset successfully loaded in the MySQL database

The process of loading the dataset into the database took exactly four minutes. The time taken for loading the dataset is quite small considering that we are working with a larger dataset. It is a satisfactory result which shows that MySQL database can efficiently and reliably handle dataset of this size. It is of high performance given the time it took which means it is very fast and manages memory very well. The scalability is one of the few reasons of its high performance, since it can handle almost any amount of data, approximately 50 million rows or more even.

# Hadoop cluster data loading performance

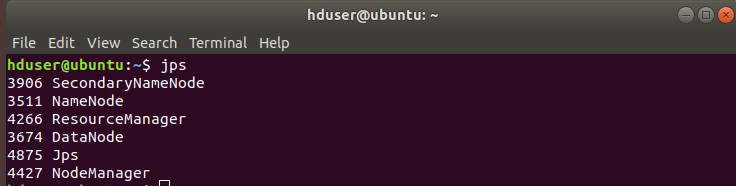
The main aim of this experiment is to understand how the Hadoop respond with regards to performance. Now that the dataset is in the database the sqoop was used by using the following command to load the data into the Hadoop. However before using the sqoop, the Hadoop needs to be launched first by using the following commands:

*Start-dfs.sh Start-yarn.sh*

# Figure 4. 3 : Commands to launch Hadoop cluster

This is to verify that the Namenode and the Datanode are running so that they can accept the data which is divided into blocks. To ensure that all the nodes are running

in the Hadoop cluster, the command *jps* is used to list all the nodes running as well as the resource manager. The list of processes is shown in Figure 4.4.



# Figure 4. 4: The running processes in the Hadoop cluster

Summarily, each process is responsible for the following:

* + - * SecondaryNameNode is a specially dedicate node in HDFS cluster its main function is to take checkpoints of the file system metadata present on namenode
      * NameNode holds the metadata for the HDFS such ad Namespace information, block information. When everything is used the information is stored in main memory
      * ResourceManager protects the data processing in the Hadoop cluster.
      * DataNode is responsible for data storage in the Hadoop File System
      * Jps (Java Virtual Machine Process Status Tool) is a command that checks the running of the Hadoop daemons like your SecondaryNameNode, NameNode, ResourceManager, DataNode and NodeManager
      * NodeManager is responsible for containers, monitoring their resource usage (CPU, memory, disk, network) and reporting this to the ResourceManager ([https://hadoop.apache.org](https://hadoop.apache.org/)/docs/YARN)

Now that the Hadoop cluster has started, running the sqoop command in Figure 4.5 was used to load the data into Hadoop.

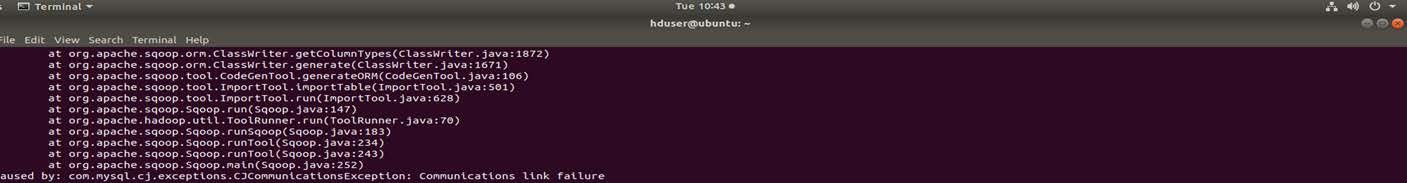
*sqoop import –connect jdbc:mysql://Ubuntu:3306/yelp2 –username root -- password Letl0tl0 –m 1 –table yelp\_user –target-dir /user/landing/sqlImport*

# Figure 4. 5: Command to Import data into HDFS using Sqoop

Running the command in Figure 4.5, one of the prerequisites for the sqoop import is to have a connection string that includes a URL navigating to the database, a valid MySQL user name, password and the database to be imported. The connect basically connects the database yelp2 with the sqoop using the jdbc connector driver, then the hostname is specified as Ubuntu along with the port number which is 3306. The table name is also specified as yelp\_user and a directory of /user/landing/sqlImport for a specific place of where exactly the dataset is to be put. Results of these processes are shown in Figure 4.6 and Figure 4.7 respectively.



# Figure 4. 6: Results of data loading into Hadoop through sqoop



**Figure 4. 7: Errors resulted by loading data into Hadoop through sqoop**

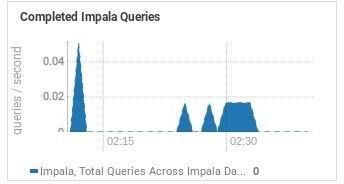
Looking at the results in Figure 4.3 thoroughly they evidently tell us that the Hadoop’s projects are not set which are the Hbase, Hcatalog, Accumulo as well as the Zookeeper. However, since these projects are not part of the study it is not necessary to have them installed.

The output further indicates that running sqoop version 1.4.7 shows the preparation taken to transport the data into the HDFS. Figure 4.7 illustrate that there is a communications link failure. The error explains that the last packet to the server was

successfully sent 0 milliseconds ago. The import tool as well has experienced a failure which we can conclude that communication failure is the main problem. Certain factors such as miscommunication between the jdbc driver and the database can be attributed as the cause of errors.

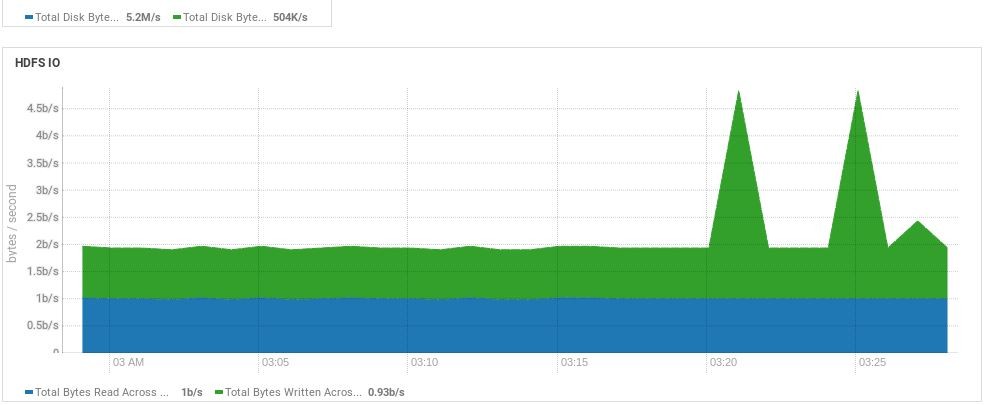
# Cloudera performance

The initial tools (localized virtual Hadoop with sqoop) were unable to produce the desired results as shown in this chapter. However, the Cloudera was able to provide a more appropriate and readily configured platform. The experiment performed on the Cloudera platform was loading the data into the HDFS through Impala. Impala has shown an impressive performance as shown in Figure 4.8. The monitoring of the performance is done by the Cloudera manager which keeps tract of every record and time.



# Figure 4. 8: The graph showing the performance of the Impala

The graph in Figure 4.8 shows that during the time between 02:15 and 02:30 there was an impressive performance of 0.02 queries per second and 0.04 queries per second. This indicates how fast the impala is as opposed to the traditional MySQL and other databases. The graph shows a fluctuation of queries in different queries which means that the shorter the query the shorter the time and the longer the query the longer the time. This indicates a direct proportionality. Once the data is in the impala it is automatically in the Hadoop cluster in the HDFS. The Figure 4.9 shows the performance of the HDFS.



# Figure 4. 9: A graph showing the performance of the HDFS

The Figure 4.9 indicates the performance of the HDFS once the data is into the Hadoop. As indicated the green region shows the total bytes written across the DataNodes. Between the time 03:00am and 03:30am a total of 2bytes per second was written across the DataNode. Furthermore, the time between 03:20am and 03:25am has shown a total of above 4.5 bytes per second of data written across the DataNode.

The blue region on the other hand indicates the total bytes read across the DataNode. As shown above in Figure 4.9 between the time 03:00am and 03:25 a total of 1 byte per second is read across the DataNode. With the read and write operations of DataNode mentioned, the processing of the data indicates a very high performance of the HDFS. It only takes a few seconds for the HDFS to execute which shows how efficient and reliable the HDFS is. This also shows how efficient the storage of the Cloudera Hadoop is.

# R studio data load performance

The final experiment done was to load the 1.4GB dataset into the R studio to study the behaviour and performance of the R studio. Initially the dataset could not load due to the memory reserved for the VMware virtualization machine. The space of the VMware was then increased.

The second attempt taken to load the data took about approximately 15 minutes. This is rather a long time for a 1.4GB data. With such performance a much larger size of data may take a while. Another attempt was taken with a 3.8GB which took over 24 hours and later the session crashed since the R studio could not handle such data. Since we have successfully loaded the dataset into the R studio, in the analysis section we look at the in-depth of dataset, i.e. we look at the insights of the dataset itself and perform some analytics.

# Analysis

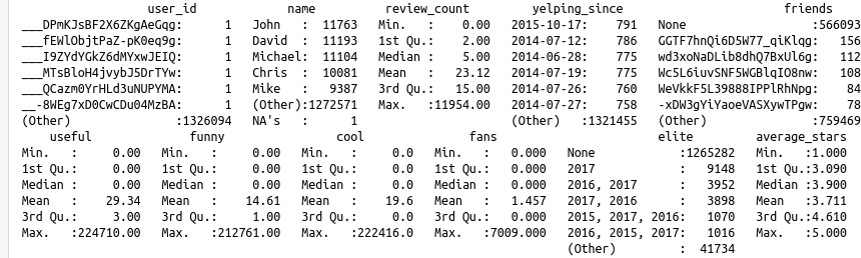
In the Section4.1, we looked at how the three tools (MySQL, Hadoop and R studio) perform when loading large dataset. The following Table 4.1 summarizes the overall performance.

# Table 4. 1: Summarized performance of MySQL, Hadoop and R studio

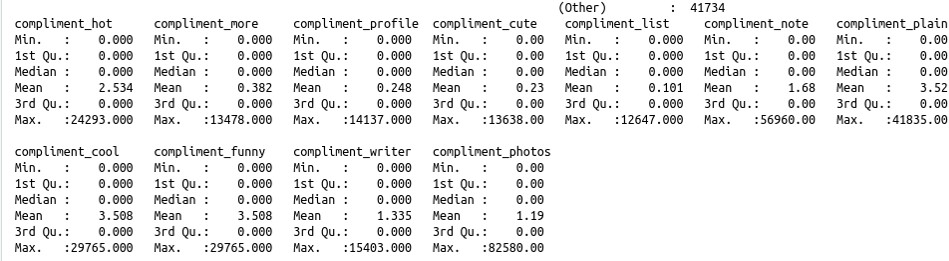
|  |  |  |  |
| --- | --- | --- | --- |
| **Processing tools** | **MySQL** | **Hadoop** | **R Studio** |
| **Data Load Time** | Exactly 4 minutes | No time was recorded due to  unforeseen events | Approximately 15 minutes |
| **Comments** | Such a small time which shows a bit of high  performance | Communications failure occurred | Indicates very slow performance however results can be found with particular size of  data. |

From Table 4.1, conclusions can be made to say that MySQL is a much better performer than the R studio although there are of different duties. By comparison of the data loading time, MySQL took far less time as opposed to the R studio.

However, with the given results of Hadoop, it has shown multiple errors which indicate a communications link failure. While experiencing these failures the Hadoop had a very fast response to that which shows that loading data into the Hadoop should literally take a few seconds. With the 1.4GB being successfully loaded into the R studio, the following results were acquired.



# Figure 4. 10: Summarized dataset



**Figure 4. 11: Summarized dataset**

The datasets in Figure 4.10 and 4.11 are about different customers who rate different restaurant by the use of various compliments. Compliments such as fun, cool, cute and hot which customers try to explain how they feel about these restaurants. Reviews are also included the column in this case is written as review\_count. The customers write how they feel about the services and have suggestions as well on how these restaurants can improve their services. This is written in a form of text.

The term yelp means a sharp loud cry especially if an accident has occurred. If you look at the dataset there is a column called yelping\_since, in this regard yelping refers to the complaints of the customers. It has been introduced by dates which indicates

the dates of when did the actions of taking the reviews of customers start. Within the reviews the customers writes if they can recommend a particular restaurant to friends hence if you look at Figure 4.7 there is a column of friends. Customers use stars to rate different customers for example if a customer rate a restaurant with five stars it shows us that the particular restaurant has impeccable service and the customer is exceedingly satisfied .

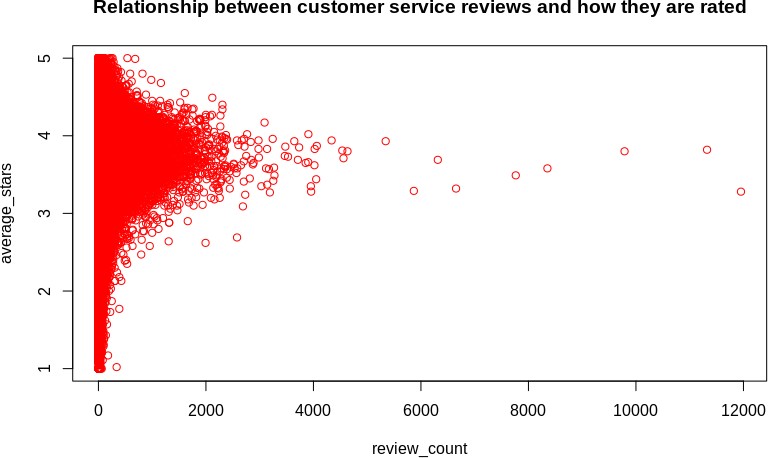
In both Figures 4.7 and 4.8 there is the first quartile, the mean and third quartile for each column. The first quartile simply means that the dataset is divided into two equal halves, then the median of the first half is computed which is what we call the first quartile. The median is generally the middle number of the whole dataset without dividing the dataset. Similarly with the third quartile which is the median of the last Half of the dataset. The mean that is the average of the all columns are computed as well. The minimum and maximum values are also given showing that for each column there is the smallest and the largest number all the way through the 1326100 observations.

The Figure 4.7 and 4.8 indicates that this dataset consist of 22 columns with observations of 1326100. In Figure 4.7 there are 11 columns, the first column which is the user\_id that uniquely identifies each user by use of different numbers. It can be their identity number or the number they acquire as they step into the restaurant. The user\_id will assist with various reasons such as order taken and what table urgently need assistance.

In overall this dataset is some sort of awareness to restaurants to show that there is always room for improvement. Restaurant need to do better to improve their services by doing so, a strong and healthy competition will emerge amongst restaurants. A good competition always yields competence of high quality which always keeps the staff on their toes.

The importance of competence is a delicate issue since restaurant provides food and this is to ensure quality to food. This because they are dealing with peoples’ lives and this is to indicate that quality and competence are relatable and goes hand in hand.

Figure 4.12 show a summarized graph of a selected columns of the dataset to give a glimpse of the behaviour.



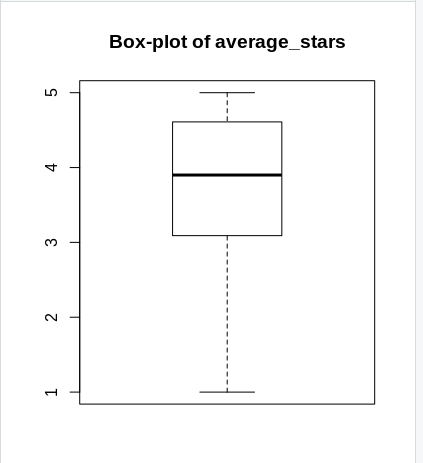
# Figure 4. 12: A graph showing the relationship between the customer service and how they are rated.

The graph depicted above describes the relationship between customer services and how they are rated. Most of the restaurants are rated between an average of three and four which shows a satisfaction of services, these are the review\_count that range amongst 4000-12000. It means that the restaurants are doing well however they must put it some work in order to improve their services and stay in the game.

This shows that should they slack they stand a chance to lose a lot of customers. Losing customers is not good for business because it may eliminate them from the competition and elimination from the competition leads to shutting down of business.

However the restaurants that are most likely to shut down are those who have been rated way lower than an average of three. If you look at the graph closely the review\_count between 0-2000 are most likely to get out of business due to the lower average\_stars acquired. As a new customer looking at the reviews as well as the ratings the most probable restaurant you would go for is the one with higher ratings.

The pattern shows satisfactory results, which means most of the restaurant are most certainly doing a great job, but there is still room for improvement. The Figure 4.10 below gives a summary of what has been explained previously on the Figure 4.9.



# Figure 4. 13: The boxplot of average stars

The boxplot proves the explanation of the customers rating. Most of the restaurants have been given a rating of between the stars of 3 and 4. The results show a satisfactory service which indicates that these various restaurants will be having frequent customers and potential customers which are possibly recommended by other customers.

The pattern of this rating brings more business to the restaurant, which means more money to the business and with more money comes with innovative ideas such as job creations that will boost the economy of a country.

# CHAPTER FIVE: CONCLUSION

The results obtained in this study have shown that the Hadoop meant for high performance. It is able to store and process data within seconds as shown in Chapter 4 through experimentations performed on Cloudera. The ultimate aim of our research investigation was achieved on the Cloudera platform that provided reliable results.

The limitation of Cloudera was space with regards to analysis. Analysis through localized virtual Hadoop was not performed as it required more resources than we have, i.e. 16GB of RAM. Hence, R programming was used in this regard.

The Cloudera was not the only platform used to check the properties of Hadoop. Initially the projects of the local Hadoop were installed individually which produced the results explained in Chapter 4. Unfortunately factors such as communication failure between the Sqoop and the database restricted us from the desired goal using this particular method.

According to Chapter 2, Hadoop is the most scalable, fault-tolerant, framework with regards to data loading. Furthermore, big data concepts were thoroughly explained with main focus being the Hadoop architecture. Comparisons between Hadoop, MySQL and R have evidently shown that MySQL is more efficient with small datasets and R platform is also limited in terms of performance for large datasets. The Cloudera has shown to be more efficient and of high performance because resources are not limited. The importance of this comparison is to see various tools playing around with big data and to test the efficiency of the traditional non-big data tools as compared to the tools built for big data such as Hadoop.

We are living in a digital world presently, where huge amounts of data heterogeneous, semi-structured, structured and unstructured data are continuously being generated at an alarmed and unusual rate. It is imperative that such big data are stored and analysed efficiently and accurately to make better and informed decisions.

**REFERENCES**

1. Acharjya, D.P. and Ahmed, K., 2016. A survey on big data analytics: challenges, open research issues and tools. *Int. J. Adv. Comput. Sci. Appl*, *7*(2), pp.1-11.
2. Aminu, L.M., 2014. Implementing Big Data Management on Grid Computing Environment. *International Journal Of Engineering And Computer Science*, *3*(09).
3. Assunção, M.D., Calheiros, R.N., Bianchi, S., Netto, M.A. and Buyya, R., 2015. Big Data computing and clouds: Trends and future directions. *Journal of Parallel and Distributed Computing*, *79*, pp.3-15.
4. Chakravarthy N, 2014. WHITE PAPER Intel® Enterprise Edition for Lustre\* Software High Performance Data Division
5. Che, D., Safran, M. and Peng, Z., 2013, April. From big data to big data mining: challenges, issues, and opportunities. In *International Conference on Database Systems for Advanced Applications* (pp. 1-15). Springer, Berlin, Heidelberg.
6. Hammad, K., Fakharaldien, M., Zain, J. and Majid, M., 2015, September. Big data analysis and storage. In *International Conference on Operations Excellence and Service Engineering* (pp. 10-11).
7. [https://hadoop.apache.org/does/stable//hdfsdesignhtm](https://hadoop.apache.org/does/stable/hdfsdesignhtml)l Accessed Date:17 September 2018
8. [https://pixshark.com/hadoop-mapreduce-architecture-diagram.htm](https://pixshark.com/hadoop-mapreduce-architecture-diagram.html)l Accessed Date: 17 September 2018
9. [https://dev.mysql.com/doc/refman/5.7/en/load-data.htm](https://dev.mysql.com/doc/refman/5.7/en/load-data.html)l Accessed Date: October 2018
10. <https://ww.tutorialspoint.com/sqoop/sqoop_import.html> Accessed Date: September 2018
11. [https://hadoop.apache.org](https://hadoop.apache.org/) > docs > YARN Accessed Date: September 2018
12. https://www.Kaggle Accessed Date: August 2018
13. <https://www.tutorialspoint.com/hadoop/hadoop_enviornment_setup.htm> Accessed Date: October 2018
14. [http://data-flair.training/forums/topic/what-is-single-point-of-failure-in-hadoop- 1-cluster](http://data-flair.training/forums/topic/what-is-single-point-of-failure-in-hadoop-1-cluster) Accessed Date: October 2018
15. <https://en.m.wikipedia.org/wiki/Sqoop>Accessed Date: September 2018
16. [https://data-flair.training/blogs/rack-awareness-hadoop-hdfs/#](https://data-flair.training/blogs/rack-awareness-hadoop-hdfs/) Accessed Date:17 September 2018
17. <http://www.edureka.co/blog/hadoop-ecosystem>Accessed Date :24 September 2018
18. [http://www.infodiagram.com/diagrams/big-data-technology-cloud-diagrams- icons-powerpoint.htm](http://www.infodiagram.com/diagrams/big-data-technology-cloud-diagrams-icons-powerpoint.html)l Accessed Date: 02 October 2018
19. Jaseena, K.U. and David, J.M., 2014. Issues, challenges, and solutions: Big data mining. *Computer Science & Information Technology (CS & IT)*, pp.131- 140.
20. Khan, N., Yaqoob, I., Hashem, I.A.T., Inayat, Z., Ali, M., Kamaleldin, W., Alam, M., Shiraz, M. and Gani, A., 2014. Big data: survey, technologies, opportunities, and challenges. *The Scientific World Journal*, *2014*.
21. Kune, R., Konugurthi, P.K., Agarwal, A., Chillarige, R.R. and Buyya, R., 2016. The anatomy of big data computing. *Software: Practice and Experience*, *46*(1), pp.79-105.
22. Maltby, D., 2011, October. Big data analytics. In *74th Annual Meeting of the Association for Information Science and Technology (ASIST)* (pp. 1-6).
23. Prasad, B.R. and Agarwal, S., 2016. Comparative study of big data computing and storage tools: a review. *International Journal of Database Theory and Application*, *9*(1), pp.45-66.
24. Ramesh, B., 2015. Big data architecture. In *Big Data* (pp. 29-59). Springer, New Delhi.
25. Raste, K.S., 2014. *Big Data Analytics-Hadoop Performance Anaylsis* (Doctoral dissertation, Sciences).
26. Tulasi, B., Wagh, R.S. and Balaji, S., 2015. High Performance Computing and Big Data Analytics a [euro]" Paradigms and Challenges. *International Journal of Computer Applications*, *116*(2).
27. Sqoop.apache.org Date accessed: September 2018
28. Venkatesh, H., Perur, S.D. and Jalihal, N., 2015. A study on use of big data in cloud computing environment. *Int. J. Comput. Sci. Inf. Technol.(IJCSIT)*, *6*(3), pp.2076-2078.
29. Zakir, J., Seymour, T. and Berg, K., 2015. BIG DATA ANALYTICS. *Issues in Information Systems*, *16*(2).