

**JIMMA UNIVERSITY**

**JIMMA INSTITUTE OF TECHNOLOGY**

**FACULTY OF COMPUTING AND INFORMATICS**

**MSC. IN INFORMATION SCIENCE (ELECTRONIC AND  
DIGITAL RESOURCE MANAGEMENT (EDRM) STREAM)**



**BIG DATA ANALYTICS SYSTEM FOR PREDOMINANT CHRONIC  
DISEASES**

**BY:-MESAY G/MARIAM**

**SEPTEMBER, 2020**

**JIMMA, ETHIOPIA**

**JIMMA UNIVERSITY**

**JIMMA INSTITUTE OF TECHNOLOGY**

**FACULTY OF COMPUTING AND INFORMATICS**

**MSC. IN INFORMATION SCIENCE (ELECTRONIC AND  
DIGITAL RESOURCE MANAGEMENT (EDRM) STREAM)**

**Big Data Analytics System for Predominant Chronic Diseases**

Thesis Submitted to Jimma Institute of Technology, Faculty of Computing and Informatics in Partial Fulfillment of the Requirements for the Degree of Master of Information Science in Electronics and Digital Resource Management.

By: Mesay Gebremariam

Principal Advisor: Worku Jimma (PhD)

Co-Advisor: Solomon Alemu (MSc)

**SEPTEMBER, 2020**

**JIMMA, ETHIOPIA**

**JIMMA UNIVERSITY**  
**JIMMA INSTITUTE OF TECHNOLOGY**  
**FACULTY OF COMPUTING AND INFORMATICS**  
**DEPARTMENT OF INFORMATION SCIENCE**

**BIG DATA ANALYTICS SYSTEM FOR PREDOMINANT  
 CHRONIC DISEASES**

**BY:**  
**MESAY G/MARIAM**

As members of the board of examining of the Msc thesis open defense examination of the above title, we members of the board (listed below), read and evaluated the thesis and examined the candidate.

Name and Signature of Members of the Examining Board

<b>Name</b>	<b>Title</b>	<b>Signature</b>	<b>Date</b>
<u>Chala Diriba</u>	Chairperson	_____	_____
<u>Worku Jimma</u>	Principal Advisor	_____	_____
<u>Solomon Alemu</u>	Co-Advisor	_____	_____
<u>Abu Santure</u>	Internal Examiner	_____	_____
_____	External Examiner	_____	_____

**Declaration**

This thesis has not previously been accepted for any degree and is not being concurrently submitted in candidature for any degree in any university.

I declare that the thesis is a result of my own investigation, except where otherwise stated. I have undertaken the study independently with the guidance and support of my research advisor. Other sources are acknowledged by citations giving explicit references. A list of references is appended.

Signature: \_\_\_\_\_

Mesay G/Mariam Mecha

This thesis has been submitted for examination with my approval as university advisor.

Principal Advisor's Signature: \_\_\_\_\_

Worku jimma (PhD)

Co-advisor signature \_\_\_\_\_

Solomon Alemu (Msc)

**To my beloved families!**

## Acknowledgment

*“He that dwelleth in the secret place of the most high shall abide under the shadow of the Almighty.” Ps 91:1*

First and for most, I would like to thank the Almighty **God** for giving me provision of knowledge, wisdom and diligence required for the successful completion of this thesis work of my master’s program. Next, my special gratefulness and appreciation goes to my advisors, Worku Jimma (PhD), for shaping and guiding my thoughts and steps on the right track through his insightful comments. Without his intellectual advice, guidance, encouragement, and regular discussion that were invaluable and inspiring in the processes of my study, this paper would not have had its end. I would also like to express my warm thanks to my co-advisor, solomon Alemu (Msc) for his comments on my study.

## Abstract

*Medical data is one of the most rewarding and yet most complicated data to analyze. Medical images, biomedical signals and handwritten prescriptions are available and can be used for pre-diagnostic tasks on the existence of chronic disease by assuming big data analytic concepts. Hence, the main objective of this study was to design a big data analytics prototype that process and visualize the huge amount of dataset by using R-studio programming software. Big data processing and visualization is a challenge that needs new way of tackling which otherwise cannot be solved with current practice of data management because data deluge and data creation frequency in varieties of formats are inevitable scenarios.*

*A big data analytics system that descriptive the occurrence of chronic disease from the big medical data was developed by using different methods and tools. In this study data computation techniques is applied and descriptive analysis were employed. The major new data management techniques are applied to ensure the quality of data and integrate data from different sources. Experimental research design was employed for this study. In addition, (descriptive) analysis approach based on a logistic activation function is employed to build the model. This study achieved as it is possible to manage big data regardless of size and nature of data.*

*The major challenge faced during conducting this study is dealing with heterogeneous data in order to generate insights for improved health-care outcomes or visualization of data. The other most challenging task was the fact that data preserved in Jimma Medical center are disorganized and distributed since it comes from various sources and having different structures and forms. The researcher strongly recommend that prototype with the capability of analyzing and visualizing heterogeneous big data should be developed. As new area of study, it is strongly recommended further studies in specific contexts.*

*Keywords: Data Visualization, data redaction, Big Data, visualization technique, Big Data analytics, chronic diseases.*

## Table of Contents

Acknowledgment .....	i
Abstract.....	ii
List of Figures .....	vi
Abbreviations and Acronyms .....	vii
CHAPTER ONE .....	1
1 Introduction .....	1
1.1 Background of the study .....	1
1.2 Statement of the problem .....	5
1.3 Research questions .....	6
1.4 General objective of the study .....	6
1.4.1 General Objective .....	6
1.4.2 Specific objective of the study.....	6
1.5 Significance of the study.....	7
1.6 Scope of the study .....	8
CHAPTER TWO .....	9
2.0 LITERATURE REVIEW .....	9
2.1 Overview.....	9
2.2 Background of Big Data and its Analysis .....	9
2.2.1 Velocity of health data: .....	10
2.2.2 Volume:.....	10
2.2.3 Variety:.....	10
2.3 Data analytics approach .....	11
2.3.1 Traditional business intelligence (BI): .....	11
2.3.2 Data mining .....	12
2.3.3 Statistical applications:.....	12
2.3.4 Predictive analysis .....	12
2.3.5 Data modeling: .....	12
2.3.1 Ensemble analysis.....	13
2.3.2 Association analysis .....	14
2.3.3 High-dimensional analysis .....	14
2.3.4 Deep analysis.....	15
2.3.5 Precision analysis.....	15



2.3.6 Divide and conquer analysis.....	15
2.4 VISUALIZATION TECHNIQUES.....	16
2.5 Tools and Framework .....	17
2.5.1 Hadoop .....	17
2.5.2 MapReduce:.....	17
2.5.3 Hadoop YARN .....	18
2.6 Challenges of Big data analytics .....	18
2.7 Big Data in Healthcare.....	21
2.8 Related Works .....	21
CHAPTER THREE.....	23
METHODOLOGY .....	23
3.1.1 Literature review.....	23
3.1.2 Development and Processing Tools .....	23
3.1.3 Visualization Tools .....	23
3.1.4 Evaluation methods.....	24
3.2 Data collection.....	24
3.3 Data Pre-Processing:.....	25
3.3.1 Data Cleaning .....	25
3.3.2 Data Integration .....	26
3.4 Classifier Selection and Parameter Modification: .....	27
3.5 Design Model: .....	28
3.5.1 System Architecture.....	28
3.5.1.1 Data Layer .....	30
3.5.1.2 Data Aggregation Layer .....	30
3.5.1.3 Analytics Layer .....	30
3.5.1.4 Information Exploration Layer .....	30
3.5.2 Framework of analytics system .....	31
3.5.3 Architecture of Data Warehouse.....	32
3.5.3 Design and Goal of the system .....	32
5.6 Data Visualization Design.....	33
3.7 Algorithms .....	34
3.7.1 Decision Tree Classifiers.....	34
3.7.2 Clustering Algorithms:.....	34

3.8 Hadoop.....	37
3.8.1 Characteristics of Hadoop .....	37
3.8.2 Hadoop MapReduce Framework .....	37
3.9 Big Data volume Reduction Methods.....	40
3.9.1 Reducer Algorithm.....	41
3.9.2 Data duplication (Redundancy Elimination) .....	41
CHAPTER FOUR.....	43
4. EXPERIMENTS AND RESULTS.....	43
4.1. Environment Setup .....	43
4.2 Requirement for big data analytics .....	44
4.2.1 Application of Sense making Models.....	44
4.2.2 Conceptual Architecture for Big Data Analytics .....	45
4.3 Experimentation .....	45
4.4 Results.....	46
4.4.1 Data Visualization.....	46
4.5 System performance evaluation .....	50
4.6 Contributions.....	50
CHAPTER FIVE.....	51
5. CONCLUSION AND RECOMMENDATIONS .....	51
5.1 Conclusion .....	51
5.2 Recommendations .....	51
5.3 Future Works.....	52
References.....	53

## List of Figures

Figure 2.1 Data Analytics Techniques .....	13
Figure 2. 2: Hadoop Architecture Framework.....	18
Figure 2 2Hadoop Architecture Framework.....	18
Figure 3.1 Big data analysis stage.....	26
Figure 3. 1 System Architecture.....	28
Figure 3. 2: Framework of analytics system.....	30
Figure 3. 3 Clustering Algorithms.....	34
Figure 3 5: Map Reduce Framework.....	36
Figure 3. Unstructured or semi-structured data set conversion.....	38
Figure 4.1 - Hardware specification.....	41
Figure 4.2: Prototype home page sample Data dashboard.....	43
Figure 4.3: sample screenshot.....	44
Figure 4.5: scatter plot screenshot.....	47

## Abbreviations and Acronyms

API = Application Programming Interface

BDA =Big Data and Analytics

BI = Business Intelligence

CPU = Central Processing Unit

DNA = Deoxyribonucleic acid

DRIP = Data Rich Information Poor

EHR = Electronic Health Record

FPGA =Field Programmable Gate Array

GPU =Graphic Processing Units

HDFS = Hadoop distribute file system

IID = Independent and Identical Distribution Theory

KPI =Key Performance Indicators

MCAR = Missing Completely At Random

RDBMS= Relational Database Management System

YARN = yet Another Resource Negotiator

## CHAPTER ONE

### 1 Introduction

#### 1.1 Background of the study

A chronic condition “is a physical or mental health condition that lasts for more than one year and causes functional restrictions or requires ongoing monitoring or treatment” (Wagner, 2008) . Chronic diseases-such as cancer, diabetes, hypertension, stroke, heart disease, respiratory diseases, arthritis, obesity, and oral diseases can lead to hospitalization, long-term disability, reduced quality of life, and death. Globally, chronic diseases have affected the health and quality of life of many citizens. In addition, chronic diseases have been a major driver of health care costs while also impacting workforce patterns, including, of course, absenteeism. More than two thirds of all deaths are caused by one or more of these five chronic diseases, namely heart disease, cancer, stroke, chronic obstructive pulmonary disease, and diabetes. Chronic diseases are characterized by high prevalence among populations, rising complication rates, and increased incidence of people with multiple chronic conditions, to name a few (Raghupathi & Raghupathi, 2018 ). Some of chronic diseases are discussed below (Raghupathi & Raghupathi, 2018 ).

**Stroke:** Stroke is a disease of the brain caused by interference to the blood supply. Stroke and heart disease are the main cardiovascular diseases.

**Cancer:** Cancer describes a range of diseases in which abnormal cells proliferate and spread out of control. Other terms used are tumors and neoplasms. There are many types of cancer and all organs of the body can become cancerous.

**Chronic respiratory diseases:** Diseases of the lung take many forms. Chronic obstructive respiratory disease and asthma are the most common forms.

Big Data, is a generic term for data sets of structured, semi-structured and unstructured data that are extremely large and complex. Traditional software, algorithm, and data repositories are inadequate to collect, process, analyze, and store and it has become an intensively studied area in recent years. With the development of the Internet, the mobile Internet, the Internet of things,

social media, biology, finance, and digital medicine, the volume of data has increased dramatically (Hermon, 2014).

Big Data not only describes the large size of data as its name suggests but also implies rapid data processing ability, novel technology and approaches for handling the data. In 21st century, Big Data went through a series of evolutionary steps, and software in suitable environment has been developed. With the growth of information exchanges, Big Data has been expanded to a certain scale, not only in its size but also in data technology. In terms of its five main characteristics, volume, variety, velocity, variability, and veracity, state-of-the-art techniques, technologies, and equipment are required to deal with Big Data in correlation analysis, clustering analysis, modeling, prediction, and hypothesis verification. Thus, advanced hardware and software are required for data acquisition, extraction, processing, analysis, and storage. Currently, infrastructure for Big Data includes servers, storage systems, cloud service, and networking equipment. Software for Big Data include parallel and distributed file systems, retrieval software, and data-mining software (Sharmila and Bhuvana, 2014).

Observing at the current and future applications of Big Data in healthcare, it is interesting to see how they further enhance and accelerate the convergence between the activities of clinicians, administrators, policy makers, insurance companies and researchers by saving costs, creating greater efficiencies based on outcome comparison, reducing risks, and improving personalized care. (Al-Shiakhli, 2019).

Recently, Big Data and business analytics approaches have been developed and implemented to analyze a large volume of data generated by different organizations. Consequently, every issues needs faster insight into growing volumes of transactional data. Analyzing data in real time helps organizations view the past and foresee the future. This is the beauty of streaming analytics and is able by knowing what occurred (descriptive), understanding why it happened (diagnostic), looking ahead to what might take place (predictive) and, ultimately, determining how to influence future occurrences or prescriptive (Fahmideh and Mahdi, 2018).

The big data opportunity is not only for achieving high efficiency in operations. There are also important opportunities for economic growth and improving the standard of living to the society. There are various ways in which big data analytics can improve health organizational outputs and

industries. These include improved health care delivery, the standard of education, national security, and enable good governance. In addition, it has a potential to assist policy-makers to gain insight in enabling policies that will grant safe playground for investors, help waste managers find the type of waste that is more generated from a particular locality and provide insight for sharing of waste collection material ( (Berman, 2013). Moreover, education monitoring agency can deploy big data and business analytics approaches to evaluate the performance of teachers and improve work attitude (Shah, Gita Basava, 2014).

Today, more advanced devices are coming to cyberspace with a lot of functionalities that provide services at different level, for instance, individual, group and community. Now, people are at a verge of simplifying life questions which could be expressed in terms of space and time. Interactions of Internet of Things (IoT) and people are generating data that cannot be left alone due to its value (Sawant, & Shah, 2013). Current infrastructure and applications are allowing human kind freedom for communication and doing activities in the format of digital data which was inconceivable some years ago. These large-scale facilities and capabilities are pouring data from different sources and directions to global data storage which is accumulated to about 1,800 EB (Exabyte) or 1.8 ZB (Zeta bytes) (Desalegn, 2016).

This data is generated in the course of building driver assistance and autonomous vehicle technologies, IoT devices including sensors in our bodies, homes, factories and cities, creating high resolution content for 360 video and augmented reality and 5G communications. It is enabled by building the edge networks and centralized data centers that help to analyze, communicate and store the resulting data. The creation of all of this digital technology is often called “digital transformation.” (Coughlin, 2018).

According to Forbes the projection is that the amount of digital data generated (what IDC calls the Data sphere) will grow from 33 ZB in 2018 to 175 ZB by 2025 as shown in the figure below. IDC says that China’s Data sphere is expected to grow 30% on average over the next 7 years and will be the largest Data sphere of all regions by 2025. By 2025 49% of the world’s stored data will reside in public cloud environments.

The continuation of data accumulation, which is expected to be 50 times in 2025, at an alarming rate within a variety of formats make it difficult for current practice of management of data.

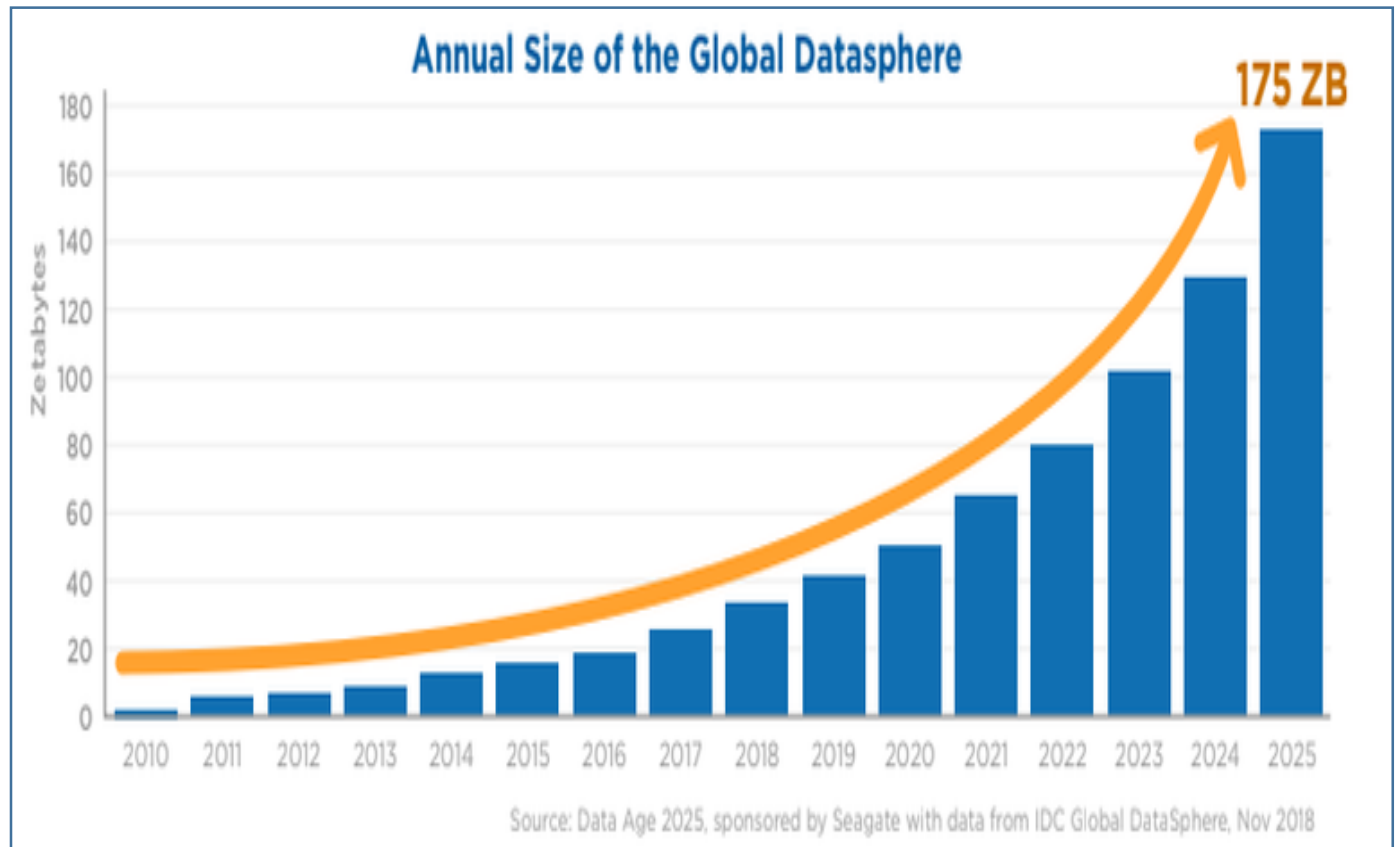


Figure 1.1 Annual size of global Data sphere (Coughlin, 2018)

The most challenging task to apply big data analytics in the health sector is the fact that data in health care are disorganized and not well recorded since it comes from various sources and having different structures and forms. This kind of data is commonly described as big data in the data science research community and Big Data analytics is the treatment to deal with this kind of data and generate insights for improved health-care decision making (Berman, 2013).

Big data processing and visualization is a challenge that needs new way of tackling which otherwise cannot be solved with current practice of data management because data deluge and data creation frequency in varieties of formats are inevitable scenarios. The approach that is employed in this study to undertake these challenges are reviewing problem areas in detail, followed by designing solution, then implementation of designed solution after that testing implemented



solution using big data sets (Alemayehu.B, 2019). Thus, the big data approach is used to store the health information which is used during disease diagnoses and the treatment. The sample big data based health informatics used to get a comprehensive understanding of big data in big medical data, especially in chronic diseases and health monitoring area. This study focuses on the full cycles of the big data processing, which includes chronic diseases medical big data preprocessing, applying big data tools and algorithm and big data visualization. It also tried to fill the gap between common big data technology and medical special needs by analyzing detail implementation of medical big data. To the best of our knowledge, this is the first survey that targets the chronic diseases and health monitoring big data technologies (Ackerman, 2012). This study addresses the challenge of understanding large amounts of data related to chronic diseases by applying visual analytics techniques and producing descriptive analytics. Therefore this study aims at developing a prototype that visualize Medical big data for Jimma Medical center.

## 1.2 Statement of the problem

The idea of Big Data has been presented to the healthcare system as a solution to a variety of healthcare related information system problems because health systems becoming increasingly complex and expensive. Effective integration of such data using data mining and medical informatics may result in lower costs and improved patient care via well informed decision making (Sharmila and Bhuvana, 2014).

To the knowledge of the researcher, in Ethiopian health system big data processing and visualization is not implemented and, data generation in a single day from each and every hospitals and health centers is so huge because of high number of patients seeking health care. Therefore, it is time to implement Big Data process and visualization in healthcare center so that the aforementioned benefits become a reality for all stake holders in healthcare system in general and patients in particular.

Traditional tools and techniques have no capabilities to serve big data in the way that accommodates three V's. (volume, velocity, and variety). Data Rich Information Poor (DRIP) is a scenario where there are vast data but inadequate useful information for a given purpose. This is because of limitation of Relational Database Management System (RDBMS), data warehouse and

data analysis tools. Jimma medical center has not the culture of analyzing the data generated from daily records. The huge amount of data generated by Jimma medical center is disorganized and incomplete for analyzing it. The data is big in size and its content needs more computerized platform to analyze for further decision making. The problems were solved to some extent by scholars such as Belay Alamayehu (2019) to predict the occurrence of cancer disease from big data. Belay suggested to conduct a system that describes and visualizes the essential information by extracting a pattern from the dataset collected from the center. This study was launched to answer the following research questions.

### 1.3 Research questions

This study attempts to answer the following research questions:

1. What are the requirements to design Big Data analytics and visualization system for chronic diseases?
2. How well will the system to be developed perform and accepted by users?

## 1.4 General objective of the study

### 1.4.1 General Objective

The main objective of this study was to design Big Data analytics and visualization system for predominant chronic diseases for Jimma Medical Center.

### 1.4.2 Specific objective of the study

1. To explore the current practices of big data analytics in the medical fields.
2. To review the major impacts of Big Data analytics on decision making process in health care system.
3. To prepare data set by collecting from medical departments of Jimma Medical Center.
4. To evaluate the system for measuring its performance.

## 1.5 Significance of the study

The beneficiaries of the system are primary health care workers (health professionals), ministry of health, policy makers, decision makers and researchers. Big Data analytics helps organizations store their data and use it to identify new opportunities. Some examples of the need of big data analytics system include ( Aldiabat, S, 2017)

Faster and better decision making across the health care organizations are looking for relevant, actionable information from a wide variety of sources of dataset to make timely strategic operational choices and provide life-saving healthcare treatment.

The data analysis plan which could reflect what categories are presented in an analytical system, detailing indicators, sources, units of analysis, and data collection techniques.

The cost reduction in computing process, storage and communication technology and mobile devices could enable the professionals to generate and access and store huge data. Such information can create value and provide knowledge, comfort, productivity, better, healthy society and in turn economic benefit. Big data analysis in healthcare provides very urbane and fast analytic tools which works on the massive and diverse kind of information.

Ministry of health could get benefit from the study, when Jimma Medical Center professionals could easily analyze the big data generated daily. The experience could be shared with other medical centers so that they can easily analyze their records as per requirement. The study could play a great role in providing essential pattern for policy makers. Health professional could easily make a decision up on the data visualized by the system.

The output of big data analytics in healthcare databases enable the health professionals to obtain the true information, and also increase the profits in Jimma Medical Center. This study also contribute its role for researchers. The output of this research can be an input for further studies to be done locally as well as globally, as per the recommendation made from this study. Results offered by big data in healthcare system, helps the users by providing the useful information about the disease predictions, hospital information and details of insurance companies who come forward to provide the financial benefits to the patients. Such useful information to the public with less expenditure improves the quality of life. The type of analytical conclusions that need to be reached

and agreed upon by multiple stakeholders in the event of joint analysis and the structure of a final report is designed to be read quickly and easily (Hermon, 2014).

## 1.6 Scope of the study

The scope of this study is specifically conducting big data process and visualization extraction form big data using available technology that facilitates knowing the landscape of data processing technologies.

The study is limited to designing a big data analytic system for chronic diseases (Diabetes, Hypertension, Cancer, Heart disease, kidney disease). Thus, dataset of other diseases is not considered in this study. Proposed system will be do data process the data set then based on the sector plot graph visualized on the system.

The other limitation of the study is that it does not include all kinds Vs. Velocity are not experimented. In addition, large data size such as Terabytes and Petabytes are not experimented in this study. These factors need well equipped labs with clusters of machines in order to conduct complete experimentation on top of adequate timespan.

## CHAPTER TWO

### 2.0 LITERATURE REVIEW

#### 2.1 Overview

To have deep understanding on this study, it is vital to review literature that have been conducted in the area. For this reason, related literature such as books, journal articles, guidelines, manuals, proceeding papers and some other sources have been revised in order to understand the domain knowledge, concepts, principles and methods that are important for big data analytics systems.

#### 2.2 Background of Big Data and its Analysis

Big data and analytics (BDA) remain to spark interest among scholars and practitioners. Organizations are increasingly aware that they may process and analyses their large data volumes to capture value for their businesses and employees. With the advent of more computational power, machine learning, particularly deep learning through neural networks has become more broadly deployable in organizations. Academic research on the topic also skyrocketed. Searching for the term 'big data', the Web of Science Core Collection yields 3347 hits in 2015, and over 4000 in both 2016 and 2017 (Singh & Singla, 2018).

Some studies have discussed how BDA influences organizational performance, arguing that firms with data driven strategies incline to be more productive and profitable than their competitors (Fahmideh and Mahdi, 2018). Scholars have argued that novel machine learning capabilities may realize the predictive value of big data, unchecking its strategic potential to transform business processes and providing the organizational capabilities to tackle key business challenges (K.Sharmila and R.Bhuvan, 2014).

Yet, very few attempts have been made to consolidate the overabundance of BDA research and explore the underlying theoretical foundations. Although some attempts have been made to review and theorize how organizational value can be derived from BDA, these attempts have mostly taken on a narrow information systems and technology. Calls to explore the organizational impact of

BDA from other functional management perspectives remain largely unanswered to date (Al-Shiakhli, 2019).

From the recent earlier years there is exponential growth in the data produced, collected, shared, by different organizations. Such Huge data cannot be managed and processed by the conventional methods is called Big data. The features of big data are volume, variety, and velocity also termed as 3 V's. The various features of big data and its relevance to medical centers' data are discussed below (Hermon, 2014).

**2.2.1 Velocity of health data:** In the manner mentioned above, with the development of technology , huge amount of healthcare data in electronic format is generated by various sources like, clinical reports, patient records, details in social media regarding healthcare benefits, medical images etc. Such data, in structured and unstructured format cannot be processed by traditional database management tools and techniques. Hence, big data analytics are required in healthcare industry to facilitate clinical decision (Berman, 2013).

**2.2.2 Volume:** refers to the incredible quantity of information produced every second through social media, mobile phones, images, videos etc. Traditional database technology and techniques cannot store and analyze such continuous increase in volume of information. Hence, other techniques are require to divide the huge data into parts and store in various locations, perform computation on these parts of data and then analyze by using appropriate software. It is great challenge to collect and analyze such information where in real world a simple click on button produces new pictures, messages in social media (Alemayehu.B, 2019).

Like the growth in volume of other organizations information, healthcare information is also growing day to day in volume. This voluminous information is from healthcare centers, educational institutes, government offices, insurance companies, social media, medical transcripts, medicine research etc (Al-Shiakhli, 2019).

**2.2.3 Variety:** Development of technology facilitated the users to create different types of data. Today, data is not only in the structured format like tables, text, numeric values but also in digital form and unstructured form like photos, videos, social media updates, etc. Big data is the upcoming technology which allows both structured and unstructured data to be gathered, stored, and use it.

In healthcare system most of the data is in the structured form of text, like prescription, medicines names, etc. Other data are in unstructured form pictures, like scanning reports, graphical images, X-Rays, etc. It means variety of data from various sources are generated in medical centers. Such information is collected stored and maintained in computers for further processing and analysis. The big data analytics provides fast process helps the researchers and developers and patient with the required information very quickly these days compared to past few decades, (Sharmila and Bhuvana, 2014).

In general, information must be true, accurate and must be updated with reference time, as the user requires the recent information for decision making. One good example is regarding the bank transaction updating details, updating of GPS data. Such databases must give the accurate results to people. Big data analytics helps to perform computation and analysis on such huge databases like GPS, banking, weather reports, DNA information etc. and give the result which can be trusted by people. Big data analytics results helps the society as well as an individual (Hermon, 2014).

## 2.3 Data analytics approach

Big Data is full of challenges, ranging from the technical to the conceptual to the operational, any of which can derail the ability to discover value and leverage what Big Data is all about. Perhaps it is best to think of Big Data in multidimensional terms, in which four dimensions relate to the primary aspects of Big Data (Singh & Singla, 2018).

According to (Al-Shiakhli, 2019) However, the complexity of Big Data does not end with just four dimensions. There are other factors at work as well, namely the processes that Big Data drives. These processes are a mass of technologies and analytics that are used to define the value of data sources, which translates to actionable elements that move businesses forward. Many of those technologies or concepts are not new but have come to fall under the umbrella of Big Data. Best defined as analysis categories, these technologies and concepts include the following:

**2.3.1 Traditional business intelligence (BI):** This consists of a broad category of applications and technologies for gathering, storing, analyzing, and providing access to data. BI delivers actionable information, which helps enterprise users make better business decisions using fact-based support systems. BI works by using an in depth analysis of detailed business data, provided

by databases, application data, and other tangible data sources. In some circles, BI can provide historical, current, and predictive views of business operations (Shu, 2016).

**2.3.2 Data mining.** This is a process in which data is analyzed from different perspectives and then turned into summary data that are deemed useful. Data mining is normally used with data at rest or with archival data. Data mining techniques focus on modeling and knowledge discovery for predictive, rather than purely descriptive purposes, an ideal process for uncovering new patterns from large data sets (Coral,& Bokelmann, 2017).

**2.3.3 Statistical applications:** look at data, using algorithms based on statistical principles and normally concentrate on data sets related to polls, census, and other static data sets. Statistical applications ideally deliver sample observations that can be used to study populated data sets for the purpose of estimating, testing, and predictive analysis. Empirical data, such as surveys and experimental reporting, are the primary sources for analyzable information (Berman, 2013).

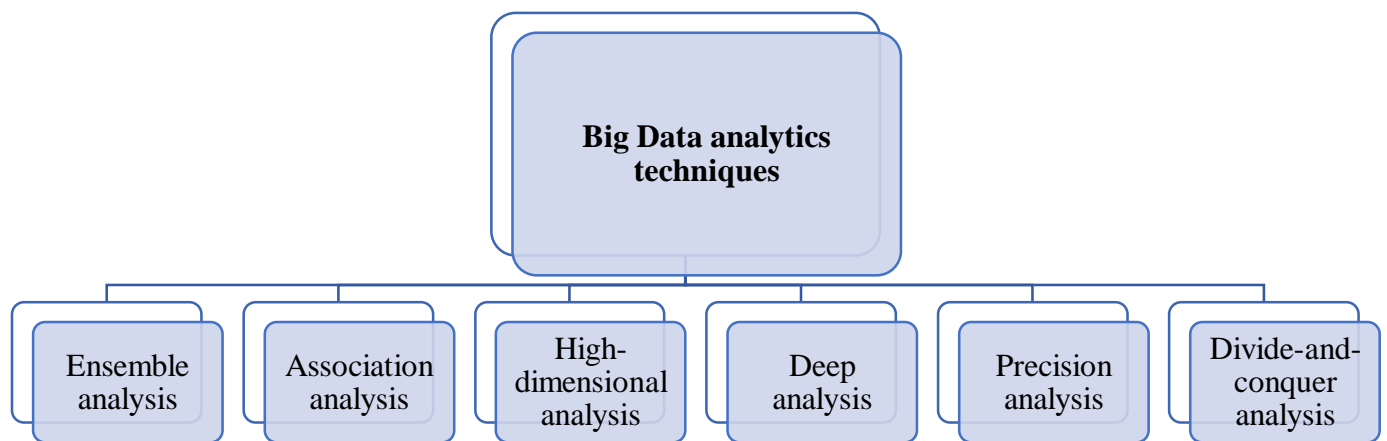
**2.3.4 Predictive analysis.** This is a subset of statistical applications in which data sets are examined to come up with predictions, based on trends and information gleaned from databases. Predictive analysis tends to be big in the financial and scientific worlds, where trending tends to drive predictions, once external elements are added to the data set. One of the main goals of predictive analysis is to identify the risks and opportunities for business process, markets, and manufacturing (Desalegn, 2016).

**2.3.5 Data modeling:** is conceptual application of analytics in which multiple “what-if” scenarios can be functional via algorithms to multiple data sets. Ideally, the modeled information changes based on the information made available to the algorithms, which then provide insight to the effects of the change on the data sets. Data modeling works hand in hand with data visualization, in which uncovering information can help with a particular business endeavor. The preceding analysis categories constitute only a portion of where Big Data is headed and why it has intrinsic value to business (Wang, & Wu, 2018).

That value is determined by the never-ending mission for a competitive advantage, encouraging organizations to turn to large repositories of corporate and external data to uncover trends, statistics, and other actionable information to help them decide on their next move. This has



helped the concept of Big Data to gain popularity with technologists and executives alike, along with its associated tools, platforms, and analytics (Vogl & Puri, 2004). Figure 2.1: Big data analytics techniques cycle below depicts Big Data analytics techniques followed by the discretion of each techniques.



**Figure 2.1: Big data analytics techniques cycle**

### 2.3.1 Ensemble analysis

Collaborative data analysis, roughly termed multi-data-set analysis or multi-algorithm analysis, is made for the whole dataset or a large volume of data. Big data is argued to be the whole dataset without any sampling purpose. What is the whole set? Approximately, it may be resampling data, labeled data and unlabeled data, prior data and posterior data. It is known that the term “ensemble” appears at least in the context of ensemble learning in machine learning, the ensemble system in

statistics mechanics, and ensemble Kalman filtering in data assimilation (Fahmideh and Mahdi, 2018).

### 2.3.2 Association analysis

Typically, big data are collected without different sampling strategies. Normally, data producers are quite different from data users, so that the cause-effect relation hidden in observation data, this is not clear to specific data users. Set theory, i.e. the theory about members and their relations in a set, is general sufficient to deal with data analysis and problem solving. In a sense, the relation among set members corresponds to the association in big data. Association analysis is critical to multi-sourcing, multi-type, and multi-domain data analysis. Typically, association analysis is exemplified with association rule algorithms in data mining (I. H & Frank, 2002) data association in target tracking, and links analysis in networks.

### 2.3.3 High-dimensional analysis

The 3<sup>rd</sup> analytic technics is measurement of an object is an intrinsic property independent of the space in which the object is embedded. In reality, the dimension is the number of perspectives from which the real world is recognized by big data analysis suggests a high variety of data. In mathematics, the dimension of a mathematical space (object) is informally defined as the minimum number of coordinates needed to specify any point within it, and the dimension of a vector space is the number of vectors in any basis for the space or the number of coordinates necessary to specify any vector. Classical physics describe the real world in three dimensions. Starting from a particular point in space, the basic directions in which we can move are up/down, left/right, and forward/backward (Shu, 2016).

The equations of classical mechanics are symmetric with respect to time. In thermodynamics, time is directional with a reference to the second law of thermodynamics, which states that an isolated system's entropy never decreases. Such a system spontaneously evolves toward thermodynamic equilibrium of the states with maximum entropy. In statistics and econometrics, there exists the multi-dimensional analysis for a sample (realization) of multiple random variables. High-dimensional statistics studies data whose dimension is larger than those dimensions in classical multivariate analysis. In many applications, even the dimension of data vectors may be larger than sample size (Singh & Singla, 2018).

### 2.3.4 Deep analysis

Today, the volume of developing data is large enough for complex artificial neural networks training. In the meantime, the high-performance computing technologies of multi cores of CPU, GPU (Graphic Processing Units), and FPGA (Field Programmable Gate Array) greatly reduce the training time of complex artificial neural networks. Under such circumstances, traditional artificial neural networks are functionally enhanced with hidden layers of latent variables, and so called deep learning is developed as compared to shallow learning. It is also believed that human cognition about the real world is getting deeper as our experience is getting rich. Deep analysis is potential in exploring complex structure properties of big data. Unobservable variables, hidden parameters, hidden hierarchies, local correlations, and the complex distribution of random variables can be found through deep analysis (Shu, 2016).

### 2.3.5 Precision analysis

In numerical analysis, accuracy is the nearness of a calculation to the true value, while precision is the resolution of the representation, often defined by the number of decimal or binary digits. Statistics prefers to use the bias and variability instead of the accuracy and precision. Bias is the amount of inaccuracy, and variability is the amount of imprecision. In practice, trueness is the closeness of the mean of a set of measurement results to the actual (true) value, and precision is the closeness of agreement among a set of results. Ideally, a measurement device is both accurate and precise, with measurements all close to and tightly clustered around the true value (Shu, 2016).

A measurement system can be accurate but not precise, precise but not accurate, neither, or both. When an experiment contains a systematic error, increasing the sample size generally increases precision but does not improve accuracy. On the other hand eliminating the systematic error improves accuracy but does not change precision.

### 2.3.6 Divide and conquer analysis

Divide-and-conquer is a general computational strategy for improving the efficiency of problem-solving and the velocity of big data computation. Through divide-and conquer analysis, a problem is recursively broken down into two or more sub-problems in the stage of dividing, until these

become simple enough to be solved directly in the stage of conquering. Upon completion, the solutions to sub-problems are combined into a solution to the original problem (Shu, 2016).

In a sense, distributed computing, such as cloud computing and distributed intelligence computing, may be considered as the computation of divide and conquer in space, and parallel computing (multi-core computing and cluster computing) may be considered as the computation of divide and conquer in time. Stream processing and real time computing somewhat are the computation of divide and conquer from the perspectives of unstructured data and time-constraints. Cloud computing and distributed intelligence computing somewhat are the computation of divide-and-conquer from the perspectives of virtualized resources and the society of mind (Alemayehu.B, 2019).

These six techniques in big data analytics are general for any data analysis, but patio temporal association's analysis is special for geographical data. Just like the four characteristics of big data given by the industrial community, six techniques in big data analytics are given in experience, and six techniques in big data analytics are not mutually exclusive and collectively exhaustive in theory ( Sowmya and Sravanthi, 2017).

## 2.4 VISUALIZATION TECHNIQUES

Visualization is the use of computer-supported, visual representation of data. Unlike static data visualization, interactive data visualization allows users to specify the format used in displaying data. Common visualization techniques are:

- *Line graph*: This shows the relationship between items. It can be used to compare changes over a period of time.
- *Bar chart*: This is used to compare quantities of different categories.
- *Scatter plot*: This is a two-dimensional plot showing variation of two items.
- *Pie chart*: This is used to compare the parts of a whole.

Thus, the format of graphs and charts can take the form of bar chart, pie chart, line graph, etc. It is important to understand which chart or graph to use for your data. Data visualization uses computer graphics to show patterns, trends, and relationship among elements of the data. It can generate pie charts, bar charts, scatter plots, and other types of data graphs with simple pull-down menus and

mouse clicks. Colors are carefully selected for certain types of visualization. When color is used to represent data, we must choose effective colors to differentiate between data elements. In data visualization, data is abstracted and summarized. Spatial variables such as position, size, and shape represent key elements in the data. A visualization system should perform a data reduction, transform and project the original dataset on a screen. It should visualize results in the form of charts and graphs and present results in user friendly way

## 2.5 Tools and Framework

### 2.5.1 Hadoop

Hadoop was produced by Doug Cutting (who also developed apache lucene) for ascendable, reliable distributed processing. Hadoop mainly consists of Hadoop distribute file system (HDFS) and Map Reduce. The Hadoop distributed file system is used to store and process the large amount of data in distributed manner. To take advantage of the parallel processing that Hadoop provides, there is a need to express query as a MapReduce job. MapReduce works by breaking the processing into two phases: the map phase and the reduce phase. Each phase has key-value pairs as input and output, the types of which may be chosen by the programmer (Raste, 2014).

The HDFS is a fault-tolerant storage system that can store huge amounts of information, scale up incrementally and survive storage failure without losing data. HDFS manages storage on the cluster by breaking files into small blocks and storing duplicated copies of them across the pool of nodes (commodity hardware/system). HDFS allows more than 1000 nodes by a single operator. HDFS offers two key advantages; Firstly, HDFS requires no special hardware as it can be built from common hardware. Secondly, it enables an efficient technique of data processing in the form of MapReduce, Collectively, multiple number of Nodes is called as ‘Racks’ and multiple number of racks are called as “clusters”.

### 2.5.2 MapReduce:

MapReduce is a data processing algorithm that uses a parallel programming implementation. It is a programming paradigm that involves distributing a task across multiple nodes by running a "map" function. The map function takes the problem, splits it into sub-parts and sends them to

different machines so that all the sub-parts can run concurrently. The results from the parallel map functions are then collected and distributed to a set of servers running "reduce" functions, which takes the results from the sub-parts and re-combines them to get the single result (Output). Making it simpler, there is a chain of inputs and outputs. The data that is to be processed becomes input for the map function (Sawant, & Shah, 2013).

### 2.5.3 Hadoop YARN

This is a framework for job scheduling and cluster resource management. YARN has the master/worker architecture. The master (Resource manager) manages all the resources on the workers and schedules work in the cluster. Furthermore, the resource manager handles all the client interactions. Hadoop architecture framework is depicted in figure 2.2.

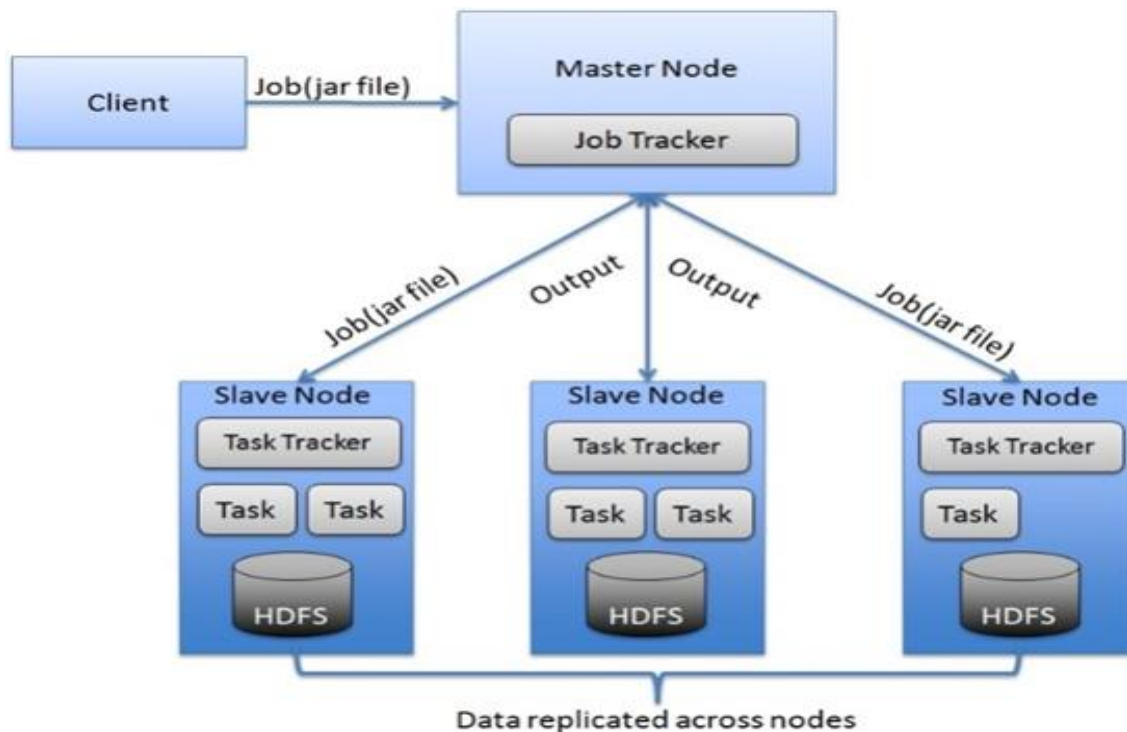


Figure 2. 1: Hadoop Architecture Framework

## 2.6 Challenges of Big data analytics

Flowing of data from different sources is accumulating large volume of data with variety of forms at high speed and velocity. Big data is different from “lots of data” or “massive data” in the way that it should incorporate all 3 V’s. (Volume, variety and velocity) in order to be treated as big

data. It is not analyzed in its totality and it is required to pass through multiple steps like data extraction, data filtration, data transformation and data analysis. The difference between small and big data is goal, location, data structure and content, data preparation, longevity, measurement, reproducibility, stakes, introspection and analysis. These dimensions help to extricate big data from small data (Cukier, 2014).

Sources of data are Call logs, mobile-banking transactions, online user generated content such as blog posts and tweets, online searches, satellite images and so on. Insight from big data narrows gap between information and time. Big data is industrial revolution of data. Data is taken as raw good with little “intent and capacity” (Leetaru, 2015). However, big data is not without challenges as it is under development stage. It has data challenges (Volume, Velocity, and Variety), process challenge (display complex analytics on mobile devices) and management challenges (data privacy, security, governance, and ethics) (Akerkar, 2014). On the other hand, benefits of big data which are by far overwhelming such as its application to medicine (e.g. flu trend analysis by Google), climate change, food safety, science, business, technology, manufacturing, financial markets, cyber security, etc.

Business firms that use big data are apart from traditional analytics implemented in a way that by focusing on data flow, depending on data scientists and process and product developers rather than data analysts, and analytics is become part of core business. Commonly used tools like Databases, Apache Hadoop Framework or Internet and Storage System provides capabilities to load, store and query large datasets in near real time. Moreover, it executes advanced analytics which is developed under information ecosystem. Big data analytics tools are considered as next generation of Information Technology processes and systems that is designed for insight but not just for automation (Thomas, 2012).

In most organizations, information is taken as boarding spring of successful business activities and similarly they give full attention for every drop of information by considering it as life blood of business activities as product or service to be sold to customers. So they are exercising information management practices as a means of managing available information for innovation and decision making. However, current trend of data overwhelm from inside as well as outside of organization is creating burden on processing capability of data as usual (Berman, 2013).

The era of big data puts thinking of data storage rather than value creation from stored datasets to deliver innovative solutions and at the same time coping changing environment in the way that enhances organization's competitive position in an industry. Even to the existent, some organizations are focusing their effort only on lighting of operation of current activities rather than enabling business as well as differentiating their services or products. Big data technologies, especially open source framework such as Apache Hadoop Framework, are creating conducive environment for processing large volume of data with low cost and high speed and it has capability to scale out as capacity of the organization grows to accommodate more data sources. It scales horizontally without need of rework in order to scale up processing capability of already in placed system (Berman, 2013).

In near real time, data is required to be processed to extract insight and achieve tangible time value of data in transit. More data does not allow us to see more rather it allows us to see new, better and different. Big data as resource or tool helps to advance society; moreover, it supports to address recurring global challenges such as energy, environment, drought, poverty and so on (Singh & Singla, 2018).

The new knowledge, tools, practices, and infrastructures produced will enable breakthrough discoveries and innovation in science, engineering, medicine, commerce, education, and national security. However, finding and using standards and measurement for big data analysis is limited due to its infant stage; so big data analysts, especially data scientists are developing a set of strategies as well as tools to align data with meaning and reality. On the other hand, in small data defining "control" is common practice; practically, groups are divided into control and test groups. But, defining "control" group in big data is impractical because data analysts have no controls over big data. In addition, experiment results are difficult to repeat with given population (Berman, 2013).

Furthermore, retesting results pass through excited and long path that require a lot of resources and time. Finally, confirmation of the result could not be as expected due to a number of factors. As a matter of fact, big data projects are done or processed without help of statistical or analytical software packages. On the contrary, human beings are better in processing large information, organizing and visualizing it as appropriate. For instance, (Domenico, 2016) "we humans have a



long-term memory capacity in the petabyte range and that we process many thousands of thoughts each day. In addition, we obtain new information continuously and rapidly in many formats (visual, auditory, olfactory, proprioceptive, and gustatory)”.

## 2.7 Big Data in Healthcare

Medical data are collected from the patient continuously which will lead to store a significant amount of data and could be difficult to handle using the existing technology. The reason is the collected data are unstructured, semi structured or sometimes may be structured. Before we analyze those medical data, we need to use the structured database to avoid some unnecessary errors which can be lead to using unstructured, semi-structured dataset.

MapReduce is the perfect to convert the open, semi-structured dataset to the structured dataset. MapReduce can deal with the massive amount of unstructured, semi-structured data and can turn those data to structure productively and a faster approach also MapReduce can handle the error very carefully. At the very recent researcher’s research about the extended version of MapReduce as i2MapReduce for the iterative computation for a significant amount of data. Some research about the task level adaptive MapReduce framework which extends MapReduce architecture by redesigning the Map and Reduce task. There are also some research works about the MapReduce, and they proposed a routing algorithm known as a joint scheduler which can improve the throughput and delay performance according to Hadoop traditional fair scheduler (I.Czarnowski and Piotr J, 2018).

## 2.8 Related Works

According M. Maier (2013) explored in spring of data sources are stretching limits of traditional data management so as to extract unused sources to gain more understandings. To realize these values, organizations need to consider architectural expansion to accommodate new technologies on top of traditional architecture. According to the research, additional requirements have to be elicited on the basis of new data behavior to design reference architecture by combining several data management components. The reference architecture is built on traditional enterprise data warehouse architecture using evolutionary approach (M. Maier, 2013).

As a result shows, Hadoop ecosystem provides platform to process unstructured data sets of Big Data in cheap, fault tolerant and high speed. The achievement of the study expounds next generation of IT in areas of data storage, processing and visualization. Especially, reliability and computational power does not need scale up in terms of hardware and processor capacities. Therefore, Big Data processing and visualization challenges are able to handle using software solutions rather than in placing specialized machines with increased hardware and processing capabilities (Desalegn, 2016).

According to Alemayehu (2019), Big Data can be used for other domains that include classification problems by making some adjustments to the multi-tier predictive model and by using of various context specific dictionaries. Big data analytics has the potential to transform the way healthcare providers use sophisticated technologies to gain insight from their clinical and other data repositories and make informed decisions. As big data analytics develops more mainstream, issues such as guaranteeing privacy, safeguarding security, establishing standards and governance, and continually improving the tools and technologies will garner attention.

Alemayehu (2019) stated that, a number of break-through approaches have emerged to address the challenges in managing, modelling and analyzing Big Data. According to this researcher, With respect to data management, the most popular current approaches employ a form of ‘divide-and-conquer’ or ‘divide-and recombine’ in which subsets of the data are analyzed in parallel by different processors and the results are then combined. Similar approaches have also been promoted, such as ‘consensus Monte Carlo’ and ‘bag of little bootstraps’, while others have studied the properties of Markov chain Monte Carlo (MCMC) subsampling algorithms. The researcher employed different techniques for analyzing and predicting cancer disease from big data preserved in health care centers. The researcher recommended that, the Hadoop platform should be further directed to improve its performance and efficiency.

## CHAPTER THREE

# METHODOLOGY

The purpose of this section is to present the methods and procedures employed in carrying out the study. Attempt will be made to describe step by step, the methods, procedures and other devices that will be used to collect, analyze and design system.

### 3.1.1 Literature review

Literature review was conducted from books, journals, conference procedures and the Internet in order to gain deeper understanding of Big Data landscape and its value proposition to society at different level. It gives the spot of current data challenges as well as its application in wide areas.

### 3.1.2 Development and Processing Tools

Apache Hadoop Framework, Open source framework, and R-studio tool is mostly customized for academic purpose and it is Hadoop Distributed File System (HDFS) which alleviates current processors limitation that processing capacity is at its maximum point. Apache Hadoop Framework provides a platform so that data sets passes a number of phases from input or raw data stage to output or result stage. Hadoop is open source platform which is used for data storage and processing of very large volumes of data at high speed with low costs. It is possible to build large scale distributed data processing system using commodity computers that lowers cost of computation source. It is also possible to run Hadoop on single desktop or laptop for testing. For this study R-studio tool was employed.

### 3.1.3 Visualization Tools

After processing data sets, the next step is converting the output or result of processing as input into visualization tools. Visualization is key companion of big data processing so that the result of huge data set processing can easily be grasped by experts as well as others. Without using visualization tools, it requires a lot of effort and time to comprehend output of big data processing. There are few big data visualization tools that are powerful and with a capability of accommodating vast data elements within a single screen (Singh & Singla, 2018).

### 3.1.4 Evaluation methods

The results of the study are evaluated with major three dimensions; namely execution time, memory requirement and presentation. Firstly, the output of investigation is evaluated in terms of time of execution that has taken to ingest, process and produce result. Secondly, memory operation to process from client command to output file. Lastly, the output of experiment is required to fit for consumption by audiences so its easiness for presentation is taken in account. The result of evaluation is expected to show better parameters value in comparison with traditional data warehouse.

## 3.2 Data collection

Medical data is a type of diversified data that cannot be forced to align a certain format or confront standards and practice of an organization. In addition, in a big data scenario, Data has to be connected with other data sets to be most valuable to produce accurate insight. A set of predominant chronic diseases medical data (structured and semi-structure unstructured) to use the analytical system and proposed system finely support structured data.

The data was in .csv format, made it more structured so as to check the reliability of Hadoop when used on datasets. In addition, in a big data situation, data is short lived in terms of value that could be extracted from it for decision making or actions. Data has to be connected with other data sets to be most valuable to harvest accurate insight. Moreover, data analytics research projects deal with behavioral aspect of data rather than pursuing veracity of data. In this study, structured data sets were used for developing analytical prototype.

The researcher has collected two different sets of patient's data for simulation to measure hypertension data and uterine contraction are considered with less hidden value and it has different features which are used to find the categorical outcome. The algorithm analyze and visualize a patient's condition in categories. The diabetic's dataset represents the more hidden state and dissimilarity among the attributes with eight features are used to identify the binary outcome which predicts whether a patient has diabetics or not. Datasets are collected from Jimma Medical center and available for public use.

### 3.3 Data Pre-Processing:

It is relatively common that the raw incoming data is often unclean in the sense that it contains nonstandard, unwanted or redundant values. In this case, data cleansing and transformation steps must be performed to prepare the data before they enter the data storage. Failures at this step will result in calculation errors at a later stage or data storage bloat with useless data.

This unit goals at detailing a thorough list of contributions on Big Data preprocessing. Classifies these contributions according to the category of data preprocessing, number of features, number of instances, maximum data size managed by each algorithm and the framework was developed. The size has been computed multiplying the total number features by the number of instances. For sparse methods, only the non-sparse cells have been considered. Data was preprocessed so that it was prepared for further analysis.

#### 3.3.1 Data Cleaning

Data documented as incomplete, wrong or unreasonable data, should be modified or deleted for improving data quality (Coral,& Bokelmann, 2017). The chronic diseases data stored in the statics department so, the multisource and multimodal nature of healthcare data results in high complexity and noise problems. In addition, there are also problems of missing values and impurity in the high-volume data. Since data quality determines information quality, which will eventually affect the decision-making process, it is critical to develop efficient big data cleansing approaches to improve data quality for making accurate and effective decisions (Wang, 2017).

A lost value for a variable is one that has not been entered into a dataset, but an actual value exists. In simple charge, missing values in a variable are replaced with a single value. However, simple imputation produces biased results for data that aren't Missing Completely At Random (MCAR). If there are moderate to large amounts of missing data, simple imputation is likely to underestimate standard errors, distort correlations among variables, and produce incorrect p-values in statistical tests. This approach should be avoided for most missing data problems (Croninger, & Douglas, K, 2005).

The study of the linear correlations enabled to fill in some new unknown values. In order to handle a dataset with missing values, most common can be followed strategies which include:

1) Remove the cases with unknowns; 2) fill in the unknown values by exploring the similarity between cases; 3) add in the unknown values by exploring the correlations between variables; and 4) custom tools that are able to handle these values (Shu, 2016).

A database also holds irrelevant attributes. Therefore, relevance analysis in the form of correlation analysis and attribute subset selection can be used to detect attributes that do not contribute to the classification or prediction task. Including such attributes may otherwise slow down and possibly mislead the learning step. Typically, data cleaning and data integration are performed as a pre-processing step. Inconsistencies in attribute or dimension naming can cause redundancies in the resulting dataset. Data cleaning was done by following the above procedures in order to increase the quality of data.

### 3.3.2 Data Integration

This stage involves integrating and transforming data into an appropriate format for subsequent data analysis. However, health Big Data are unbelievably large, distributed, unstructured and heterogeneous, making integration and transformation all the more problematic. Integrating unstructured data is a major challenge for BDA. Even with structured data integration there are many issues (Bisandu, 2016).

Challenges: functional, metadata and instance In the process of data integration or aggregation, datasets are matched and complex on the basis of shared variables and attributes. Advanced data processing and analysis techniques allow to mix both structured and unstructured data for eliciting new insights; however, this requires “clean” data. Data fusion techniques are used to match and aggregate heterogeneous datasets for creating or enhancing a representation of reality that helps data mining. Mid-level data fusion methodologies that merge structured and machine-produced data basically work well. On the other hand, high level data fusion tasks for merging multiple unstructured analogue sensor inputs remains challenging (Desalegn, 2016).

Metadata is structured data that describes the characteristics of a resource. In the relational database model, the column names are used as metadata to describe the characteristics of the stored data. There are two major problems in metadata integration. First, different database systems use different metadata to describe content. For example, one system might use ‘sex’ while another might use ‘gender’ in referring to a patient. A computer does not recognize that ‘sex’ and ‘gender’ are semantically similar. Second, there are problems in mapping simple metadata to composite

metadata. For example, a computer cannot automatically map a metadata ‘Patient Name’ in one system into composite metadata ‘First Name’ + ‘Last Name’ in the other system.

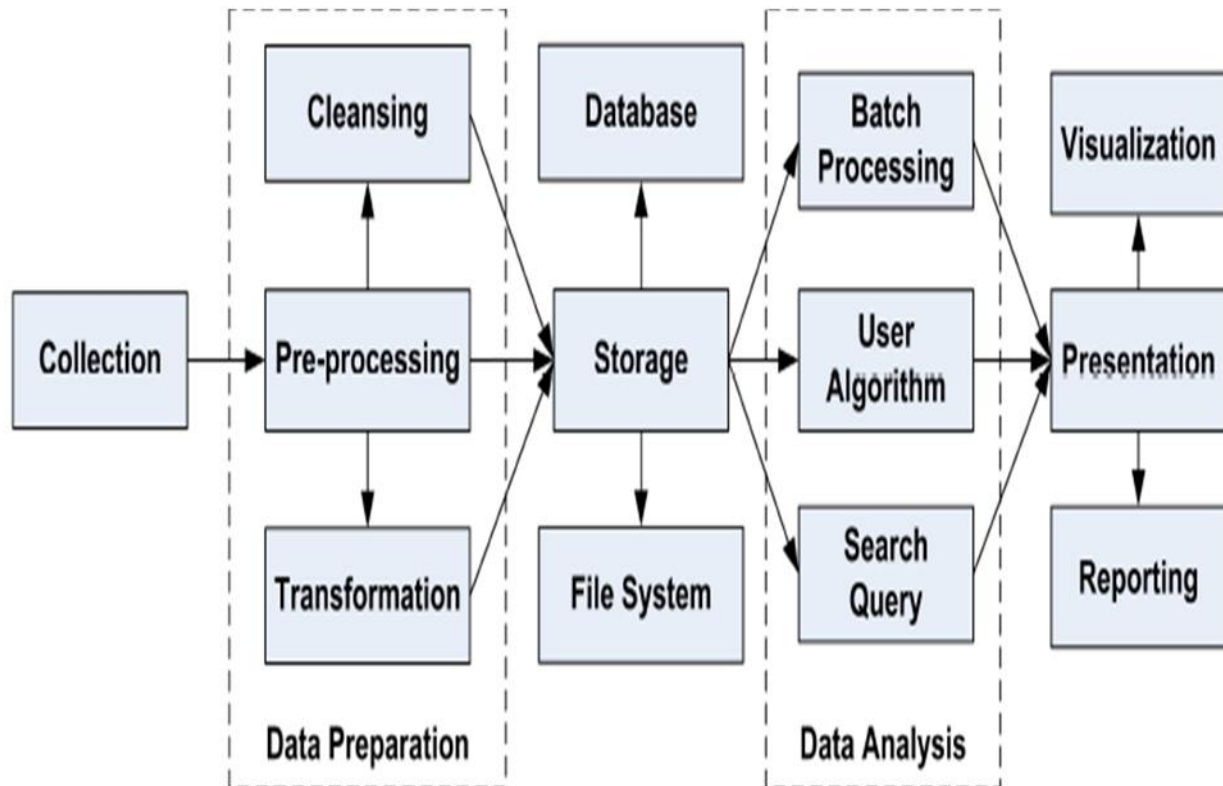


Figure 3.4 Big data pre-processing stage (I.Czarnowski and Piotr J, 2018)

### 3.4 Classifier Selection and Parameter Modification:

Descriptive analytics (DA) more of statistics is the term given to the analysis of data that helps describe, show or summarize data in a meaningful way. Its benefits include a strong statistical foundation and a probabilistic model useful for explaining the data. There is a perception that DA is slow, unstable, and unsuitable for large learning or classification tasks.

Through fast estimated numerical methods, regularization to avoid numerical instability, and an efficient implementation, it would show that DA can outperform modern algorithms like Support Vector Machines (SVM) on a variety of learning tasks. These new implementation, which uses a

modified iteratively re-weighted least squares estimation procedure, can compute model parameters for sparse binary datasets with hundreds of thousands of rows and attributes, and millions or tens of millions of nonzero elements in just a few seconds (K.Sharmila and R.Bhuvan, 2014).

Descriptive analysis provides the early detection of a disease which can minimize the cost of hospitalization and reduce the mortality. A proper estimation of early disease prediction can provide an accurate analysis and fast action facility to a diseases. Thus, the researcher build analytical model pipeline to analyze and describe disease more accurately. Classifier was selected based up on scholar's recommendation and parameters of the datasets were modified.

## 3.5 Design Model:

### 3.5.1 System Architecture

The system compute and analyze the chronic diseases data based on the big data uploaded. Those data will be sent to the High-Performance Computer system for data aggregation and primary analysis. In the system, the data collecting phase or data layer of data analyze scheme where all data are being gathered.

The next layer is data aggregation layer where some programming model such as SQL framework will be applicable for creating a structured database with some log file and metadata. Those metadata and database will make a good user interaction in data management. After creating the log file and database, those are integrated and sent to another analytical layer high-performance computer system which we called a data warehouse.

The log file is used as metadata for providing fast searching capability and analysis. In application layer, data warehouse sends the database to the data center for data storage which is located in the database. Database gives the facility of fast computing to diagnose a disease and predict the future disease and enables many more information and knowledge's by visualization tools. The system architecture is depicted in figure 3.2



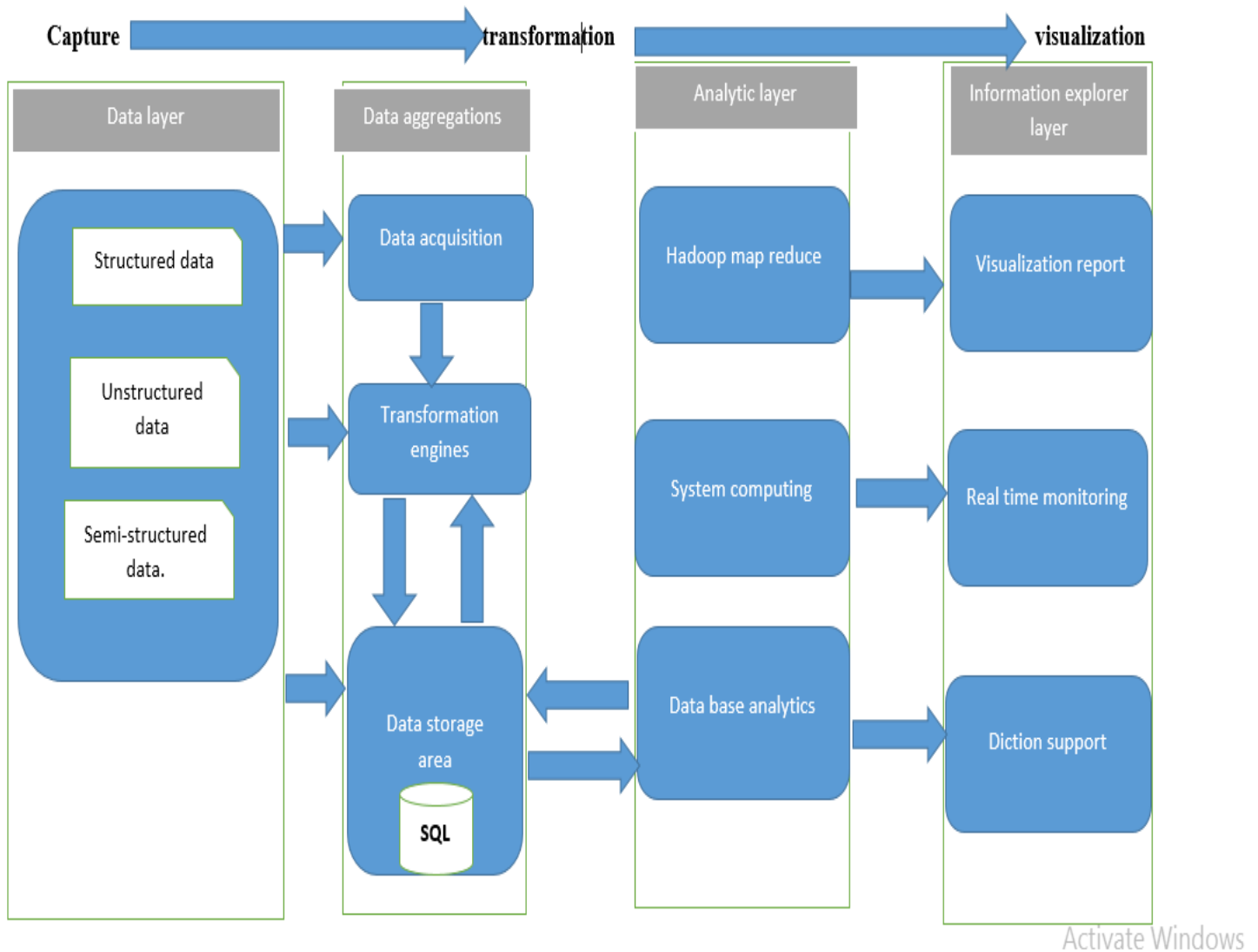


Figure 3.5 System Architecture

As showed in the above figure 3.2, there are three main phases of the prototype. The capture phase, data transformation phase and visualization phase are the broad phases of the prototype. Under each phases there are different tasks performed and to be done by the system. Under capture phase three types of data were collected (structured, semi-structured and unstructured data). Transformation phase includes data aggregation and analytical layer. Data were acquired, transformed and stored in organized way after capturing from the medical center. Then data could

be analyzed under analytical phase and report in different format is visualized by the prototype. This is discussed in detail below

#### 3.5.1.1 Data Layer

Healthcare data has multiple sources such as structured and unstructured different types of file format. At the same time, the healthcare data is collected with different formats such as structured data, semi-structured data and unstructured data, which lead to the challenges for data collection and preprocessing.

The Data Layer of the future of architecture has goals to offer services facilitating data collection and pre-processing for compatible health records, back-end data of healthcare system and flowing data generated by wearable medical sensors.

#### 3.5.1.2 Data Aggregation Layer

The main tasks of the Data Aggregation Layer include data extraction, transformation and loading into storage system. With the support from Data Layer, necessary operations including data moving, cleaning, splitting, translation, merging, and sorting can be performed. Afterward, the healthcare big data with standard format can be loaded into a storage system which may be relational databases, NO-SQL databases, distributed file systems and etc.

#### 3.5.1.3 Analytics Layer

Under this study, with the support from the Data Aggregation Layer, the Analytic Layer focuses on basic statistical analysis work. Usually, work on this layer includes On-Line massive healthcare data analytical processing, streaming data processing, database construction & optimization, indexing, etc

#### 3.5.1.4 Information Exploration Layer

This layer consists of visualization/reporting, real-time monitoring and clinical decision support. As we know, the healthcare big data could be massive and complex, which makes it difficult to understand and observe. Therefore, powerful techniques for efficiently visualizing and summarizing the healthcare big data become vital which means for patients, t analysis results not only about the historical data but also the current signs are vital.

For this purpose, real-time monitoring based on transient vital signs of patients is needed. Thanks to the recent development of big data technologies, there is a way to enable real-time monitoring by utilizing streaming-like techniques. Besides, according to the further investigation on historical clinical data, it is feasible to provide better clinical decision support for doctors. So far, some

artificial intelligent algorithms such as Bayesian model, logistic regression model, decision tree, support vector machine, random forest and others can be integrated with domain knowledge for clinical decision purposes.

### 3.5.2 Framework of analytics system

As said previous in this paper, the main goal was to develop a good descriptive analysis. According to some research, the process of data preprocessing is considered the most crucial phase in the whole data analytics process and it can take more than a half of a total time consumed in solving the data mining problems. That is the reason why more consideration and effort was done to build a rich dataset that can be used afterwards in order to gather useful information and gain knowledge from Big Data analytical system. The framework of the system is depicted in figure

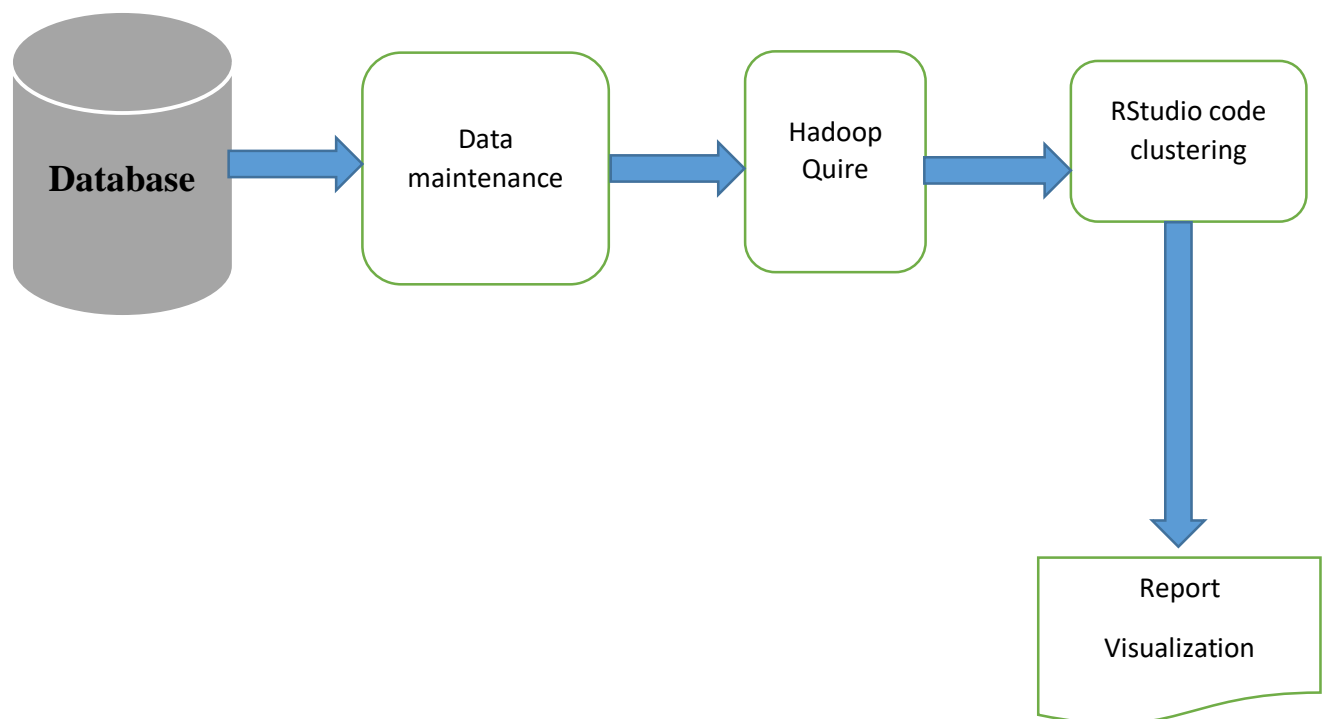


Figure 3 6: Framework of analytics system

### 3.5.3 Architecture of Data Warehouse

Proposed interactive decretive system holds the structured database in the data warehouse and enables more functionality. After analysis of data in stored them in HDFS, the database with metadata are sent to the data warehouse. Data warehouse keeps the database with the modification of metadata and sends the database to the cloud server for fast computation facility with higher accuracy with privacy.

The data warehouse provides a two-way interaction between Hadoop and Rstudio tool. The primary purpose of a data warehouse is to keep track of all the data in the database with metadata along with log files. It will make an interactive environment for the policy and decision developer. The user can search and get the information quickly from the data warehouse, even if they are not familiar with any types programming languages. This warehouse can able to take user's question and able to convert the question as the query language. Thus, it provides the fast searching capabilities.

The architecture of data warehouse holds two unit: metadata with log files and structured chronic database. Nodes are work under metadata to do faster search and access the chronic database. Thus, metadata offers the exploratory information extraction through its log file system and distributed database. The top unit of metadata holds the knowledge graph which maps with information base and log file.

The log file contains some intent keyword such as age, sex, etc. and the knowledge of database table attribute and the relationship of every table to one another. The other part of a data warehouse is a database which makes the communication between the cloud servers. When a user wants to search something, data warehouse takes the question and convert into the query with the convenient word and make a map with information base and knowledge graph and send to the cluster node.

### 3.5.3 Design and Goal of the system

Applying big data technology masses require planning and management works to be performed. As number of clusters increases, clusters is organized into racks and number of nodes in a cluster depends on file size and storage space of each node. Complexity of design for Hadoop ecosystem is the right relative to file size to be stored, storage space of nodes, size of clusters and number of frames to group number of clusters. Even though Master nodes and Job tracker nodes are not part

of clusters as well as racks, they play a major role in controlling, managing, scheduling jobs etc. of all clusters and frames being outside of both clusters and frames.

The bigger file size that decreases the number of nodes in a cluster which directly affects the efficiency of Map task processing capability by slowing because total file breakdown into chunks become heavy. On the contrary, as file size becomes smaller which is least size of 64 MBs the number of nodes in a cluster will be higher. Similarly, performance of Map task becomes faster than larger file sizes; however, Reduce task gets much load in processing or aggregating data from a number of nodes (Sawant, & Shah, 2013).

The goal of design is developing Hadoop ecosystem environment to test big data technology stack on unstructured and structured data sets processing so as to differentiate from traditional relational data technology stack. In doing so, test environment will be setup using Apache Hadoop library on the top of windows operating system which is natural operation platform for Hadoop ecosystem. In general, the main approach that is employed to conduct this research is experimenting a set of big data technology stack by performing data processing and visualization.

## 5.6 Data Visualization Design

If vast set of data produces submerge information to make decisions, without data visualizations tools the values of data would be difficult to be realized (J. G. Wolff, 2014.). The trouble of seizing out meaning or insight from big data will increase by many folds without data visualization tools in place. So, the importance of data visualization is similar hand and glove for big data situations because it does not make sense just long processing data for the sake of analysis.

Data visualization is dependent on a number of factors to be effective which could hinders its utilization as well as applicability for desired purpose unless properly considered in detail at the time of design (K. B. Carter, 2014).

The main parts of data visualization elements are screen size, screen resolution, data nature and machine capacity. Screen size creates a room to accommodate more data elements and at the time it creates comfort to explore as well as navigate. Screen resolution, on the other hand, provides an ability to see or visualize clearly all data sets in terms of inter data elements and intra data elements as well. Data nature puts burden for visualization tools as it is more unstructured, variety in type and huge in amount. Finally, machine capacity plays major role in processing and presenting to end user. In a nut shell, data visualization constraints can be expressed in terms of a formula below.

$$\text{MaxRequestSize}=30*1024^2$$

In this study, data visualization is considering all the above factors to ensure data presentation through a computers (desktop and laptop). As it is well known that smart devices have limited capabilities in terms of screen size and machine capacity; however, data visualization is encompassing this limitation by adding interactivity by hovering cursor over data elements. On the additional word, computers are defector standards for any data visualization design.

## 3.7 Algorithms

### 3.7.1 Decision Tree Classifiers

Decision tree is a well-known managed learning method used for classification and regression. A decision tree of a pair  $(x; y)$  denotes a function which takes the input attribute  $x$  (Boolean, discrete, continuous) and outputs a simple Boolean  $y$ . This is basically a model which is used to map the observations regarding an item to conclusion about the item's target value. This can be used to visually and explicitly represent decisions. The ultimate goal of this method is analyzing and describing the value of a target variable based on simple decision rules concluded from the data features (Hind Bangui, 2018).

### 3.7.2 Clustering Algorithms:

Clustering is a popular concept which group's organization of unlabeled data based on similarity. So as a result, similar kind of data belongs to one group and other reside in another group. There are mainly three types of clustering algorithms available, out of which K-means is the most widely used technique.

The most widely used partition algorithm is the iterative K-means approach. The objective function that the K-means optimizes is hence, the K-means algorithm minimizes the intra-cluster distance (Omran, 2007).

$$J_{\text{K-means}} = \sum_{k=1}^K \sum_{\forall z_p \in C_k} d^2(z_p, m_k)$$

The K-means algorithm starts with K centroids (initial values for the centroids are randomly selected or derived from a priori information). Then, each pattern in the data set is assigned to the closest cluster (i.e. closest centroid). Finally, the centroids are recalculated according to the associated patterns. This process is repeated until convergence is achieved as depicted in figure 3.4.

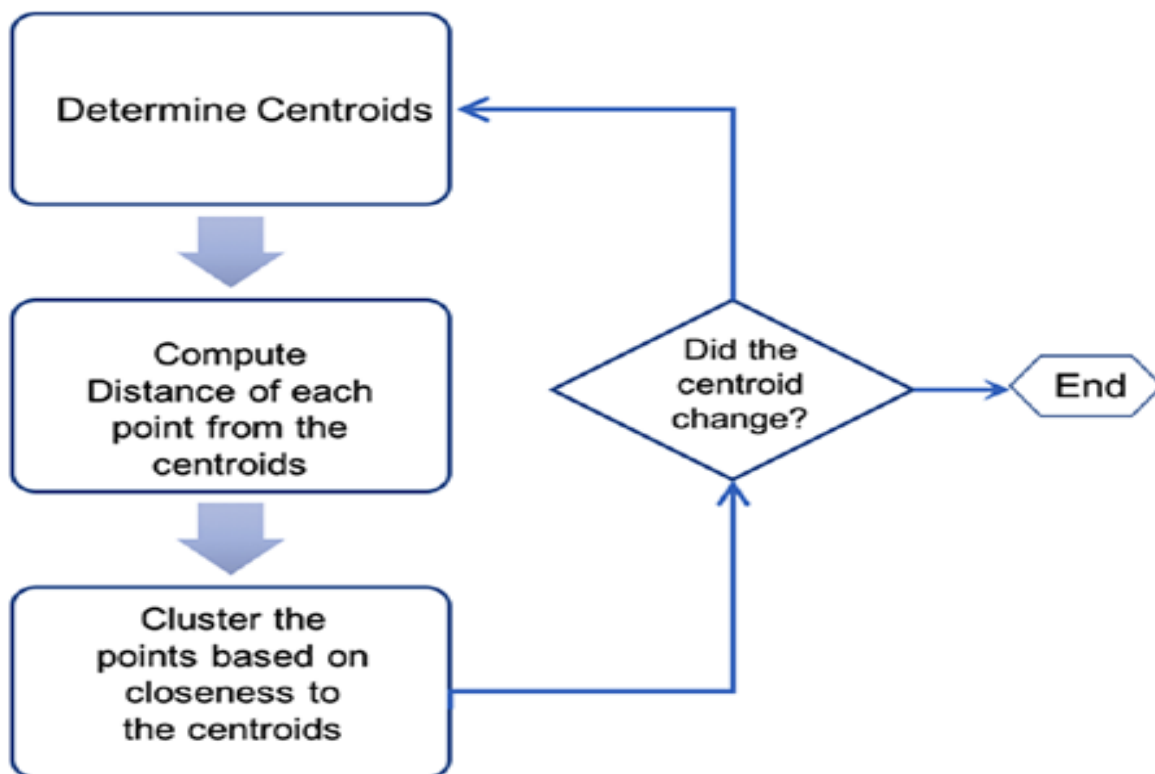


Figure 3.7 Clustering Algorithms

Hence, K-means has a hard membership function. Furthermore, K-means has a constant weight function, thus, all patterns have equal importance (Omran, 2007).

The K-means algorithm has the following main advantages.

- It is very easy to implement, and
- Its time complexity is  $O(Np)$  making it suitable for very large data sets. However, the K-means algorithm has the following drawbacks
- The algorithm is data-dependent,
- It is a greedy algorithm that depends on the initial conditions, which may cause the algorithm to converge to suboptimal solutions, and
- The user needs to specify the number of clusters in advance.

The K-medoids algorithm is similar to K-means with one major difference, namely, the centroids are taken from the data itself. The objective of K-medoids is to find the most centrally located patterns within the clusters. These patterns are called medoids.

According to (Raste, 2014) Clustering is an essential data mining used as a Big Data analytics method. The principle of this technique is to create groups or subsets that contain the objects with similar characteristic features. Consequently, the cluster analysis makes data manipulation simple by finding structure in data and classifying each object according to its nature. Besides, it is divided into two categories: single machine clustering techniques, which use resources of just one single machine, and multiple machine clustering techniques, which run in several machines and have access to more resources.

In this section the researcher tried to categorize the majority of available clustering algorithms according to their applicability in Big Data as follows: Hierarchical algorithm: The goal of hierarchical clustering is to build a hierarchical tree to show the relation of clusters in two different manners, which are agglomerative ("bottom-up") method and divisive method.

Agglomerative method starts with one-point (singleton) clusters and recursively adds two or more appropriate clusters until it achieves a K number of clusters. On the other hand, divisive method divides the data to a single cluster, which contains all data objects, into smaller appropriate clusters until a stopping criterion is achieved (Hind Bangui, 2018).



## 3.8 Hadoop

Apache Hadoop is an open-source software framework that supports massive data storage and processing. Instead of relying on expensive, proprietary hardware to store and process data, Hadoop enables distributed processing of large amounts of data on large clusters of commodity servers. Because of the great success of Google's distributed file system and the MapReduce computation model in handling massive data processing, its clone, Hadoop, has attracted substantial attention from both industry and scholars alike. In fact, Hadoop has long been the mainstay of the big data movement (add Citation)

### 3.8.1 Characteristics of Hadoop

Hadoop has many advantages, and the following features make Hadoop particularly suitable for big data management and analysis:

**Scalability:** Hadoop allows hardware infrastructure to be scaled up and down with no need to change data formats. The system will automatically redistribute data and computation jobs to accommodate hardware changes.

**Cost Efficiency:** Hadoop brings massively parallel computation to commodity servers, leading to a sizeable decrease in cost per terabyte of storage, which makes massively parallel computation affordable for the ever-growing volume of big data.

**Flexibility:** Hadoop is free of schema and able to absorb any type of data from any number of sources. Moreover, different types of data from multiple sources can be aggregated in Hadoop for further analysis. Thus, many challenges of big data can be addressed and solved.

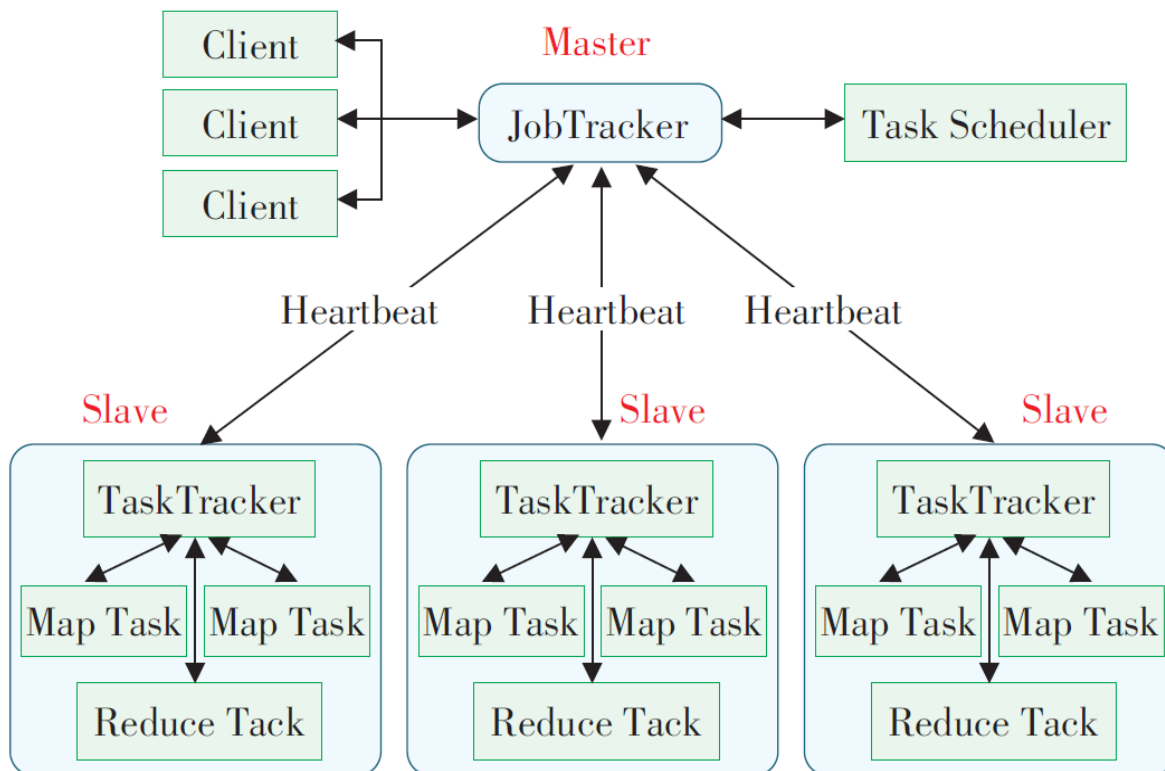
**Fault tolerance:** Missing data and computation failures are common in big data analytics. Hadoop can recover the data and computation failures caused by node breakdown or network congestion.

### 3.8.2 Hadoop MapReduce Framework

In data collection phase, HPC collects healthcare data in a continuous time domain. The processing of those extensive data is a most challenging task for getting a better output. Hadoop provides Hadoop Distributed File System (HDFS) to store those data in a structured database. MapReduce algorithm enables to make the structured distributed database from raw unstructured and semi-structured data. MapReduce has two primary class: the mapper class and the reduce class. Mapper class take the data record as input and put a key with a corresponding value to the data record.

Then Map function shuffle those record, sort them and send to reducer class. Reducer class merge the same keyed data as per the algorithm instruction and provide the structured database which store in HDFS.

### Hadoop MapReduce Framework



The same keyed data as per the algorithm instruction and provide the structured database which store in HDFS.

**INPUT:** Unstructured or semi-structured data set.

**OUTPUT:** Create a structure data set.

#### Mapper function to generate the pair of key and values

1. Input the data set and take it in InData.

2. **for** each line column and each row: **do**
3. Data = Strip and split the line.
4. **if** number of column in Data! = length *Data:head*: **then**
5. Skip that row from Data.
6. **else**
7. Store those data generating a pair of key and values.
8. Go to the next row in Data.
9. **end if**
10. **end for**

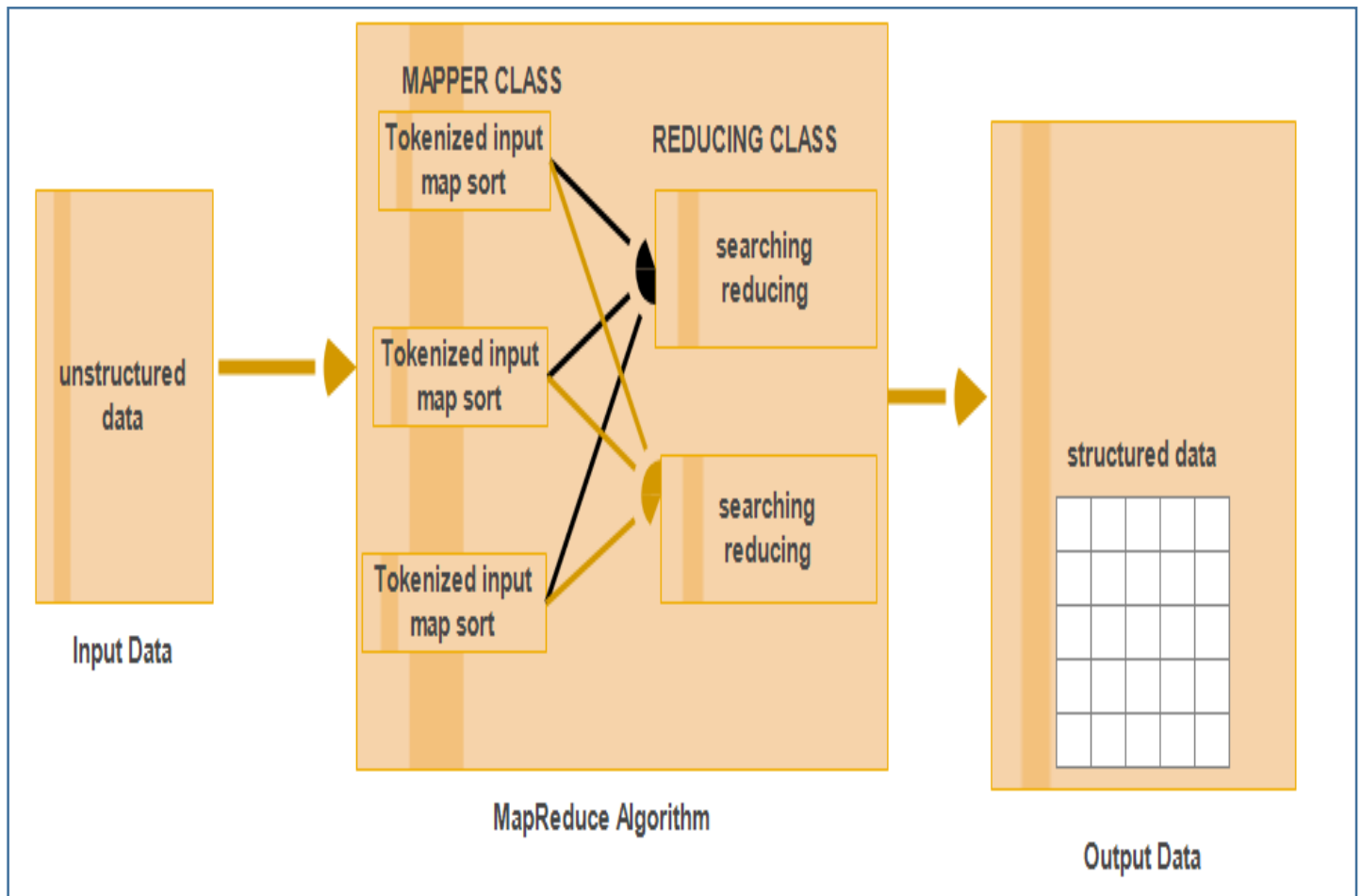


Figure 3.6 Unstructured or semi-structured data set conversion

### 3.9 Big Data volume Reduction Methods

This study more focus on the volume of data redaction and visualization. Data reduction methods for big data vary from pure dimension reduction techniques to compression-based data reduction methods and algorithms for preprocessing, cluster-level data duplication, redundancy elimination, and implementation of graph theory concepts.

Big data, so far, does not have a recognized definition, even if it is generally accepted that the concept refers to datasets that are too large to be processed using conventional data processing tools and techniques. Modern information systems produce data in huge quantities that are difficult to be measured. It means that already we have found ourselves in the “big data era,” and the question of how to solve large scale machine learning problems is open and requires a lot of research efforts. Dealing with huge datasets poses a lot of the processing challenges (I.Czarnowski and Piotr J, 2018).

The big data sources including modern information systems and databases contain fundamentally complex data characterized by the well-known ‘V’ properties: huge volume, high velocity, much variety. Since traditional techniques used for analytical processing are not fit to effectively deal with the massive datasets, searching for new and better techniques, methods, and approaches suitable for big data mining is a hot area for the machine learning community. Considering the above facts and observing current trends in the machine learning research, it can be observed that among main contemporary challenges, the most important one is a search for improvements with respect to scalability and performance of the available algorithms ( Sowmya and Sravanthi, 2017).

Among techniques for dealing with massive datasets are different parallel processing approaches aiming at achieving a substantial speed-up of the computation. Examples of such techniques are Hadoop and MapReduce techniques which have proven suitable for the computation and data intensive tasks (I.Czarnowski and Piotr J, 2018) .

This topic contributions the data reduction methods being applied in big data systems. The methods any optimize the storage program of data or reduce data redundancy and duplication. In addition, some of the methods only reduce the volume by compressing the original data and some of the methods reduce the velocity of data streams at the earliest before entering in big data storage systems.

### 3.9.1 Reducer Algorithm

In the current research, Reducer algorithm theatres a great role in aggregating values of a key by summing or combining set of values from a single or multiple Map functions. Reduce function depends on two major libraries from Hadoop are shuffle and sort functions which take intermediate output from Map function as input to shuffle and sort the same keys together so that Reduce function can easily combine or aggregate values of each key. A single Reduce function, most of the time, is implemented for all Map functions output aggregation (I.Czarnowski and Piotr J, 2018).

The application of Reducer algorithm in this research considers single Reduce function in order to aggregate outputs of Map functions. Intermediate outputs from shuffle and sort functions, these functions are libraries of the framework, is directly processed by Reduce function. The single file output which is generated by Reduce function will be taken to visualization to present result in convenient for human interpretation.

The reducer algorithm follows the following step by step procedure:

- 1: MapReduce library shuffles and sorts intermediate result
- 2: For each word, its value is aggregated
- 3: Hadoop writes key/values of aggregated word to Hadoop Distributed File System file
- 4: Output file is saved to local file system

### 3.9.2 Data duplication (Redundancy Elimination)

Data redundancy is the key problem for data analysis in big data environments. Three main reasons for data redundancy are: (1) addition of nodes, (2) expansion of datasets, and (3) data replication. The addition of a single virtual machine brings around 97% more redundancy, and the growth in large datasets comes with redundant data points.

In addition, the storage mechanism for maximum data availability (also called data replication) brings 100% redundancy at the cluster level. Therefore, effective data duplicated and redundancy elimination methods can cope with the challenge of redundancy. The workload analysis shows that the 39 higher throughput improves performance about 45% but in some extreme cases the performance degrades up to 61%. The energy overhead of duplication is 7%; however, the overall

energy saved by processing duplicated data is 43%. The performance is degraded to 5%, whereas energy overhead is 6% for pure solid state drive (SSD) environments. However, in hybrid environment the system's performance is improved up to 17%.

## CHAPTER FOUR

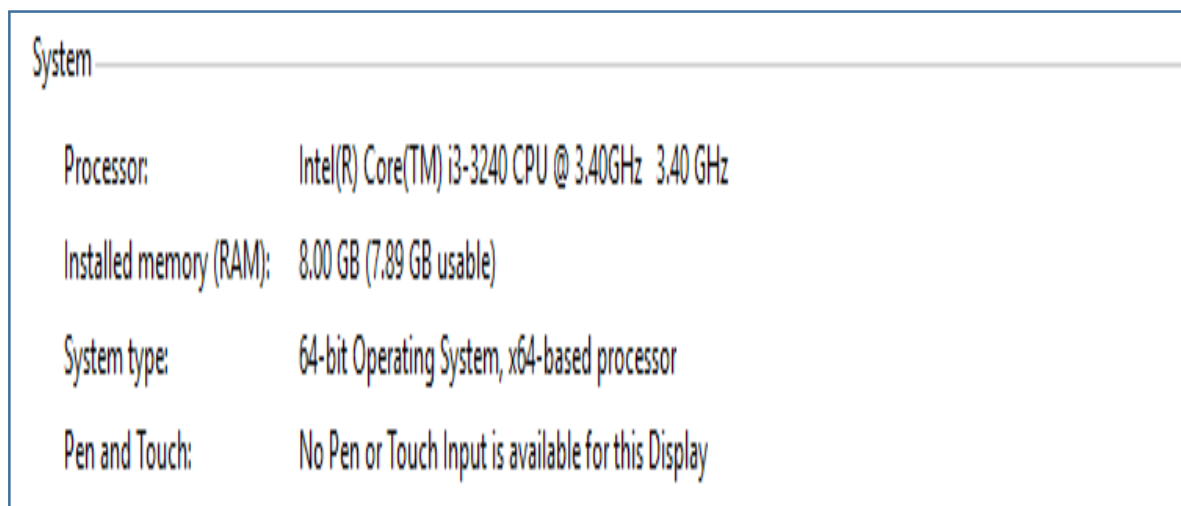
### 4. EXPERIMENTS AND RESULTS

#### 4.1. Environment Setup

Experiment was conducted with R programming language using RStudio (Version 1.0.143 2009-2017 R-Studio, Inc.). The pre-processing stage is carried out to establish a structured database in windows environment which is installed on virtual machine. MapReduce algorithm is implemented in Java JDK Integrated Development Environment. All the simulations are done on 64-bit Intel core i3 3.40 GHz machine.

This section provides details about the experimental setup and obtained results that are used to evaluate the efficiency of the proposed discovery hadoop framework for the purpose of classifying diabetics as either hypertension or benign based on dataset data. The first subsection explains the experimental setup including the performance metrics, and the later subsections present the results obtained using the experimental setup for the task of laceration clustering.

A hadoop cluster was setup using open stack on a fundamental server whose specification is shown in figure 4.1. Three virtual machines were configured on the server. The hardware and software configuration for each of the virtual machines is shown in previous chapter. Each virtual machine is assigned 1vcpu core i3 8GB RAM and 1TB of hard disk storage.



System	
Processor:	Intel(R) Core(TM) i3-3240 CPU @ 3.40GHz 3.40 GHz
Installed memory (RAM):	8.00 GB (7.89 GB usable)
System type:	64-bit Operating System, x64-based processor
Pen and Touch:	No Pen or Touch Input is available for this Display

Figure 4.1 - Hardware specification

Hadoop-2.7.4 was used with a single VM configured as the NameNode and the remaining four VMs as DataNodes. The NameNode was not used as a DataNode. The replication level of each data block was set to 3. Two typical Hadoop MapReduce applications were run as Hadoop YARN jobs. The application available as part of the Hadoop distribution was used to generate different sizes of input data.

Software	Operating System	Windows 10
	JDK	Openjdk 1.8
	Hadoop	Hadoop 3.2.0
Hardware	Cpu	3.40GHz
	Processor	Intel®
	Hard disk	25GB
	Memory	4GB

## 4.2 Requirement for big data analytics

A big data analytics approach was used to capture from the stakeholders their views on current practice in selected data intensive and cognitively complex processes, and the initial vision on what could be improved from both users and technologists. A Decode specific requirement elicitation strategy was designed and deployed to tackle the seemingly diverse cases. Common characteristics were extracted to identify common interests for technological innovation. This step mobilized ideas from both users and technologists.

### 4.2.1 Application of Sense making Models

In addition to data collection from the ground, theoretical models for sense making were identified for a deeper understanding of sense making behavior in each of the use cases. We considered an individual sense making model which provides a detailed view of data-driven analysis when trying to make sense of large volume of data. We supplemented it by a collaborative sense making model which presents the triggers of collaboration and characteristics of building shared understanding. The models provide a common framework for comparison in order to identify the commonalities and differences in sense making activities within different context. This step provided focus for



users and technologists in positioning the benefits of proposed technical solutions and when these could be used.

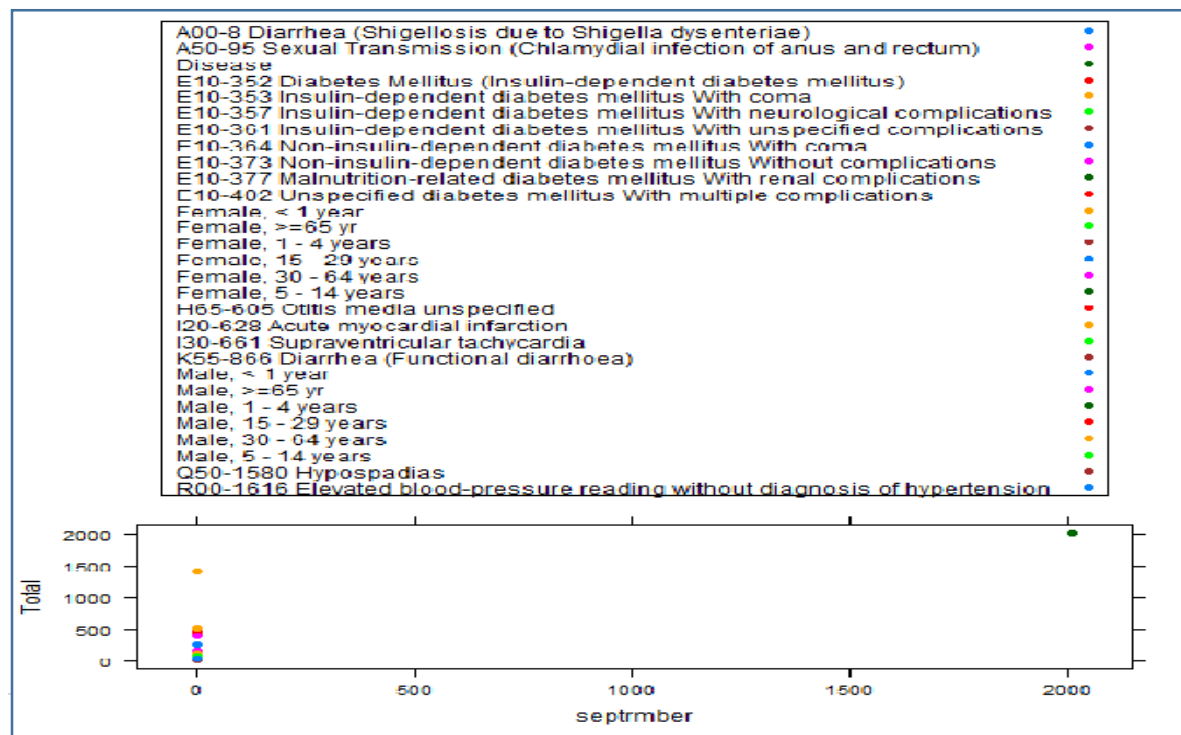
#### 4.2.2 Conceptual Architecture for Big Data Analytics

Lastly, a conceptual architecture was developed as a high level specification of how the various tools might work together for each of the use cases in a big data processing and visualization platform. In designing the architecture, we followed the design research process proposed by Prefers and his colleagues and aimed to create useful artefacts that solve relevant design problems in organizations. Usage situations were produced to walk through how the platform may be used. This step provided a high level blue print which could be used as a communication tool between the users and the technologists on requirements.

### 4.3 Experimentation

Dataset is chopped into three sub dataset so that one of dataset or splits is being processed on each of DataNodes. Totally, three DataNodes are used to process the data sets; however, the reducer is aggregating output of mappers using single node, one of three DataNodes, in order to return final result to specified location. Mapper firstly ingests a split using RecordReader library which provides every single line of statements as key/value pairs.

Hadoop ecosystem is coming up with great advantages for current limitation of computation by enhancing processing speed and storage capacity.



Visualization process is final result that display for the end user or specific group. Totally charts or graphs show results of processing in different forms but they are conveying the same information by helpful capability to figure out the content or value of a single element of data. Horizontal Bar chart, Treemap, Pie Chart, Highlight Table, Stacked Bar Chart, Circle Views Chart, Bubble Chart, Box-and-Whisker plot, Heat Map and Packed Bubbles are the format through which results are displayed. Chart are used to present MapReduce framework processing results. Each of them has given interactivity as well as elegant format of information presentation capabilities by enhancing information consumption for all audiences.

More importantly, such information presentation for advanced users creates a room for further exploration and analysis

## 4.4 Results

Huge dataset is extensively produced in every medical sector. Health professional is willing to extract useful information from the transactions in order to make the best decision; researchers are expecting to extract the useful information from the experimental results and thus to develop new theories and products; doctors need to extract useful information from data model to determine the direction of disease. Thus, how to realize the parallel data mining algorithms to improve the executing speed is becoming a significant problem. It requires the efforts from all sectors to achieve the optimum state of data mining.

### 4.4.1 Data Visualization

Even though experimentation has generated a single raw file which is output of Reduce function, the need of data visualization of the same file is a must so that the result of processing could easily be understood to grasp the information in appropriate and consumable format. However, the availability of visualization platforms for big data is just handful, i.e., there are very few companies as providers of visualization components.

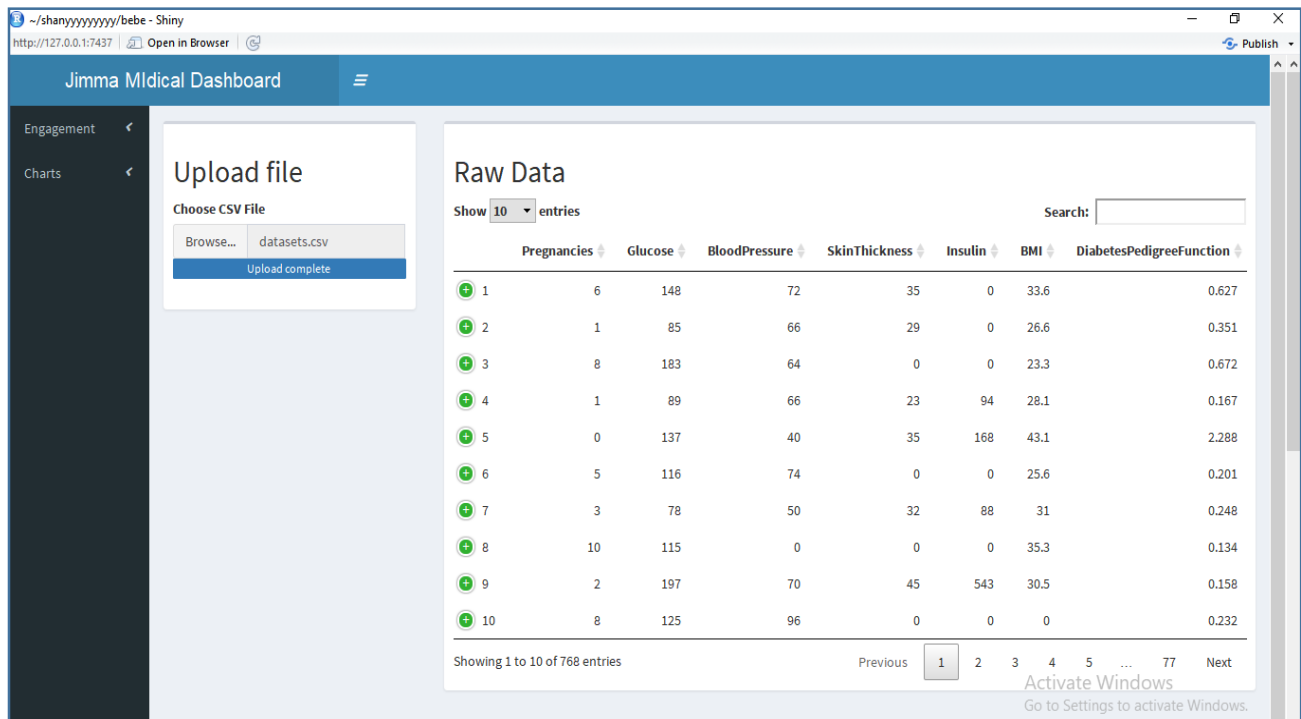


Figure 4.3: sample screenshot row data selection

In figure 4.3 it shows how to use the raw data that displays all kinds of columns and rows for system users. Actually, big data technology sets are now emerging on top of traditional business intelligence technologies so big data visualization tool sets are at their infant stage. As shown below charts, the output file of MapReduce framework processing is converted into interactive charts using the shiny visualization platform.

Shiny is one of the few great visualization tools that embraces the whole output of big data processing results without breaking them down into a set of files so as to visualize. In addition, its charts are consumable through all devices regardless of their screen size or pixel density because it provides interactivity by allowing a mouse cursor to hover over an element of interest.

All graphs show results of processing in different forms but they are conveying the same information by accommodating the capability to figure out the content or value of a single element of data. Horizontal Bar chart, Tree map, Pie Chart, Highlight Table, Stacked Bar Chart, Circle Views Chart, Bubble Chart, Box-and-Whisker plot, Heat Map and Packed Bubbles Chart are used to present MapReduce framework processing results.

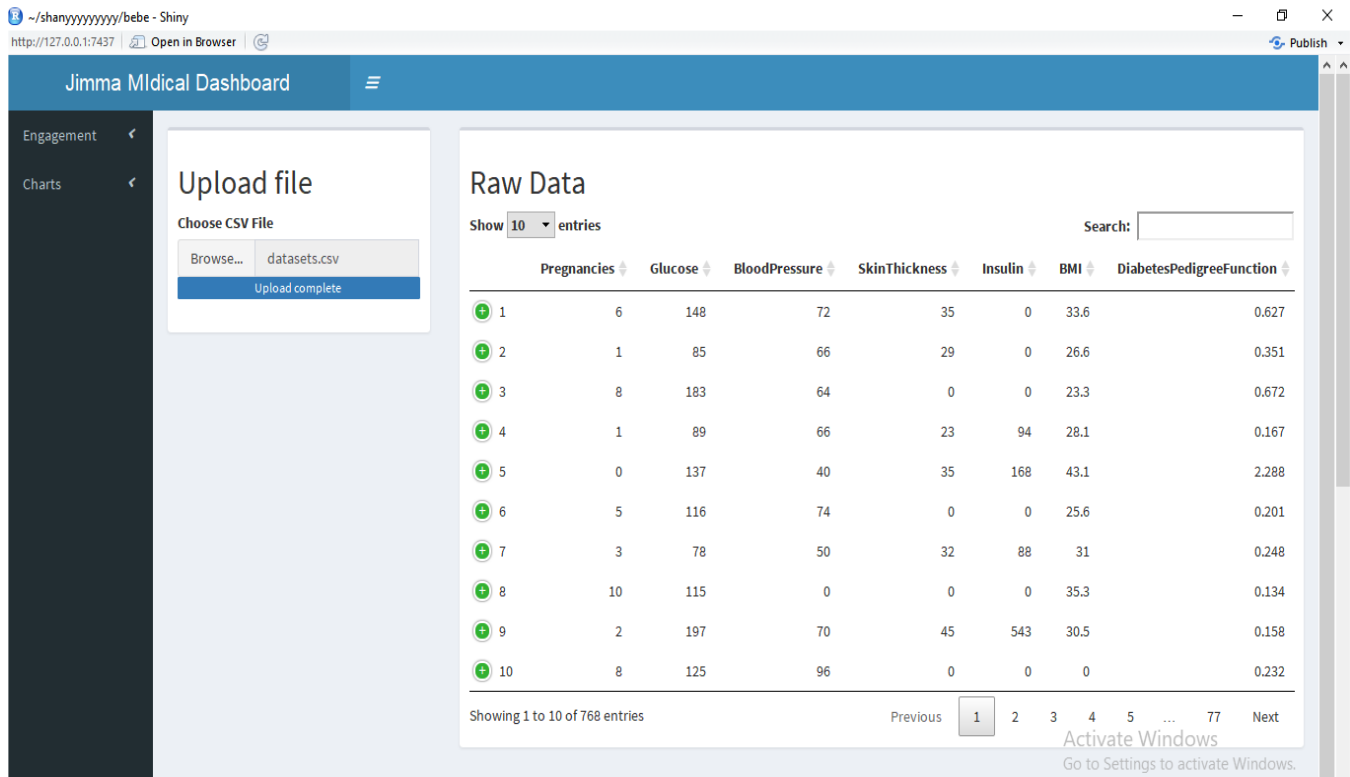


Figure 4. 4: sample screenshot upload dataset

In figure 4.4 simply show the upload dataset columns and rows the user you can show different number of row at the same time it select 3,5,10 based on user interest the Plot description is a scatter plot is described as a 2-dimensional plot which displays the joint variation of two data items. A scatter plot is also called a scatter chart, scatter diagram, scatter graph. For a scatter plot, observations are being represented by each marker and the marker position usually indicate value for the observations. A scatter plot shows data in Cartesian coordinate in a graphical display which displays the relationship that exist between two variables in which one is represented as a vertical distance and the other as horizontal distance.

The moment all the data are plotted on a scatter plot, it is capable to determine in a visual form if the data points are related or not. Scatter plots can help you know how the data points are scattered or spread across the graph and also you will know they are closely related (Singh & Singla, 2018).

A scatter plot displays the variables and how strong they are related and it is also possible to know how far the data are scattered (Fig. 4.5 and Fig. 4.6).

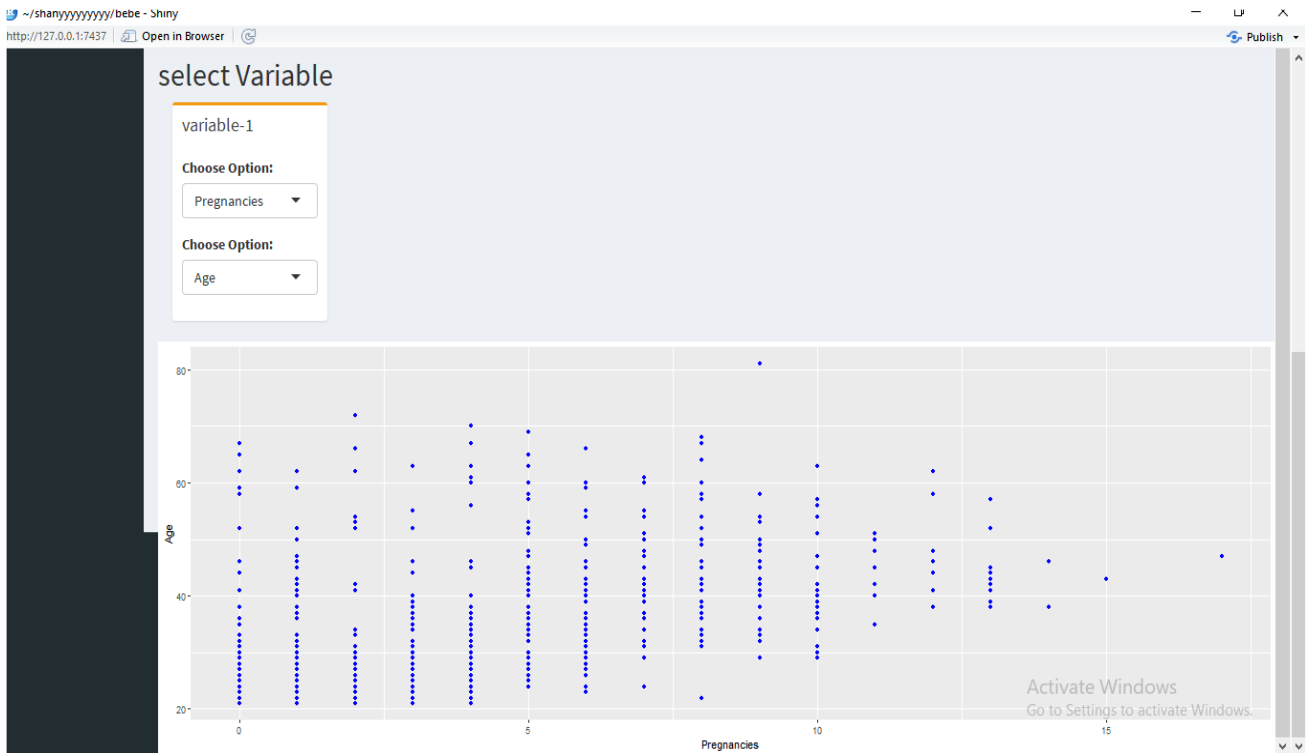


Figure 4.5: scatter plot screenshot

As scatter plot indicates, the count of diabetics' data set is shown by pregnancy of a diabetics by representing its relative value from other attributes. For instance, attributes are listed on vertical axis and values of each ages is placed on horizontal axis and the plot is interactive enough to display specific value of flying over it.

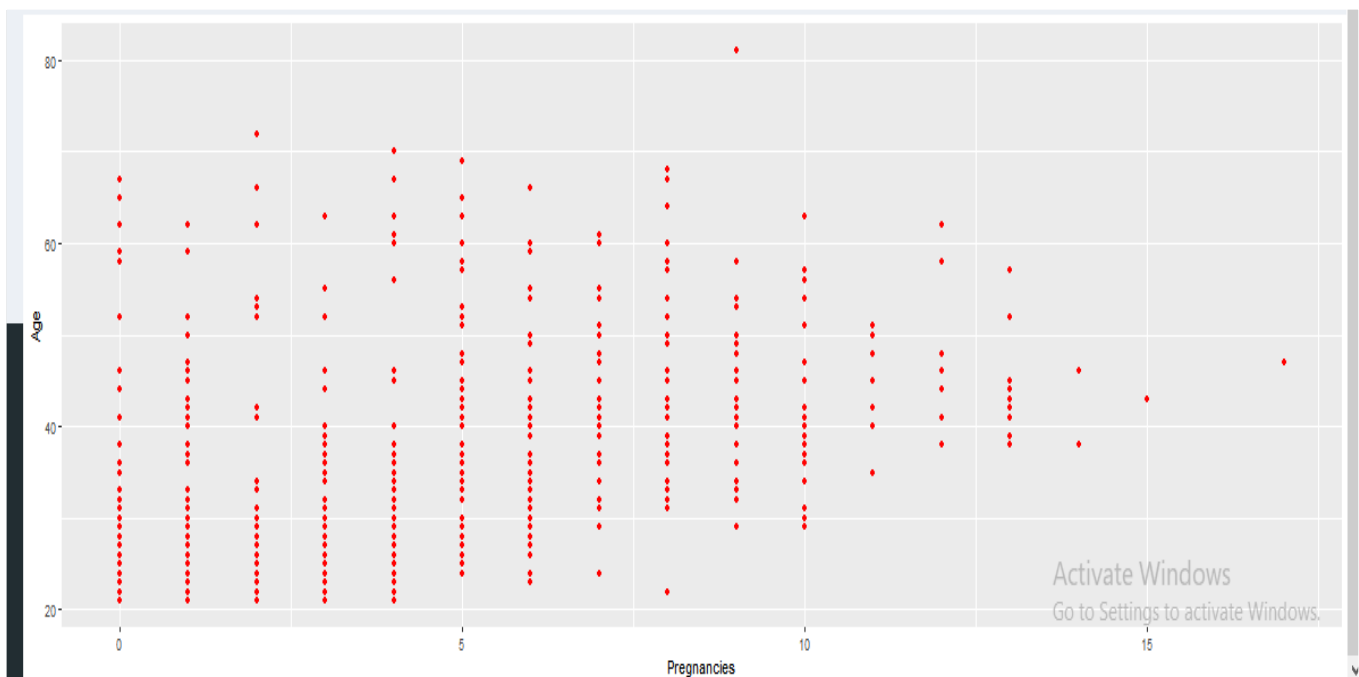


Figure 4.6 sample screenshot

All kinds of visualized data must be need domain experts to interpret and understand a relationship between the diabetic's and chronic data's. The importance of data interpretation is undeniable. Dashboards visualization not only bridge the information gap between traditional data interpretation methods and technology, but they can help medicine and prevent the major drawbacks of interpretation.

As a digital age solution, they combine the best of the past and the present to allow for informed decision making with maximum data interpretation.

#### 4.5 System performance evaluation

Somesysfunction ("my function (with, arguments)")

- > Start time: 2020-08-30 00:00:00 # output of somesysfunction
- > "Result" "of" "dataset" # output of dataset
- > End time: 2020-08-30 00:00:02 # output of somesysfunction
- > Total Execution time: 02 seconds

#### 4.6 Contributions

- ✓ The researcher well collect data and prepare a dataset for predominate chronic and the other contribution of this study is Time reduction high speed of tool r-studio and analytics can be easily identify new dataset and analyze.
- ✓ It also improve processes of dataset visualization can easily their based on descriptive models.
- ✓ The main contribution of this thesis work is designing a good and general user interface for big data analytical system for predominate chronic dataset.

## CHAPTER FIVE

### 5. CONCLUSION AND RECOMMENDATIONS

#### 5.1 Conclusion

Big data analytics process and visualization is new study area and very challenge-full, that requests new method of tackling which can be solved with up-to-date practice of data management because data overflow and data creation frequency in varieties of formats are inevitable scenarios. The approach that is employed in this study to undertake these challenges are reviewing problem areas in detail, followed by designing solution, then implementation of designed solution after that testing implemented solution using big data sets. As a result shows, Hadoop ecosystem provides platform to process data sets of Big Data in cheap, fault tolerant and high speed. The success of the study develop next generation of data science in areas of data storage, processing and visualization. Particularly, consistency and computational power does not need scale up in terms of hardware and processor capacities. Therefore, Big Data processing and visualization challenges are capable to handle using hardware and software solutions rather than in placing dedicated machines with increased hardware and processing capabilities.

This study has mainly focus points to be raised for real-world study in the area. These are data redaction and visualization which indicates hint of light that sheds for upcoming challenges how to confront and extract insights from big data sets. This study also achieved as it is possible to manage big data regardless of size and nature of data. Apart from these, the points that require further investigation and study are fully distributed environments or clustered machines to exploit full potential by processing Terabytes and Petabytes of data sets of big data in general and its specific application for decision making by implementing revealed insights. The heterogeneity and disorganization of the data collected were also challenge.

#### 5.2 Recommendations

Big Data has huge potential and benefits at every level of societies which can be considered as an eye opener to new discoveries and innovations. Now, it is not only possible to study populations as whole without looking for samples and its representativeness but also it becomes common to forecast or trend analysis of unimaginable situations. So, it is important to study further impacts

of structured and semi-structured data sets by accommodating data along with power of parallelism computation in fully distributed setup.

As new area of study, it is strongly recommended further studies in specific contexts to identify wealth of benefits and cautions. The researches strongly recommend that study should be carried on multimedia data. Other researchers should also conduct study and develop the prototype which has the capability to analyze and visualize heterogeneous data format such as image with respect to audio, or video. The implementation of other clustering and classification algorithms in Mahout should gain more attention in the future.

### 5.3 Future Works

This research work explores different points that can be further improved for better functioning of the system.

- Applying the full system on different datasets.
- Applying the system on several machines.
- learning different feature selection methods by combining our approach with other techniques



## References

- Ackerman, M. ). (2012). Big data. *The Journal of Medical Practice Management*, 153-160.
- Adler & Jha. (2013). Healthcare's "Big Data" challenge. *American Journal of Managed Care*,, 537-538.
- Akerkar. (2014). *Big Data Computing*; CRC Press,.
- Alemayehu.B. (2019, May). BIG DATA ANALYTICS TO PREDICT CANCER BASED ON DIAGNOSED CLINICAL DATA.
- Al-Shiakhli, S. (2019). Big Data Analytics: A Literature Review Perspective.
- Berman, J. J. ( 2013). *PRINCIPLES OF BIG DATA Preparing, Sharing, and Analyzing Complex Information*,.
- Bisandu, D. B. (2016, November ). Design science research methodology in Computer Science and Information Systems. *International Journal of Information Technology*.
- Coral, C. (2018). Analytical Framework for a Systemic Analysis of Drivers and Dynamics of Historical Land-Use Changes: A Shift Toward Systems Thinking. *In Balancing Individualism and Collectivism ). Springer, Cham*, (pp. 259-269.
- Coral,& Bokelmann. (2017). The Role of Analytical Frameworks for Systemic Research Design, Explained in the Analysis of Drivers and Dynamics of Historic Land-Use Changes. *Systems*, 5(1), 20.
- Coughlin, T. (2018, Nov 27). Retrieved from [https://www.forbes.com](https://www.forbes.com/sites/tomcoughlin/2018/11/27/175-zettabytes-by-2025/#6a22383d5459):  
<https://www.forbes.com/sites/tomcoughlin/2018/11/27/175-zettabytes-by-2025/#6a22383d5459>
- Croninger, & Douglas, K. (2005). Missing data and institutional research. New directions for institutional research. 33-39.
- Cukier, K. (2014). Big data is better data.
- Dashboards as a Service : . (2009, October). *Journal of Service Researc*, Volume 12, 175-189.

- Desalegn, T. (2016). Big Data Processing and Visualization in the Context of Unstructured data set.
- Domenico. (2016). *Trends in Big Data Analysis*.
- Fahmideh and Mahdi. (2018). Big data analytics architecture design — An application in manufacturing systems Big data analytics architecture design — an application in manufacturing systems. (July 2019).
- Guba, E. G. (2007., April). THE ELEMENTS OF A PROPOSAL.
- Hermon, R. (2014). Big data in healthcare : What is it used for ?
- Hind Bangui. (2018). Exploring Big Data Clustering Algorithms for Internet of Things Applications. In IoTBDS. 269-276.  
<https://www.newgenapps.com/blog/6-reasons-why-choose-r-programming-for-data-science-projects>. (n.d.).
- I. H & Frank. (2002). Data mining: practical machine learning tools and techniques with Java implementations. *Acm Sigmod Record*, 76-77.
- I.Czarnowski and Piotr J. (2018, November 5). An Approach to Data Reduction for Learning from Big Datasets. 13 pages.
- J. G. Wolff. (2014.). Bg Data and the SP Theory of Intelligence,.
- K. B. Carter. (2014). Actionable Intelligence A Guide to Delivering Business Results with Big Data Fast, .
- K.Sharmila and R.Bhuvan. (2014). ROLE OF BIG DATA ANALYTIC IN HEALTHCARE USING DATA. *Volume-1(1)*, 68-70.
- Leetaru, K. (2015). Do Big Data results Depend on what Data we Look At? .
- M. Maier. (2013). "Towards a Big Data Reference Architecture,".
- M. Sowmya and N. Sravanthi. (2017). Big Data: An Overview of Features, Tools, Techniques and Applications. *International Journal of Engineering Science and Computing, Volume 7 (Issue No.6)*.

- Omran, M. (2007, November). An overview of clustering methods.
- Raghupathi, W., & Raghupathi, V. (2018 ). An Empirical Study of Chronic Diseases in the United States: A Visual Analytics Approach to Public Health. *Int J Environ Res Public Health*.
- Raste, K. S. (2014). BIG DATA ANALYTICS – HADOOP PERFORMANCE ANALYSIS.
- Sawant, & Shah. (2013). Big Data Application Architecture. *In Big Data Application Architecture Q & A*, pp 9-28.
- Shah, Gita Basava. (2014). Design an Efficient Big Data Analytic Architecture for Retrieval of Data Based on. (September 2017).
- Sharmila and Bhuvana. (2014). ROLE OF BIG DATA ANALYTIC IN HEALTHCARE USING DATA. *Volume-1(1)*, 68-70.
- Shu, H. (2016). Big data analytics: six techniques. *Geo-spatial Information Science*,. 119-128.
- Singh & Singla. (2015). Big data: tools and technologies in big data. *International Journal of Computer Applications*,.
- Thomas. (2012). How ‘Big Data’ is Different; MITSloan Management Review,.
- Vogl & Puri. (2004). Tools and methods for data collection in ethnobotanical studies of homegardens. *Field methods*, 285-306.
- Wagner, E. (2008). Chronic disease management: what will it take to improve care for chronic illness? *effective clinical practice*.
- Wang, & Wu. (2018). Chronic diseases and health monitoring big data. *IEEE reviews in biomedical engineering*, 277-288.
- Wang, L. (2017). Heterogeneous data and big data analytics. *Automatic Control and Information Sciences*,. 8-15.