# DATA ANALYSIS OF SKIN DISEASE USING BIG DATA ANALYTICS AND HADOOP

*Dissertation proposal submitted in fulfilment of the requirements for the Degree of*

**MASTER OF TECHNOLOGY**

**in**

**COMPUTER SCIENCE AND ENGINEERING**


By

**AMAN SINGH**

**11306569**

Supervisor

**MR.SANJEEV KUMAR**




**School of Computer Science and Engineering**

Lovely Professional University

Phagwara, Punjab (India)


December 2017

# DECLARATION STATEMENT

I hereby declare that the research work reported in the dissertation proposal entitled "DATA ANALYSIS OF SKIN DISEASE USING BIG DATA ANALYTICS AND HADOOP" in partial fulfilment of the requirement for the award of Degree for Master of Technology in Computer Science and Engineering at Lovely Professional University, Phagwara, Punjab is an authentic work carried out under supervision of my research supervisor MR.SANJEEV KUMAR. I have not submitted this work elsewhere for any degree or diploma.

I understand that the work presented herewith is in direct compliance with Lovely Professional University's Policy on plagiarism, intellectual property rights, and highest standards of moral and ethical conduct. Therefore, to the best of my knowledge, the content of this dissertation represents authentic and honest research effort conducted, in its entirety, by me. I am fully responsible for the contents of my dissertation work.

*Signature of Candidate*

**Aman Singh**

**11306569**

# SUPERVISOR'S CERTIFICATE

This is to certify that the work reported in the M. Tech dissertation proposal entitled "**DATA ANALYSIS OF SKIN DISEASE USING BIG DATA ANALYTICS AND HADOOP"**, submitted by **Aman Singh** at **Lovely Professional University, Phagwara, India** is a bonafide record of her original work carried out under my supervision. This work has not been submitted elsewhere for any other degree.

Signature of Supervisor

(Sanjeev Kumar)

**Date:**

# ACKNOWLEDGEMENT

I would like to express our deepest appreciation to all those who provided me the possibility to complete this dissertation proposal. I would like to express my special gratitude to my dissertation mentor, Mr. Sanjeev Kumar, whose contribution in stimulating suggestions and encouragement, helped me to coordinate my dissertation and especially in writing this proposal. Without great support of LOVELY PROFESSIONAL UNIVERSITY, it would be not possible to complete this proposal. The University has always been very responsive in providing necessary information, and without their generous support plan would lack in accurate information on current developments.

I perceive this opportunity as a big milestone in my career development. I will strive to use gained skills and knowledge in the best possible way, and will continue to work on their improvement, to attain desired career objective.

**Aman Singh**

# TABLE OF CONTENTS

**CONTENTS** **PAGE NO.**

# LIST OF FIGURES

# ABSTRACT

Analysis of structured data has been tremendous success in past. However, analysis of large unstructured data in the form of video and images format remain a challenging area. Skin disease datasets are mostly images format and there are large no of data generate in the hospital. Analysis of the particular skin disease in the city and there is huge demand to store, process and carefully study this large amount of data to make it usable.

The main objective of this project is to demonstrate by using Hadoop concepts and big data Analytics. How data will be analysis of skin disease using big data Analytics and Hadoop.

Analysis of skin disease using big data analytics and Hadoop, skin disease datasets to analysis and outcomes are how many people to have their disease and describe the prediction of that disease and symptoms also. There are using to HDFS (Hadoop Distributed file system) to store the dataset of skin disease into Hive database. Hive uses a SQL-like interface to query data stored in various database and file systems that integrate with Hadoop. HDFS (Hadoop distributed file system) is a primary Hadoop application and a user can directly interact with HDFS using various shell-like command supported by Hadoop.

# CHAPTER 1

# INTRODUCTION

Analysis of data, also known as Data Analysis or Data Analytics is process of inspecting cleaning, transforming and modeling the data with the goal of discovering useful information, suggestions conclusion and supporting the decision-making. Data Analysis has multiple facts and approaches. Analysis of skin diseases datasets using big data analytics and Hadoop ,Dataset will be images or text formats to analysis that which disease are more effect to human in present time and there are technique to be used for analysis of the datasets for skin diseases, and process of data analysis .

## 1.1 Process of Data Analysis

Data Analysis is a process for obtaining raw data and converting it into information useful for the decision- making users, and test the hypothesis and disprove theory. “Procedures for analyzing the data, technique for interpreting the results of such procedure, ways of planning to easier the analysis of data. There are few points to process of data analysis.

### 1.1.1 Data collection

Data is collected from variety of sources and data will be create it, the data may also be collected forms sensors and environment such as traffic cameras, satellites, recording devices, and healthcare data also.

### 1.1.2 Data processing

Data firstly obtained must be processed or organized for analysis. They may involve placing data into rows and columns in a table formats (structured data) for further analysis.

### 1.1.3 Data cleaning

This is a very important process because if anything wrong happen in this step then whole performance and integrity of data may be challenged. In this step the input is the real data set collected from the real world data which is continues and new. There are chances of inaccurate data, duplicate data, formatting issue, un-structured data, redundant records, naming conventions and incomplete data. The increase in such data leads towards incorrect and non-desirable results.

In this process our primary goal is to remove duplicacy and incomplete data from the data base. This is because when we apply any query on the database than that time it takes a lot of time to result an output. So to increase the probability of its correctness we need to remove its incompleteness and duplicacy of the data. This is achieved by applying various algorithms, tools and techniques which complies a pure complete and unique data in the database. And then when we apply query on that unique filtered data that time we get correct output as a result which will be used for taking various decisions for the organization and various predictions in-favor of the organization.

### 1.2 Skin Disease Analysis

Skin is the largest organ of your body and your skin separates the inside of your body from the outside of the world .It protects you from bacteria and viruses, and regulates your body temperature. There are many skin diseases as Acne, Hives, Allergies, Melanoma, and skin cancer.

In Skin diseases analysis to use the dataset of particular diseases and analysis of that diseases and quantitative and qualitative analysis technique will be used. How many people suffering from the particular disease.

### 1.3 Data Mining Techniques

Data mining is an enduring procedure within which the growth is spelled out by unearthing trends or patterns. It is the most beneficial technique in a tentative analysis scenario where there are no prearranged concepts about what will constitute a conclusion. The knowledge results obtained from data mining processes are used to assist human decision makers in efficient and effective decision making and finally to solve their complex problems.

#### 1.3.1 Prediction

This data analysis methodology evaluates the numeric and well-ordered future values grounded on patterns of a dataset. Prediction attributes are always continuous and not categorical. Predictive analysis is a statistical method which takes data as input and extracts meaningful information from the dataset. The extracted information is then used to predict the behavioral trends and patterns. In simpler terms, Predictive analysis tends to learn from the dataset to predict the future trend for the sake of better decision making.

### 1.3.2 Regression

This method is used to disclose the relationship between independent variable and dependent variable. This technique aims to describe the relation between a dependent variable and one or more independent variable. It describes the change in the value of a dependent variable when any one of the independent variable is changed while others remain constant. Regression technique follows basic mathematical statistical approach. Other mathematical methods used under regression are linear regression and logistic regression.

### 1.3.3 Visualization

Visualization is termed as the modest and easier way to epitomize data in an understandable data presentation. The input is multifaceted data and subsequent easy and clear uncovered patterns of complex data that can be effortlessly recited by the users through the progression of data mining.

This is closely related to the pattern detection capabilities of a human. This is done with the help of certain factors like size, vision and previous experiences.

## 1.4 Big data

"Big data is the term for a collection of data sets for large no data and complex that it becomes difficult to process it using traditional database management tools or data processing application."

### 1.4.1 Big data Attributes

The three Vs-volume, velocity and variety-are commonly used to describe different aspects of the big data. See figure 1.1 these three attributes make it easy to define the nature of the data.

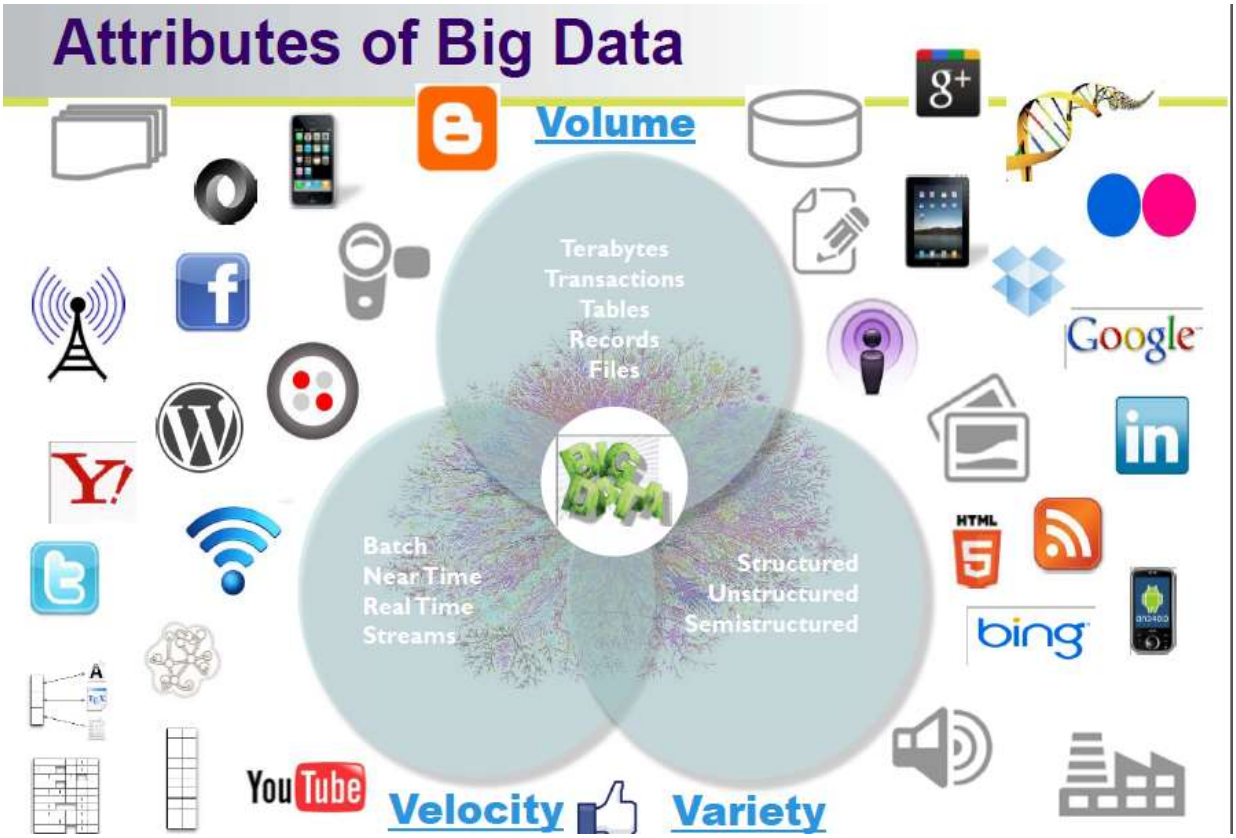


Figure 1.1 Attributes of Big data

#### 1.4.1.1 Volume

Volume is the most challenging aspect of big data. Big enterprises already have large amount of data, it could be in the form of system logs, record keeping etc. Data warehouse based solutions may not necessarily have the ability to process and analyze the data due to lack of parallel processing architecture. A lot can be derived from text data, locations or log files. For example, email communication patterns, transaction-based data, and security investigation. Big data technology offer a solution to create a value from this massive and previously difficult to process data.

#### 1.4.1.2 Velocity

Data is following at a large speed in to organization. The speed at which data is generated and processed to meet the demands and challenges that lie in the path of growth and development. With the invention of the smartphone era is there is even further location based data generate and it's becoming important to be able to take advantage of this huge amount of data.

#### 1.4.1.3 Variety

All the data generated with the social and digital media is rarely structured data. Unstructured text document, video, audio data, images, are example of unstructured data. Manage the complexity of data in many different structure, ranging from relational, to logs, to raw text.

## 1.5 Need of Big Data Analytics

The above mentioned attributes of big data, data is massive, comes at a speed and highly structured that it does not suitable conventional relational database structure.so much insight hidden in this data. Lot of data means lot of hidden insights. The ability to quickly analyze big data means the possibility to learn about customer, market trends and equipment monitoring performance analysis and much more. Benefits of using big data technology may come at a downside of a loss of privacy.

## 1.6 Hadoop

Hadoop is an open-source software framework for storing and processing big data in a distributed fashion on large clusters of commodity hardware. Essentially, it accomplishes two tasks:

Massive data storage

Faster processing.

- With the ability to economically store and process any kind of data (not just numerical or structured data), organizations of all sizes are taking cues from the corporate web giants that have used Hadoop to their advantage (Google, Yahoo, Etsy, eBay, Twitter, etc.), and they are asking "What can Hadoop do for me?"

### 1.6.1 Hadoop component

#### 1.6.1.1 HDFS(Hadoop Distributed File System)

HDFS is the storage system for a Hadoop cluster. When data lands in the cluster, HDFS breaks it into pieces and distributes those pieces among the different servers participating in the cluster.

Each server stores just a small fragment of the complete data set, and each piece of data is replicated on more than one server.

#### 1.6.1.2 Map-Reduce

Map Reduce is most effective technique used for data cleaning. This technique applies for the huge amount of the data, in this we write algorithms for mapping and reducing the data in any programming language like python etc. Map Reduce map the data and make key value pair from it [10]. These key value pairs contains duplicate data or attributes in it. So to reduce that duplicacy of the data the Reducer comes in to play. Reducer job is to take those key value pairs and then combines or aggregate them to produce the final result. The Reducer performs its job only when Map is done with its job. This helps to optimize the data and provide only the data as output which is accurate, consistent and beneficial for the user.

#### 1.6.1.3 Hive

A data warehousing and SQL-like query language that presents data in the form of tables. Hive programming is similar to database programming. (It was initially developed by Facebook.)

#### 1.6.1.4 HBase

A nonrelational, distributed database that runs on top of Hadoop. HBase tables can serve as input and output for MapReduce jobs.

#### 1.6.1.5 Pig

A platform for manipulating data stored in HDFS. It consists of a compiler for MapReduce programs and a high-level language called Pig Latin. It provides a way to perform data extractions, transformations and loading, and basic analysis without having to write MapReduce programs

### 1.6.2 HDFS architecture

The Hadoop distributed file system (HDFS) is the storage component of the core Hadoop infrastructure. HDFS provides a distributed architecture for extremely large scale storage.it is important to mention the difference between scales up and scale out. The core concept of HDFS is that it can be made up of dozens, hundreds, or even thousands of individual computers, where the system's files are stored in directly attached disk drives. Each of these individual computers is a self-contained server with its own memory, CPU, disk storage, and installed operating system (typically Linux, though Windows is also supported). Technically speaking, HDFS is a user-space-level file system because it lives

on top of the file systems that are installed on all individual computers that make up the Hadoop cluster.

HDFS is a system that allows multiple commodity machines to store data from a single source. HDFS consists of a Name Node and a Data Node. HDFS operates as master slave architecture as opposed to peer to peer architecture. Name Node serves as the master component while the Data Node serves as a slave component. Name Node comprises of only the Meta data information of HDFS that is the blocks of data that are present on the Data Node

# CHAPTER 2

# REVIEW OF LITERATURE

**Priyanka K, Prof Nagarathna Kulennavar, A Survey on Big Data Analytics in Health Care International Journal of Computer Science and Information Technologies, Vol. 5 (4), 2014**

This paper characterizes huge information examination and its attributes, Comments on its points of interest and difficulties in human services huge information investigation in medicinal services is developing into a promising field for giving knowledge from expansive informational collections and enhancing results while lessening costs. Its potential is awesome; however there remain difficulties to Overcome. Issues, for example, ensuring protection, shielding security, setting up guidelines and administration, and constantly enhancing the instruments and Technologies will collect consideration. Enormous information examination and Applications in social insurance are at an incipient phase of Development, yet quick advances in stages and apparatuses can quicken their developing procedure.

**Dr. Smitha Rao, Ms. Suma. S. N, Ms. Sunitha. M, Security Solution for Big Data Analytics in Healthcare. IEEE, 2015**

In this paper discuss security in enormous information investigation in medicinal services. Huge Data examination gives tremendous upper hand to organizations, helping the organizations, tailor their items to buyer needs. Huge information in medicinal services alludes to electronic wellbeing informational indexes that are identified with tolerant social insurance and prosperity. In human services division the protection and security issues of huge information are a noteworthy worry as information is bound by worldwide directions like the Health Insurance Portability and Accountability Act (HIPAA), The Health Information Technology for Economic and Clinical Health (HITECH), HCSC, FTC (Federal Trade Commission) and so forth...

**Eiman Al Nuaimi, Hind Al Neyadi, Nader Mohamed, Jameela Al- Jaroodi, Applications of big data to smart cities. Journal of Internet Services and Applications, 2015 .**

This paper explains the concept of smart city. There are various definitions of smart city across The globe, but the idea behind it is all the same – to improve the Quality of life of citizens. This Paper talks about the applications of Big Data, how it helps to implement the concept of smart City. This paper further targets to understand the benefits that a big data can incorporate in the Smart city and various challenges that it might face.

**Gunasekaran Manogaran, Chandu Thota, Daphne Lopez, V. Vijayakumar, Kaja M. Abbas and Revathi Sundarsekar. Big Data Knowledge System in Healthcare. Springer International Publishing AG 2017.**

This paper examine about a major information based learning administration framework to build up the clinical choices. The proposed information framework is produced in light of assortment of databases, for example, Electronic Health Record (EHR), Medical Imaging Data, Unstructured Clinical Notes and Genetic Data. What's more, the qualities and difficulties for enormous information in social insurance, and proposes a major information based learning framework..

**Ashwin Belle, Raghuram Thiagarajan, S.M.Reza Soroushmehr, Fatemeh Navidi, Daniel A. Beard and Kayvan Najarian1. Big Data Analytics in Healthcare. Biomed Research International · January 2015**

This paper discuss a devices to aggregate, oversee, break down and absorb huge volumes of unique, organized and unstructured information delivered by current medicinal services frameworks. Huge Data investigation has been as of late connected towards helping the procedure of care conveyance and malady investigation .and Analytical strategies are utilizing as a part of social insurance.

**Dharavath Ramesh Pranshu Suraj,and Lokendra Saini, Big data Analytics in Healthcare: A Survey Approach. IEEE Access, 2016**

This paper discuss Big Data Analytics and utilization of enormous information (e.g. human services, business insight) to examination of the application moreover. What's more, we can utilize the Big Data investigation for compelling basic leadership in medicinal services area utilizing the current machine learning calculations with some Modification to it.

**Matthew Herland, Taghi M Khoshgoftaar and Randall Wald. A review of data mining using big data in health informatics, Journal of Big Data, 2014.**

This paper examine about an essential objective of Health Informatics is to take in genuine restorative information from all levels of human presence to help propel our comprehension of drug and medicinal practice. Breaking down Big Data of this degree has just been conceivable to a great degree as of late, because of the expanding ability of both computational assets and the calculations which exploit these assets. Research on utilizing these instruments and procedures for Health Informatics is basic, since this area requires a lot of testing and affirmation before new strategies can be connected for settling on true choices over all levels. The way that computational power has achieved the capacity to deal with Big Data through productive calculations (and also equipment progresses, obviously) lets information mining handle the Big Volume, Velocity, Variety, Veracity, and Value of the information created by Health Informatics

**Maryam M Najafabadi,Flavio Villanustre,Tagi M Khoshgoftaar,Naeem Seliya,Randall wald and Edin Muharemagic. Deep learning application and challenges in big data analytics, Journal of Big Data, 2015**

This paper explores how deep learning can be utilized for addressing some important problems in Big Data Analytics and including extracting complex patterns from massive Volumes of Big Data, Semantic indexing, data tagging, fast information retrieval, and simplifying discriminative tasks And which can contain useful information about problem such as national intelligence ,cyber security, fraud detection, marketing and medical informatics.

**Emad A Mohammed, Behrouz H Far, Christopher Naugler. Applications of the MapReduce programming framework to clinical big data analysis: current landscape and future trends, Biodata Mining, 2014.**

This paper reviews the existing applications of the MapReduce Programming framework and its implementation platform Hadoop in clinical big data and related medical health informatics fields. Various genetic algorithms for example have been implemented using the MapReduce framework in top of Hadoop Distributed File System. Applications of MapReduce can be extended to all kinds of data in clinical field - Publicly available clinical datasets, Biometrics datasets, Bioinformatics datasets, and biomedical signal datasets. Along with these applications,

there are quite a few challenges - Technology straggling, Data dispersion, Security concerns and privacy issues, Standards and regulation. Despite these challenges, Big Data has the potential to revolutionize the Healthcare sector.

**Lei Xu, Chunxiao Jiang, Jian Wang, Jian Yuan, Yong Ren. Information Security in Big Data: Privacy and Data Mining, IEEE Access, October 2014**

This paper discuss about a major challenge of big data and data mining – Privacy. It talks about PPDM (Privacy Preserving Data Mining). The privacy issues and the methods to protect. Privacy with regards to four types of users – data provider, collector, miner and decision maker are also discussed. Also, some light on the game theory has been shed to understand the model of common grounds for all the users.

**Min Chen, Shiwen Mao, Yunshan Liu, Big Data: A Survey. Springer, January 2014.**

The paper talks about the big data and related technologies - cloud, IoT, Hadoop and data centers. It discusses about the 4 phases of big data – data generation, acquisition, storage and analysis. Applications of big data is also discussed in this paper- medical, enterprises, social networks, smart grid, etc.

**Harshawardhan S. Bhosale, Prof. Devendra P. Gadekar,A review paper on big data and Hadoop. International Journal of Scientific and Research Publications, Volume 4, Issue 10, October 2014**

This paper talk about the big data and Hadoop components and tools. Describes innovative technique and technologies to store, manage and analyze petabyte-or larger –sized datasets with highly velocity and different structure. Hadoop is core platform for structuring big data and solves the problem of making it useful for analysis purpose. Hadoop is open source software project that enables the distributed processing of large data set. The paper also focuses big data processing problems. The paper describes Hadoop which is an open source software used for processing of big data.

**Bijesh Dhyani Anurag Barthwal. Big Data Analytics using Hadoop, International Journal of Computer Applications (0975 – 8887) Volume 108 – No 12, December 2014**

This paper talk about basic understanding of Big Data and it is useful to organization from the performance perspective. The paper also describe the difference in the challenge faced by the other organization as compared to medium or large scale operation and therefore the differences in their approaches and treatment of big data. Hadoop is a popular tool for big data

implementation. The paper talks about with the overall architecture of Hadoop along with the details of Hadoop component.

**S Ramya silpa,chidvila v, A Review of skin cancer. International research journal of pharmacy 2013.**

This paper talk about skin cancer and part of the skin diseases. There are two tow type of skin cancer malignant melanoma and non-malignant melanoma. Skin cancer mainly occur due to exposer of sunlight.so there is a need to detect and treat disease at early stage. Overall 80% skin cancer of are BCC, 16% are SCC and 4% are melanoma. UV-A and B are mainly responsible.

# CHAPTER 3

# PROBLEM DEFINITION

As this is era of technology and within India the technology is still in developing stage so in upcoming days the concept of smart city and digitalization is on peak which creates a huge or very huge amount data. These large dataset to analysis of that to use the big data and Hadoop concepts. Big data is used for the large datasets and analysis of dataset to use the Hadoop.

In this time big data is more useful for any area as an example of healthcare, social media, crime prediction, etc. And there are analysis of the datasets to using Hadoop environment and there using language of R and python for implementation in Hadoop. Hive will be using to the analysis of the datasets of skin diseases.

Analysis of skin disease using big data analytics and Hadoop .there are many skin disease to analysis of the particular diseases as example skin cancer, allergy, etc. Firstly define the particular skin diseases and analysis of that disease. How many people are suffering from that diseases in particular state? And there are given to the symptoms and prescription for particular skin disease. And there are using to Hadoop technology and there are using to the hive component to use the analysis of the dataset.

It is very necessary to have a look on this field for the growth of technology, nation as well as human. Data analysis become necessary part of the companies now. And there many companies to work on the data analysis of dataset. And there are many companies to work on Hadoop technologies and big data analytics.so provide the data of skin diseases and analysis of that particular disease and work on it very deep and thoroughly.

# CHAPTER 4

# SCOPE OF STUDY

In this time value of data is ten times more than that of its infrastructure.so there are many datasets to analysis of the data but they are not able to provide the better analysis of the data. There are some field where we still have need to go and touch those little corners will increase the reliability of user on data. . The some of the gaps and challenges we find during the research are explains below:

Analysis of the data set using big data analytics and Hadoop. There are slow response to analysis of the dataset. And there may be using the different tool for the analysis of dataset due to fast response.

Another gap comes into consideration when we apply data analysis of the skin disease dataset. There are text format dataset and images format dataset to analysis of that dataset then output will be different.

Another gap is when it comes to using the big data analytics. Big data is applicable when datasets are large and there are many datasets .it does not applicable for large no of datasets .most of the datasets variety, volume, and velocity are not follows.

After the analysis of the datasets using Hadoop technologies to implement .there some output will be different for the other tools .and different tools will be given the fast response of analysis of datasets.

# CHAPTER 5

# OBJECTIVES OF STUDY

5.1 To Analysis of skin diseases datasets using big data analytics and Hadoop.

5.2 To Analysis of the different types of diseases and apply the big data technique and also describe the symptoms and perception of the particular skin diseases. With the help of graphs to explain them.

5.3 To Analysis of the skin diseases datasets and describe how many people suffering from the different skin disease in the particular state?

# CHAPTER 6

# PROPOSED RESEARCH METHODOLOGY

As our research is in database field where we choose a topic that is Data Analysis. This is very important to describe some research methodologies. As this is very huge field which contains various tools and techniques for obtaining a correct output. But the methods and techniques tools are we are using in our research are much summarized and they have great impact in the field of data analysis. The first tool that we are going to use is Hadoop, this is a software which provides very large number of packages for analysis the data. This tool helps to make dataset more correct and useful without writing any complex coding. This tool makes our dataset worth able for data Analysis.

The other technique we are going to use is big data analytics, because analysis of the datasets to use large amounts of data to apply the big data for the analysis purpose.it is a main role of the analysis of datasets. And there are using to hive commands for analysis of dataset .it provide the simple and understandable syntax for the programmer.

So by using above mention methods and technology we are going to analysis of the skin data in which first we are going to collect from the different -different sources, as hospitals, clinic and static dataset.

# CHAPTER 7
# EXPECTED OUTCOME

In the data analysis of skin diseases using Big Data Analytics and Hadoop describe the expected outcomes from the analysis of datasets .firstly how many people are suffering from that diseases? .And there are also provide the symptoms and prescription of that diseases

To Analysis of the skin disease Dataset as example skin cancer, hive, etc. Analysis of the particular datasets of skin diseases and outcomes are which skin diseases are more harmful and apply technique of Big Data Analytics. Also describe how many people suffering for the different types of skin diseases in state.

.

# CHAPTER 8

# SUMMARY AND CONCLUSIONS

The proposed research work aims at construction and training of the data analysis of skin diseases using big data analytics and Hadoop .there are using to big data technique for analysis of datasets. By the end of the proposed research work, to Analysis of the datasets of different skin diseases and given the symptoms and prescription of the particular disease. There are using Hadoop platform and hive component to analysis of the datasets.

Data analysis of skin diseases using big data analytics and Hadoop. There are using process of data analysis as Data collection, Data cleaning, Data processing. In big data analytics technique will be used as 3Vs –variety, volume, velocity and Hadoop platform will be using to implementation, analysis of the skin diseases datasets. Skin disease are many types as example Allergy, skin cancer, rashes, etc., analysis of the dataset in city.

# LIST OF REFERENCES


[1] S Ramya silpa,chidvila v, A Review of skin cancer. International research journal of pharmacy 2013.

[2] Bijesh Dhyani Anurag Barthwal. Big Data Analytics using Hadoop, International Journal of Computer Applications (0975 – 8887) Volume 108 – No 12, December 2014

[3] Harshawardhan S. Bhosale, Prof. Devendra P. Gadekar,A review paper on big data and Hadoop. International Journal of Scientific and Research Publications, Volume 4, Issue 10, October 2014

[4] Min Chen, Shiwen Mao, Yunshan Liu, Big Data: A Survey. Springer, January 2014.

[5] Lei Xu, Chunxiao Jiang, Jian Wang, Jian Yuan, Yong Ren. Information Security in Big Data: Privacy and Data Mining, IEEE Access, October 2014

[6] Emad A Mohammed, Behrouz H Far, Christopher Naugler. Applications of the MapReduce programming framework to clinical big data analysis: current landscape and future trends, Biodata Mining, 2014.

[7] Matthew Herland, Taghi M Khoshgoftaar and Randall Wald. A review of data mining using big data in health informatics, Journal of Big Data, 2014.

[8] Priyanka K, Prof Nagarathna Kulennavar, A Survey on Big Data Analytics in Health Care International Journal of Computer Science and Information Technologies, Vol. 5 (4), 2014

[9] Maryam M Najafabadi,Flavio Villanustre,Tagi M Khoshgoftaar,Naeem Seliya,Randall wald and Edin Muharemagic. Deep learning application and challenges in big data analytics, Journal of Big Data, 2015

[10] Ashwin Belle, Raghuram Thiagarajan, S.M.Reza Soroushmehr, Fatemeh Navidi, Daniel A. Beard and Kayvan Najarian1. Big Data Analytics in Healthcare. Biomed Research International · January 2015

[11] Eiman Al Nuaimi, Hind Al Neyadi, Nader Mohamed, Jameela Al- Jaroodi, Applications of big data to smart cities. Journal of Internet Services and Applications, 2015 .

[12] Dr. Smitha Rao, Ms. Suma. S. N, Ms. Sunitha. M, Security Solution for Big Data Analytics in Healthcare. IEEE, 2015

[13] Dharavath Ramesh Pranshu Suraj,and Lokendra Saini, Big data Analytics in Healthcare: A Survey Approach. IEEE Access, 2016

[14] Priyank Jain, Manasi Gyanchandani and Nilay Khare, Big data privacy: a technological perspective and review.journal of Big Data 2016.

[15] Allard J. van Altena, Perry D. Moerland, Aeilko H. Zwinderman and Sílvia D. Olabarriaga, Understanding big data themes from scientific biomedical literature through topic modeling,journal of Big Data 2016.

[16] Jasleen Kaur Bains, Big Data Analytics in Healthcare- Its Benefits, Phases and Challenges, International Journal of Advanced Research in Computer Science and Software Engineering Volume 6, Issue 4, April 2016.

[17] Gunasekaran Manogaran, Chandu Thota, Daphne Lopez, V. Vijayakumar, Kaja M. Abbas and Revathi Sundarsekar. Big Data Knowledge System in Healthcare. Springer International Publishing AG 2017.