Indian Health Care Analysis using Big Data Programming Tool
Mohit Dayal* and Nanhay Singh

Ambedkar Institute of Advanced Communication Technologies and Research, New Delhi

Abstract

Big Data in layman terms means lots of data. This enormous amount of data is used for analysis and research purposes. This data accounts for trends in various domains. Organizations are using Big Data to predict the future in turn making them smarter and efficient. Applications from Big Data are innumerable, from retail industry where Big Data helps retailers gain insights into the customer to needs and habits, to Banking, HealthCare & Hospitality. Government agencies are increasingly incorporating Big Data analytics to curb crime and maintain law and order through social media traffic analysis and other means. Therefore, to get actionable data and perform analytics requires specialized tools. There are thousands of Big Data tools available in the market. There are open source tools like Hadoop, a term which has become synonymous to Big Data. In this paper, we authors analyzed the health care dataset against different research queries using Pig Latin Script, over the last few decades; quality of health care services in India has been improved tremendously because of the improved health care services, increased number of private and government hospitals and increased number of doctors with recognized medical qualification. In spite of significant growth government has to take strict measures to improve the overall health care facilities in India because considerable amount of gaps do exist between the demands to that of quality supply of healthcare services.

© 2016 The Authors. Published by Elsevier B.V.

Keywords: Adaptive Learning Rate; Fuzzy Decision Tree; Momentum Factor; User Localization; Wireless Signal Strength.

1. Introduction

Big Data is a voluminous collection of data that cannot be managed by traditional database management systems. Big Data in an umbrella term used for the humongous amount of data generated from myriad of sources such the Web, mobile devices, sensors, enterprise applications and digital repositories. The data can be structured as well as unstructured. The data ranges from terabytes to exabytes of data. The relational database management systems (RDBMS) have proven to inefficient to handle such huge volumes of data. Another important factor which renders the conventional database systems unsuitable is that the majority of data being generated is unstructured; the RDBMS systems are only adept to handle structured data. Hence new tools and schemes for data analysis and management were in order. Big Data can be characterized by the 4 V’s.

1. Volume- is the voluminous collection of data. The sheer volume of data generated these days by real time applications and other data sources such as twitter feeds, photos, videos on social media, click streams of web, sensor-enabled equipment is so mammoth it runs to petabytes and exabytes of data. Big Data technology enables us to store this amount of data on distributed systems.

*Corresponding author. Tel.: +91-9999954597.
E-mail address: mohitdayal.md@gmail.com
2. Velocity- is the rate at which data is arrived. For example the social media portals generate process able data at a very high rate, the amount of credit card transactions that happen in a second. All these require data to be analyzed at a very high rate.$^8$

3. Variety- refers to the myriad of sources from which data is accumulated. Data is in different formats structured, semi-structured, unstructured. A full 90% world data has been generated over the past two years and majority of it is unstructured.$^9$

4. Veracity- refers to the trustworthiness of data. There are uncertainty surroundings the data being generated these days with data being incomplete and inconsistent. The Big Data analytics empowers us to work with such kind of data.

5. Value- Refers to techniques of deriving value from data. There is an intrinsic value that the data may possess and must discover for analysis. This makes ‘value’ the most important ‘V’ of Big Data. Modern technologies have made it possible to find the value from data.$^{10}$

To deal with these characteristics is an expensive turnout. Apache Hadoop is the knight in the shining armour here. Apache Hadoop is a platform that is responsible for providing with the solutions that are cheaper. Hadoop’s major components include a distributed file system which is called the HDFS (Hadoop Distributed File System) and a layer for implementation of the processing paradigm-MapReduce.$^6$ Hadoop is an open source system. Hadoop uses a cluster of commodity machines as its nodes, forming a network which is used as a single, logical, storage and computational platform among multiple users or groups.$^3$ Hadoop’s MapReduce performs tasks in parallel, automatically and in a synchronized fashion.

2. Research Methodology

The following figure describes the research methodology adopted. The first step involves identification of problem and then adopting appropriate approach and strategy to solve the problem.

According to proposed framework first relevant data was collected then the data from the files will be loaded onto the HDFS using copy command. Data cleansing is one of the most important component required in data analytics. It includes data cleansing, data extraction, removing duplicates and converting it into standardized schemas. Storing and processing is done used HDFS.$^{12}$ HDFS being reliable and scalable is highly recommended for storing and processing large amount of data. HDFS replicates data over multiple nodes and thus does not require RAID (Redundant Array of Integrated Devices) storage.$^4$ The task of storing, accessing and modifying data is performed using two different components Job Tracker and Task Tracker. The Job Tracker assigns the MapReduce tasks to Task Trackers. The Task Trackers send their status of being active and ready to take up the job by sending heartbeats to the Job Tracker.$^7$ The analytical tool here is the scripting language; Pig Latin. Overall development time and testing time is much less than that of map reduce program. Writing fewer codes without having prior knowledge of JAVA and reduced testing time are major advantages of Pig latin over MapReduce program. At the end results were analyzed using graphs to make useful decisions.$^5$

![Fig. 1. Research Methodology.](image-url)
3. Query Formulation

In this paper, we authors analyzed the health care dataset against different research queries, over the last few decades; quality of health care services in India has been improved tremendously because of the improved health care services, increased number of private and government hospitals and increased number of doctors with recognized medical qualification. The main objective is to provide the healthcare services to all and to improve the accessibility of healthcare amenities to all the stratum of the society.

The data set was in .csv format which was first cleaned by removing the missing values and other inconsistent values. Fully distributed hadoop cluster was used to store and process the voluminous data.

4. Empirical Observations

4.1 Aggregate number of hospitals from 1950–2015

Input: health care dataset  
Output: Aggregate number of hospitals from 1950–2015

Creating Pig Script

1. Using Pig command enter into Pig’s interactive shell  
2. \( A = \) Load the healthcare data set;  
3. \( B = \) Sort and group number of hospitals from 1950–2015;  
4. \( C = \) Find out the sum of hospitals in each group year;  
5. Dump \( C \);

Above graph clearly depicts that there is tremendous growth of hospitals from the year 1950–2015. This indicates that the ease of availability of healthcare facilities is increasing every year. In spite of significant growth government has to take strict measures to improve the overall health care facilities in India because considerable amount of gaps do exist between the demands to that of quality supply of healthcare services.

4.2 Aggregate number of physician from 2005–2015

Input: Health Care dataset  
Output: Aggregate number of physician from 2005–2015
Fig. 3. Aggregate Number of Physician from 2005–2015.

Creating Pig Script

1. Using Pig command enter into Pig’s interactive shell
2. $A =$ Load the healthcare dataset;
4. $C =$ Find out the sum of physician in each group year;
5. Dump $C$;

Above graph highlights that more numbers of doctors with recognized medical qualification have been registered from the year 2005 to 2015. India’s doctor – patient ratio is still alarming despite of the fact that there has been tremendous growth in the number of doctors and nurses.

4.3 Male-female life expectancy in various states

Input: Health care dataset
Output: Male-female Life Expectancy in Various states

Creating Pig Script

1. Using Pig command enter into Pig’s interactive shell
2. $A =$ Load the healthcare dataset;
3. $B =$ Group and generate the life expectancy of male and female in each state;
4. Dump $B$;

From the above graph it could be concluded that the highest male and female expectancy is in Kerala while least in Madhya Pradesh. Therefore government has to come up with different measures to improve the overall healthcare facilities in different state to match up with the international standards of health.

4.4 Percentage of people happy with the healthcare standards in different state

Input: Health care dataset
Output: Percentage of people happy with the healthcare standards in different state
Creating Pig Script

1. Using Pig command enter into Pig’s interactive shell
2. \( A = \) Load the healthcare data set;
3. \( B = \) Sort and group number of happy people with the healthcare standards in each state;
4. \( C = \) Find out the percentage of happy people in each state;
5. Dump \( C \);

From the above graph it could be highlighted that citizen of Kerala are most satisfied with the healthcare facilities provided in their states while least in Uttar Pradesh. This indicates that the ease of availability of healthcare facilities is more in Kerala as compare to Uttar Pradesh. Government must take strict measures to address all the issues to achieve the objective: To provide the healthcare services to all.
5. Limitation

Following are the main limitations:

1. Apart from using Pig Latin, MapReduce programming framework can be used to analyze the health care data set.
2. Pseudo distributed hadoop mode was used to implement the data set.
3. The main limitation includes few publications and Non English publications were excluded therefore as authors we could not claim that the work has not been printed in other languages.

6. Conclusions

In this paper, we authors analyzed the health care dataset against different research queries using Pig Latin Script, over the last few decades; quality of health care services in India has been improved tremendously because of the improved health care services, increased number of private and government hospitals and increased number of doctors with recognized medical qualification. The main objective is to provide the healthcare services to all and to improve the accessibility of healthcare amenities to all the stratum of the society. In spite of significant growth government has to take strict measures to improve the overall health care facilities in India because considerable amount of gaps do exist between the demands to that of quality supply of healthcare services.

Acknowledgements

We authors would like to extend our deepest thanks and appreciation to Ambedkar Institute of Advanced Communication Technologies and Research, New Delhi in helping us to accomplish this work.

References

