HEALTHCARE ANALYSIS USING HADOOP

B. Durga Sri[1], K.Nirosha[2], M. Padmaja[3]
[1,2]Assistant Professor, [3]Student, MLR Institute of Technology

Abstract
This paper gives vision of healthcare analytics and delineate the significance for the solution of healthcare for offering magnificent consequences. The main purpose of our project is to create software that analyzes the data regarding the population effected with the diseases with respective to gender (female/male) and it also analyses the death rate and it clearly gives us yearly prospective of disease in a country with respect to the death rate and female and male rates. The procedure after obtaining different healthcare analysis based upon different factors like region,year.,death rate, male deathrate,female death rate etc.; we need to apply different algorithms to check with which datamining algorithms is efficient. Health care organizations provides finer effects when the big data analytics is applied effectively along with the clinical information in and mainly reciprocation of taking apt decisions in accurate time. The real-time analysis is performed using Hadoop in big data analytics for examining vast volume of data for anticipating or speculating the emergency situation in advance.

Keywords: Hadoop, Big Data, Cloudera, Hue, Impala

I. Introduction

The advancements in treatment and disease prevention to tools that help physicians target diagnoses improves quality and standardizes the extent of healthcare. It distinctly retaliates the way of big data generated by the systems, security concerns, data features that helps in gaining a valid perception on these dataset. Large amount of data is generated for record keeping acceptance and patient related data. Now – a-days in digital world, it is obligatory that these data should be digitalized. To upgrade standard of health care by reducing the costs, its certain that infinite volume of data generated should be analyzed effectively to counter-part new challenges.

Uses of Big data analytics in Healthcare analytics

(i) Supplying the patient centric services:
To provide quicker aid to the people for the verification by providing proof based medicine discovering medicine at the prior stages based on the clinical data available, reducing drug doses to avoid side effects and issuing structured and efficient medicine based on the genetic-setup.

(ii) Discovering diseases at the prior stages:
The spreading of diseases can be avoided by estimating the viral diseases before spreading based on the live analysis. Tribulating from a disease in a particular location can be recognized by surveying the social-logs of the patients. This assists the healthcare professional to suggest the victims by taking obligatory obstructive measures.

(iii) Examining the hospital standard:
Surveiling to check whether hospitals are arranged according to the rules and norms set by the Indian medical council. The regular checkup assists government in taking essential measures against ineligible hospitals.

(iv) Enhancement of treatment methods:
Examining the effects of medicine repeatedly based on the analysis dosages medication for quicker solace. The patient's crucial signs to impart lively concern to the patients. The data

triggers by the patients suffered from the similar symptoms, assists and aids doctor to issue new medicines based on the analysis report.

2.Literature Survey

The enhancement in the quality of care, healthcare outcomes and minimization in costs is possible due to the optimization and escalation in the digitalization for the providers and payers. The digital information of healthcare organizations can produce valuable insights by the tools and technologies. The precise computation in the risk and outcomes is done by the analysis of internal and external information of patients in the organization. The prediction model for the readmission of survey in bigdata analytics on healthcare system is presented by Kiyana Zolgaghar.The cohesive delivery networks can be devised using healthcare information exchange by various providers and payers

Joseph M .Woodside has depicted regarding the analytical techniques. The substantial analytical techniques are applied to huge amount of existing patient data and medical data to gain a depth understanding of results.

3. PROPOSED SYSTEM

The proposed system depicts that the healthcare analysis and multilevel system id obligatory for decisive making system. The diseases are analyzed using Hadoop by different factors. The large amount of data (Giga bytes) can be analyzed and results can be obtained in the proposed system.

The pivot on main elements of Weblog analysis, which are request, status, link, date, etc; The online services limitations confinement can be resolved y Bigdata. Large amount of data can be analyzed easily in Bigdata. The data can be analyzed easily for the upcoming enhancement using tools such as Cloudera, Impala, Apache Hadoop. The pace of tools can be known by a part of Hadoop system such as hive and hue. The process of execution can be known by executing with high speed.

4. REQUIREMENT ANALYSIS

After analyzing the requirements certain amount of tasks that have to be performed, the next step is to analyze the problem and understanding the context. There are majorly two phases that are present. The first activity in the phase is studying the existing system and the other is to understanding the requirements of the new system.

Understanding the properties and requirements of a system is more difficult and requires creative thinking and understanding of existing systems.

Requirements which we are using in this project include both Hardware and software requirements for the process.

(a)Hardware Requirements

- Dual Quad-core CPU
- 4-8 GB of memory per processor core
- 1 Gigabit Ethernet

(b)Software Requirements:

- Hadoop 2.x
- MySQL
- VMware
- HDFS
- Hive
  - **Hue**
  - Impala.

5. METHODOLOGY

In this project, we explain how we prepared our datasets. After that, we provide how we analyzed the data using some statistical analysis. Then, we introduce how we constructed some of the models to achieve our purpose.

a)How HDFS used in project:

HDFS is a java based file system that provides scalable and reliable data storage and it was designed to span large clusters of commodity servers. The quality and quantity of enterprise data is available in HDFS. It is scalable, fault-tolerant, distributed storage system that works closely with wide variety of concurrent data applications.

HDFS cluster is comprised of Name node manages the clusters in metadata and Data Nodes that stores the data. These Attributes like

- Permissions
- Modification
- Access time
- Name space

5.2 Hive is used in project

The data which is present in MySQL is imported to Hive using Sqoop command. Steps that are involved in hive are,

- Start installation
- Preparing to use a MySQL streaming Result set.
- Beginning code generation.
- Transferring of Data
- Retrieving of Records
- Execute SQL
- Loading Uploaded data into Hive.

5.3 Cloudera used in Project

Cloudera is revolutionizing enterprise data management by offering the first unified platform for Big data.

CDH contains the main, core elements of Hadoop that provides reliable, scalable data for processing of large datasets. It provide security, high availability, and integration with Hardware and software.

It contains single Hardware and Software system

- Single Management Model
- Single Security Model
- Common Storage
- Single Meta Data

5.4 Hue used in project

Hue includes web Applications that let you browse HDFS and H Base files, write apache Hive and Impala queries, Export data with Apache Sqoop, Submit Map Reduce programs.

Hue is used to get results in different types of charts like pie charts, Bar charts in the result sets. This is used to analyze the results easily.

5.6 Sqoop used in project

Sqoop is a command line interface application for transferring data between relational data base and Hadoop.

It Supports incremental loads of a single table or a free from SQL query as well a saved jobs which can run multiple times to import updates made to a database.

It supports both Sqoop import and Export.

6.IMPLEMENTATION:

Collection of Data sets according to customer complaints like product, Sub-product, Issue ,Sub-Issue, State, Company Name, Zip-code, Date-received and so on.

These are different types of attributes which are collected from the various Organization. For e.g., for each complaint there is a unique Zip-code is present and the complaint-Id is present and Data-received and State.

To implement customer complaint analysis the following steps are to be followed.

Step1: Collect the data sets of customer complaint Analysis.

Step 2: Format the data file in windows by removing the header.

Step 3: Convert the database into CSV format and copy the file into Cloudera.

Step4: Create the Data base with the file name and use the database.

Example: create database table name;

Use table name;

Step 5: Creating the tables un Sequential Query Language i.e., SQL

Step 6: Loading the data into MYSQL.

Step 7: Importing Data from MYSQL to Hive using Sqoop.

Step 8: Check the tables in database.

Step 9: Execute the commands in Impala for getting results.

7. RESULTS:

After execution of Impala command:

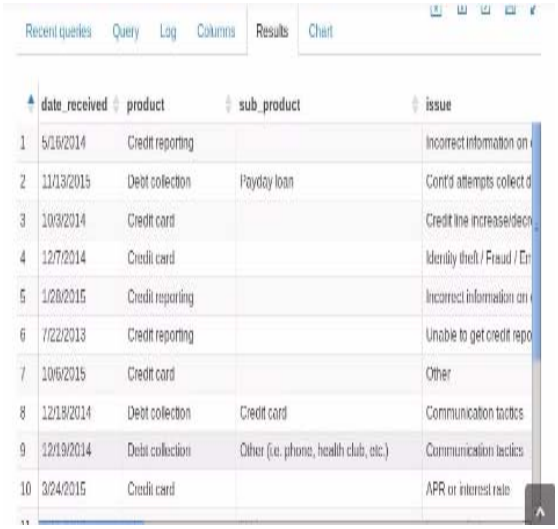
Enter into the Browser and then select Hue Impala in the query Editor and write then write the command for Execution

For example when the limit is 20 then the query is shown as

Select * from table limit 20;

Example: select * from customer complaints limit 20;

From the above figure 7.1 shows the results that are present in customer complaints with the help of Impala tool with respect to the product



	date_received	product	sub_product	issue
1	5/16/2014	Credit reporting		Incorrect information on
2	11/13/2015	Debt collection	Payday loan	Cont'd attempts collect d
3	10/3/2014	Credit card		Credit line increase/decr
4	12/7/2014	Credit card		Identity theft / Fraud / En
5	1/28/2015	Credit reporting		Incorrect information on
6	7/22/2013	Credit reporting		Unable to get credit repo
7	10/6/2015	Credit card		Other
8	12/18/2014	Debt collection	Credit card	Communication tactics
9	12/19/2014	Debt collection	Other (i.e. phone, health club, etc.)	Communication tactics
10	3/24/2015	Credit card		APR or interest rate

Fig.7.1 Results in Impala

To see the results in the form of charts such as Bar-graphs or pie-charts we can analyze the data with the help of X-axis and Y-axis.

From the above figure.7.2 it shows the results in Bar-graphs. The results is based on Zip-code and date-received. Date-received is taken on X-axis where as zip-code is given on y-axis.

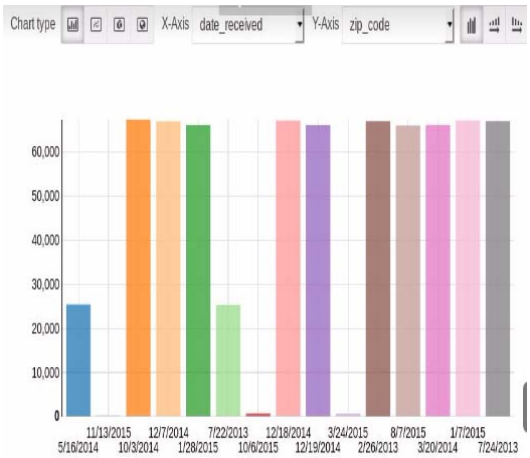


Fig.7.2 results in Bar graphs

Similar to the Bar-graphs we can create a pie-charts with the following Attributes.

From the above fig.7.3 the results are shown in the form of pie-chart with the attributes date-received and zip-code. Whereas zip-code is unique to each and every product.

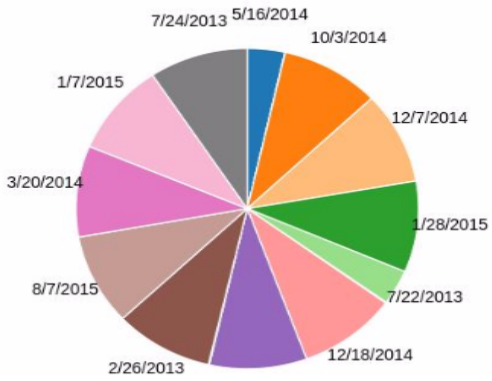


Fig.7.3 Results in pie-charts

8. CONCLUSION:

Health care analysis is important to find and there's no better way to collect direct feedback regarding the patients and improve the product or service. However, the way of handling a patient's health using the healthcare analysis is simpler task to provide a finer health quality to the patients As earlier loading large amount of data is very difficult. By using Big data complexity of loading large amount of data can be reduced. The proposed tool enables agencies too easily and economically clean, characterize and analyze the data to identify actionable patterns and trends.

9. REFERENCES

[1] "Big data analytics in healthcare : promise and potential" published by VijuRaghupathi and Wullianallur Raghupathi in 2014

[2]"Hadoop Based Analytics on Next Generation Medicare System" published by Gopal Tathe , Pratik Patil , Sangram Parle

[3] Big data for Better Health Planning" published by Jigna Ashish Patel and Priyanka Sharma

[4] Map Reduce Algorithms for Big Data Analysis" published byKyuseok Shim