International Journal of Applied Engineering Research ISSN 0973-4562 Volume 13, Number 7 (2018) pp. 5311-5315
© Research India Publications. http://www.ripublication.com

Data Analysis of Heart Disease Dataset using Hadoop and Impala with MYSQL

Niha Beera [1], Nysha Chaparala [1], Jaya Lakshmi Gundabathina [2]

[1] Student Scholar, [2] Assistant Professor
Department of Information Technology, VR Siddhartha Engineering College,
Vijayawada, Andhra Pradesh

Abstract

Analyzing and working with medical large data might be very complex using traditional methods similar to relational database management systems or desktop software parcels for classification and prediction. Heart disease is one of the major cause of death and disability in the world, killing million people a year. In recent times several work has been done to predict the possibility of heart disease, but only some factors measure the probability. In this paper we will analyze the possibility of heart disease among various symptoms in different countries. One of the software tools widely used for storage and processing of medical big data sets is Hadoop. IMPALA is used to build a predictive model on the heart disease dataset to analzse the possibility of patient heart disease.

Keywords: hadoop, impala, mysql, heart, hive, sqoop

INTRODUCTION

A. Importance of Health care:

Similar to confidentiality, health explore has high importance to society. It can supply significant information about disease trends and threat factors, outcomes of cure or public health intervention, practical ability, patterns of care, and health care costs and use. Medical trials can provide significant information about the efficiency and unpleasant property of medical interventions by calculating the variables that could impact the consequences of the study, but advice from real-world medical knowledge is also crucial for comparing and humanizing the use of drugs, vaccines, medical procedure, and diagnostics.

B. Heart diseases:

Heart disease is familiar both in the people at large but also in the people of working age. It is expected that heart disease, including stroke and high blood pressure, is in charge for more costs than any other disease or injury. The cost in industrial terms of cardiovascular disease (CVD) is yet, harder to enumerate but is likely to be similarly high.

C. Hadoop:

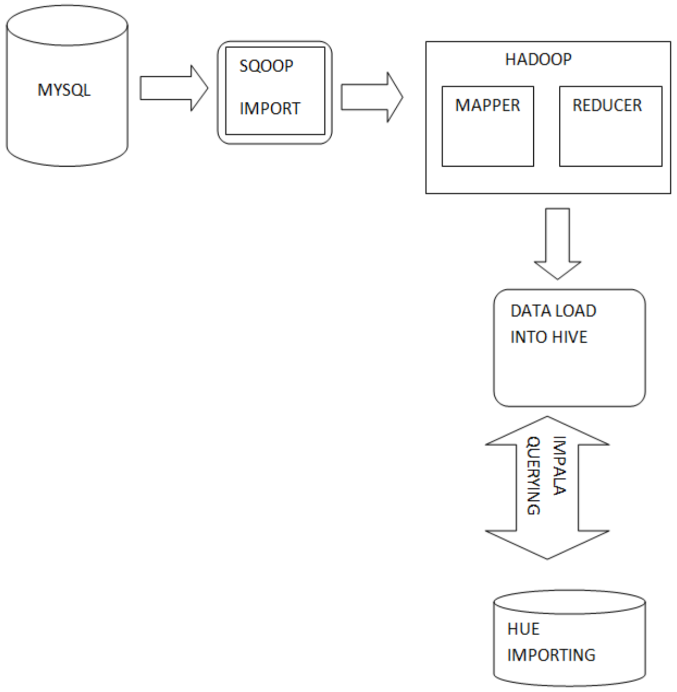
Hadoop is an open source, java based programming structure that chains the dealing out and storage space of tremendously big data sets in a scattered computing atmosphere. The Apache Hadoop software records is a structure that allows for the scattered dealing out of large data sets crossways clusters of computers using easy encoding models. It is calculated to level up from single servers to thousands of equipment, each offering local; computation and storage

LITERATURE SURVEY

This introduces features of the big data, health care data plus a number of major issues of big data. Big data in physical condition heed is used near guess the diseases, analyze the symptoms, recover the diagnosis, offer medicine perfectly for the patients to improve from early heart diseases[1].This summarises a little of the new works done in data mining linked to cardiovascular diseases. Big Data algorithms can be efficiently used to mine related information from the huge amounts of information generated by the healthcare business[2]. The discusses about CAD prediction and study of by unstructured data gives useful information for the patients on disease growth and direct it. possibility and efficiency of a disease and care supervision copy in the main health care organization for patients with heart breakdown[3]. The objective is to give a learn of dissimilar information mining techniques that can be working in automatic heart disease predict systems. A variety of techniques with data mining classifiers are defined in this work which has emerged in new years for ordered and efficient heart disease analysis. The study shows that different technologies are used in all the papers by means of taking different amount of attributes[4]. Every year, the American Heart Association, in conjunction with the Centers for Disease Control and Prevention, National Institutes of Health and other government agencies, compiles up-to-date statistics on heart disease, stroke and other vascular diseases in the *Heart Disease and Stroke Statistical.* Prevalence is an estimate of how many people have a specific disease, condition or risk factor at a given point in time[5]. The prediction of heart disease survivability has been a challenging research problem for many researchers. Therefore, the main objective of this manuscript is to report on a research project where we took advantage of those available technological advancements to develop prediction models for heart disease survivabilityThe survey of the papers related to heart disease and also the survey of many categories of heart disease such as coronary heart disease, coronary artery disease, heart failure, ischemic heart disease, cardiovascular disease, congenital heart disease, valvular heart disease and hypoplastic left heart syndrome are presented in this paper. [6]. Coronary artery disease often leads to myocardial infarction, which may be fatal. Risk factors can be used to predict CAD, which may subsequently lead to prevention or early intervention. Patient data such as co-morbidities, medication history, social history and family history are required to determine the risk factors for a disease[7]. Heart disease diagnosis is a challenging task which can offer automated prediction about the heart disease of patient so that further treatment can be made easy. Due to this fact, heart disease diagnosis has received immense interest globally among medical community. Based on this perspective, several researches have been conducted in the literature recently[8].

Machine Learning Techniques	Author	Year	Dataset description	Resources of Data Set	Tool	Accuracy
K-MEANS AND Navie BAYES	Rahul Patil	2016	Different types of disease symtomps	DIFFERENT SECTORS OF SOCIETY IN INDIA	HADOOP	95%
FUZZY C MEANS AND ID3	Mukesh Borana	2016	Different types of disease symtomps	UCI Data Repository	HADOOP	85.1%
C4.5 AND ID3 NAVIE BAYES	Gemson Andrew Ebenezer J.1	2015	Heart and diabetes dataset	UCI Data Repository	HADOOP	90%
SVM AND RANDOM FORST	Ashfaq Ahmed K	2013	Heart Disease,liver and cancer dataset.	DIABETIC RESEARCH INSTITUTE IN CHENNAI	MATLAB	65%

PROPOSED ARCHITECTURE



MYSQL:

MYSQL is an open source relational database management system.The tables are formed in the MYSQL and the information is loaded into them. To enter into the MYSQL, use the command "mysql –u root –p". Create a database in it to store the datasets. We have four countries datasets and all are loaded into it. at present the queries are printed to generate the tables and to fill the data into it.

SQOOP:

SQOOP is a process planned to move data between hadoop and relational database servers. It is used to introduce information from relational databases such as MYSQL to hadoop HDFS and send abroad from Hadoop file system to relational databases. Sqoop is used to load the data from MYSQL to HDFS.

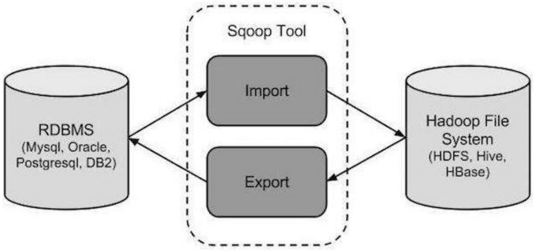


Figure 3.2 Architecture of Sqoop

HDFS:

The HDFS is a dispersed, scalable, and moveable file system printed in Java for the Hadoop structure. HDFS stores huge files across numerous apparatus. HDFS stores huge files across multiple machines. The Mapreduce program is inherent in it and it repeatedly generates the code in the HDFS. The data is then stored in the HDFS.

HIVE:

Hive is a information store infrastructure tool used to procedure the planned data in Hadoop. It reside on peak of Hadoop to sum up Big Data, and makes querying and analyzing easy. In hive the tables are created and all the datasets which are stored in HDFS are overloaded to hive.

HUE:

Hue is an interface in the Hadoop which contains Impala,Hive etc. The tables are automatically loaded from the hive to the hue . Through Hue the queries are written in the Impala to analyze the datasets.

IMPALA:

Apache Impala is an open source parallel processing SQL query engine for information stored in a computer cluster organization Apache Hadoop. Impala is included with Hadoop to utilize the similar file and data format, metadata, safety and reserve management frameworks used by MapReduce, Apache Hive. Impala being real-time query engine best suitable for analytics and for information scientists to perform analytics on information stored in Hadoop File System. In Impala queries are written to generate the graph and to analyse the possibility of heart disease in various countries depending on the symptoms.

DATASET DESCRIPTION:

The dataset consists of 14 attributes. They are-

1. age - age in years

2. sex - sex (1 = male; 0 = female)

3. cp - chest pain type

Value 1: typical angina

Value 2: atypical angina

Value 3: non-anginal pain

Value 4: asymptomatic

4. trestbps - resting blood pressure

5. chol - serum cholestoral in mg/dl

6. fbs - fasting blood sugar > 120 mg/dl (1 = true; 0 = false)

7. restecg - resting electrocardiographic results

Value 0: normal

Value 1: having ST-T wave abnormality - Value 2: showing probable or definite left ventricular hypertrophy by Estes' criteria

8. thalach - maximum heart rate achieved

9. exang - exercise induced angina (1 = yes; 0 = no)

10. oldpeak - ST depression induced by exercise relative to rest

11. slope - the slope of the peak exercise ST segment

Value 1: upsloping

Value 2: flat

Value 3: downsloping

12. ca - number of major vessels (0-3) colored by flourosopy

13. thal - 3 = normal;

6 = fixed defect;

7 = reversable defect

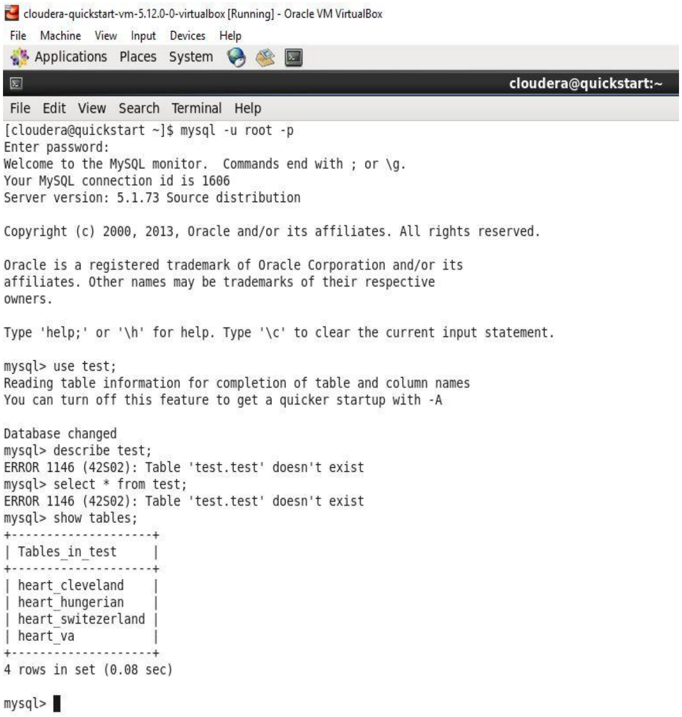
14. num - diagnosis of heart disease

Value 0: < 50% diameter narrowing

Value 1: > 50% diameter narrowing

RESULTS AND OBSERVATIONS:

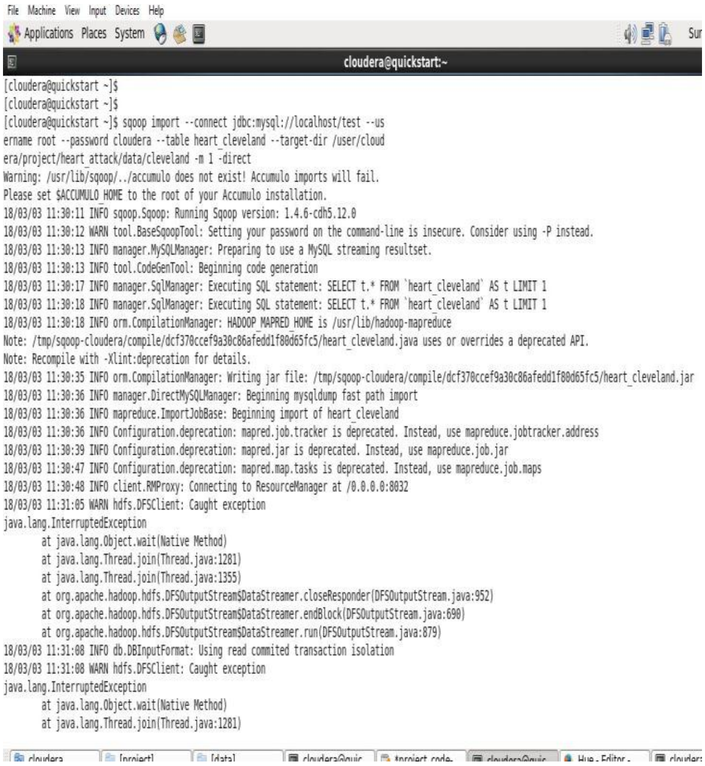
In MYSQL, tables are created and the data is loaded into it.


cloudera-quickstart-vm-5.12.0-0-virtualbox [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
Applications Places System
cloudera@quickstart:~
File Edit View Search Terminal Help

```
[cloudera@quickstart ~]$ mysql -u root -p
Enter password:
Welcome to the MySQL monitor.  Commands end with ; or \g.
Your MySQL connection id is 1606
Server version: 5.1.73 Source distribution

Copyright (c) 2000, 2013, Oracle and/or its affiliates. All rights reserved.

Oracle is a registered trademark of Oracle Corporation and/or its
affiliates. Other names may be trademarks of their respective
owners.

Type 'help;' or '\h' for help. Type '\c' to clear the current input statement.

mysql> use test;
Reading table information for completion of table and column names
You can turn off this feature to get a quicker startup with -A

Database changed
mysql> describe test;
ERROR 1146 (42S02): Table 'test.test' doesn't exist
mysql> select * from test;
ERROR 1146 (42S02): Table 'test.test' doesn't exist
mysql> show tables;
+--------------------+
| Tables_in_test     |
+--------------------+
| heart_cleveland    |
| heart_hungerian    |
| heart_switezerland |
| heart_va           |
+--------------------+
4 rows in set (0.08 sec)

mysql> 
```

Figure 5a

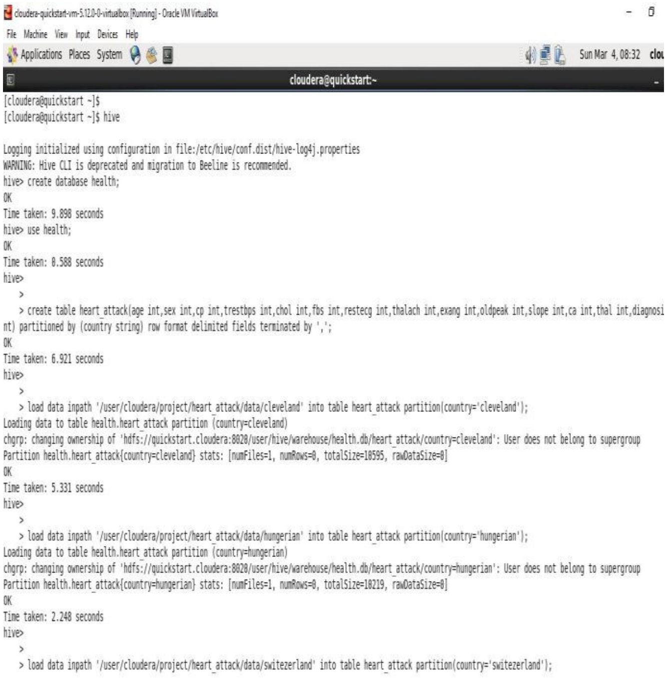
Sqoop is the tool worn to introduce information from MYSQL to HDFS. It is used to introduce data from relational databases such as MySQL, Oracle to Hadoop HDFS, and export it from Hadoop file system to relational databases.



```
[cloudera@quickstart ~]$
[cloudera@quickstart ~]$
[cloudera@quickstart ~]$ sqoop import --connect jdbc:mysql://localhost/test --us
ername root --password cloudera --table heart_cleveland --target-dir /user/cloud
era/project/heart_attack/data/cleveland -m 1 -direct
Warning: /usr/lib/sqoop/../accumulo does not exist! Accumulo imports will fail.
Please set $ACCUMULO_HOME to the root of your Accumulo installation.
18/03/03 11:30:11 INFO sqoop.Sqoop: Running Sqoop version: 1.4.6-cdh5.12.0
18/03/03 11:30:12 WARN tool.BaseSqoopTool: Setting your password on the command-line is insecure. Consider using -P instead.
18/03/03 11:30:13 INFO manager.MySQLManager: Preparing to use a MySQL streaming resultset.
18/03/03 11:30:13 INFO tool.CodeGenTool: Beginning code generation
18/03/03 11:30:17 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `heart_cleveland` AS t LIMIT 1
18/03/03 11:30:18 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `heart_cleveland` AS t LIMIT 1
18/03/03 11:30:18 INFO orm.CompilationManager: HADOOP_MAPRED_HOME is /usr/lib/hadoop-mapreduce
Note: /tmp/sqoop-cloudera/compile/dcf370ccef9a30c86afedd1f80d65fc5/heart_cleveland.java uses or overrides a deprecated API.
Note: Recompile with -Xlint:deprecation for details.
18/03/03 11:30:35 INFO orm.CompilationManager: Writing jar file: /tmp/sqoop-cloudera/compile/dcf370ccef9a30c86afedd1f80d65fc5/heart_cleveland.jar
18/03/03 11:30:36 INFO manager.DirectMySQLManager: Beginning mysqldump fast path import
18/03/03 11:30:36 INFO mapreduce.ImportJobBase: Beginning import of heart_cleveland
18/03/03 11:30:36 INFO Configuration.deprecation: mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
18/03/03 11:30:39 INFO Configuration.deprecation: mapred.jar is deprecated. Instead, use mapreduce.job.jar
18/03/03 11:30:47 INFO Configuration.deprecation: mapred.map.tasks is deprecated. Instead, use mapreduce.job.maps
18/03/03 11:30:48 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
18/03/03 11:31:05 WARN hdfs.DFSClient: Caught exception
java.lang.InterruptedException
        at java.lang.Object.wait(Native Method)
        at java.lang.Thread.join(Thread.java:1281)
        at java.lang.Thread.join(Thread.java:1355)
        at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.closeResponder(DFSOutputStream.java:952)
        at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.endBlock(DFSOutputStream.java:690)
        at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.run(DFSOutputStream.java:879)
18/03/03 11:31:08 INFO db.DBInputFormat: Using read commited transaction isolation
18/03/03 11:31:08 WARN hdfs.DFSClient: Caught exception
java.lang.InterruptedException
        at java.lang.Object.wait(Native Method)
        at java.lang.Thread.join(Thread.java:1281)
```

Figure 5b

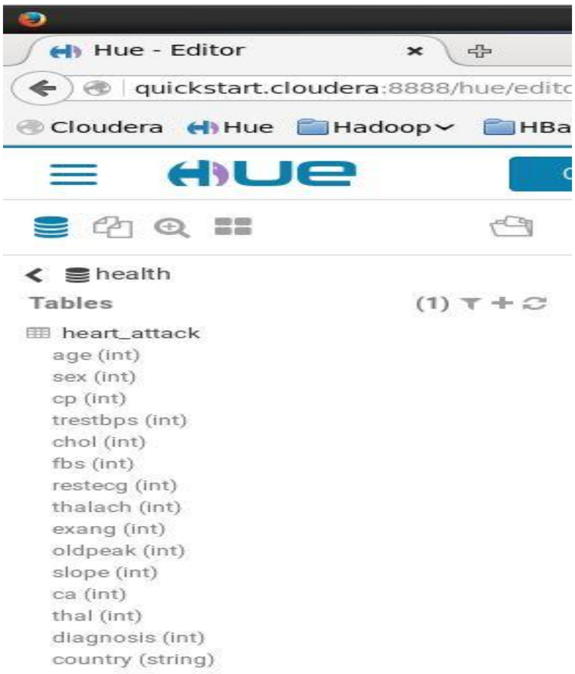
Enter the Hive. It provides a database query interface to Apache Hadoop to create table and load the data into the table.


cloudera-quickstart-vm-5.12.0-0-virtualbox [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
Applications Places System
Sun Mar 4, 08:32
cloudera@quickstart:~

```
[cloudera@quickstart ~]$
[cloudera@quickstart ~]$ hive

Logging initialized using configuration in file:/etc/hive/conf.dist/hive-log4j.properties
WARNING: Hive CLI is deprecated and migration to Beeline is recommended.
hive> create database health;
OK
Time taken: 9.898 seconds
hive> use health;
OK
Time taken: 0.588 seconds
hive>
    >
    > create table heart_attack(age int,sex int,cp int,trestbps int,chol int,fbs int,restecg int,thalach int,exang int,oldpeak int,slope int,ca int,thal int,diagnosi
nt) partitioned by (country string) row format delimited fields terminated by ',';
OK
Time taken: 6.921 seconds
hive>
    >
    > load data inpath '/user/cloudera/project/heart_attack/data/cleveland' into table heart_attack partition(country='cleveland');
Loading data to table health.heart_attack partition (country=cleveland)
chgrp: changing ownership of 'hdfs://quickstart.cloudera:8020/user/hive/warehouse/health.db/heart_attack/country=cleveland': User does not belong to supergroup
Partition health.heart_attack{country=cleveland} stats: [numFiles=1, numRows=0, totalSize=10595, rawDataSize=0]
OK
Time taken: 5.331 seconds
hive>
    >
    > load data inpath '/user/cloudera/project/heart_attack/data/hungerian' into table heart_attack partition(country='hungerian');
Loading data to table health.heart_attack partition (country=hungerian)
chgrp: changing ownership of 'hdfs://quickstart.cloudera:8020/user/hive/warehouse/health.db/heart_attack/country=hungerian': User does not belong to supergroup
Partition health.heart_attack{country=hungerian} stats: [numFiles=1, numRows=0, totalSize=10219, rawDataSize=0]
OK
Time taken: 2.248 seconds
hive>
    >
    > load data inpath '/user/cloudera/project/heart_attack/data/switezerland' into table heart_attack partition(country='switezerland');
```

Figure 5c

Enter the Hue. It is an interface which is used to analyze the dataset using Impala through Hue. The dataset and tables are directly loaded into the Hue from Hive because hive and impala are integrated with each other to store the tables into two components.



health
Tables (1)
heart_attack
age (int)
sex (int)
cp (int)
trestbps (int)
chol (int)
fbs (int)
restecg (int)
thalach (int)
exang (int)
oldpeak (int)
slope (int)
ca (int)
thal (int)
diagnosis (int)
country (string)

Figure 5d

In Impala we will write the queries to analyze the dataset and to predict the possibility of heart disease in different countries based on the symptoms. The query is written on partitioning the dataset upon different symptoms. The graphs are generated on both the genders such as males and females.

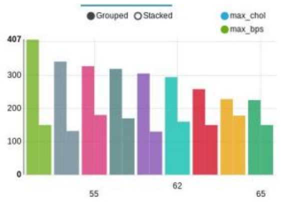


Figure 5e

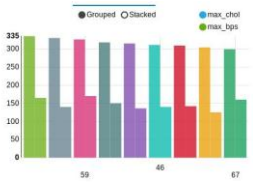


Figure 5f

The above graphs are generated on the country Cleveland.

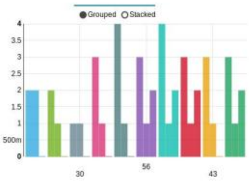


Figure 5g

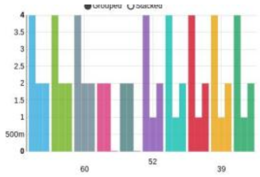


Figure 5h

The above graphs are generated on the country Hungerian.

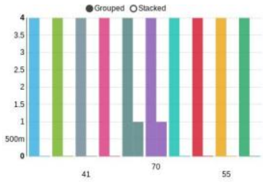


Figure 5i

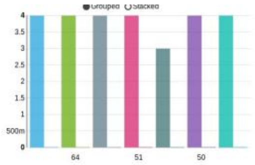


Figure 5j

The above graphs are generated on the country Switzerland.

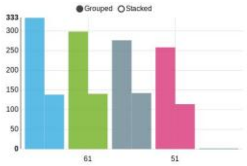


Figure 5k

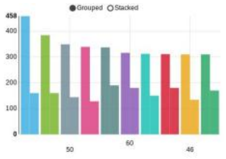


Figure 5l

The above graphs are generated on the country vatican city.

CONCLUSION

Hadoop with mysql and impala are used to store the huge amount of information and also for predicting the possibility of heart disease among various symptoms. The dataset which consists of symptoms are taken as the input. Queries are generated based on the input data, if the symptom values are more then there is more probability of heart disease. Based on that the graphs are generated after analyzing the symptoms to show the maximum probability.

REFERENCES

[1] Al Mamoon I, Sani AS, Islam AM, Yee OC, Kobayashi F, Komaki S (2013) A proposal of body implementable early heart attack detection system, 1-4.

[2] Ghadge P, Girme V, Kokane K, Deshmukh P (2015) Intelligent heart attack prediction system using big data. International Journal of Recent Research in Mathematics Computer Science and Information Technology 2: 73-77.

[3] Wanaskar UH, Ghadge P, Girmev V, Deshmukh P, Kokane K (2016) Intelligent Heart attack prediction system using big data. International Journal of Advanced Research in Computer and Communication Engineering 5: 723-725.

[4] Ciccone MM, Aquilino A, Cortese F, Scicchitano P, Sassara M, et al. (2010) Feasibility and effectiveness of a disease and care management model in the primary health care system for patients with heart failure and diabetes (Project Leonardo). Vasc Health Risk Manag 6: 297-305.

[5] Mozaffarian D, Benjamin EJ et al.,” Heart disease and stroke statistics- update” a report from the American Heart Association, pp.4-9, 2016.

[6] Prediction of Heart Disease using Classification Algorithms by Hlaudi Daniel Masethe, Mosima Anna Masethe, USA.

[7] Jitendra Jonnagaddala et al., “Coronary artery disease risk assessment from unstructured electronic health records using text mining” Journal of Biomedical Informatics, volume 58,pp.203-210,2015.

[8] Hui Yang and Jonathan M. Garibaldi, “A hybrid model for automatic identification of risk factors for heart disease” Journal of Biomedical Informatics,volume 58, pp.171-182,2015.