International Journal of Computer & Mathematical Sciences
IJCMS
ISSN 2347 – 8527
Volume 4, Issue 10
October 2015

Application of Big Data Analytics to UCI Diabetes Data Set Using Horton works Hadoop

Muni Kumar N
Asst. Professor, Department of CSE,
SKIT, Srikalahasti, India.

Xiaolin Qi
27-3-6-2 Haiyi, Kaifa District,
Dalian, China 116600

Abstract *— Diabetes Mellitus has become a major public health problem in India. Latest statistics on Diabetes reveal that, 63 million people in India are suffering from diabetes and this figure is likely to go up to 80 million by 2025. Type 2 Diabetes (T2D) is strongly associated with Morbidity and mortality, and carries a heavy financial burden[1].A better understanding and analysis of the determinants that progresses pre-diabetes to diabetes using Diabetic datasets and Big Data analytics may improve patient outcomes and lower the cost of inpatient care. The term 'Big Data' refers to the massive volumes of both structured and unstructured data which couldn't be stored and processed using traditional database management systems. With the rapid increase in the diabetic patients in India and number of determinants for the diabetes, the data will grow enormously and becomes Big Data which couldn't be processed by conventional DBMS. This paper discusses about Hadoop, an open source framework that allows the distributed processing for massive datasets on cluster of computers. Further, it also addresses the characteristics, the critical computing and analytical ability of Big Data in processing huge volumes of transactional data in real time situations.The objective of this paper is to analyze and extract the useful knowledge using big data analytics with Hadoop installed on top of Hortonworks Data Platform (HDP) Sandbox. This paper also presents the step by step execution of the Hartonworks Hadoop for the extraction of useful information based on the query related to the various determinants of Diabetes from UCI Diabetes dataset.*

Keywords *—* ***Diabetes Mellitus, Big Data Analytics, Type 2 Diabetes, Hortonworks Hadoop, UCI Diabetes Data Set.***

I. INTRODUCTION

After combating gigantic problem of communicable diseases, like many developing nations, India is also facing the new problem of chronic non communicable diseases such as Diabetes because of rapid urbanization and adaptation of modern life styles [12]. After Hypertension, Diabetes Mellitus (DM) is one of the most daunting challenges posed by chronic non-communicable disease. Diabetes has been proved to be the leading cause of morbidity and mortality in developed countries, and is gradually emerging as an important health problem in developing countries as well[12].Insulin dependent diabetes mellitus (IDDM) is a chronic disease that appears when hormone insulin has not been produced enough in patient's body [8]. In parallel with the rising prevalence of obesity worldwide, especially in younger people, there has been a dramatic increase in recent decades in the prevalence of metabolic consequences of obesity, in particular pre-diabetes and type 2 diabetes mellitus (DM2). It is predicted that by 2030 almost 10% of the world's population will have diabetes mellitus[11]. As obesity and DM2 are associated with a wide range of serious chronic health complications affecting renal, neurologic, retinal, cardiac and vascular systems with consequent decreased life span, the anticipated impact on global health and health care costs is enormous. The International Diabetes Federation estimated that in 2012, more than 371 million people worldwide had DM and that treating DM accounted for atleast $471 billion which is around 11% of total health care expenditures in adults.

For predicting diabetes risk, numerous patient level historic, clinical, biochemical and genetic risk factors for development of DM2 have been identified[11] and a range of predictive models have been proposed to more precisely estimate risk. A variety of models have been developed by incorporating simple clinical parameters such as age, weight, body mass index (BMI), family history of DM2 and blood pressure; basic laboratory measures such as glucose and lipid levels or more complex inflammatory, biochemical and genetic markers. Existing tools in the market do provide an estimate of absolute risk for DM2 using categorical rather than continuous variables and do not include HbA_{1c}, which is superseding glucose testing.

None of the tools available today for DM2 risk prediction have seen widespread adoption in clinical practice, and many lack adequate external validation in different settings. Thus, estimation of an individual's absolute risk for developing DM2 remains a challenge. The potential approach recently becoming available is the use of very large clinical databases from diverse settings to develop, refine and validate practical tools to predict individual absolute risk for developing DM2. The rapid increase in computer storage and database analysis capacity, along with the advent of electronic medical records in recent years has facilitated the aggregation of a vast amount of patient-level clinical data.

Historically, the health care industry in general, has generated huge volumes of data, driven by record keeping, compliance and regulatory requirements and patient care, which is considered as big data. In particular, the test data corresponding to the Diabetic Mellitus have huge number of determinants or attributes which constitute Big data. With the advent of technology and the changing life-styles of people, more and more are getting affected to Diabetes and the awareness towards the treatment also increased to consult the doctor during the early stages to prevent any harmful effects in the later stages. This part of the work requires the role of the data scientist to mine/analyse the big data and discover the associations, understand patterns and trends to improve healthcare, decrease the diabetic affected patients, increase life expectancy and lower costs involved by proper diagnosing during the pre-diabetic stages.

This paper is organized as follows: in section II, we discuss about the literature survey, section III discusses the characteristics of Big Data, section IV focuses on the Hadoop for Big Data, section V discuses about the Hortonworks Hadoop, section VI discusses about the Data Set considered for the Experiment, section VII shows the step by step procedure to use Hortonworks Hadoop for the query processing as part of big data analytics and section VIII concludes the work.

II. LITERATURE SURVEY

India has a high prevalence of diabetes mellitus and the numbers are increasing at an alarming rate. In India alone, diabetes is expected to increase from 40.6 million in 2006 to 79.4 million by 2030 and the projected estimate of the people with diabetes worldwide is 354 million [16]. This statistics clearly indicates that, out of 4 diabetic people in the world, one will be Indian. Other studies have shown that the prevalence of diabetes in urban Indian adults is about 12% and the Type 2 Diabetes is 4-6 time higher in urban than in rural areas. This growth in the urban areas is because of the increase in the rates of obesity which have tripled in the last two decades due to the change in life-style and lack of physical activity. Type 2 Diabetes (T2D) is strongly associated with morbidity and mortality and carries a heavy financial burden[1].

Diabetes Mellitus

Diabetes mellitus [19] is a chronic, lifelong condition that affects your body's ability to use the energy found in food. There are three major types of diabetes: type 1 diabetes, type 2 diabetes, and gestational diabetes. All types of diabetes mellitus have something in common. Normally, your body breaks down the sugars and carbohydrates you eat into a special sugar called glucose. Glucose fuels the cells in your body. But the cells need insulin, a hormone, in your bloodstream in order to take in the glucose and use it for energy. With diabetes mellitus, either your body doesn't make enough insulin, it can't use the insulin it does produce, or a combination of both.Since the cells can't take in the glucose, it builds up in your blood. High levels of blood glucose can damage the tiny blood vessels in your kidneys, heart, eyes, or nervous system. That's why diabetes -- especially if left untreated – can eventually cause heart disease, stroke, kidney disease, blindness, and nerve damage to nerves in the feet.A periodic test called the A1C blood test estimates glucose levels in your blood over the previous three months. It's used to help identify overall glucose level control and the risk of complications from diabetes, including organ damage.

Big Data and Hadoop

When the data is huge and could not be handled by the conventional database management system, then it is called big data. The big data can be in three forms, unstructured, semi-structured and structured form. The unstructured for of big data is difficult to handle and it requires Apache Hadoop which contains better tools and techniques to handle huge amounts of data. The Hadoop project contains a distributed file system called Hadoop Distributed File System (HDFS) and the Map Reduce algorithms [2]. The origin of big data could be from several places including logs, social media and live streaming. The data thus generated should be stored on the storage manager, the storage should be managed in such a way that it can be retrieved in an effective way, the effective means that, the retrieval should take less time, less CPU instructions, Less band width etc. HDFS stores data on different nodes. The storage will be in the form of blocks. The default size of each block is 64 MB. The Hadoop system consists of Name Node, Secondary Node, Data Node, Job Tracker and Task Tracker.

III. CHARACTERISTICS OF BIG DATA

Big data generally refers to the social network data from the micro-blogging sites like Twitter, LinkedIn and social media platforms like Facebook, Traditional enterprise including transactional data, web store transactions etc. and machine generated / sensor data like call data records, smart meters, manufacturing sensors, trading systems, traffic data, air data etc. which keeps on increasing without the human intervention. Big data is not only driven by the exponential growth of data but also by changing user behaviour and globalization. Globalization provides competition among the participants in the market. As a result, organizations are constantly looking for opportunities to increase their competitive advantage by using better analytical models. The following are the four characteristics of big data.

Volume: Data volume has been increasing exponentially: up to 2.5 Exabytes of data is already generated and stored every day. This is expected to double by 2015. The Big data generated vast amounts of data being by organizations or individuals from Terabytes to Exabytes and Zettabytes of data.

Velocity: Big data grows rapidly, which generated unprecedented quantities need to be stored, transmitted, and processed quickly. Velocity is the speed at which the data is being generated like streamed data from various smart devices into social media and also camera streamed data which stores the data in motion from huge number of closed circuit cameras.

Variety: In Big data, the variety and heterogeneity of data sources and storage has increased, fuelled by the use of cloud, web and online computing. Variety makes big data really big. Big data comes from a great variety of sources and generally has three types: structured, semi-structured and unstructured. Structured data inserts a data warehouse already tagged and easily sorted but unstructured data is random and difficult to analyse. Semi structured data does not conform to fixed fields but contains tags to separate data elements.

Veracity: Big Data Veracity refers to the biases, noise and abnormality in data. Is the data that is being stored, and mined meaningful to the problem being analyzed. Veracity in data analysis is the biggest challenge when compared to other characteristics like volume and velocity.

IV. HADOOP FOR BIG DATA

Hadoop is an open source framework which employs a simple programming standard that allows distributed processing of massive data sets on clusters of computers. The entire technology incorporates shared utilities, a distributed file system (DFS), analytics and information storage platforms, plus an application layer which manages the activities like workflow, distributed processing, parallel computation and configuration management[7].

HDFS

The basic idea of Hadoop is to make use of the Distributed file system for storing and processing the data. This HDFS splits the file into blocks and these blocks are allocated in the Hadoop cluster nodes. The input data in HDFS is given once and it is processed by MapReduce and the outcomes are sent to HDFS. The HDFS data is safeguarded by duplication mechanism among the nodes which gives reliability and avaialability regardless of node failures.

In Hadoop, there are two types of HDFS nodes:

(1) Data Node (2) Name Node

Data Node stores the data blocks of the files

Name Node contains the metadata, with record blocks and a list of DataNodes in the cluster.

MapReduce

MapReduce is the programming paradigm that allows for massive scalability across hundreds or thousands of servers in the Hadoop cluster. MapReduce is the heart of Hadoop where the processing is carried out by assigning the tasks to various clusters.

V. HORTONWORKS HADOOP

Hadoop Distributed File System (HDFS) [21] is a Java-based file system that provides scalable and reliable data storage that is designed to span large clusters of commodity servers. HDFS, MapReduce, and YARN form the core of Apache Hadoop.

What HDFS Does?

HDFS was designed to be a scalable, fault-tolerant, distributed storage system that works closely with MapReduce. HDFS will “just work” under a variety of physical and systemic circumstances. By distributing storage and computation across many servers, the combined storage resource can grow with demand while remaining economical at every size.

These specific features ensure that the Hadoop clusters are highly functional and highly available:

- **Rack awareness** allows consideration of a node's physical location, when allocating storage and scheduling tasks
- **Minimal data motion.** MapReduce moves compute processes to the data on HDFS and not the other way around. Processing tasks can occur on the physical node where the data resides. This significantly reduces the network I/O patterns and keeps most of the I/O on the local disk or within the same rack and provides very high aggregate read/write bandwidth.
- **Utilities** diagnose the health of the files system and can rebalance the data on different nodes
- **Rollback** allows system operators to bring back the previous version of HDFS after an upgrade, in case of human or system errors
- **Standby NameNode** provides redundancy and supports high availability
- **Highly operable.** Hadoop handles different types of cluster that might otherwise require operator intervention. This design allows a single operator to maintain a cluster of 1000s of nodes.

How HDFS Works?

An HDFS cluster is comprised of a NameNode which manages the cluster metadata and DataNodes that store the data. Files and directories are represented on the NameNode by inodes. Inodes record attributes like permissions, modification and access times, or namespace and disk space quotas.[21]

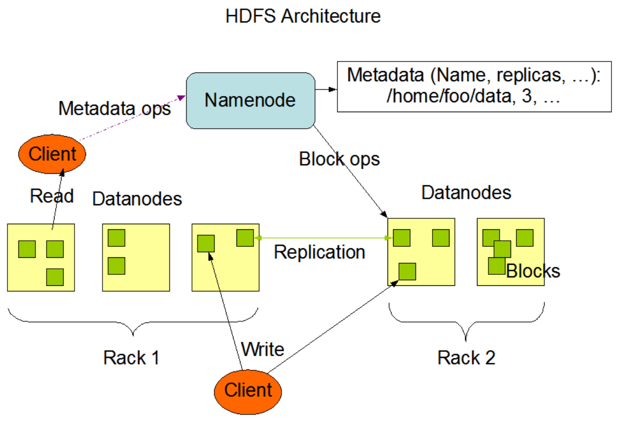


Fig 1. HDFS Architecture

The file content is split into large blocks (typically 128 megabytes), and each block of the file is independently replicated at multiple DataNodes. The blocks are stored on the local file system on the datanodes. The NameNode actively monitors the number of replicas of a block. When a replica of a block is lost due to a DataNode failure or disk failure, the NameNode creates another replica of the block. The NameNode maintains the namespace tree and the mapping of blocks to DataNodes, holding the entire namespace image in RAM.

The NameNode does not directly send requests to DataNodes. It sends instructions to the DataNodes by replying to heartbeats sent by those DataNodes. The instructions include commands to: replicate blocks to other nodes, remove local block replicas, re-register and send an immediate block report, or shut down the node.

VI. DATA SET DESCRIPTION

The dataset [20] represents 10 year of clinical care at 130 US hospitals during 1999 – 2008 and integrated delivery networks. It includes 55 attributes representing patient and hospital outcomes. The data set consists of 101767 records, and the information was extracted from the database for encounters that satisfy the following criteria.

1. It is an inpatient encounter (a hospital admission)
2. It is a diabetic encounter, that is, any kind of diabetes was entered in to the system as a diagnosis.
3. The length of the stay was at least 1 day and at most 14 days.
4. Laboratory tests were performed during the encounter.
5. Medications were administered during the encounter.

The data set contains the data such attributes as patient number, race, gender, age, admission type, time in hospital, medical specialty of admitting physician, number of lab test performed, HbA1c test result, diagnosis, number of medication, diabetic medications, number of outpatient, inpatient, and emergency visits in the year before the hospitalization, etc.

The data set was collected from the UCI Machine Learning Repository

through the link https:// archive. ics.uci.edu /ml/datasets/ Diabetes+ 130-US+ hospitals +for +years+ 1999-2008# accessed on 18-12-2014.

The data has been prepared and donated to UCI Machine Learning Repository to analyze factors related to readmissions as well as other outcomes pertaining to patients with diabetes.

VII. EXPERIMENT

This experiment is conducted on data set described in Section –VI with 1,01,767 instances and 55 attributes in each instance. The configuration of the system used is

System	Sony VAIO VPCEG25EN
RAM	4 GB
Disk	50 GB
Processor	Intel Core i3-2330M, 2.20 GHz
CPU	64-bit
Operating System	Linux CentOS 64 Bit
Installation Process	Installed using Hortonworks Data Platform Sandbox 2.1
Hadoop	Hortonworks Hadoop

Hortonworks Data Platform Sandbox 2.1 [21] is installed using VM Ware Player as a virtual machine with the above configuration. The dataset was downloaded from the above link and stored in the local system. The dataset was diabetes_data.csv file with 1,01,767 instances.

The steps carried out in the experiment are:

Step 1: Open the VMWare Player and Play the Hartonworks_Hadoop_2.1 Virtual Machine. Click on Play Virtual Machine, this step will start the operating system and the required services for the Hadoop.

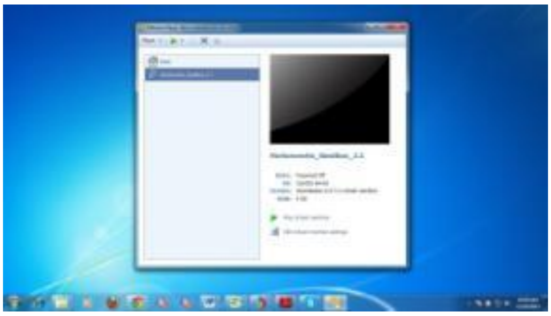


Fig 2: Start-up screen to Play virtual machine

Step 2: After successfully starting all the required services, we get the following black screen and it provides the ipaddress of the installed node to open in the browser.

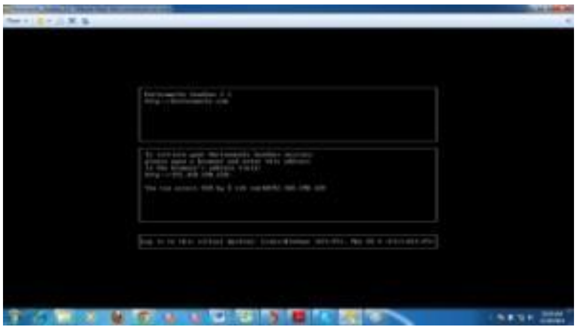


Fig 3: Screen to show Hadoop services started

Step 3: Open any web browser and type the ipaddress provided in the above screen to get the HUE Page and it displays the default username and password as hue/1111

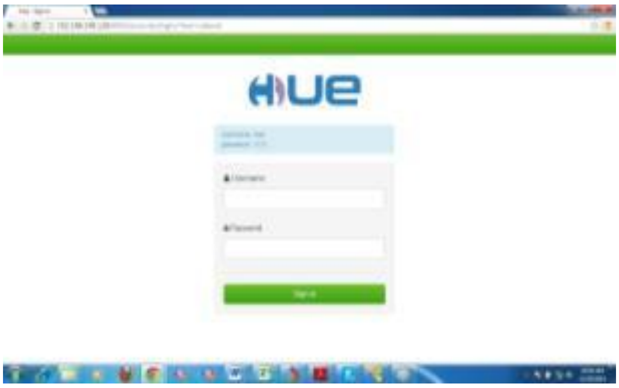


Fig. 4 Login screen for HUE dashboard

Step 4: After successful login to the HUE system, we get the dashboard as shown below. Which shows the list of components in Hadoop and their versions.

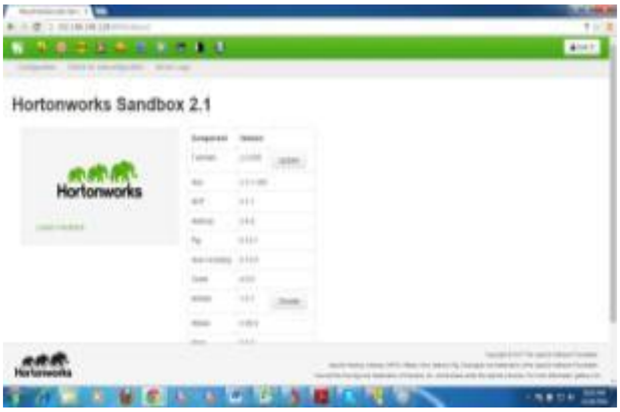


Fig. 5: Hue Dashboard

Step 5: Now click on the Beeswax (Hive UI), this opens the query editor screen as shown.

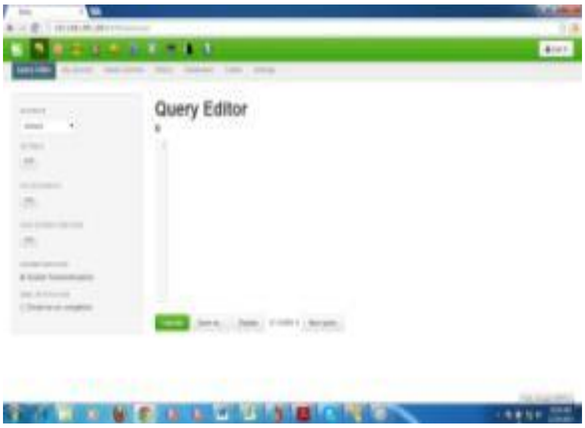
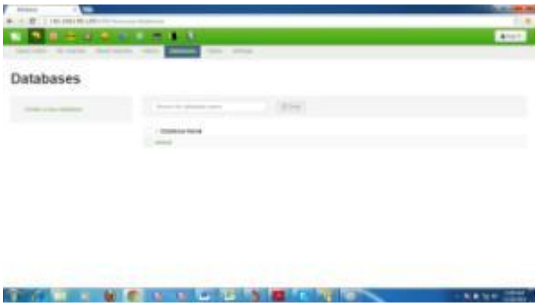


Fig 6: Bees Wax (Hive UI) – Query Editor

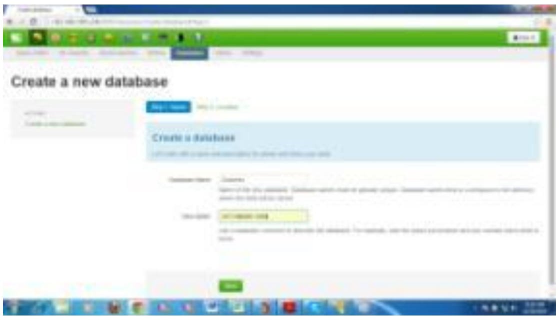
Step 6: As the database is not available, we will create a new database. Click on the Databases to get the below screen.



Databases

Fig 7: Databases Screen

Step 7: Now provide the name of the Database and an optional Database description. Then click next. Choose the location of the database, check the use default location and click on Create Database

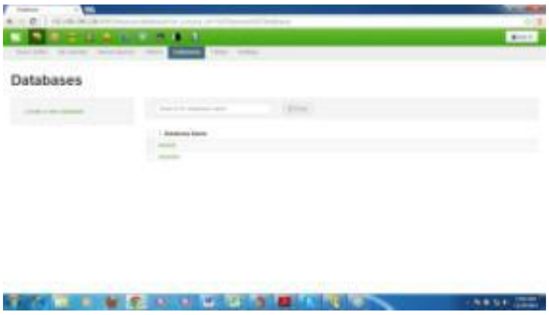


Create a new database

Create a database

Fig 8: Database creation screen

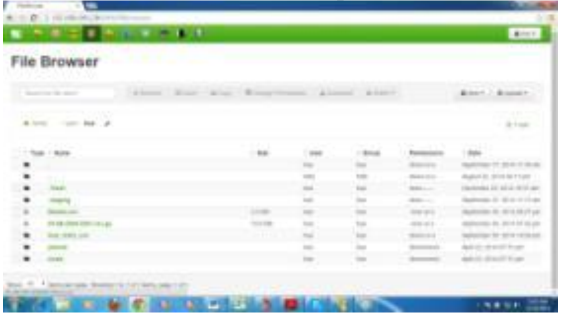
Step 8: Now, the newly created database appears in the list of databases.



Databases

Fig 9: List of Databases

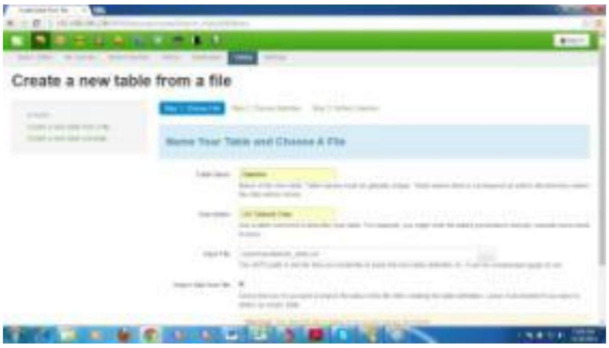
Step 9: Now, the database is ready to import data in to the database from the csv file, first we have to upload the file to HDFS. To do that, click on the file browser, that opens the following screen. Click on upload, point the location of the dataset in the local system and make sure the file uploads to HDFS.



File Browser

Fig. 10 File Browser

Step 10: Now, Go to the Database in Beeswax (Hive UI), click on tables. Choose the database, Select the action, Create a new table from a file. Provide the table name, optional table description and the input file path in HDFS.

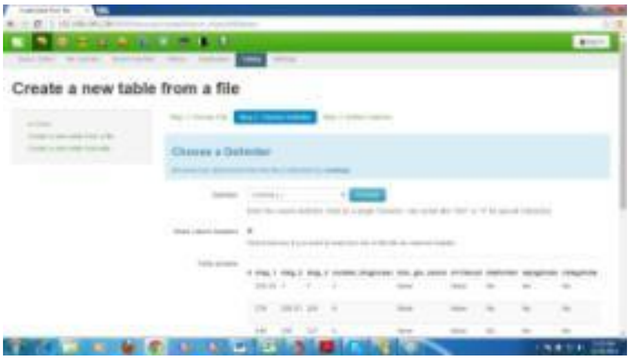


Create a new table from a file

Name Your Table and Choose A File

Fig.11. Table creation from a file

Step 11: Click Next, Choose Delimiter and the table preview will be shown.

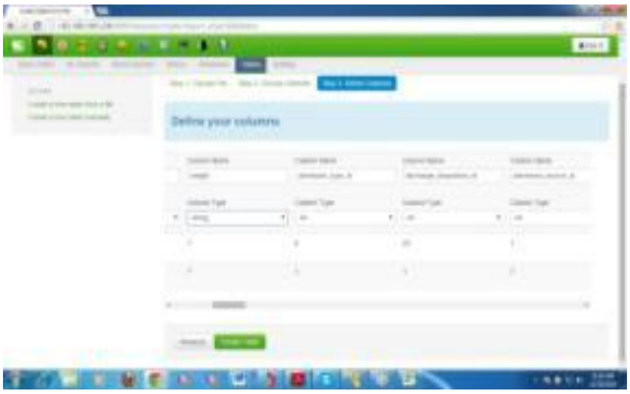


Create a new table from a file

Choose a Delimiter

Fig 12: Table creation from a file

Step 12: Click Next, define the columns and the column types. Now the table is ready for writing queries.



Define your columns

Fig.13. Defining Columns for table

Step 13: Now, Click on Bees Wax (Hive UI), in query editor, Write the following query to get the total number of records in the created table

Select count(*) from Diabetes;

Click on Execute.

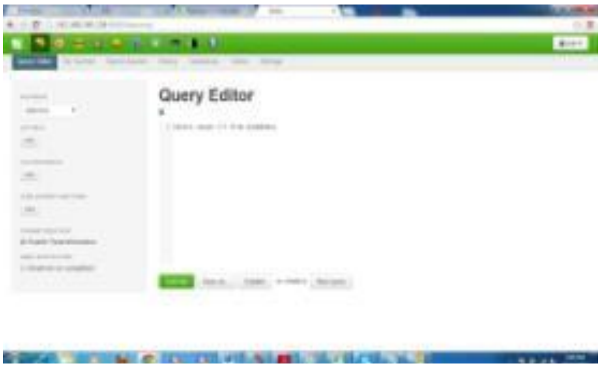


Fig.14. Query Editor

Step 14: As the query execution starts, the job will be created and the log file is written with all the actions happening in the background and the status of the Query will "Waiting for query.... Unsaved Query"

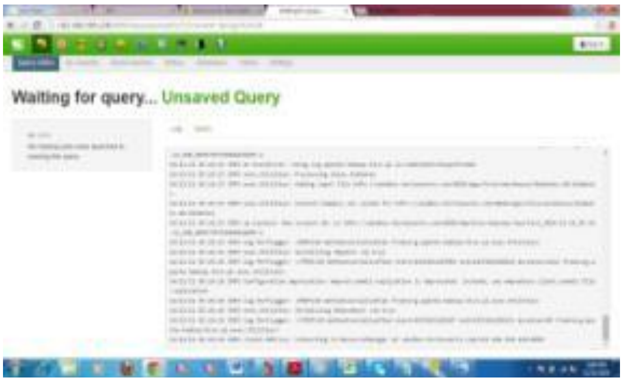


Fig.15. Waiting for query

Step 15: After the query gets executed. The results are displayed as shown below.

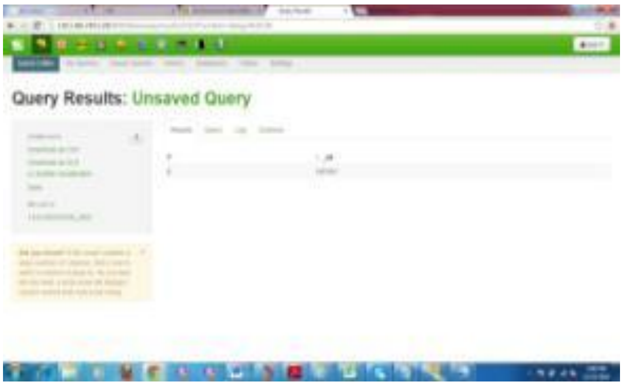


Fig.16. Query Results

Step 16:

The following is the log file consists of the details about the map reduce operations and the amount of time consumed for the processing of the query. As per the log information available for this query, the job is given at 14/12/21 03:26:03 and the results are obtained at 14/12/21 03:29:15, a total of 3 minutes and 12 seconds took for the complete execution

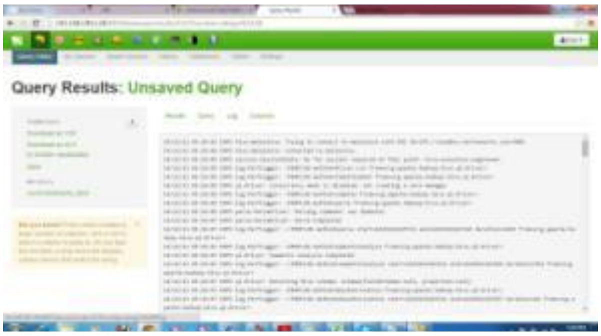


Fig.17.Query results log file

From the above log file, we observed that,

Job 0: Map: 1 Reduce: 1

Cumulative CPU: 12.93 sec

HDFS Read: 19159631

HDFS Write: 7

SUCCESS

Total Map Reduce

CPU Time Spent: 12 seconds 930 msec

Step 17:

Now, consider the other query

select * from diabetes where age = '[30-40)'; which get the list of patient records whose 30<= age < 40.

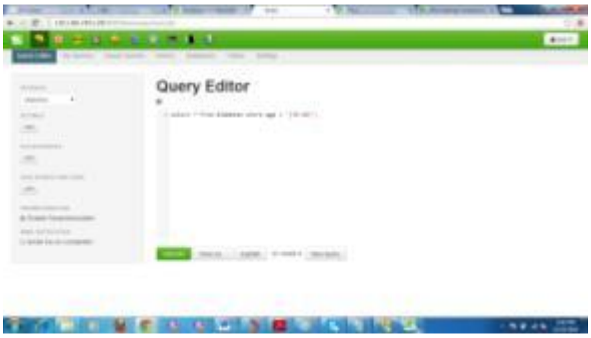


Fig.18. Query Editor

Step 18: The following results are obtained after the query execution.

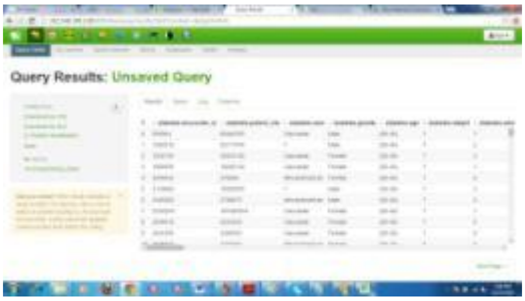


Fig.19. Query Results

Step 19: The log file for the above query shows

Job 0: Map: 1

Cumulative CPU: 3.78 sec

```
HDFS Read: 19159631
 HDFS Write: 711015
SUCCESS
Total MapReduce
CPU Time Spent: 3 seconds 780 msec
```

Query Results: Unsaved Query

Fig.20. Query Results log file

VIII. CONCLUSION

This paper discussed about the current statistics of Diabetic Mellitus in India and projected statistics by 2025 and showed that for every 4th persons having diabetes in the world will be an Indian as 25% of the world diabetic patients will be from India due to changes in the life styles. Also this paper discussed about the need for Big Data analytics for the analysis of Diabetic Mellitus datasets to predict and forecast the disease in pre-diabetic stages, so that better diagnosis can be made to reduce the diabetic patients. Further, this paper discussed about the Big Data and its characteristics and introduced Hortonworks Hadoop for the analysis of huge datasets using Map Reduce operation. Also this paper showed step by step procedure to create a database, create table, upload bulk data from file in to HDFS and then to table and execute the queries using the Hadoop Query Editor. Finally, the log file for the queries are analysed and the time taken for the query is identified.

ACKNOWLEDGMENT

We would like to take this opportunity to express our profound gratitude and deep regard to our colleagues and professors who encouraged and extended their co-operation in completion of this publication. Also, we would like to thank our family members for their constant support and encouragement in completing this paper.

REFERENCES

[1] Anderson JP et al, “Identification of Determinants Of Progression to Type 2 Diabetes Using Electronic Health Records and ‘Big Data’ Analytics”, ISPOR 19th Annual International Meeting , May 31- June 4, 2014, Canada

[2] Pal, A.; Agrawal, S., "An experimental approach towards big data for analyzing memory utilization on a hadoop cluster using HDFS and MapReduce," *Networks & Soft Computing (ICNSC), 2014 First International Conference on* , vol., no., pp.442,447, 19-20 Aug. 2014

[3] Sankaranarayanan, S.; Perumal, T.P., "Diabetic Prognosis through Data Mining Methods and Techniques," *Intelligent Computing Applications (ICICA), 2014 International Conference on* , vol., no., pp.162,166, 6-7 March 2014

[4] Dede, E.; Sendir, B.; Kuzlu, P.; Weachock, J.; Govindaraju, M.; Ramakrishnan, L., "A Processing Pipeline for Cassandra Datasets Based on Hadoop Streaming," *Big Data (BigData Congress), 2014 IEEE International Congress on* , vol., no., pp.168,175, June 27 2014-July 2 2014

[5] Strack, B.; Jonathan P.; et al, “Impact of HbA1c Measurement on Hospital Readmission Rates: Analysis of 70,000 Clinical Database Patient Recrods,” Research Article, Hindawi Publishing Coporation, Volume 2014, Article ID 781670, 2014.

[6] Elizabeth B; Andrea V; et al, “Opportunities and Challenges in Using Epidemiologic Methods to Monitor Drug Safety in the Era of Large Automated Health Databases,” Curr Epidemiol Rep (2014) 1: 194-205.

[7] Kotiyal, B.; Kumar, A.; Pant, B.; Goudar, R.H., "Big data: Mining of log file through hadoop," *Human Computer Interactions (ICHCI), 2013 International Conference on* , vol., no., pp.1,7, 23-24 Aug. 2013

[8] Motka, R.; Parmarl, V.; Kumar, B.; Verma, A.R., "Diabetes mellitus forecast using different data mining techniques," *Computer and Communication Technology (ICCCT), 2013 4th International Conference on*, vol., no., pp.99,103, 20-22 Sept. 2013

[9] NirmalaDevi, M.; Appavu, S.; Swathi, U.V., "An amalgam KNN to predict diabetes mellitus," *Emerging Trends in Computing, Communication and Nanotechnology (ICE-CCN), 2013 International Conference on* , vol., no., pp.691,695, 25-26 March 2013

[10] Velu, C.M.; Kashwan, K.R., "Visual data mining techniques for classification of diabetic patients," *Advance Computing Conference (IACC), 2013 IEEE 3rd International* , vol., no., pp.1070,1075, 22-23 Feb. 2013

[11] Harry Glauber; Eddy Karnieli; , “Preventing Type 2 Diabetes Mellitus: A Call for Personalized Intervention,” The Permanente Journal/Summer 2013/Volume 17 No.3 pp.74-79

[12] Hemant D Mahajan, “Health Profile of Diabetic Patients in an Urban Slum of Mumbai, India,” Innovative Journal of Medical and Health Science 3: 3 May – June. (2013) pp.102-109

[13] Ramani, R.G.; Balasubramanian, L.; Jacob, S.G., "Automatic prediction of Diabetic Retinopathy and Glaucoma through retinal image analysis and data mining techniques," *Machine Vision and Image Processing (MVIP), 2012 International Conference on* , vol., no., pp.149,152, 14-15 Dec. 2012

[14] Revolution Analytics White Paper, “ Advanced ‘Big Data’ Analytics with R and Hadoop,” www.revolutionanalytics.com, 2011

[15] Ramachandran A, et al. , “Current Status of Diabetes in India and Need for Novel Therapeutic Agents,” Review Article, Supplement to JAPI – June 2010. Vol 58. Pp. 7-9, 2010

[16] Mehta SR.; Kashyap AS.; Das S; “Diabetes Mellitus in India: The Modern Scourge”, Review Article, MJAFI, Vol. 65, No.1, 2009 pp.50-54

[17] Kayaer, K. and Yildirim, T., Medical diagnosis on Pima Indian diabetes using general regression neural networks. In: Proceedings of international conference on artificial neural networks neural information processing, pp. 181-184.

[18] http://www.diabetesfoundationindia.org/ Accessed on 17-12-2014

[19] http://www.webmd.com/diabetes/types-of-diabetes-mellitus Accessed on 17-12-2014

[20] Beata Strack, Jonathan P. DeShazo, Chris Gennings, Juan L. Olmo, Sebastian Ventura, Krzysztof J. Cios, and John N. Clore, “Impact of HbA1c Measurement on Hospital Readmission Rates: Analysis of 70,000 Clinical Database Patient Records,” BioMed Research International, vol. 2014, Article ID 781670, 11 pages, 2014.

[21] http://hortonworks.com/products/hortonworks-sandbox/ Accessed on 17-12-2014