ISSN (Print) : 0974-6846
Indian Journal of Science and Technology, *Vol 9(32), DOI: 10.17485/ijst/2016/v9i32/100184, August 2016* ISSN (Online) : 0974-5645

A Systematic Review of Type-2 Diabetes by Hadoop/Map-Reduce

Munaza Ramzan, Farha Ramzan and Sanjeev Thakur

Amity University, Noida - 201313, Uttar Pradesh, India; munaza.ramzan43@gmail.com, Id:frimz.r@gmail.com, sanjeevthakur3@amity.edu

Abstract

Objectives: The goal is to study about the disease called Diabetes Mellitus (Type_2) and how to cure it for better healthcare. **Methods/Statistical Analysis:** To explore dataset of the Type-2 Diabetes and R environment statisticallythe Hadoop/Map-Reducer algorithm will be used.Findings:Analyzing the different parameters of the disease, apart from medical diagnosis and causal agents, our review shows that by using Big Data we can predict the other factors which result in demographic variations of diabetes, geographical distribution of disease and its causes and other factors needed for better outcomes of healthcare. **Application/Improvements:** Big data is mostly used for data analytics in business, more research is needed to make use of this technology in other fields where the data generated is huge.

Keywords: Big-data, Diabetes-Mellitus (DM), Hadoop

1. Introduction

In the new era of digital world, the three attributes of big data Volume, velocity and variety are growing exponentially. For example: on a daily basis in 2013 an average of approx. 2.8 quintillion bytes of data were created[1]. The source of this data explosion includes smart phones, different types of sensors, individual Archives, Social-Networking sites, business Enterprises, digital audio video recorders, Software Logs, data from healthcare industry etc. Due to such data explosions it becomes challenging for researches to analyze and interpret the large and complex datasets manually or by using different data management applications thus leading to an issue known as "Big-Data Problem". The Big-Data problem is not confined to one sector but is a common phenomenon in sector of Science, Medicine, commerce and engineering. Healthcare industry is an important sector in terms of Big Data application; a wide variety of data sets differing in volume, variety and velocity are generated. The data from health industry as suggested by IDC digital universe study (2011) is growing at an enormous rate, according to their study in year 2005 around 130 Exabyte's of data was created and stored, in 2010 data generated grew rapidly to 1227Exabytes and in 2015 it is expected to grow at the rate of 45% i.e. around 7910 Exabytes. The application of Big-data analytics in health-care industry is emerging into a promising field and is providing insights from very large datasets to help in improving outcomes cost effectively[2].

Diabetes a prolonged metabolic disorder characterized by high level of sugar in blood which can either be due to inadequate insulin production by beta cells of pancreas or improper response of body's cells to insulin or both. In 2014, around 382 million people throughout the world were diagnosed with diabetes. As diabetes is a lifestyle disorder, treatment and prevention can be done through diet control, weight reduction, exercise and smoking cessation along with along with drugs (type2) and insulin (type 1) treatment[5].

2. Types of Diabetes

2.1 Type-1 Diabetes

This is an immune disorder characterized by insufficiency of beta cells of pancreas to produce insulin. Although, can occur at any age group but usually seen more in people who are in early adulthood or teenage years. Approximately 10% of all diabetes cases are of this type.

* *Author for correspondence*

2.2 Type-2 Diabetes

This is the common form of diabetes which is characterized by either inadequate insulin production by beta cells of pancreas or insulin produced is defective because of which cells in the body are not able to react with it. This form of diabetes is prevalent worldwide and comprises 90% of all diabetic cases. Obesity is a major predisposing factor for type 2 diabetes. This is a progressive disorder, which gradually becomes life threatening over a period of time and patients usually end up using insulin as a treatment option. A possible method of preventing type-2 diabetes is by adopting healthy life style habits like weight loss, healthy diet, exercise, and monitoring blood glucose levels on regular intervals.

2.3 Gestational-diabetes

This type of diabetes occurs in pregnant women. This type of gestational diabetes is temporary and usually disappears after pregnancy. Pregnant women need 3-4 times more insulin to combat high glucose levels in the blood. A woman who has gestational diabetes has 50%-60% chances of developing type 2 diabetes later in her life.

3. Symptoms of Diabetes

a) Polyuria(frequent urination)
b) Polydipsia (Increasingly thirsty)
c) Polyphagia(Hungry)
d) Weight Gain
e) Unusual Weight Loss
f) Fatigue
g) wounds that don't heal
h) Numbness and Tingling in hands and feet[6]

4. EmergingMedical Information Technologies

There are wide varieties of techniques and emerging models that are making a major stir in the medical information technologies and their applications. Some of the trends are observed like:

Health sensing: To capture various aspects of Physiological, Cognitive and Physical Health. The wide variety of consumer devices and medical Sensors were increasing sharply.

Big-data analysis in Health-care: With the growing digitization of health-care industry, enormous amount of health-care data has been generated and the size is increasing at a bizarre rate. To deliver best evidence based, generalized and patient centric care, the big data from industry has to be analyzed to discover deep knowledge and values.

Health-Care Cloud Computing: Cloud computing supports the analysis of big data in healthcare and enables healthcare providers to efficiently use computing resources, manage and improve services in a cost effective manner. How effective a health-care service would be, depends on the efficiency of health problem detection, and allocation of medical resource, which in turn is decided by how effectively healthcare information is collected, managed and utilized. This paper aims at documenting how to implement emerging technologies on creating new ways to access and use health-care information and at the same time help in improving the quality, safety, and efficiency of the health-care services.

Big-data in health-care includes the patient-oriented data (physician notes, Lab reports, radiological reports, case history, and Diet plans), list of health care providers in a hospital, national health register data, medicine and surgical instrument manuals and their expiry data[8]. Big-data technology is widely used by Health-care organizations to capture all the patient related information for better patient care coordination, outcomes of services, and management.

5. Literature Review

Mu-Hsing-Kuo[1] discusses how to improve healthcare services by improving Cloud computing tools and identify novel and innovative solutions to allocate resources for treating patients.

Elsevier, Software tools and techniques for big-data analytics in health-care[2], Discuses about how techniques are applied on the data which is growing both in terms of in the volume of data, rate of generation (velocity), and variety.

Big-Data[3], the data from health industry as suggested by IDC digital universe study (2011) is growing at an enormous rate, according to their study, in year 2005 around 130Exabytes of data was created and stored, in 2010 data generated grew rapidly to 1227Exabytes and in 2015 it is expected to grow at the rate of 45% i.e. around 7910Exabytes.

Kumar et al.[4] discusses about the big data and its characteristics, methods and challenges and suggests how to overcome the underlying problems being faced by the health care industry. Also presents the big ideas to fix the healthcare system in India.

Aljumah et al.[5] presents a study that concludes that treatment plan for elderly patients should be assessed on basis of their needs and lifestyle habits. Also predictions on the effectiveness of different treatment methods for young and old age groups were elucidated.

Rajesh etal.[6] disuses about the reliable prediction methodology to diagnose diabetes and interpret the data patterns so as to get meaningful and useful information for the healthcare providers.

Sharmila etal.[7] surveyed various progress made in the area of big-data technology, its latest adoption in Hadoop platform, algorithms used in such platform, and listed out the open challenges in using such algorithms in health-care data-sets.

Archenaa etal.[8] discusses about the big-data use cases in health-care and government. Also gives an insight of how uncover additional values from the data generated by health-care and government.

Ren etal.[9]discuses about how to detect diseases at prior stages based on the historical medical data, minimizing drug usage to avoid the side effects and give effective and efficient medicine based on genetic make-ups.

Dean etal.[10] discuss about how to address the basic needs of a citizen quickly and reliably using big-data analytics.

Yean etal.[11] discusses how big-data analytics helps the government to assess the educational needs for children who are in the age to be admitted to the school.

Hamilton etal.[12] discusses how to minimize the unemployment rate by predicting the job needs before based the literacy rate by analyzing the students data that are graduating every year.

Yang etal.[13] in analyzing genomic data of patients, The Big Data technologies such as the Apache Hadoop project provides distributed and parallelized data processing and analysis.

6. Methodology for Analysis

Our analysis is solely based on Apache Hadoop. To store and process enormous amount of data in a distributed manner, The Hadoop distributed file system is used. Our query is expressed as a MAP Reduce job for parallel processing. Query processing using MapReduce works by breaking the processing of data into the map phase and the reduce phase having their key value pairs as input and output. Programmer decides which type to be used depending on query. The environments like Hive, pig and R were used to analyze the data better[7].

6.1 Dataset Collection

The medical data sets for analysis were collected from the GHDx (Global Health Data Exchange). The GHDx is a data catalogue produced and supported by IHME (Institute for Health Metrics and Evaluation).

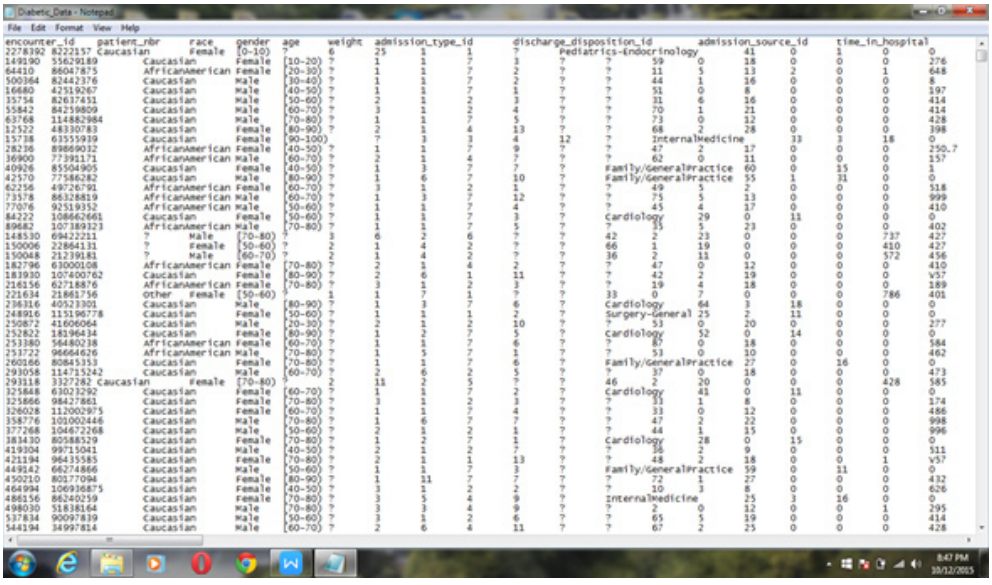

Figure 1. Screenshot of the data.

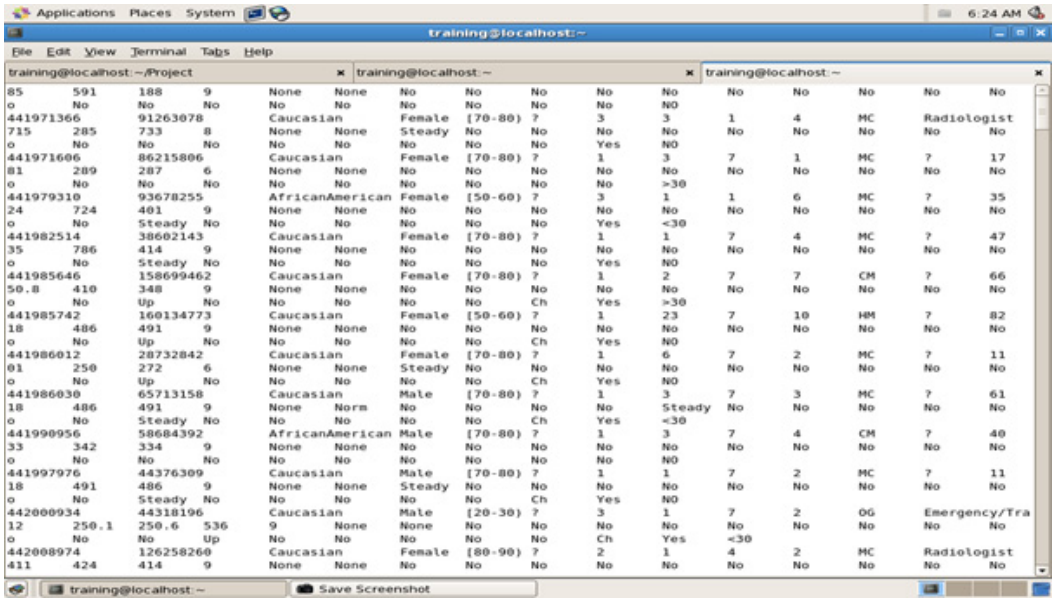

Figure 2. Screenshot of data in HDFS.

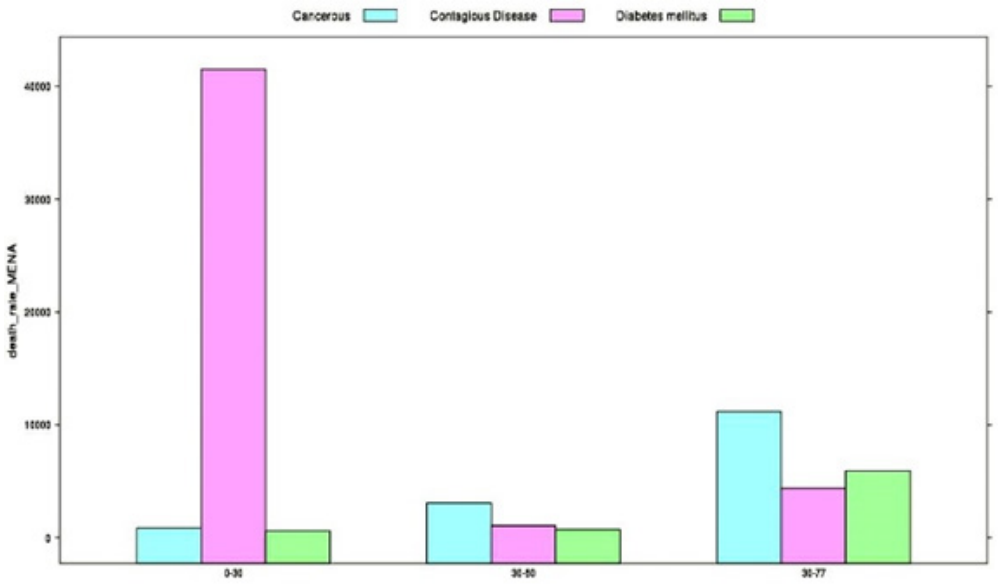

Figure 3. Death rate caused in particular age group of MENA.

This dataset, delimited with commas (.csv), provides the different attributes for the treatment of type-2 diabetes by race, age, and gender and also the age adjusted lower and upper limit in both genders. The data was in .csv format, we made it more unstructured so as to check the reliability of Hadoop when used on unstructured data-sets. Then again saved as in tab delimited format. Figures 1 and 2 are the screenshot of the dataset after all manipulation process.

The Type-2 Diabetes Mellitus has been monitored according to the following parameters:

a. North Africa and Middle East (Egypt, Iran, Iraq, Saudi Arabia, Morocco etc.)
b. South Asia (India, Pakistan, Afghanistan etc.)
c. East Asia (Mongolia, China, South Korea, Japan etc.)
d. Age Group (0-30, 30-60 and 60-77)
e. Year (1990, 2005 and 2010)
f. Race and Gender(Male, Female and Both)

7. Syntax for Plotting

]$ R> res <- read.Table("Data-File-Name", sep=",",header=FALSE)>barplot(Y_Axis_Data_Frame,names.arg=c(X_Axis_Data_Frame), col=c("Red", "Blue")).

8. Graphs

The following are the graphs with the corresponding variables and constants:

Result 1

Region: MENA (Iraq, Iran, Algeria, Arab countries, Sudan etc.)

Years: All

Gender: All

We have seen the deaths caused in particular years and in particular genders of MENA. Figure 3 bar plot gives a notion of the death rate caused in particular age group. As indicated the population of the age group "0-30" is most effected. While in the third age group cancer is the dominatingdisease category, causing more deaths.

Result2

Region: East Asia

Gender: All

Years: All

As indicated by the earlier graphs of East Asian countries the cancerous disease, which has caused most deaths in the region, is affecting the population of the age group "30-60" and "60-70". While as the contagious diseases is affecting the age group of "0-30" followed by

"60-77" and a little in "30-60". The death rate caused by Diabetes mellitus is almost constant for the first two age groups but is gaining hike in the last one (Figure 4).

Result3

Region: South Asia

Gender: All

Years: All

Figure 5 is the death rate in South Asian countries by three diseases. We have already analyzed that contagious diseases have caused most death rates in South Asian countries, with male population affected worse. This graph indicates that in the age group of 0-30 yrs. maximum deaths have been caused by communicable diseases. And next to this age group is "60-77", again the pink bar has the more shares of deaths in this age group, followed by the cancerous diseases. Age group "30-60" has not been affected as are the other two.

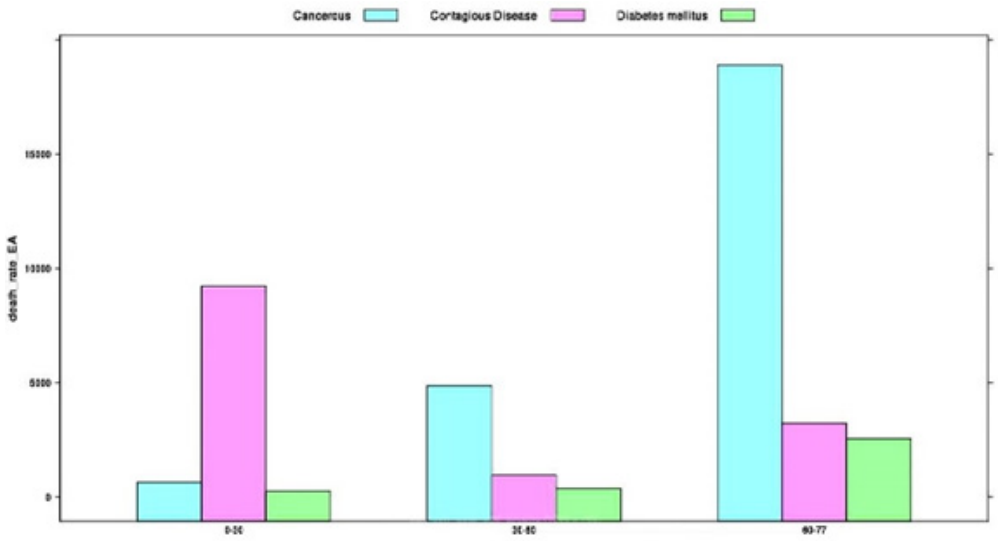

Figure 4. Death rate caused in particular age group of East Asia.

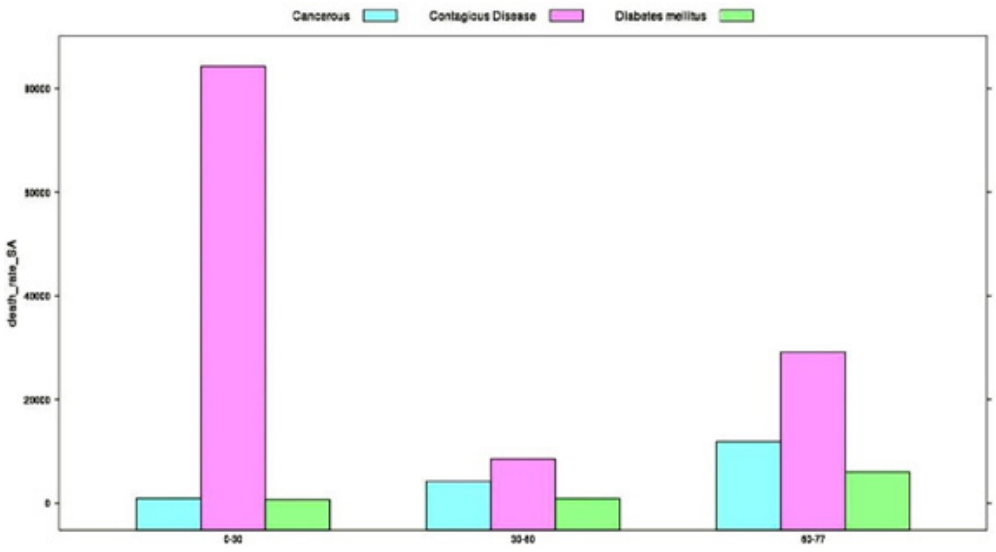

Figure 5. Death rate caused in particular age group of South Asia.

9. Conclusion

Big-data technology for data analytics is an emerging technology. Although, big data is mostly used for data analytics in business, more research is needed to make use of this technology in other fields where the data generated is huge. For our research, we used data from healthcare sector for analytics using Big Data. In this research the data of Diabetes Mellitus from different parts of the world was taken for analysis using Hadoop. Analyzing the different parameters of the disease, apart from medical diagnosis and causal agents, using Big Data we can predict the other factors which result in demographic variations of diabetes, geographical distribution of disease and its causes and other factors needed for better outcomes of health-care.

10 References

1. Yang JJ, Li J, Mulder J, Wang Y, Chen S, Wu H. Emerging information technologies for enhanced healthcare. Computers in Industry.2015 May, 69:3–11.
2. Lizhe Wang, Ranjan R, Kolodziej, Zomaya A. Software tools and techniques for big data computing inhealth-care clouds.Future Generation Computer Systems. 2015 Feb; 43:38–9.
3. Big-Data[Internet]. [Cited 2015 Apr 05]. Available from: https://www.ida.gov.sg/~/media/Files/Infocomm%20Landscape/Technology/TechnologyRoadmap/BigData.pdf.
4. Kumar NM, Manjula R. Role of big data analytics in rural health care – a step towardsSvasthBharath**.** International Journal of Computer Science and Information Technologies. 2014; 5(6):7172–78.
5. Aljumah A, Ahamad M, Siddiqui M. Application of data mining: diabetes health care in young and old patients. Journal of King Saud University-Computer and Information Sciences. 2015; 25(2):127–36.
6. Rajesh K,Sangeetha V. Application of data mining methods and techniques for diabetes diagnosis. The International Journal of Engineering and Innovative Technology. 2012; 2(3):224–29.
7. Sharmila K, Vethamanickam SA. Survey on data mining algorithm and its application in healthcare sector using Hadoop platform. International Journal of Emerging Technology and Advanced Engineering. 2015; 5(1):567–71.
8. Archenaa J, Anita EAM. A survey of big data analytics in healthcare and government. *Procedia Computer Science*. 2015; 50:408–13.
9. Ren Y, Werner R, Pazzi N, Boukerche A.Monitoring patients via a secure and mobile healthcare system. IEEE Wireless Communications.2010 Feb; 17(1).
10. Dean J, Ghemawat S. MapReduce: Simplified data processing on large clusters. ACM; 2008.
11. Yean J et al. White paperby U.S.General Services Administration, Big data bigger opportunities.
12. Hamilton B etal.Cognizant white paper. Big data is the future of healthcare;2010.
13. Patel JA, Sharma P. Big data for better health planning. 2014 International Conference on Advances in Engineering and Technology Research (ICAETR);2014.