Indian Journal of Science and Technology, *Vol 9(40), DOI: 10.17485/ijst/2016/v9i40/101618, October 2016*

ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645

Diagnosing Diabetic Dataset using Hadoop and K-means Clustering Techniques

K. Sharmila* **and S. A. Vetha Manickam**

Department of Computer Science, Vels University, Chennai - 600117, Tamil Nadu, India; sharmilasenthil@ymail.com, drsavm@gmail.com

Abstract

Objectives: The articles display how enormous measure of information in the field of social insurance frameworks can be dissected utilizing grouping method. Removing helpful data from this gigantic measure of information is profoundly compound, exorbitant, and tedious, in such territory information mining can assume a key part. Specifically, the standard information digging calculations for the examination of colossal information volumes can be parallelized for speedier preparing. **Methods/Statistical Analysis:** This paper concentrate on how grouping calculation to be specific K-means can be utilized as a part of parallel handling stage in particular Apache Hadoop bunch (MapReduce paradigm huge) so as to dissect the gigantic information quicker. **Findings:** As an early point, we complete examination keeping in mind the end goal to evaluate the adequacy of the parallel preparing stages as far as execution. **Applications/Improvements:** Based on the final result, it shows that Apache Hadoop with K-means cluster is a promising example for versatile execution to anticipate and analyze the diabetic infections from huge measure of information. The proposed work will give an insight about the big data prediction of diabetic dataset through Hadoop. In future this technology has to be extended on cloud so as to connect various geographic districts around Tamil Nadu to predict diabetic related diseases.

Keywords: Apache Hadoop, K-means, MapReduce

1. Introduction

Diabetes is a clinical disorder that is portrayed by hyperglycemia, because of insufficiency of insulin in the human body[1]. The disorder has turned out to be very standard in today's life independent of age gathering. The malady is endless and does not have a particular cure. It shifts from individual to individual contingent upon the manifestations and levels of glucose in human body. Diabetes has influenced more than 246 million individuals worldwide with a greater part of them being ladies. As per the WHO report, by 2025 this number is required to ascend to more than 380 million[2].

As of late the social insurance industry has produced monstrous volume of information. Big Data is the hottest trend in the business and IT world right now[4] for large dataset. Enormous information can be described by ***4Vs***: Volume, Variety, Velocity and Veracity. Volume measures the measure of information accessible, Variety is a measure of the extravagance of the information representation-content, pictures video, sound, and so forth, Velocity measures the pace of information creation, gushing, and accumulation and Veracity alludes to the reliability of the information[6–8].

Medicinal services are a standout amongst the most essential zones for creating and created nations to facilitate the extremely valuable human asset. The dispersed preparing of colossal information sets crosswise over gatherings of frameworks is encouraged by utilizing figuring models of the Apache Hadoop Framework for a huge dataset. The Size of the "term" large depends upon the particular individual who is handling that data[5]. The structure is unsurprising to enlarge from desolate servers to a large number of frameworks, each showing calculation and capacity. This intense eventual fate of Hadoop system pulls in assortment of organizations and associations to utilize it for both examination and creation.

Hadoop is an open source programming system for

** Author for correspondence*

capacity and substantial scale handling of information sets on a MapReduce programming model for tremendous scale datasets on bunches of item equipment. It has a MapReduce programming model for enormous scale information preparing and records execution, adaptability and adaptation to internal failure.

The dataset used in this study is “The Pima Indians Diabetes Data Set” which was taken from the UCI Machine Learning Repository[3]. The original owner of this data set is the National Institute of Diabetes and Digestive and Kidney Diseases. Several constraints were placed on the selection of this dataset from larger database. In particular, all patients selected are females at least 21 years old of Pima Indian heritage.

2. Disease- Diabetes Mellitus

Diabetes Mellitus is the coercive effect of insulin on the glucose metabolism. It is 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. This causes sugar to build up in our blood leading to complications like heart disease, stroke, neuropathy, poor circulation leading to loss of limbs, blindness, kidney failure, nerve damage, and death.

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[16].

Overweight, obesity and diabetes mellitus are strongly associated with prehypertension. Indian Diabetic Risk Score (IDRS) detects that a person who is having normal blood pressure but with high Indian diabetic risk score is likely to become hypertensive or diabetic in near future[14]. Prevalence of both obesity and diabetes is rising in the recent years. So it could be treated withgreat care.

3. Methodology

3.1 Hadoop

Hadoop is the open-source disseminated information handling stage from Apache. It has the ability to associate PCs with various processor centers with a scale running from hundreds to thousands. Boundless volumes of information can be effectively dispersed crosswise over bunches of PCs utilizing Hadoop.

Hadoop has been worked with the capacity to oversee limitless information sets whose size can without much of a stretch lie between couple of gigabytes to a great many peta bytes[9–11]. Hadoop gives its answer as a Distributed File System which parts the information and stores it in a few distinct machines. This empowers parallel preparing of the issue and productive calculation is conceivable.

Hadoop utilizes two principle segments to carry out its employment: 1. Map/Reduce. 2. Hadoop Distributed File System.

3.1.1 Map/Reduce

MapReduce is an information parallelization programming model- the dataset is part into free subsets and conveyed, then the same guidelines are connected to every subset in simultaneously, to process and creating gigantic measures of datasets[11].

Hadoop’s usage of Map/Reduce depends on programming models to process vast information or datasets by partitioning them into little pieces of assignments. It is an instrument actualized for overseeing and handling limitless measures of information in parallel taking into account division of a major work thing in littler autonomous assignment units. Programs which are Map Reduces are modified to oversee boundless measures of information in parallel. It comprises of two capacities:

The Map( ) capacity which dwells on the expert hub and afterward isolates the information or undertaking into littler subtasks, which it then appropriates to specialist hubs that procedure the littler assignments and pass the answers back to the expert hub. The subtasks are keep running in parallel on various PCs.

The Reduce ( ) capacity gathers the aftereffects of all the subtasks and consolidates them to deliver an amassed last result- which it returns as the response to the first huge inquiry.

3.1.2 Hadoop Distributed File System (HDFS)

Apache Hadoop accompanies a disseminated record framework called HDFS, which remains for Hadoop Distributed File System. HDFS is scalable, reliable and manageable solution or working with huge amount of data. HDFS has been deployed in cluster of 10 to 4000 data nodes[12]. It is perfect for putting away a lot of information (terabytes and petabytes). The outline of HDFS is firmly connected to the Google File System or the GFS. Hadoop DFS stores every record as a succession of hinders; all pieces in a document aside from the last square are the same size in Figure 1.

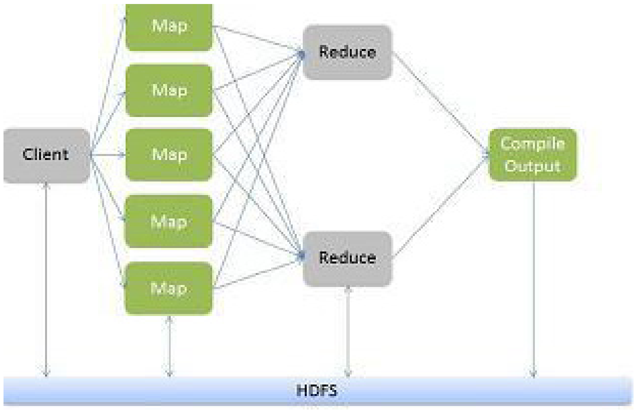

Figure 1. Map reduce.

3.2 K-Means

Clustering is a main task in data mining and a general technique for statistical data analysis used in many fields, including machine learning, pattern recognition, image analysis, information retrieval, and bio-informatics[14].

Clustering involves separating data points into groups according to how "similar" their attributesare[13]. A commonly used algorithm for clustering is K-means. Its calculation is the most surely understood and normally utilized grouping strategy[15]. K-means bunching is regularly utilized for various characterization applications. It takes the information parameter, k, and segments an arrangement of n items into k bunches so that the subsequent intra-group likeness is high though the entomb bunch closeness is low. Group closeness is measured by mean estimation of the articles in the bunch, which can be viewed as the clusters "center of gravity".

4. Results and Discussion

As we talked about before the social insurance industry has produced huge volume of information. So as to examine the enormous dataset, the Big Data instrument to be specific HadoopMapReduce has been brought with grouping systems since it works with the idea of parallelization as they run a few calculations simultaneously.

The underneath Table1 portray the dataset from Pima Indians Diabetes Database of National Institute of Diabetes and Digestive and Kidney Diseases which has 8 number of characteristics with 768 examples.

Table 1. Dataset description

Dataset	No. of Attributes	No. of Instances
Pima Indians Diabetes Database of National Institute of Diabetes and Digestive and Kidney Diseases	8	768

Table 2. Attribute description

	Attribute	Relabeled values
1	Number of times pregnant	Preg
2	Plasma glucose concentration	Plas
3	Diastolic blood pressure	Pres
4	Triceps skin fold thickness	Skin
5	2-Hour serum insulin	Insu
6	Body mass index	Mass
7	Diabetes pedigree function	Pedi
8	Age	Age
9	Class Variable	Class

The Table 2 dataset is utilized to discover for diabetic treatment which is important to test the examples like, plasma glucose focus, serum insulin, diastolic circulatory strain, diabetes family, Body Mass Index (BMI), age, number of times pregnant. The example disclosure of prescient examination must incorporate the bunching of comparative examples of utilization.

The K-means bunching sort of information mining with MapReduce has been connected to the Pima Indians Diabetes Database of National Institute of Diabetes and Digestive and Kidney Diseases where the class variable result has been bunched into two gatherings in particular Cluster1 (Diabetic) and Cluster 2 (Non-Diabetic). The below table shows the Clustered Output.

Table 3. Clustered distribution

No. of Instances	Cluster 1(Diabetic)	Cluster 2(Non-Diabetic)
768	268	500

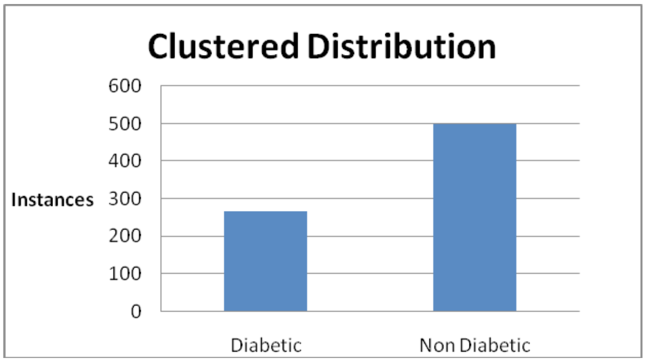

Figure 2. Basic clustered distribution.

According to our work, the output clustering of diabetic and non-diabetic using MapReduce function with K-Means time was noted to calculate it's elapsed and CPU time. Basically the real time (the time from start to finish of the call) is termed as elapsed time, the user and sys time predicts the total time taken for analysis by CPU to execute process. The CPU taken time is 15.842 sec for clustering the dataset.

Table 4. Time taken for clustering

Dataset	No. of Instances	Time taken for Clustering(S)
Pima Indians Diabetes Database of National Institute of Diabetes and Digestive and Kidney Diseases	768	15.842

Below are the algorithm steps to cluster the data. The grouping with MapReduce will function as takes after:

In the guide step

Step 1: Read the bunch focuses into memory from a sequence file.

Step 2: Iterate over every bunch community for every info key/esteem pair.

Step 3: Measure the separations and recovery the closest focus which has the most minimal separation to the vector.

Write the bunch focus with its vector to the document framework.

In the lessen step (we get related vectors for every middle).

Step 1: Iterate over every worth vector and figure the normal vector. (Whole every Vector and partition every part by the quantity of vectors we got).

Step 2: This is the new focus; spare it into a Sequence File.

Step 3: Check the merging between the bunch focus that is put away in the key item and the new focus.

Step 4: If it they are not equivalent, increase a redesign counter.

Run the subject of until nothing was redesigned any longer.

5. Conclusions

Constant examinations are directed to assess the viability of two parallel preparing stages as far as execution to acquire learning. An extensive investigation was made in the assessment and demonstrates that the information size was an exceptionally essential specialist in accomplishing execution when parallelizing utilizing Hadoop bunch. The assessment comes about clearly demonstrates that parallelism with Hadoop group is the most versatile and the most proficient approach to actualize investigation of huge information mining forms with expanding information set size. The final product demonstrates that Apache Hadoop group is a promising model for adaptable execution when the dataset size is huge.

6. Future Work

For future work, we can assemble model utilizing classifier to yield certainty for each of the tests and measure the classifier result per test. Likewise, we could incorporate more components to attempt to enhance the results.

7. References

1. Sadhana S, Shetty S. Analysis of diabetic data set using hive and R. International Journal of Emerging Technology and Advanced Engineering. 2014; 4(7):626–9.
2. 2. Iyer AS, Jeyalatha J, Sumbaly R. Diagnosis of diabetes using classification mining techniques. IJDKP. 2015; 5(1):1–14.
3. Koklu M, Unal Y. Analysis of a population of diabetic patients databases with classifiers. World Academy of Science, Engineering and Technology International Journal of Medical, Health, Pharmaceutical and Biomedical Engineering. 2013; 7(8):167–9.
4. Kumar P, Rathore VS. Efficient capabilities of processing of big data using hadoop map reduce. International Journal of Advanced Research in Computer and Communication Engineering. 2014; 3(6):7123–6.
5. Rajendran PK, Asbern A, Kumar KM, Rajesh M, Abhilash R. Implementation and analysis of mapreduce on biomedical big data. Indian Journal of Science and Technology. 2016 Aug; 9(31). DOI: 10.17485/ijst/2016/v9i31/83451.

6. Arun k, Jabasheela L. Big data: Review, classification and analysis survey. IJIRIS. 2014; 1(3):17–23.
7. Raghupathi W, Raghupathi V. Big data analytics in healthcare: Promise and potential. Health Information Science and System. 2014; 2(3):1–10.
8. Augustine DP. Leveraging big data analytics and hadoop in developing India's healthcare services. International Journl of Cpmputer Application. 2014; 89(16):44–50.
9. 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.
10. Kumar SNM, Eswari T, Sampath P, Lavanya S. Predictive methodology for diabetic data analysis in big data. Science Direct, Elsevier, Procedia Computer Science. 2015; 50:203–8.
11. Sharmila K, Vethamanickam SA. Application of mapreduce in diabetic dataset using hadoop platform. International Journal of Applied Engineering Research. 2015; 10(60):15–20.
12. Greeshma L, Pradeepini G. Big data analytics with apache hadoop mapreduce framework. Indian Journal of Science and Technology. 2016 Jul; 9(26). DOI: 10.17485/ijst/2016/v9i26/93418.
13. Can A. Benchmarking of data mining techniques as applied to power system analysis. Department of Information Technology, Uppsala University; 2013.
14. Nagarajan S, Chandrasekaran RM. Design and implementation of expert clinical system for diagnosing diabetes using data mining techniques. Indian Journal of Science and Technology. 2015 Apr; 8(8):771–6.
15. Zhao W, Ma H, He Q. Parallel K-means clustering based on mapreduce. Springer-Verlag Berlin Heidelberg. 2009; 5931:674–9.
16. Ramzan M, Ramzan F, Thakur S. A systematic review of type-2 diabetesbyhadoop/map-reduce. Indian Journal of Science and Technology. 2016 Aug; 9(32). DOI: 10.17485/ijst/2016/v9i32/100184.