Big Data Analysis using Apache Hadoop and Spark

K. Sharmila, S. Kamalakkannan, R. Devi, C. Shanthi

Abstract*: Big Data have increased immense interest in the past few years. Nowadays analyzing Big Data is very common constraint and such chuck really turns into a big challenge to analyze the mass amount of data to get impact and different patterns of information on aconvenient way.Processingthe big data information in a single machine or evento storethese Big Data has become another big challenge of the Big Data. The elucidation for the above constraints is to give out data over large clusters so that Big Data to be analyzed and for storinginformation should beovercome. The article will explore perceptions of Big Data Analysisusing emerging tools of Big Data such as ApacheHadoop and Spark and its performance.

***Index Terms*: Apache Hadoop, Big Data, Spark..**

I. INTRODUCTION

Enormous data for the most part incorporates datasets with sizes a far from the capacity of normally utilized programming devices to catch, precise, oversee, and process information inside a bearable slipped by time. Huge Data "estimate" is an always moving focus, starting at 2012 extending from a couple of dozen terabytes to numerous petabytes of information. Enormous Data is a lot of procedures and advancements that require new types of combination to uncover huge concealed qualities from huge datasets that are differing, complex, and of a large-scale.

Velocity, Volume, Veracity, Variability, and Variety are considered as Five V's of Bigdata.Velocity relies upon the speed how quick information is developing and preparing. Volume decides the extent of information is treated like huge information. Veracity decides the nature of caught and prepared information. Variability relies upon the irregularity in information, if information is increasingly conflicting different strategies are utilized to deal with this. Variety is the kind of information, for example, organized, unorganized and semi-organized.Apache Hadoop utilizing a MapReduce programming model is a software and open source frame used to save access and process datasets. Apache Hadoop has HDFS which is utilized for putting away information in dispersed condition and MapReduce which is a Hadoop programming model.Apache Spark is an open-source motor for huge scale information handling with fast, simple to utilize, and complex investigation. It keeps running on crown of existing Hadoop Distributed File System (HDFS) framework to give raised and additional usefulness. Apache Spark is utilized for performing a quick and constant examination at an extremely quick speed which is impossible by Hadoop.

II. LITERATURE REVIEW

Big Data and the five V's, i.e., Velocity, Volume, Veracity, Variability, Variety are portrayed by Vibhavari Chavan et al.,[4] .The working of Hadoop HDFS and MapReduce work are explained by the creators in this paper. To investigate the enormous measure of information Apache Hadoop was utilized as a working model. Information is put away in HDFS and to process a lot of information it has utilized the idea of key-esteem sets. Wei Huang et al.,[5] have talked about the function and design of Spark alongside Hadoop YARN.YARN is designed to work in a heterogeneous domain for Apache tempest and Tez. Abhishek hattacharya also, ShefaliBhatnagar [6]seek to justify the characteristics of Apache Spark and its standing as very efficient software pertaining to the current scenario of Big Data. PriyaDahiyaet.al., [7] demonstrated the engineering and working of Hadoop and Spark and furthermore draws out the contrasts among them and the difficulties looked by MapReduce amid the preparing of enormous datasets and how Spark takes a shot at Hadoop YARN. The difference between Hadoop and Spark is proposed by AnkushVerma et al.,[8] This article demonstrates the working of the two, and a relative report was done on them. Sparkle has beaten the restrictions which are available in the customary framework. Better execution is a significant factor for keeping up a huge measure of information.

III. RELATED TECHNOLOGIES

A. *Hadoop*

Hadoop, is an open source configuration used for huge datasets to process in a distributed environment. It uses a Master-Slave Architecture. In the area of Big Data, Hadoop has become a well known and excellent platform for processing.

Revised Manuscript Received on 30 July 2019.
***** Correspondence Author

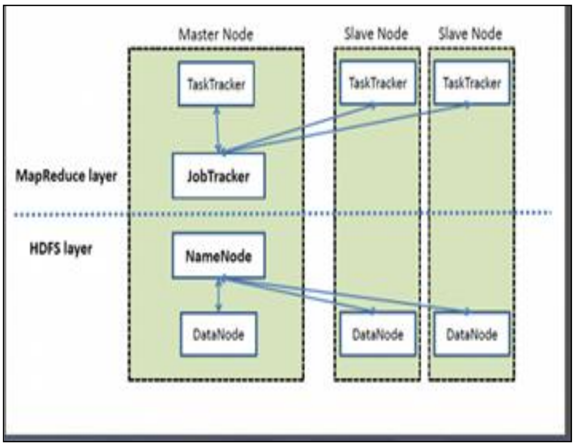
Dr K.Sharmila*, Associate Professor, Department of Computer Science, Vels Institute of Science, Technology &Advanced Studies, Chennai.

.Dr.S.Kamalakkannan, Associate Professor, Department of Computer Science, Vels Institute of Science, Technology &Advanced Studies, Chennai.

R.Devi, Associate Professor, Department of Computer Science, Vels Institute of Science, Technology &Advanced Studies, Chennai.

Dr.C.Shanthi, Associate Professor, Department of Computer Science, Vels Institute of Science, Technology &Advanced Studies, Chennai.

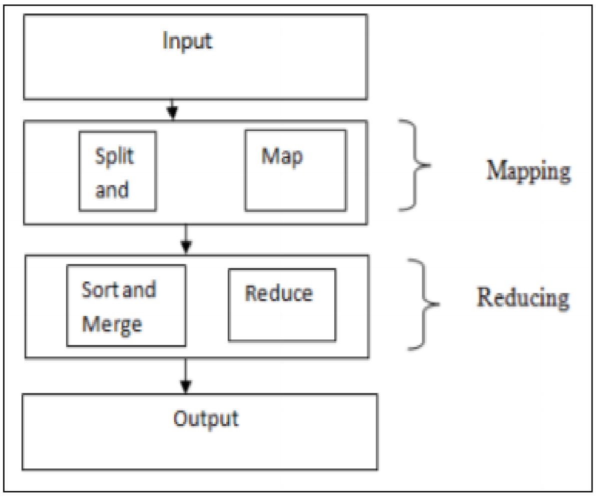
It has reliable storage and high performance for Big Data. The Hadoop system has Map Reduce and Hadoop Distributed File System (HDFS)as components. [9].



Structure of Hadoop

B. *MapReduce*

Hadoop framework uses MapReduce as a computational model and software structure to process the data. They are fit for handling monstrous information in parallel on enormous groups of calculation hubs. A MapReduce work typically parts the information dataset into autonomous lumps which are handled by the guide undertakings in a totally parallel way.The system sorts the yields of the maps, which are then contribution to the lessen assignments [10].



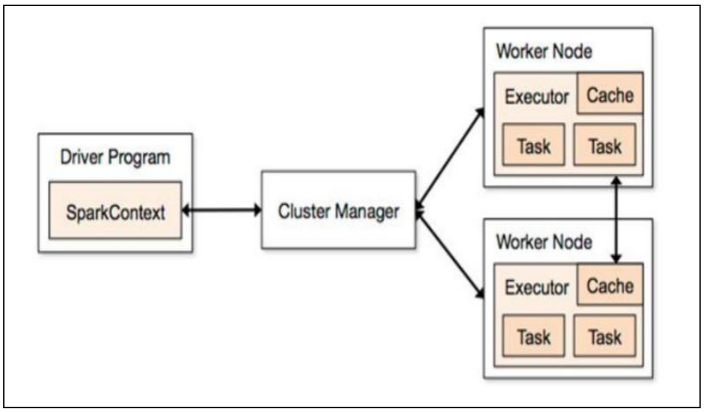
MapReduce Architecture

C. *Hadoop Distributed File System (HDFS)*

HDFS depends on Google File System (GFS) to save information in different hubs by part the colossal information into little parts. HDFS is a dispersed document framework that keeps running over the neighborhood record frameworks of the hubs and can store in enormous amounts of huge records appropriate for gushing information get to. HDFS has two hubs to be specific, DataNodes which go about as an ace and NameNode which go about as a specialist [4]. These hubs are utilized for performing capacities like perusing, compose, make and erase.

D. *Apache Spark*

Apache Spark, an open source processing structure centered in information examination. It is build on the top of Hadoop HDFS. Spark,[11] programming model is enthused by the parallel concept of Map Reduce. Keeping favorable properties of Map Reduce for example, parallelization, fault tolerance, information dispersion and burden adjusting, spark adds support for iterative classes of algorithms, interactive applications and algorithms containing common parallel primitive methods like join and match.Sparkkept running over Hadoop and utilized for gushing of information which is progressively. The Machine Learning, SQL inquiries, chart information handling and spilling information which are bolstered by Spark are used to examine the Big data. Spark is utilized as an option in contrast to conventional MapReduce who neglected to work appropriately for continuous information. Resilient Distributed Datasets (RDD) and Directed Acyclic Graph (DAG) are the two key ideas utilized in Apache Spark.



Apache Spark Architecture

IV. HADOOP VS SPARK

The framework Hadoop, a parallel processing which is traditionally used to run map/reduce jobs. For ongoing stream information preparing, the spark has intended to keep running over Hadoop and it is an option in contrast to the conventional group map/lessen model that can be utilized for quick intelligent inquiries. Both traditional MapReduce and Spark. Spark are supported by Hadoop. Spark information is put away in-memory though Hadoop information is put away on a disk. For Hadoop, adaptation to non-critical failure is accomplished by replication though Spark utilizes a piece of alternate information stockpiling model, resilient distributed datasets (RDD).

V. MACHINE LEARNING TECHNIQUES FOR DISEASE DIAGNOSIS

Machine learning, however is a quickest developingfield in computer science and health informatics, is having most noteworthy difficulties in it. Therapeutic is an irksome task and assumes a imperative job in sparing human lives so it wants to be executed exactly and proficiently A suitable and definite PC based mechanized choice emotionally supportive network is required to achieve medical tests in a diminished cost. Thusly Machine learning methods are utilized in diagnosing different sicknesses for which different information mining classifiers have developed as of late.

The classification accuracy depends on the exact features that have been utilized and the type of classifier selected [12].

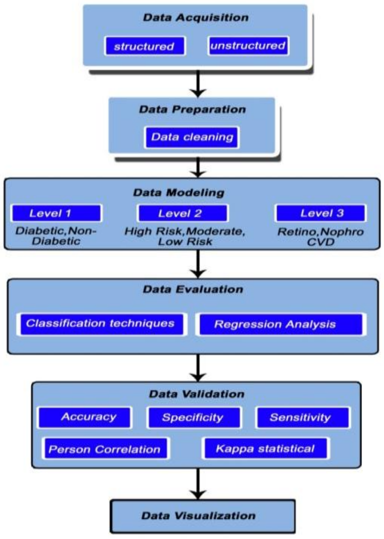
VI. EXPERIMENTAL RESULTS

A. *Dataset*

Diabetic data sets are used in this study which was collected from different districts with some questionnaire and medical report of a patient. Data collection was carried out using the following parameters such as Unhealthy diet, Physical inactivity, HBA1C, Age, Obesity, Residence (Urban/Rural), Genetic factors (Family history-pedigree), No of years the patients having diabetics, LDL,HDL values and Creatine Value as CSV files. Among these parameters they are divided as risk factor attributes to classify diabetic and non diabetic, and for prediction of diabetic related diseases.

B. *Methodology*

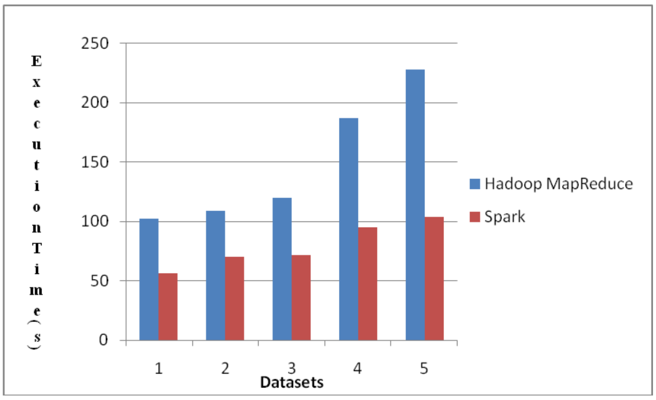
In this article a systematic evaluation of Hadoop Map Reduce is compared with another Big Data framework, Apache Spark based on the performance. For this purpose, we evaluated the diabetic dataset algorithm on various datasets of different sizes. Data The process of dataset acquisition comes under data collection where the data has been collected in structured and unstructured format. Next the collected data has been moved to data cleaning to preprocess the data that is to handle the missing data, to handle inconsistent data, and to handle noisy data. After preprocessing, the next segment is to model the data to analyze the Big Data by techniques of Machine Learninglike K-means (unsupervised) also,SVM(supervised) algorithms along with MapReduce concept which is known as MRK-SVM algorithm. Here the data is clustered into two groups namely diabetic and non-diabetic using K-means. From diabetic data three clusters are formed with Diabetic-High, Diabetic-Medium and Diabetic-Low. From Diabetic-High data it is classified using SVM to predict whether the patients have chances to get diabetic complications such as Retinopathy, Nephropathy and Cardio Vascular Diseases. The obtained results has validated using the metrics such as Accuracy, Sensitivity, Specificity, Kappa statistics. Finally the results have been visualized.



Methodology of the work

C. *Results*

1) As mentioned earlier, the rationale of this study was to assess the Map Reduce-HDFS performance with Spark. HDFS performances under the same setup using MRK-SVM algorithm. The tests were conducted for five various datasets having different sizes namely 650000,843000, 857229, 1650000, 2000000, and 6000000and the processing time was noted. From the study, it is clear that when the dataset size was 6 million or more the standalone mode used for this research was not able to process the job successfully, due to system resource constraints.



Performance comparison of HadoopMapReduce and Spark

The above figure represents the performance assessment ofHadoopMap Reduce and Spark in graphical manner. The figure evidently shows that Apache Spark is take less time to execute than Map Reduce.

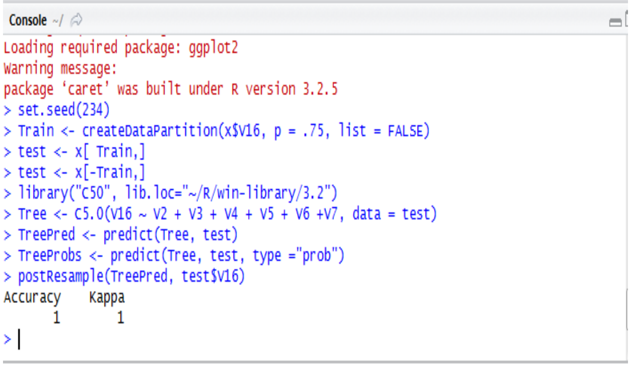
2)The research study demands an appropriate techniquefor data analysis hence the data set has been analyzed to predict the number of people having the chances for getting the diabetic related diseases such as CVD, Nephropathy and Retinopathy in the near future for Diabetic-High risk patients based on the attributes collected such as No of years having diabetic, LDL and HDL value, and Creatine. The study predicted the percentage of patients to have Nephropathy (62.9%), followed by Retinopathy (67.9%) and Cardio Vascular Disease (46%) according to the attribute in the datasets.

3) Evaluation of the study

The dataset taken to predict the diabetic related diseases was loaded into RStudio to show the statistical analysis of it because R is a very good statistical tool. The performance measures of the models such as accuracy (ACC) and Kappa statistics, Sensitivity and Specificityare assessed using C5.0 and it shows 100 % in all these measures for this dataset. So it is clear that if the data set has been properly preprocessed and then model the data the result will more accurate.

Retrieval Number: A2128058119/19©BEIESP
DOI: 10.35940/ijrte.A2128.078219
Journal Website: www.ijrte.org

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication



```
Loading required package: ggplot2
Warning message:
package ‘caret’ was built under R version 3.2.5
> set.seed(234)
> Train <- createDataPartition(x$V16, p = .75, list = FALSE)
> test <- x[ Train,]
> test <- x[-Train,]
> library("C50", lib.loc="~/R/win-library/3.2")
> Tree <- C5.0(V16 ~ V2 + V3 + V4 + V5 + V6 +V7, data = test)
> TreePred <- predict(Tree, test)
> TreeProbs <- predict(Tree, test, type ="prob")
> postResample(TreePred, test$V16)
Accuracy    Kappa
       1        1
> |
```

Performance measures of the Model.

VII. CONCLUSION

In this paper two programming model MapReduce and Apache Spark has been displayed for examining their execution with HadoopMapReduce and Apache Spark both can adapt to each kind of information organized, unstructured or semi-organized. Diabetes mellitus is a medical syndrome which was noted that the diabetic'spatients in the world by 2025 may attain up to 60 million, and India's role to it would be 30 million which was estimated by World Health Organizations.So it is necessary to predict earlier and do the medical help prior. Hence this technique will predict more accurately and to predict fast MapReduce and Spark has been compared. The two systems on different datasets of various sizes, execution of MapReduce and Apache Spark has been thought about. Apache Spark gives outlying enhanced performance in terms of execution time as compared to MapReducedue to the fact that it maintains all its current operations inside. In future, the particular work can be complete to predict the diabetic related diseases asconstant information.

REFERENCES

1. Bobade, V. B. (2016). Survey paper on big data and Hadoop. Int. Res. J. Eng. Technol.(IRJET), 3(01).
2. Samuel, S. J., RVP, K., Sashidhar, K., &Bharathi, C. R. (2015). A survey on big data and its research challenges. ARPN J EngApplSci, 10(8), 3343-7.
3. Amp Lab web page : https:// amplab.cs.berkeley.edu/projects/spark-lightning-fastcluster-computing.
4. Chavan, V., &Phursule, R. N. (2014). Survey paper on big data. Int. J. Comput. Sci. Inf. Technol, 5(6), 7932-7939.
5. Huang, W., Meng, L., Zhang, D., & Zhang, W. (2016). In-memory parallel processing of massive remotely sensed data using an apache spark on Hadoop YARN model. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 10(1), 3-19.
6. Bhattacharya, A., &Bhatnagar, S. (2016). Big Data and Apache Spark: A Review. no, 5, 206-210.
7. PriyaDahiya ,Chaitra.B , UshaKumari ,(2017). Survey on Big Data using Apache Hadoop and Spark, International Journal of Computer Engineering In Research Trends, 4(6):pp:195-201.
8. Verma, A., Mansuri, A. H., & Jain, N. (2016, March). Big data management processing with HadoopMapReduce and spark technology: A comparison. In 2016 Symposium on Colossal Data Analysis and Networking (CDAN) (pp. 1-4). IEEE.
9. Pradeepa, A., &Thanamani, A. S. (2013). Hadoop file system and fundamental concept of MapReduce interior and closure rough set approximations. International Journal of Advanced Research in Computer and Communication Engineering, 2(10), 5865-5868
10. Pol, U. R. (2016). Big Data analysis: comparison of hadoopMapReduce and apache spark.International Journal of Engineering Science,6389.
11. Apache Spark, http://spark.apache.org/
12. Gagankumar and RohitKalra(2016). A survey on Machine Learning Techniques in Health Care Industry. International Journal of Recent Research Aspects ISSN: 2349-7688, Vol. 3, Issue 2, June 2016, pp. 128-132.