# Analytics For Healthcare Using Hadoop Mapreduce, Apache Spark And In Cloud Services

**Dr.K.Sharmila, Dr.T,Kamalakannan**

**Abstract**: Decision making and knowledge discovery from voluminous big data is a challenging problem. Extracting useful information from the enormous amount of data is highly complex, difficult and time consuming. Therefore standard data mining algorithms are essential for the analysis of big data with different platform. This investigation focuses on benchmarking of parallel processing platforms and Cloud computing environment. Cloud computing facility has emerged as service oriented computing model to deliver infrastructure, platform and applications as services from the providers to the consumers. This study utilized the services provided by Amazon Web Services as an effective metaphor for the management of large scale data processing in elastically scalable computing and for storage. This paper also discusses about the framework of MapReduce integrated with K-means and SVM machine learning techniqes algorithm on standalone environment and spark to predict the diabetic related diseases from real-time data set collected in various districts of Tamil Nadu. Ultimately, the present study has established that parallelization using Apache Hadoop with spark shows a better performance compared with a standalone model in a single machine. With the expansion of Information and communication technology, the health care industry also is producing extensively large data day by day. In developing countries like India, the accumulation of data is large and there exist various problems. This type of Big Data analysis will hopefully help the diabetes patients and physicians to predict the disease and to treat them at an early.

**Index Terms**: AWS, Big data, Cloud computing, Diabetic Mellitus , Hadoop MapReduce, K-means, SVM algorithm, spark.

## 1. INTRODUCTION

Knowledge discovery and decision making from rapidly growing voluminous big data is a challenging problem. Many challenges are to be faced due to the growing health records and complications associated with health industry. Thus it is essential to make the size and quality of the data into vital nominal value with possible solutions. Diabetic Mellitus (DM) is a major health hazard in developing countries like India. The acute nature of DM is associated with long term complications and numerous health disorders. In this paper, machine learning algorithm on Hadoop Map Reduce platform in standalone and spark was used to analyse the big data and also to predict the diabetes related diseases like Nephropathy, Retinopathy and Cardio Vascular Diseases. Based on the analysis, this investigation provides an efficient predictive methodology, MRK-SVM hybrid algorithm, to identify diabetes types and predict complications associated with it at an early stage. For big data analysis, a real time dataset was prepared by collecting records from five districts in Tamil Nadu using a replica method and big data analysis was carried out using Hadoop MapReduce, Spark and in Cloud environment. The results were statistically analysed by Rstudio identification number.

## 2 LITERATURE REVIEW

The review work makes the researcher to starttheir work with essentials isses and challenges. Inmon [1] described Traditional Data as an integrated and non-volatile collection of data which helps analysts in decision-making process. Traditional databases lack in providing the solutions for unstructured, large volumes of rapidly changing data.


- *DrK.Sharmila is currently working as , Associate Professor, Department of Computer Science, Vels Institute of Science, Technology &Advanced Studies, Chennai.*
- *Dr. T.Kamalakannan is currently working as Associate Professor, Department of Computer Science, Vels Institute of Science, Technology &Advanced Studies, Chennai.*


Wei Fan and Albert Bifet [2] had focused on Big Data mining from which useful information could obtain. In the past, data mining operations on large amount of data was not possible. But currently, with the help of softwares like Apache Hadoop, it is possible. The authors concluded that in addition to Apache Hadoop, there are Big Data tools like R, MOA, Strom, Vow pal Wabbit which are a few open source softwares to deal with Big Data. Stephen Kaisler et al.[3] had discussed about fundamental issues like storage, management, and processing. Sadhana and Savitha Shetty[30] attempted to analyze the facts regarding diabetic dataset and proposed a prediction model. Kiran Kumara Reddi & Indira [4] have described the current huge data is a group of structured, semi-structured, unstructured homogenous and heterogeneous data. Therefore, they suggested transfer of huge amount of data over the network and recommended new algorithms to transfer Big Data. Raghupathi and Raghupathi [5] have discussed about the importance of platform and tools in the application of healthcare industry which can accelerate their processing time. Huang et al.(2015) are of the opinion that “available techniques at present cannot do Big Data analysis and processing the voluminous healthcare data. On the other hand, the currently available state of the art technique, Cloud platform provides a scalable, parallel and distributed processing framework and MapReduce for fast healthcare data processing” [6]. Ashwin Belle et al.(2015) [7] have reported that “the fast growing field of Big Data analytics has started to play an important role in the evolution of healthcare practices and research. It has provided tools to accumulate, manage, analyze, and assimilate large volumes of structured and unstructured data produced by current healthcare systems. Aditya et al.[8] have reported about the Big Data problem and its optimal solution using HDFS for storage and parallel processing large data sets using Map Reduce framework. They have done prototype implementation of Hadoop cluster, HDFS storage and Map Reduce framework for processing large data sets. Spark [9] is an open source project developed by UC Berkeley AMPLab. With the realization of RDDs [10], a distributed memory abstraction that lets programmers perform in-memory computations on large clusters, Spark provides RDD transformations and

actions for the users to use Spark easily. Apache Spark is the fast general purpose big data analytics engine and it is very suitable for any kindof big data analysis. Only following two scenarios, can hinder the suitability of Apache spark are Low Tolerance to Latency requirements and shortage of Memory resources [11]. Sanjay et al.[12] are of the opinion that the current trend in the medical field using Cloud computing technique will help patients and physicians. The use of Cloud is progressing slowly due to the challenges like security in Cloud-computing model. Sahoo et al.(2016) [13] have designed a data collection mechanism and correlation analysis for the data collected. Also, they have also designed a prediction model to foresee the future health condition of the patients based on their current health status. Prajesh P Anchalia et al.[14] have described the K-means implementation on Hadoop platform. The key to the execution of the algorithm and the steps involved in the implementation have been described. Burbidgeet al.[15] have stressed the importance of support vector machine classification algorithm in structure–activity relationship analysis. They compared support vector machine algorithm with various machine learning techniques in this field and found that support vector machine was significantly better than others.

## 3 METHODOLOGY

The methodology used in this study are shown in figure below, namely the first step is data acquisition followed by data preparation and then data modelling which is followed by data evaluation and the final step is data visualization.

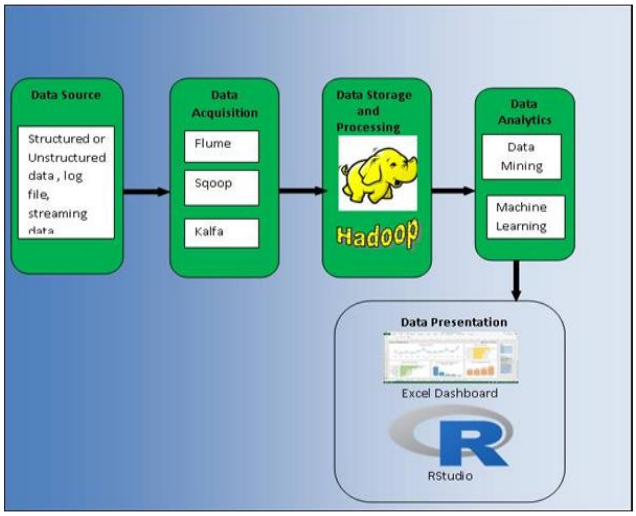


***Figure 1:*** *Methodology for prediction of dieseases using Hadoop and spark*

The process of dataset acquisition comes under data collection that is the source from where the data set has been collected. The raw structured or unstructured voluminous input data are obtained from Clinical systems and external sources such as government, laboratories, pharmacies, insurance companies, etc., in various formats. The data are obtained from people residing at various geographic locations namely Chennai, Kanchipuram, Thanjavur, Salem, and Thiruvarur.

The next stage is data preparation is a segment which is regarded as one of the key factors for excellent model quality. Real databases usually contain noisy data, missing data, and inconsistent data. The importance of data pre-processing is that real data could be dirty and could drive to the extraction of useless rules and can generate a smaller data so that it helps to improve the effectiveness. There are several steps in data pre-processing namely Data cleaning, Data integration, Data transformation and Data reduction. Data modeling is the next segment to transform the data for analysis when the data has been pre-processed, and the problems have been fixed. When the dataset used for this study is fully pre-processed, the machine learning techniques K-means (unsupervised) and SVM(supervised) algorithms were used along with MapReduce concept to predict the diabetic related diseases.

**The data modeling was done using the following methods**

1. Hadoop MapReduce in standalone.
2. Apache Spark in standalone.

The techniques used in data modeling are

1. K-means clustering.
2. MRK-SVM.

The data analytics stage which is after processing, to model the data to analyse the Big Data by Machine Learning techniques such as K-means (unsupervised) and 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 has 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 has been visualized.

## 4 RESULTS

### 4.1. Hadoop in standalone

The study used here is a new hybrid algorithm, which uses the concept of MapReduce along with K-means and SVM. The input data moved from local machine to Hadoop HDFS, where the mapper function starts to execute the input data from HDFS. In the mean time, the hybrid algorithm first divides the input data into clusters by applying the K-means clustering. The clusters produced using K-means are C1, …, C10. From the clustered output, the cluster class label 9 and 10, is given as input to SVM to build the model for the prediction.

***Table1:*** *Execution time for processing six types of datasets through MRK-SVM in standalone mode*

| Processing time of data in standalone mode Hadoop |
|---|

| Dataset | Districts | No of Records | Process time |
|---|---|---|---|
| 1 | Salem | 650000 | 102 |
| 2 | Thiruvarur | 843000 | 109 |
| 3 | Thanjavur | 857229 | 120 |
| 4 | Kanchi | 1650000 | 187 |
| 5 | Chennai | 2000000 | 228 |
| 6 | All 5 districts | 6000000 | Not able to process |

In the above table, the six datasets of different sizes 650000,843000, 857229, 1650000, 2000000 and 6000000 were processed and the processing time was noted. From the study, it is clear that when the dataset size was 6 million or more the standalone mode was not able to process the job successfully, due to system resource constraints.

### 4.2 Hadoop with spark.

In this study, the dataset is loaded into HDFS, which would then be available in Spark Driver which manages the job and schedules the task. The executors are responsible for executing work, in the form of tasks, as well as for storing output data. Using the MRK-SVM hybrid algorithm, the six datasets of different sizes 650000,843000, 857000, 1650000, 2000000 and 6000000 were processed and the processing time was noted.As discussed in Hadoop implementation, when the dataset size was 6 million or more it was not able to process the job successfully, due to system resource constraints.

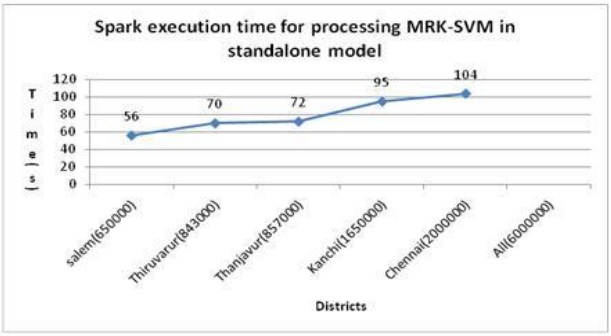


***Figure 2:*** *Spark execution time for processing MRK-SVM in standalone model*

### 4.3 Cloud computing

In Cloud computing mode, a Hadoop cluster interacts with Web Services like S3 and EC2 in a distributed environment. The Virtual computing machines for large dataset processing with single or multiple instances (machines) is provided by EC2 service available in a Cloud environment. The enormous amount of Big Data is stored in S3 storage service. A Cloud instance is an essential server working in a Cloud network. Single hardware is executed with the software and operates on top of multiple computers in a Cloud instance. when the dataset size is around 6 million in Hadoop standalone environment and Hadoop with spark standalone is not possible. Therefore, it is essential to move for elastic Hadoop parallelization in the distributed Cloud computing environment for both storage and processing which gives a better results than standalone since MapReduce in Cloud environment had decreased the response time and cost of processing large datasets which was the conclusion by Daneshyar Samira and Patel Ahmed [170]. the response time in Cloud computing mode is reduced with the increasing number of instances, because, the dataset is distributed and processed in Cloud computing in a distributed environment. In the present study, the response time was in a decreasing manner based on the increasing number of instances used. Similarly, the running speed is also improved with the number of instances increased as concluded by Samira and Patel. In this study, the running speed of time was in decreasing order 226, 195, 163, and 131(sec) for the instances 1,2,3,4 respectively.

***Table 2:*** *EMR processing time with multiple instances for the dataset(6000000)*

| EMR processing time with multiple instances for the dataset(6000000) | |
|---|---|
| EC2 Instances | Process time(s) |
| 1 | 216 |
| 2 | 184 |
| 3 | 150 |
| 4 | 114 |

| **EMR processing time with multiple instances for the dataset(6000000)** | |
|---|---|
| **EC2 Instances** | **Process time(s)** |
| 1 | 216 |
| 2 | 184 |
| 3 | 150 |
| 4 | 114 |

| **EMR processing time with multiple instances for the dataset(6000000)** | |
|---|---|
| **EC2 Instances** | **Process time(s)** |
| 1 | 216 |
| 2 | 184 |
| 3 | 150 |
| 4 | 114 |

The problem faced by the standalone has been overcome by moving to the cloud web service AWS which offers service on demand. From the investigation, it is clear that the execution time is decreased with the increasing number of instances when the data size is 60,00,000 which was the problem in standalone.

### 4.4 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 Specificity are 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[16].

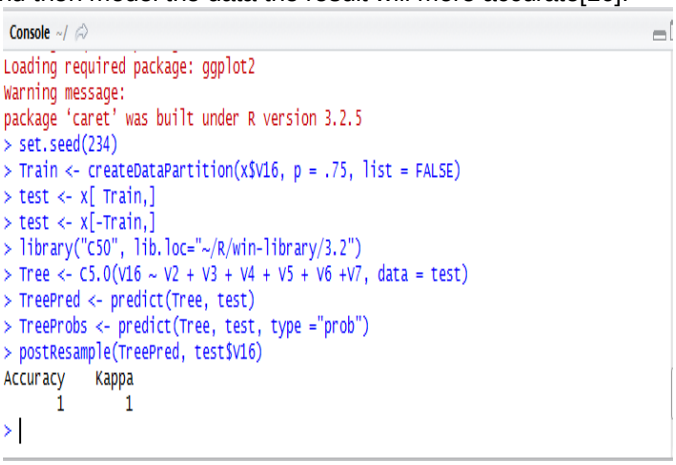


***Figure 3**:Performance measures of the Model*

## 5.DISCUSSIONS

Big Data handles massive amount of data collected over time, which forms a difficult task to analyze and handle using common database management tools [17]. Even though Big Data can yield extremely useful information, it urges new challenges both in data organization and processing [18]. To face such challenges Apache Hadoop, an open source, reliable, scalable and distributed computing platform is gaining popularity for its maximum performance. Samira Daneshyar and Ahmed Patel, [19] are of the view that the MapReduce framework is an effective technique to process and analyse large amount of data. It rapidly processes vast amount of data in parallel and distributed mode operating on a large cluster of machines. Recently, few hybrid algorithms have been proposed for data mining. Among them, K-SVM hybrid algorithm working to select the most informative samples using K-means clustering algorithm, and the SVM classifier is built through training on those selected samples. Experimental results showed that the new hybrid algorithm, K-SVM reached the goal of reducing the scale of training set, and greatly reduced the training and predicting time and meanwhile assures the generalization ability of the K-SVM algorithm[20]. However, they have concentrated only on clustering and classification and have not focused on the Big Data analysis. In this study, in order to reduce the processing time, MapReduce, k-means and SVM were integrated to develop a new hybrid algorithm, MRK-SVM [21] by which the present investigation was carried out. From the present investigation, it is concluded that Spark is faster than MapReduce to a certain extent. Considering the characteristics of MapReduce and Spark, Spark uses RAM memory and can use disk for processing, whereas MapReduce is strictly disk-based. While Hadoop and Spark might seem like competitors, they do not perform the same tasks and in some situations can even work together. While it is reported that Spark can function more than 100 times faster than MapReduce in some case scenario, it does not have its own storage system which is an important criteria in distributed storage but in Hadoop it is possible. The technical challenge in handling the Big Data, which is in the increasing order of vast quantities of data, is alarming. With the increase in size and accumulation of data every day, handling, managing and analyzing the vast amount of data becomes a difficult task. In this part of the investigation, the job was not successful using MRK-SVM, when the dataset size is around 6 million in Hadoop standalone environment and Hadoop with spark standalone is not possible. Therefore, it is essential to move for elastic Hadoop parallelization in the distributed Cloud computing environment for both storage and processing.

## 6.CONCLUSION

In this paper Hadoop MapReduce and Hadoop with Spark framework has been used which shows the later one is better than the former. From the present investigation, it is concluded that Spark is faster than MapReduce to a certain extent. Considering the characteristics of MapReduce and Spark, Spark uses RAM memory and can use disk for processing, whereas MapReduce is strictly disk-based. While Hadoop and Spark might seem like competitors, they do not perform the same tasks and in some situations can even work together. While it is reported that Spark can function more than 100 times faster than MapReduce in some case scenario, it does not have its own storage system which is an important criteria in distributed storage but in Hadoop it is possible. And in case of large dataset the storage problem faced by the standalone has been cleared by cloud computing services AWS. In future, large data set can be stored and processed without any constraints in memory by using cloud computing services because now-a-days data size has been increased voluminously and also real time data processing has to be carried out in healthcare .

## REFERENCES

[1] Inmon W. H., Building the Data Warehouse, 3rd edition, John Wiley & Sons, 2002.

[2] Wei Fan and Albert Bifet. "Mining Big Data: Current Status, and Forecast to the Future", SIGKDD Explorations. 14(2).

[3] Stephen Kaisler, Frank Armour, J. Alberto Espinosa, William Money. "Big Data: Issues and Challenges Moving Forward", 46th Hawaii International Conference on (pp. 995-1004).IEEE, 2013.

[4] Kiran kumara Reddi & Dnvsl Indira “Different Technique to Transfer Big Data :survey” IEEE Transactions. 52(8):2013.

[5] Wullianallur Raghupathi, Viju Raghupathi. Big data analytics in healthcare: promise and potential. Health Information Science and Systems, 2(3): 2-10. 2014.

[6] Huang T, L. Lan, X. Fang, P. An, J. Min, and F. Wang, "Promises and challenges of big data computing in health sciences ", Big Data Res., vol. 2, no. 1, pp. 2-11, 2015.

[7] Ashwin Belle, Raghuram Thiagarajan, S. M. Reza Soroushmehr Fatemeh Navidi, Daniel A. Beard, and Kayvan Najarian, Big Data Analytics in Healthcare BioMed Research International Volume 2015, Article ID 370194.

[8] Aditya B. Patel, Manashvi , Birla, Ushma Nair. Addressing big data problem using Hadoop and Map Reduce, http://ieeexplore.ieee.org/document/6493198/

[9] Matei Zaharia, Mosharaf Chowdhury, Michael J. Franklin, Scott Shenker, Ion Stoica. Spark: Cluster Computing with Working Sets.HotCloud 2010. June 2010.

[10] Matei Zaharia, Mosharaf Chowdhury, Tathagata Das, Ankur Dave, Justin Ma, Murphy McCauley, Michael J. Franklin, Scott Shenker, Ion Stoica. Resilient Distributed Datasets: A Fault-Tolerant Abstraction for In-Memory

Cluster Computing. NSDI 2012. April 2012.
[11] Spark MLib, Apache Spark performance,https://spark.apache.org/mllib/.
[12] Sanjay P. Ahuja1, Sindhu Mani1 & Jesus Zambrano1, A Survey of the State of Cloud Computing in Healthcare, Network and Communication Technologies. Canadian Centre of Science and Education.1(2): 12-19, 2012.
[13] Mukaka. M, "A guide to appropriate use of correlation coefficient in medical research," Malawi Med. J., vol. 24, no. 3, pp. 69-71, 2012.
[14] Prajesh P Anchalia, Anjan K Koundinya, Shrinath N K. "MapReduce Design of K-means Clustering Algorithm", IEEE, 2013.
[15] Burbidge, R.Trotter, M. Buxton B. and Holden, S. “Drug design by machine learning: support vector machines for pharmaceutical data analysis”, Computers and Chemistry, 26,5-14. 2001.
[16] K. Sharmila, S. Kamalakkannan, R. Devi, C. Shanthi, " Big Data Analysis using Apache Hadoop and Spark", International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-8 Issue-2, July 2019.
[17] Este, A.Gringoli F.and Salgarelli, L.“Support Vector Machines for TCP traffic classification”, Computer Networks, 53, 2476-2490. 2009.
[18] L Qu J.and M. J. Zuo, “Support vector machine based data processing algorithm for wear degree classification of slurry pump systems”, Measurement, 43, 781-791. 2010.p:
[19] Abdullah A. Aljumah, Mohammed Gulam Ahamad, Mohammad Khubeb Siddiqui, “Application of data mining: Diabetes health care in young and old patients” in Journal of King Saud University – Computer and Information Sciences, 25:127-136.2013.
[20] M.G. Jaatun, G. Zhao, and C. Rong (Eds.) Parallel K-means Clustering Based on MapReduce—.:Cloud COM. LNCS, 674–679, 2009.
[21] Stephen Kaisler, Frank Armour, J. Alberto Espinosa, William Money. "Big Data: Issues and Challenges Moving Forward", 46th Hawaii International Conference on (pp. 995-1004).IEEE, 2013.