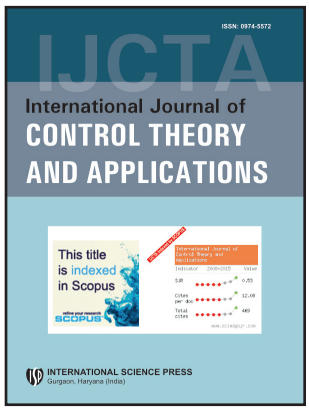



**International Journal of Control Theory and Applications**

ISSN : 0974-5572



**Volume 10 • Number 25 • 2017**

# Performance Analysis of Diabetic Dataset Using MapReduce Framework in Cloud and Standalone Computing

**K. Sharmila[a] and S.A. Vethamanickam[b]**

[a]*Asst Prof. & Research Scholar, Dept. of Computer Science, Vels University, Chennai, India*
[b]*Research Advisor, Chennai, India*

***Abstract:*** As we expand to a more digital society, the amount of data being created and collected is growing and accelerating drastically. The computer industry is being challenged to extend methods and techniques for these data to process the large datasets at optimal response times. The technical challenges in dealing with the increasing order to handle vast quantities of data is alarming and on the rise. Hadoop, an open source implementation of MapReduce model, is an effective tool for handling, processing and analyzing data generated on these days by cloud applications. Clouding Computing is a virtual pool of resources provided to users as service through a web interface. This paper specifies the performance of MapReduce to process large amounts of data in both standalone mode and cloud computing mode using a new hybrid algorithm MRK-SVM which is a combination of K-means, SVM and MapReduce for a diabetic dataset.

***Keywords:*** MapReduce, Hadoop, Cloud Computing, K-means, SVM.

## 1. INTRODUCTION

With the increase in size of the data every day, there is a need to handle, manage and analyze for the Business Applications and future prediction. To handle such large volume of semi-structured and unstructured data, Google's Map Reduce technique has proven to be an efficient tool[7].

Hadoop was inspired in 2006 by Doug Cutting (named by his son's stuffed elephant)[4] now being used by major companies, including Amazon, IMB, Yahoo, Facebook and a growing number of other companies. For the question how to process large amounts of distributed data quickly with good response times and replication at minimum cost, Hadoop is the best method for the huge data processing to perform parallel and distributed computing in a cloud computing environment with MapReduce framework. It was originally conceived on the basis of Google's Map Reduce, in which an application is broken down into numerous small parts[8].

MapReduce is a popular computing framework and its open-source implementation of Hadoop is widely used for big data processing in cloud environment. MapReduce, first proposed by Google in 2004, is an efficient programmable framework for handling large data in a parallel, distributed manner in a cluster of many systems.

The main reason of using MapReduce in cloud computing and standalone mode is a key point of MapReduce that hides how parallel programming work away from the developer.

Cloud computing provides massive data processing and data-storage services by using MapReduce as a computational platform in Amazon cluster. In a cloud computing environment offered by Amazon, developers enable to write their MapReduce systems and run them in a fully parallel and distributed platform. They will be charged only for the consumed time used on each working machine in the cluster. Therefore they are enabled to perform parallel data processing in a distributed environment with affordable cost and reasonable time.

Amazon web services platform provides a facility to implement a MapReduce system by offering its services to store, process and analyze large-scale datasets. Amazon services include, Amazon Elastic Compute Cloud (EC2) and Amazon Simple Storage Service (S3). Amazon EC2 is a web service platform that provides resizable compute capacity in a cloud[1]. Amazon S3 provides a simple web services interface that can be used to store and retrieve any amount of data, at any time, from anywhere on the web[2].

Diabetic Mellitus (DM) is one of the Non Communicable Diseases (NCD), is a major health hazard in developing countries such as India. It was estimated that 61.3 million people aged 20-79 years live with diabetes at 2011 in India. This number was expected to raise to 101.2 million by 2030. As diabetes is a lifestyle disorder, treatment and prevention can be done through early detection in order[21].

This study presents performance analysis of diabetic dataset from various districts collected which is replicated further for many times to make it as big data and processed by taking advantage of MapReduce framework in cloud computing mode (AWS) as a parallel and distributed computing environment.

## 2. RELATED WORK

Ruchi Mittal Ruhi Bagga had evaluated and analyzed the performance of WordCount MapReduce application using Hadoop on Amazon EC2 using different Ubuntu instances. The performance has been evaluated both on single node and multi-node clusters. Multi-node clusters include both the homogeneous and heterogeneous clusters[20].

Samira Daneshyar and Ahmed Patel comprehensively discussed the evaluations of MapReduce system in both standalone and cloud computing mode. The implementation of data processing was very fast with the highest end supercomputers available on Amazon Web Services with an extremely low budget in a cluster. They concluded that analysis using cloud computing and MapReduce together improves the speed of processing and decreases the response time and cost for processing of large datasets[5].

Samira Daneshyar and Majid Razmjoo conducted a comprehensive review and analysis of a MapReduce framework as a new programming model for massive data analytics and its open source implementation, Hadoop. They proposed a framework to process a large dataset in a cloud environment in parallel and distributed fashion, as well proposed Amazon Web Services as one instance of using MapReduce framework. At the end they presented an experimentation of running a MapReduce system in a cloud environment to validate the proposed framework's working schema and the main requirements of the MapReduce framework[6].

## 3. MAPREDUCE MODEL

Predictive modeling which is used to predict diabetic related diseases has been executed through Hadoop on AWS. Hadoop is a successful implementation of the MapReduce model[12]. Hadoop is used by Yahoo servers, where hundreds of terabytes of data are generated on at least 10,000 cores[10] which is an open-source, Java

based implementation of MapReduce. Among the two main components of Hadoop, Map Reduce is used for processing, running parallel computations on data and retrieval of data[6].

MapReduce libraries have been written in several programming languages include, LISP, Java, C++, Python, Ruby and C[13]. The advantage of using MapReduce for data processing is it allows developers to create their own MapReduce system with No. need to have specific knowledge of distributed programming[18].

A MapReduce programming model drives from three fundamental phases:

1. *Map phase*: partition into M Map function (Mapper); each Mapper runs in parallel. The outputs of Map phase are intermediate key and value pairs.
2. *Shuffle and Sort phase*: the output of each Mapper is partitioned by hashing the output key. In this phase, the number of partitions is equal to the number of reducers; all key and value pairs in shuffle phase share the same key that belongs to the same partition. After partitioning the Map output, each partition is stored by a key to merge all values for that key.
3. *Reduce phase*: partition into R Reduce function (Reducer); each Reducer also runs in parallel and processes different intermediate keys[5].

## 4. AWS

A Major web company, Amazon web services platform offers a service called Amazon Elastic MapReduce to store and process massive datasets by running MapReduce system on Amazon cloud. It utilizes a hosted Hadoop framework running on the web-scale infrastructure of Amazon Elastic Compute Cloud (Amazon EC2) and Amazon Simple Storage Service (Amazon S3). EC2 is a web service platform that provides resizable compute capacity in a cloud[14]. Amazon Web Services are a collection of remote infrastructure services that make a cloud computing environment with MapReduce. It includes a number of Amazon Web Services which is listed below[6]:

1. Amazon S3: the Simple Storage Service (S3) stores the input and output dataset.
2. AWS management console: it manages and accesses a collection of remote services offered over the internet by amazon.com.
3. Amazon Elastic MapReduce: EMR introduces the Hadoop framework as a service.
4. Amazon EC2 Cluster: Elastic Compute Cloud (EC2) is used for running a large distributed processing in parallel.

First uploads MapReduce program and input dataset on Amazon S3 service. Input dataset is transmitted into HDFS to be used by EC2 instances. A job flow is started on Amazon EC2 cluster .In this service, according to number of instances one machine works as master node and the others work as slave nodes to distribute the MapReduce process. All machines are terminated once the MapReduce tasks are completed. The final result is stored in an output file and the output can be retrieved from Amazon S3[1][2][14].

## 5. PROPOSED WORK MODEL

Some research has been directed at implementing and evaluating performance of the MapReduce model[11][15][16][17]. In our experiment, a MapReduce program was written in python to process a large dataset. The program determines the occurrences of diabetic related diseases in the given dataset. In the experiment, the dataset used was collected from various districts of Tamilnadu with 13 parameters. The present investigation uses simplified diabetes risk score for identifying diabetic patients with related diseases risks using the following parameters as risk factor namely.

- Family history of diabetes
- Unhealthy diet
- Physical inactivity
- HBA1C
- Age
- Obesity
- LDL, HDL, No. of years with diabetes, Creatinine (For second Level)

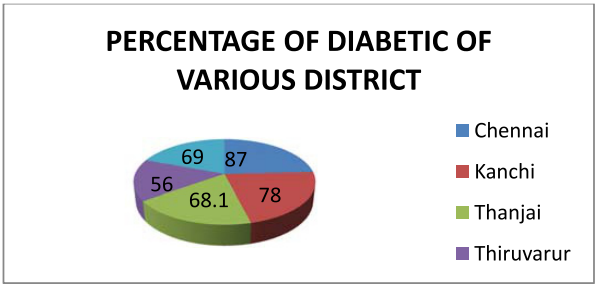
The model used is an hybrid algorithm MRK-SVM designed for the big dataset, by combining clustering (K-means) and classification (SVM) techniques applied with MapReduce in Hadoop which is used to predict high risk diabetic patients, having the possibility of getting diseases like Nephropathy, Retinopathy and Cardio Vascular Diseases[9].

Using the risk factors taken into consideration, the dataset has been clustered as diabetic and non diabetic at the first level, from which the diabetic risk patient is further classified as for having the possibility of getting diseases like Nephropathy, Retinopathy and Cardio Vascular Diseases.

Since the dataset has been collected from five different districts, the first level of clustering the dataset into Diabetic and non-diabetic has been listed using the model MRK-SVM.

**Table 1**
**District wise Percentage of Diabetic affected people**

| *District wise Percentage* | |
|---|---|
| *District* | *Percentage* |
| Chennai | 87 |
| Kanchi | 78 |
| Thanjai | 68.1 |
| Thiruvarur | 56 |
| Salem | 69 |

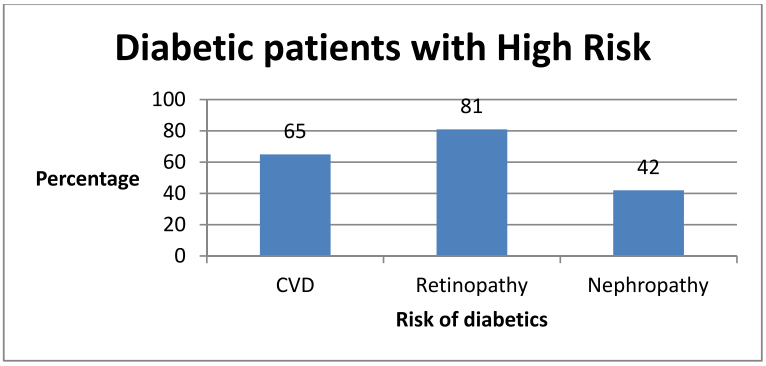



From the obtained results, diabetic patients from each districts has been tested with the second level parameters to find the diabetic patients having the risk of getting diseases like Nephropathy, Retinopathy and Cardio Vascular Diseases.

We executed the MapReduce program in standalone and in a cloud computing environment hosted by Amazon Web Services. For standalone mode the MapReduce program will run via Hadoop single node cluster and Hadoop cluster of machines for cloud computing mode will run via Hadoop AWS; these two modes are described in following sections.

**Table 2**
**Diabetic patients with High Risk**

| *Diabetic patients with High Risk* | |
|---|---|
| *Diseases* | *Dataset* |
| CVD | 65 |
| Retinopathy | 81 |
| Nephropathy | 42 |

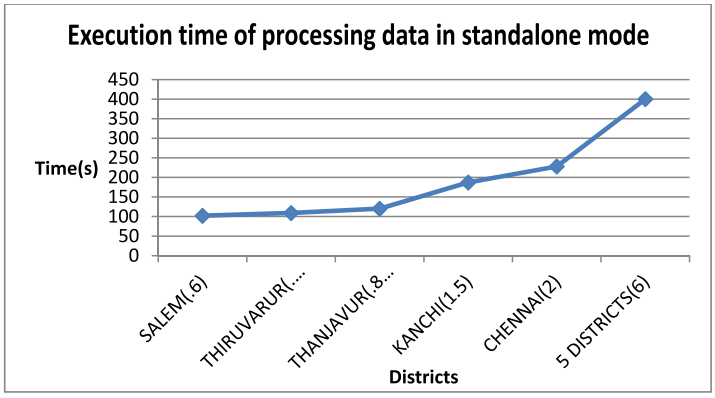



## 6. STANDALONE MODE

In standalone mode, a local machine or laptop is used, in order to run the MapReduce program, a single node Hadoop cluster must be initiated. In the experiment, the system was tested with five different sizes of dataset, collected from various districts which is made as big dataset using replica method and their execution time of processing data is shown in Table 3.

**Table 3**
**Execution time of processing data in standalone mode**

| *Execution time of processing data in standalone mode* | | |
|---|---|---|
| *Dataset* | *Districts* | *CPU Time (s)* |
| 1 | Salem (.6) million | 102 |
| 2 | Thiruvarur (.80) million | 109 |
| 3 | Thanjavur (.85) million | 120 |
| 4 | Kanchi (1.5) million | 187 |
| 5 | Chennai (2) million | 228 |
| 6 | 5 Districts (6) million | 400 |



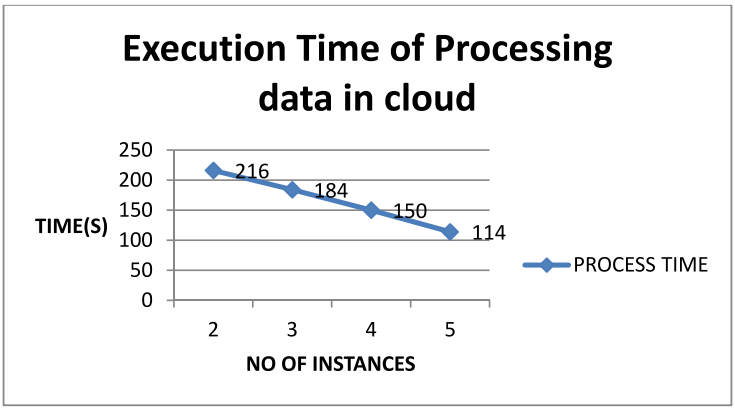
In standalone mode, there is only one node that everything runs in a single processor with storage using a local machine's standard file system. By comparing the processing of these datasets (millions), the time taken increases in each test by increasing the size of dataset. It shows that the size of dataset is a key parameter for time taken to process the datasets. The time taken of processing these datasets is high due to non-distributed environment to process large datasets and it has the limitations of the single processor in this standalone environment.

## 7. CLOUD COMPUTING MODE

In cloud computing mode, we have to to set up a Hadoop cluster in order to interact with Amazon Web Services such as Amazon S3, Amazon EC2 for running the MapReduce program in a fully distributed environment. The Amazon EC2 service provides virtual computing machines for large dataset processing and the S3 service is a storage service which stores large datasets in clouds[5].

To run the MapReduce system in cloud computing mode, a Hadoop cluster in Amazon EC2 is to be launched with homogenous cluster. The cluster has one master instance and one or more worker instances. After starting the cluster, the MapReduce program is run in the cluster and their execution time of processing data in a distributed manner is shown in Table below.

| *Amazon EMR Process Time* | |
|---|---|
| *EC2 Instances* | *Process Time for All Dist* |
| 2 | 216 |
| 3 | 184 |
| 4 | 150 |
| 5 | 114 |




According to the results obtained, the time taken for processing decreases when the dataset is distributed and processed in cloud computing environment by increasing the number of instances in the cluster.

By comparing the processing of these datasets it shows that the time taken to process in cloud computing mode is smaller than the time taken to process in standalone mode. In cloud computing mode Hadoop provides a distributed environment to run the MapReduce system in parallel over the cluster of machines. It means the processing of large datasets in cloud computing mode is faster than the processing of large datasets in standalone mode.

## 8. CONCLUSIONS

In conclusion, this paper conducted a comprehensive analysis of a big dataset using MapReduce framework as a new programming model for massive data analytics and its open source implementation, Hadoop. MapReduce

is an easy, effective and flexible tool for large-scale fault tolerant data analysis. It has proven to be a useful for the programmers to develop easily high performance system for running on cloud platforms and to distribute the processing over as many processors as possible. Thus We proposed a framework to process a large dataset in a cloud environment in parallel and distributed fashion, as well proposed Amazon Web Services as one instance of the using MapReduce framework.

## Future Work

Our future work aims in the direction of developing methods which can suggest some combinations for heterogeneous cluster for enhancing their performance in this scenario.

## REFERENCES

[1] Amazon, EC2, (accessed January 2012), Available online at http://aws.amazon.com/ec2

[2] Amazon, Elastic MapReduce, (accessed January 2012), Available online at http://aws.amazon.com/elasticmapreduce.

[3] J. Dean and S. Ghemawat, (2004). "MapReduce: simplified data processing on large clusters", Google Inc. In OSDI'04: Proceeding of the 6th conference on Symposium on Operating Systems Design & Implementation, San Francisco, CA.

[4] Hadoop MapReduce, (accessed February 2012), Available online at http://wiki.apache.org/hadoop/MapReduce.

[5] "Evaluation Of Data Processing Using Mapreduce Framework In Cloud And Standalone Computing", Samira Daneshyar and Ahmed Patel, International Journal of Distributed and Parallel Systems (IJDPS) Vol. 3, No. 6, November 2012.

[6] "Large-Scale Data Processing Using Mapreduce In Cloud Computing Environment", Samira Daneshyar and Majid Razmjoo, International Journal on Web Service Computing (IJWSC), Vol.3, No.4, December 2012.

[7] "Finding Insights & Hadoop Cluster Performance Analysis over Census Dataset Using Big-Data Analytics", Dharmendra Agawane, Rohit Pawar,, Pavankumar Purohit, Gangadhar Agre Guide: Prof. P B Jawade, Vol. 2, Issue 3, May-June-2016, Available at: www.knowledgecuddle.com/index.php/IJRAE.

[8] "Application Of Mapreduce In Diabetic Dataset Using Hadoop Platform", K. Sharmila & Dr. S.A. Vethamanickam, International Journal of Applied Engineering Research, ISSN 0973-4562 Vol. 10 No.60 (2015),© Research India Publications; httpwww.ripublication.comijaer.htm.

[9] "MRK-SVM: An Effective Technique for Big Data In Health Care Sector", K. Sharmila, Dr. S.A. Vethamanickam, International Journal of Scientific & Engineering Research, Volume 7, Issue 6, June-2016 ISSN 2229-5518.

[10] Yahoo! launches worlds largest hadoop production application. http://tinyurl.com/2hgzv7.

[11] J. Dean and S. Ghemawat. Mapreduce: Simplified data processing on large clusters. OSDI '04, pages 137–150, 2008.

[12] "Big Data Analytics Options on AWS", Amazon Web Services, January 2016.

[13] R.W. Moore, C. Baru, R. Marciano, A. Rajasekar and M. Wan, (1999) "Data-Intensive Computing", Morgan Kaufmann Publishers Inc. San Francisco, USA, ISBN:1-55860-475-8, pp. 105-129.

[14] Amazon Web Services, (accessed February 2012), Available online at http://aws.amazon.com.

[15] B. He, W. Fang, Q. Luo, N. Govindaraju, and T. Wang. Mars: a MapReduce framework on graphics processors. ACM, 2008.

[16] M. Isard, M. Budiu, Y. Yu, A. Birrell, and D. Fetterly. Dryad: distributed data-parallel programs from sequential building blocks. In EuroSys '07: Proceedings of the 2nd ACM SIGOPS/EuroSys European Conference on Computer Systems 2007, pages 59–72. ACM, 2007.

[17] C. Ranger, R. Raghuraman, A. Penmetsa, G. Bradski, and C. Kozyrakis. Evaluating mapreduce for multi-core and multiprocessor systems. High-Performance Computer Architecture, International Symposium on, 0:13–24, 2007.

[18] J. Dean, (accessed April 2011), "Experience with MapReduce, An Abstraction for Large-Scale Computation", Google Inc. Available online at www.slideshare.net/rantav/introduction-to-mapreduce.

[19] 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.

[20] "Performance Analysis of Multi-Node Hadoop Clusters using Amazon EC2 Instances", Ruchi Mittal, Ruhi Bagga, International Journal of Science and Research (IJSR) ISSN (Online): 2319-7064 Index Copernicus Value (2013): 6.14 | Impact Factor (2014): 5.611.

[21] "Predictive Methodology for Diabetic Data Analysis in Big Data", Dr Saravana kumar N M, Eswari T, Sampath P & Lavanya S, Procedia Computer Science 50 (2015) 203–208.