Hadoop MapReduce with Machine Learning Techniques for Big Data Computing On Cloud Environment

Ms. Priti Sadaria[1], Dr. Achyut C. Patel[2]

[1]Department of CS & IT, Atmiya University

[2]Smt. M. T. Dhamsania Commerce College, Saurastra University

[1]priti.sadaria@atmiyauni.ac.in, [2]acp2809@gmai.com

Abstract

Now a day no field remains untouched with Information Technology. Health care industries are using Information Technology for different aspects. Mining of valuable information by analyzing this quickly rising data for building a useful model which can be relevant in real life is really a difficult task. Knowledge discovery and decision making from such huge data is a novel trend that is Big Data Computing. Machine learning techniques can be used to make predictive analytics. Cloud computing provides computing services over the internet which includes servers, storage, databases, software and analytics for big data processing. Extracting useful information from this massive amount of data is highly difficult, expensive, and time consuming and therefore the problem can be solved by processing huge amount of data by applying machine learning techniques on Hadoop platform in Cloud environment.

Keywords: *Hadoop, MapReduce, Machine Learning, Big Data, Cloud*

1. Introduction

Processing Big Data with low cost setup is possible by using components of Hadoop frame work, which is MapReduce with HDFS. It provides mechanism of processing data in more simplified form and it gives highest performance with more accurate result. This paper includes MapReduce with Hybrid - SVM technique which is a machine learning technique used on Cloud for processing Big Data.

1.1 Cloud Computing

Cloud Computing is the way which provides facility to access servers and databases to achieve services over the Internet. Following list shows some Cloud services providers:

- Amazon Web Services (AWS)
- Google Cloud

- Microsoft Azure
- IBM Cloud
- Salesforce
- Oracle
- VMWare etc.

Among all of above, Amazon Web Services (AWS) provides admirable services to access database and servers all over the world means internet. Cloud Computing Services can be organized in three different way, Public cloud, Private cloud and Hybrid cloud.

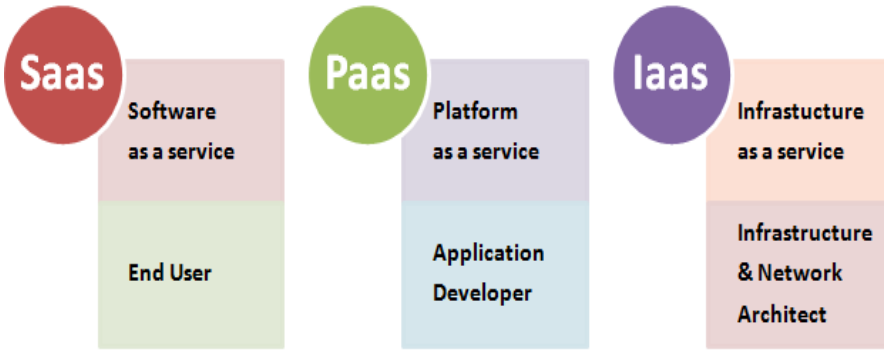

Figure 1. AWS Services

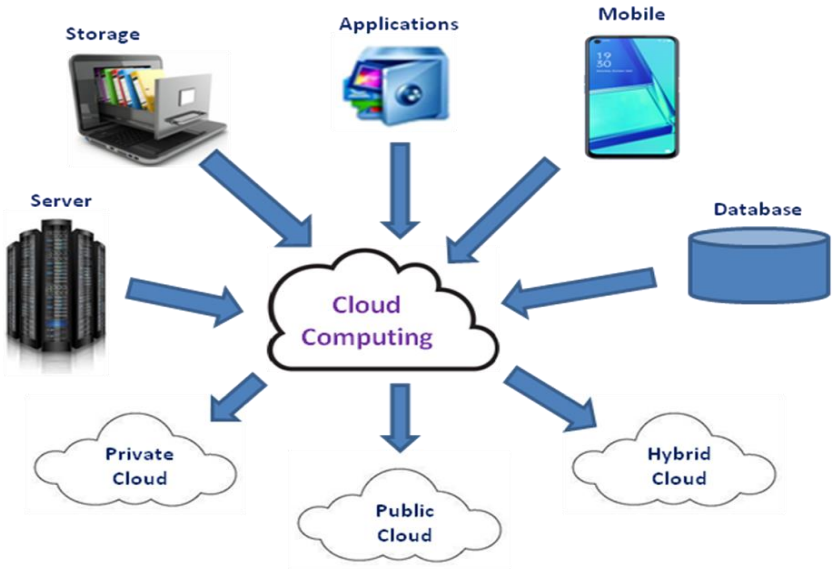

Figure 2. Cloud Computing

1.2 Amazon Web Services

AWS provides different types of services at lower cost. Main services provided by AWS are Compute Services, Storage Services, Database Services, Migration Services and Messaging Services etc. In computing environment, MapReduce infrastructure service provides Amazon EC2 and Amazon S3 for processing and storing huge dataset. Here input and output dataset is

accumulated in Amazon S3. Elastic MapReduce (EMR) and Elastic Compute Cloud (EC2) services are used for interaction purpose between developer and Cloud environment.

1.3 Hadoop

Hadoop is an open source environment which can be useful for processing huge datasets. Processing huge dataset on a single computer is not convenient even though the computer is large and storage capacity is huge. Hadoop provides facility to analyze gigantic datasets in parallel manner by using the clustering concepts. Amazon EMR simplifies the task of processing huge dataset and makes processing fast and gives profitable approach. Amazon EMR uses Hadoop environment for distributing data and such distributed data processed parallel manner among Amazon EC2 instances.

Hadoop uses MapReduce and HDFS where jobs are divided into smaller task and distributed among different nodes in Amazon EMR cluster. Hadoop Distributed File System stores data into local storage disks in large blocks so it can be used with Amazon S3 to store the input and output data file.

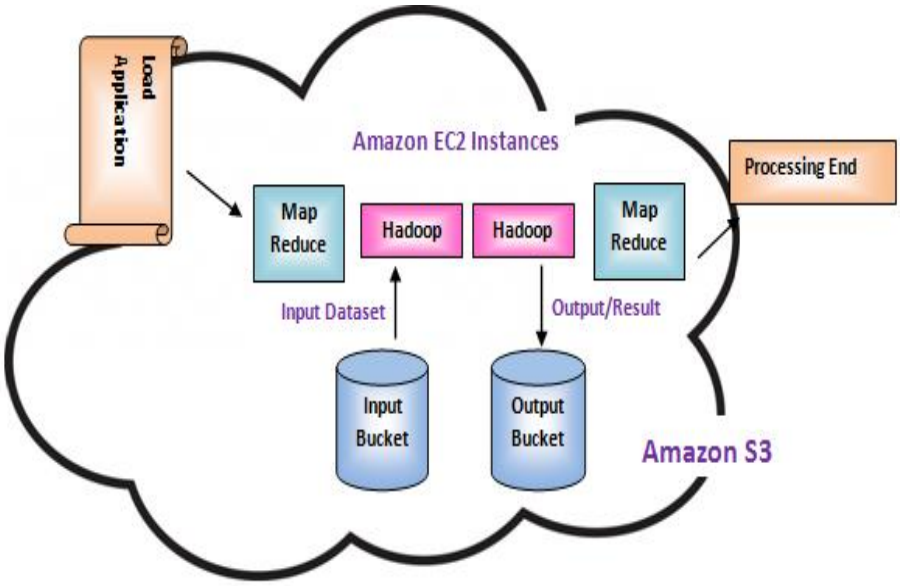

Figure 3. Amazon Web Service with Hadoop MapReduce

2. Implementation of work

The main aim of the research is to scrutinize big data by using Hadoop MapReduce with AWS. The data which is used here is gathered from different district of Gujarat. The proposed model used Hadoop MapReduce with Cloud computing to predict chronic diseases related to diabetes. In research the Hybrid - SVM model is used classification technique SVM with K-means clustering with Hadoop MapReduce for prediction purpose.

2.1 Cloud Computing with AWS

Hadoop cluster with EC2 and EMR has been loaded with main instance for processing Hadoop MapReduce in Cloud environment. First cluster is required to be start and the coding and input datasets were loaded on S3 and later on it has been moved to Hadoop platform which is accessible via Cloud environment. Once processing is completed, the result file has been send to the output directory of S3. Here the execution or processing time has been also noted. Among form the various datasets, the dataset which size is 6000000 can be proficiently analyzed in AWS but it cannot be processed in Hadoop standalone mode.

The Table 1 shows the execution time in AWS environment for different datasets.

Table 1. Execution time for processing Dataset in AWS Cloud Environment

Dataset	Districts	No of Records	Processing Time
1	Anand	650000	72
2	Rajkot	843000	86
3	Baroda	857229	104
4	Ahmedabad	1650000	155
5	Surat	2000000	180
6	Total of 5 district	6000000	216

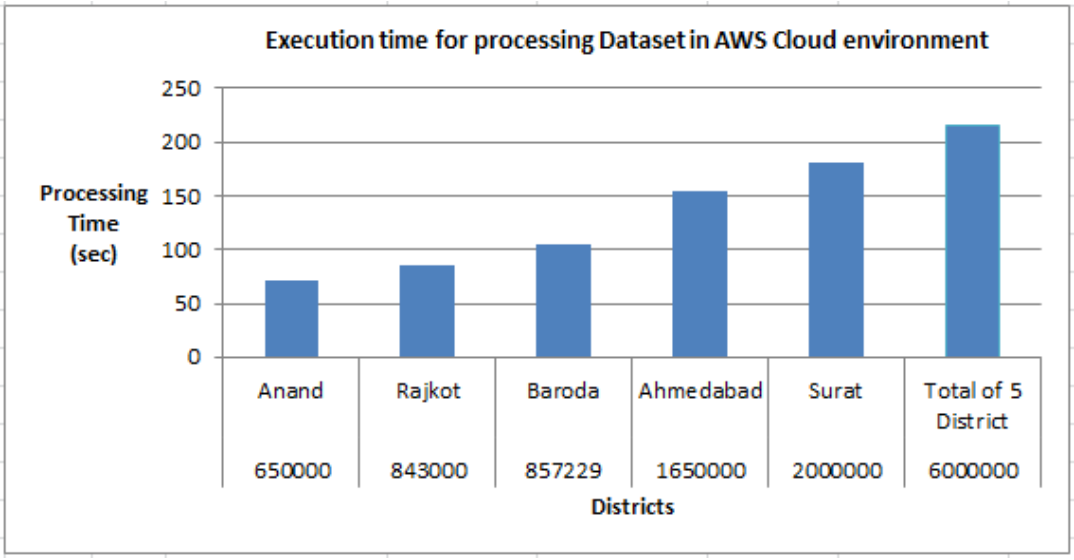

Figure 4. Execution time for processing Dataset in AWS Cloud Environment

2.2 Distributed processing in Cloud environment

For processing more than one instance, several computing machines are used in Cloud environment for execution of MapReduce program. In the research, the data set with size 6000000 processed with four instances in Hadoop MapReduce environment is distributed manner. Here the processing time for individual instance was noted down and it concluded that as the number of instances increase, the processing time is decreased. Table 2 shows the analyzed result for processing time and number of instances.

Table 2. Execution time for multiple instances in EMR Cloud Environment

EC2 Instance	Execution Time
1	216
2	184
3	150
4	114

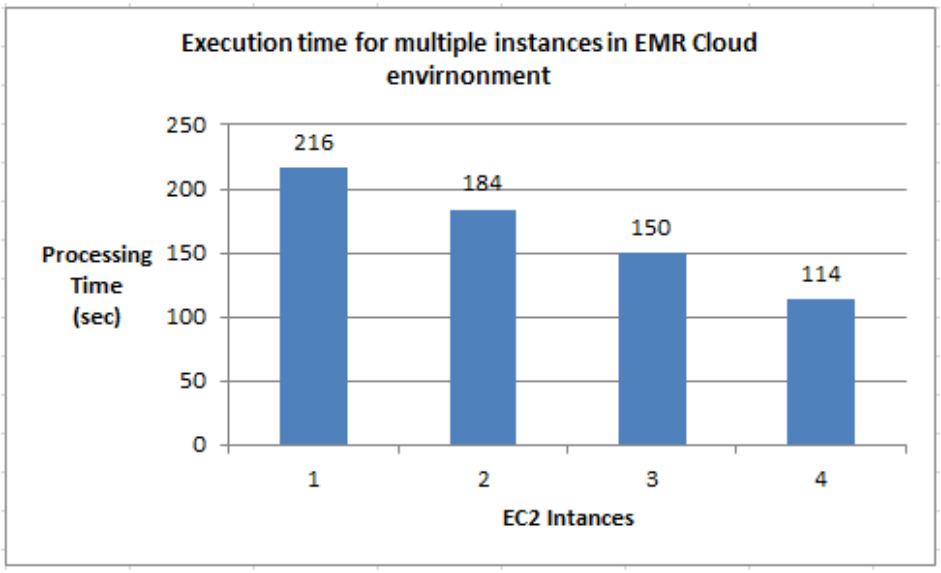

Figure 5. Execution time for multiple instances in EMR Cloud Environment

3. Evaluation

The evaluation has been done by using speed of processing and response time.

3.1 Response Time

The response time is more significant for evaluation in Hadoop MapReduce environment. The time taken between a starting of request and actual beginning of cluster is known as

response time. In fact response time very with the number of running instances and the size of EC2 cluster.

Table 3. Response time for multiple EC2 instances

Number of EC2 instances	Response Time
1	10
2	11
3	14
4	17

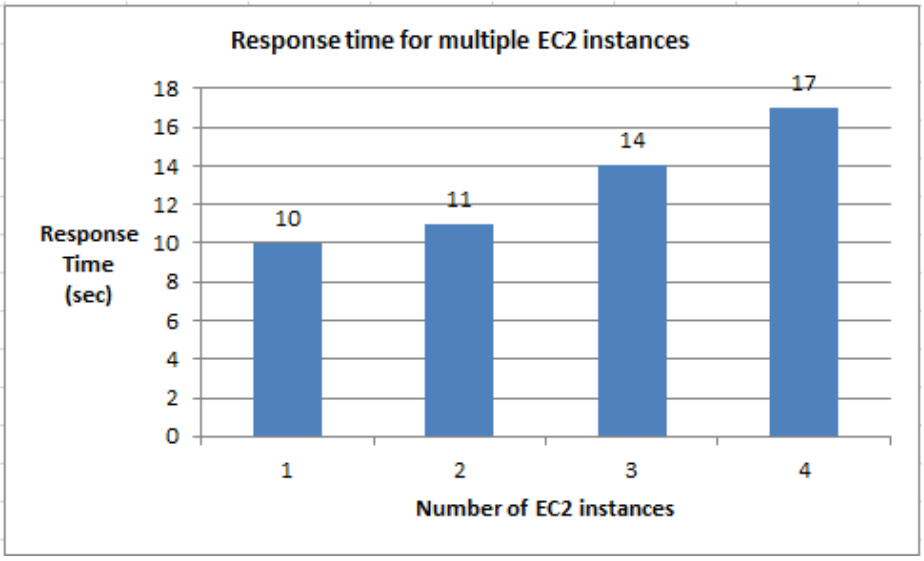

Figure 6. Response time for multiple EC2 instances

3.2 Speed of Analysis

The speed of analysis is determined from the response time and processing time. The following formula shows the calculation of speed of analysis.

Speed of analysis = Response Time + Processing time

SA (in seconds) = RT (in seconds) + PT (in seconds)

Table 4 shows EC2 instances and processing time in Cloud environment.

Table 4. Nos. of EC2 instances and processing time in Hadoop on AWS

Number of EC2 instances	Processing time (seconds)
1	226
2	195
3	164
4	131

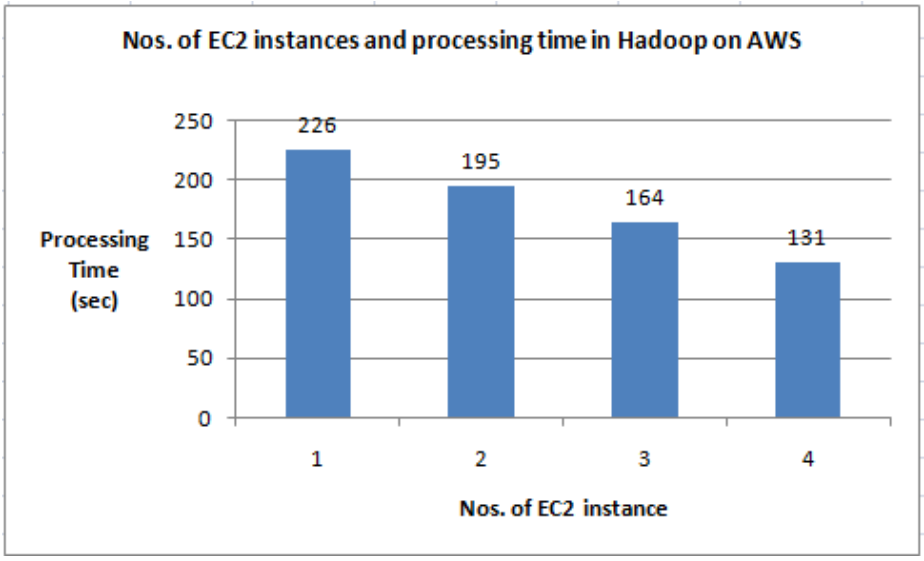

Figure 7. Nos. of EC2 instances and processing time in Hadoop on AWS

Figure 7 shows that with increasing the number of EC2 instances on AWS, the processing speed is decreased.

The Table 5 shows the comparison of processing time for single instance in Hadoop standalone mode and Hadoop on AWS.

Table 5. Processing time in Hadoop standalone & Hadoop on AWS for single instance

Dataset	No of Records	Processing Time (Hadoop standalone) (seconds)	Processing Time (Hadoop on AWS) (seconds)
1	650000	102	72
2	843000	109	86
3	857229	120	104
4	1650000	187	155
5	2000000	228	180
6	6000000	Not able to process	216

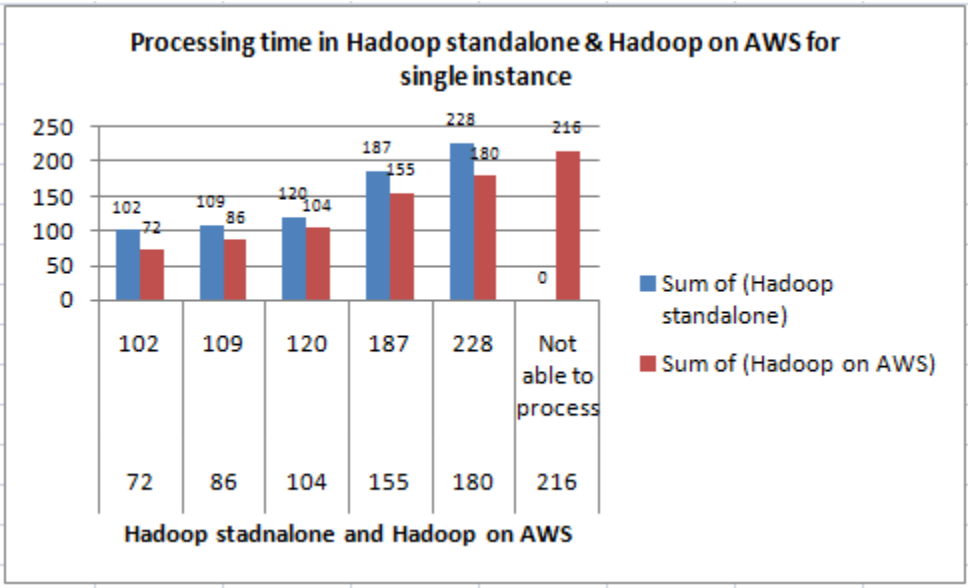

Figure 8. Processing time in Hadoop standalone & Hadoop on AWS for single instance

It is concluded from the Figure 6.9 that as compare to Hadoop standalone mode, Hadoop on Cloud with AWS takes fewer processing time for analyzing big dataset because in Cloud environment dataset is processed in distributed form on cluster of various machine.

4. Conclusions

With reference to above research, it can be concluded that for processing big dataset Hadoop MapReduce with Cloud provides best environment because it process data very quickly in distributed and parallel manner. At low cost high speed data processing can be done by using computer with higher configuration and high speed. The most important aspect of the research is

that it overwhelms the problem faced by the standalone mode for processing big dataset by usage of clusters of machines is AWS Cloud environment.

5. Reference

1. *Muni kumar N, Manjula R, " Role of Big Data Analytics in, Rural Helath Care – A Step Towards Svasth Bharath", International Journal of Computer Science and Information Technologies*
2. *Viceconti M, Hunter P, Hose R. Big data, big knowledge: big data for personalized healthcare. IEEE J Biomed Health Inform. 2015*
3. *https://www.toptal.com/spark/introduction-to-apache-spark*
4. *Repu Daman, Manish M. Tripathi, Saroj K. Mishra - "Cloud Computing for Medical Applications & Healthcare Delivery:Technology, Application, Security and Swot Analysis" ACEIT Conference 2016*
5. *Zhanquan Sun ―Study on Parallel SVM Based on MapReduce in conference on world comp. 2012.*
6. *Wullianallur Raghupathi, Viju Raghupathi, "Big data analytics in healthcare: promise and potential", Health Information Science and Systems, 2(3): 2-10. 2014.*