International Journal of Research
Available at https://edupediapublications.org/journals
e-ISSN: 2348-6848
p-ISSN: 2348-795X
Volume 05 Issue 15
May 2018

# Mapreduce Programming Paradigm in Cloud Environment for Large-Scale Data Mining

**1.Mr. Bhaludra R Nadh Singh, Research Scholar, Acharya Nagarjuna University, Guntur, Andhra Pradesh, India, 2. Dr. B. Raja Srinivasa Reddy, Research Guide, Acharya Nagarjuna University, Guntur, Andhra Pradesh, India.**

## Abstract

Massive amount of data is produced of late. Such data is known as big data with characteristic such as volume, velocity, variety and value. Such data when mined can provide comprehensive business intelligence. However, the problem with big data is that it needs storage measures in peta bytes and huge computational power to process it. Cloud computing is the perfect solution to storage and mine big data. MapReduce is the new programming paradigm to process big data. It is associated with distributed programming frameworks like Hadoop. Scalable mining of big data is the problem addressed. However, the mining services that can scale up to big data are to be investigated. In this paper we investigated different approaches and proposed an algorithm for processing big data. We built a prototype application to demonstrate proof of the concept. The empirical results revealed that the proposed method is useful in mining massive amounts of data.

**Index Terms –** Big data, cloud computing, Hadoop, MapReduce, data mining

## 1. INTRODUCTION

There has been enormous growth of data of late. It is known as big data which comes from different sources. The sources generally include social media, sensor networks, and satellites and so on. In addition to this, enterprises produce huge amount of data every day. Handling big data is an important problem. Big data cannot be stored and processes in the local systems. Therefore it needs the help of cloud computing where large pool of computing resources is provisioned in pay per use fashion without time and geographical restrictions.

With cloud computing resources, big data can be stored and managed efficiently. However, there are many mining methods that need to be understood to have better strategy to gain business intelligence.

There are many existing methods to handle big data. MapReduce programming is used to process massive amounts of data. The existing methods focused on specific approaches that are meant for mining data. In this paper we proposed and implemented three kinds of machine learning or data mining approaches that can handle or mine massive amount of data. They are known as sampling method, ensemble method and MapReduce based method. The three approaches are evaluated with big data. The sampling approaches are found in [7] and [14] while the ensemble approaches are explored in [4]. The MapReduce approaches are found in [10] and [12]. The following are main contributions of the paper.

1. Three kinds of mining approaches are explored to handle big data. These methods are meant for processing big data with machine learning by training classifiers and utilising them further to classify big data.
2. A prototype application is built to demonstrate proof of the concept. The empirical results revealed that the application is able to support different kinds of methods in data mining in the context of big data and cloud eco-system.
3. We evaluated the three approves and found the one which shows superior performance. The three approaches are compared to have useful insights and help in making well informed decisions.

The remainder of the paper is structured as follows. Section 2 presents review of literature. Section 3 provides sampling approach for processing big data. Section 4 presents ensemble method while section 5 describes MapReduce approach. Section 6 presents experimental results. Section 7 concludes the paper besides providing directions for future work.

## 2. RELATED WORK

This section provides review of literature on big data and mining of it in the distributed computing frameworks. Chang et al. [1] explored a distributed storage system that is meant for storing and managing structured data. Cao et al. [2] on the other hand explored data

mining techniques on KDD dataset for context-aware query suggestions. As big data needs MapReduce programming frameworks, Chu et al. [3] studied machine learning approaches on MapReduce environments with multi-core. An ensemble data mining approach that helps producing better performance is explored in [4]. It focused on a scalable and accurate approach. Genetic Algorithm (GA) based large scale data mining is investigated in [5] for exploring machine learning techniques. Chang et al. [6] explored parallel algorithms that work on large volumes of data. Rich-media data with large scale is studied to obtain business intelligence. Parallel algorithms are able to exploit parallel processing power of computing infrastructure.

For discovering knowledge from large databases, Domingo et al. [7] studied adaptive sampling methods for scalable extraction of business intelligence. Gu and Grossman [8] explored high performance data cloud with design and implementation. Data mining approaches with high performance data clouds is the focus in [9] for obtaining knowledge. MapReduce programming is studied on the large clusters in [10]. The distributed file system used by Google is explored in [11] for simplified data processing. Distributed computing and machine learning are studied with MapReduce programming framework in [12]. Data mining techniques are explored in [13] for scalable online collaborative filtering. Data mining applications with sampling is investigated in [14]. In the literature it is found that there are different approaches found in the data mining with large volumes of data. In this paper we studied three approaches to understand the performance with respect to machine learning on big data in cloud computing.

## 3. ENSEMBLE APPROCH FOR LARGE-SCALE DATA MINING

Ensemble approach is followed to handle mining big data in a better way. According to this approach the dataset (big data training set) is divided into multiple groups. Each file is randomly assigned to a group. An individual classifier is built on each group and performs parallel classification of testing datasets.

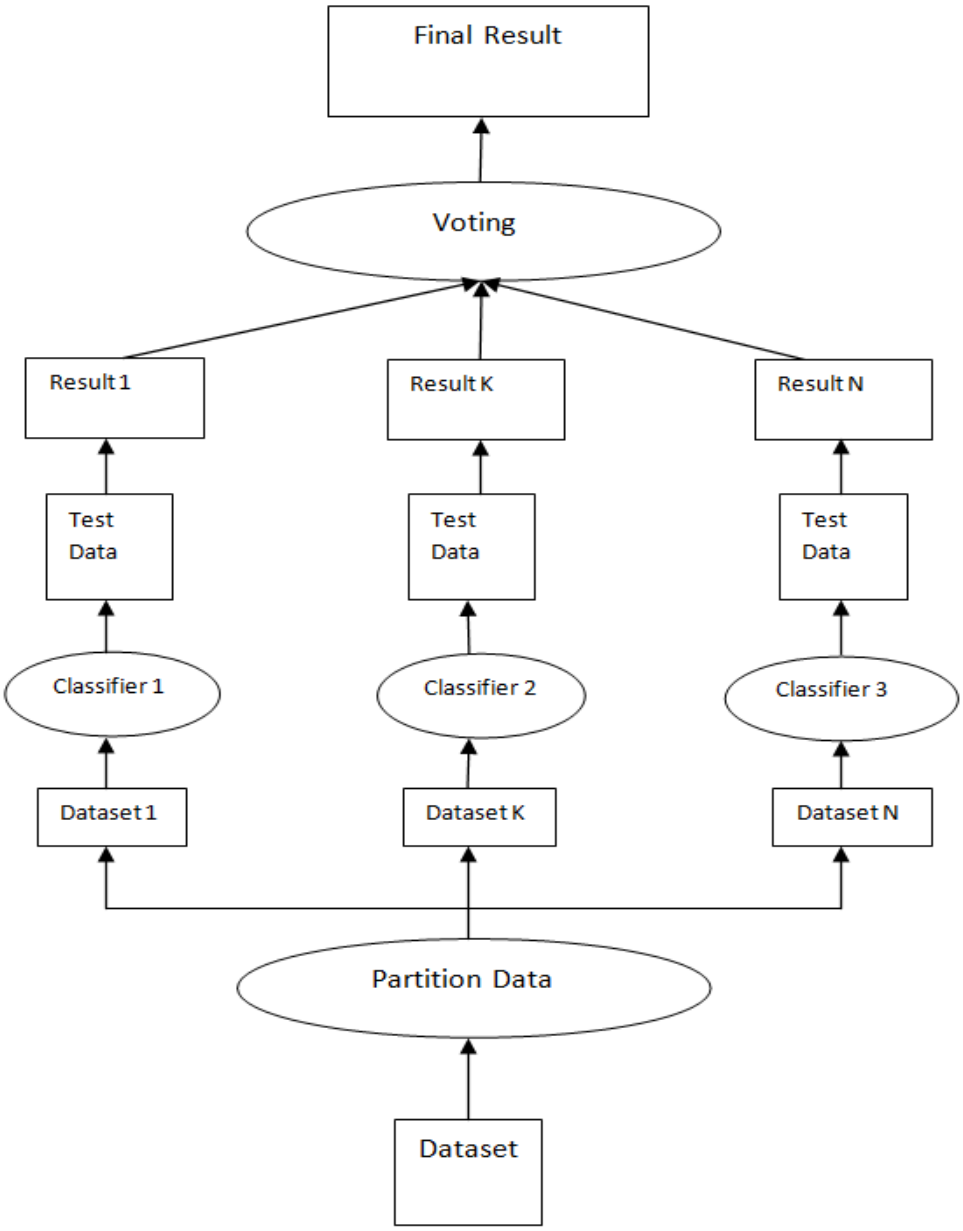


**Figure 1:** Ensemble approach

As presented in Figure 1, it is evident that the dataset is partitioned into multiple parts. Each part is given to a classifier. Once a classifier is built, it can be used to classify testing data and the ensemble of multiple classifiers is expected to improve accuracy of machine learning process.

## 4. SAMPLING MODEL

There are many sampling methods in data mining. Out of them random sampling is widely used approach. The sampling model proposed in this paper makes use of random number in order to choose data from training set and then sample dataset is used to train a classifier named Naive Bayes.

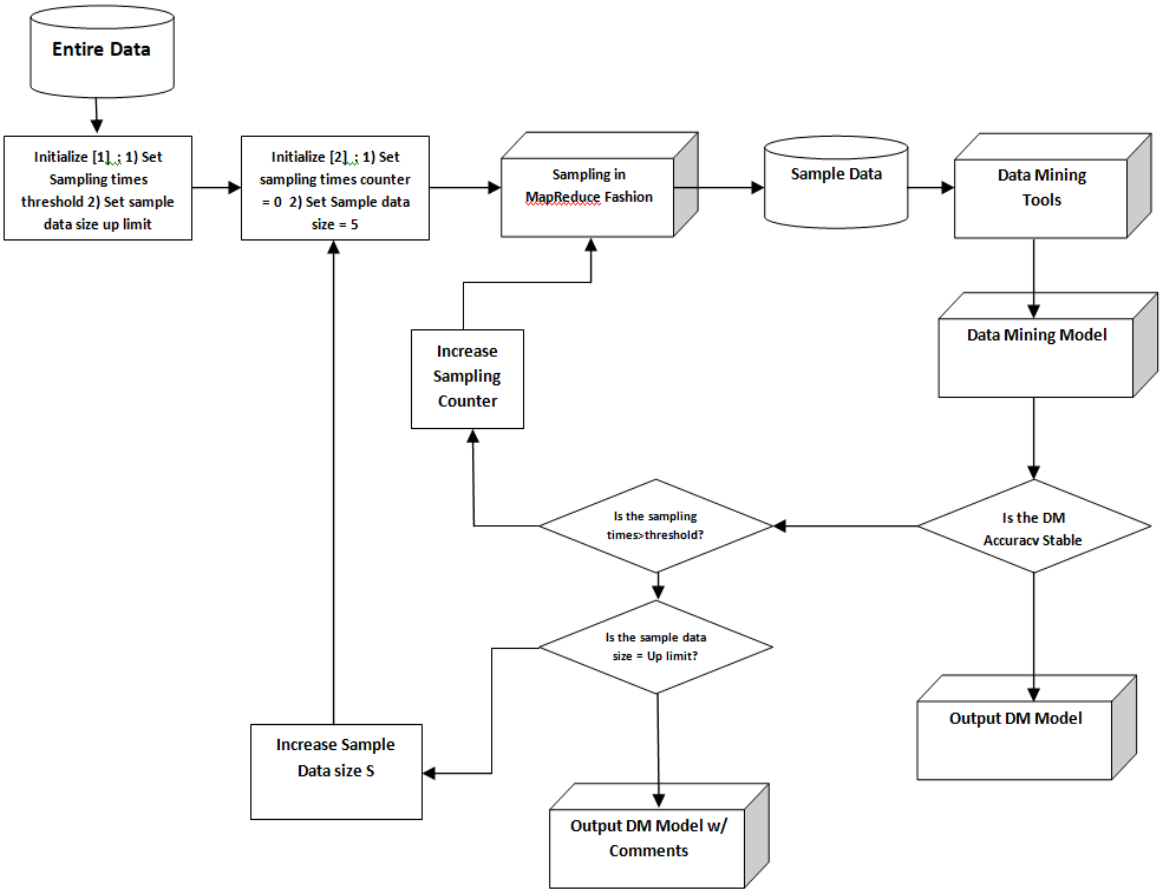


Figure 2: MapReduce sampling framework

As presented in Figure 2, it is evident that the entire dataset is subjected to sampling in MapReduce fashion. Sampled data is subjected to data mining tools and data mining models. Once the stable accuracy is produced, the output data mining model is finalized. If not the sampling process is continued.

## 5. MAP REDUCE CLASSIFIER FRAMEWORK FOR LARGE-SCALE DATA MINING

The MapReduce classifier framework provides the required model to classify big data. It is based on MapReduce programming framework such as Hadoop. In this model training set

with huge amount of data is given as input. Then the mappers and reducers are trained in order to provide output model. Once output model is produced, the testing set is given as input to have classification of the data and produce final output.

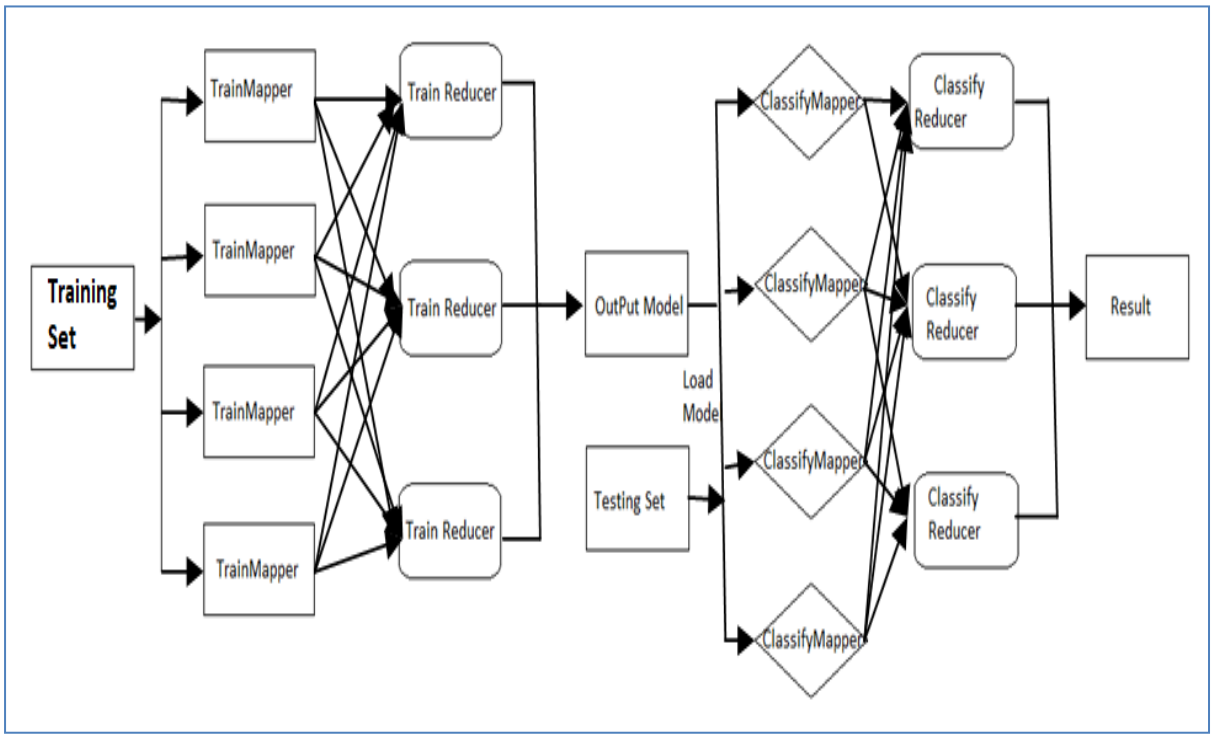


**Figure 3:** MapReduce classifier framework for processing big data

As shown in Figure 3, it is evident that the Map and Reduce phases are involved in the training and testing as well. Once training is made with the training dataset, the output is used to classify the testing data.

## 6. EXPERIMENTAL RESULTS

Experiments are made using different approaches and the accuracy of data mining or machine learning operations such as classification (supervised learning). Hadoop MapReduce environment is used for experiments.

| Round | Classification Accuracy | | | |
|---|---|---|---|---|
| | 1K | 2K | 5K | 10K |
| 1 | 64.2 | 65.5 | 66.6 | 67 |

| | | | | |
|---|---|---|---|---|
| 2 | 64.2 | 65.5 | 66.6 | 67 |
| 3 | 64.2 | 65.5 | 66.6 | 67 |
| 4 | 64.2 | 65.5 | 66.6 | 67 |
| 5 | 64.2 | 65.5 | 66.6 | 67 |
| 6 | 64.2 | 65.5 | 66.6 | 67 |
| 7 | 64.2 | 65.5 | 66.6 | 67 |
| 8 | 64.2 | 65.5 | 66.6 | 67 |
| 9 | 64.1 | 65.2 | 66.5 | 66.8 |
| 10 | 64.2 | 65.5 | 66.6 | 67 |

**Table 1:** Shows mining accuracy with sampling data mining

As shown in Table 1, it is evident that the classification accuracy with different rounds is recorded for different data sizes such as 1k, 2k, 5k and 10k. The statistics show the sampling accuracy on different sampling sizes.

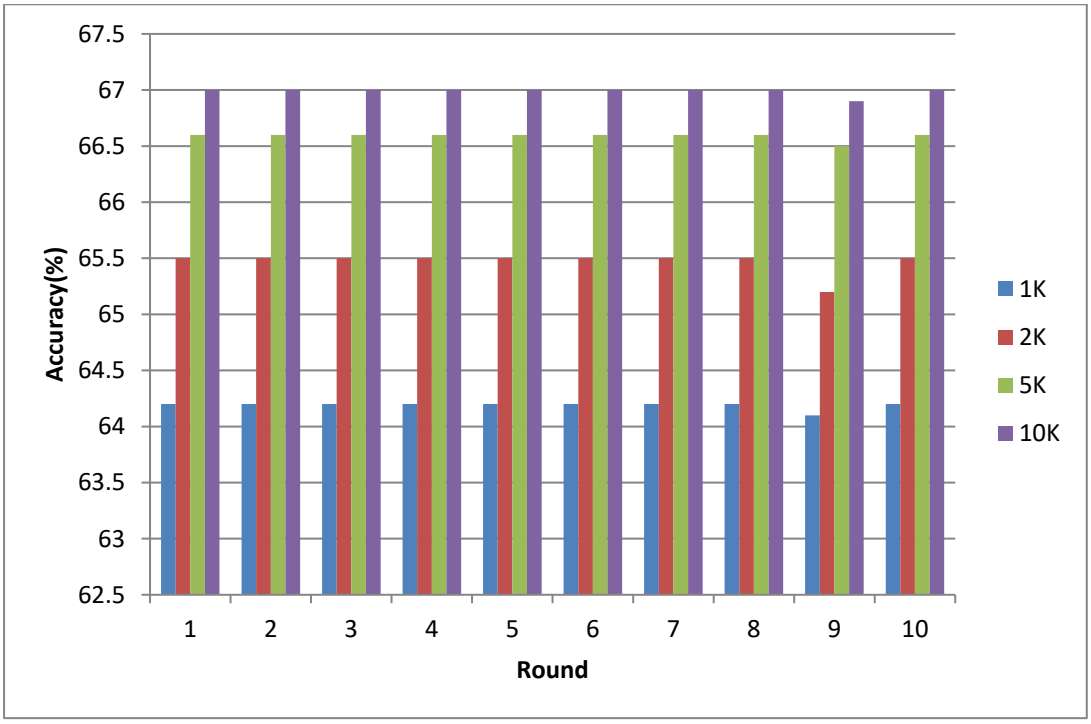


**Figure 4:** Shows performance of classification accuracy on different sampling sizes

As shown in Figure 4, it is evident that the horizontal axis shows the rounds while the vertical axis shows classifier accuracy (%) with different sampling sizes. There is liner relationship between accuracy and the sampling size.

| **Round** | **Accuracy** | | |
|---|---|---|---|
| | **5 sub models** | **10 sub models** | **15 sub models** |
| 1 | 69.1 | 68 | 66 |
| 2 | 69.2 | 68.5 | 67.5 |
| 3 | 68 | 66 | 65.5 |
| 4 | 69.1 | 67.5 | 67 |
| 5 | 69.2 | 68 | 67.2 |
| 6 | 69.05 | 68 | 66.5 |
| 7 | 68.5 | 67 | 66 |
| 8 | 69.3 | 68.3 | 67 |
| 9 | 69.15 | 68.5 | 66.5 |
| 10 | 69 | 68.5 | 67 |

**Table 2:** Shows mining accuracy with ensemble data mining approach

As shown in Table 2, it is evident that the classification accuracy with different rounds is recorded for different sub models. The statistics show the classification accuracy on different sub models of ensemble data mining.

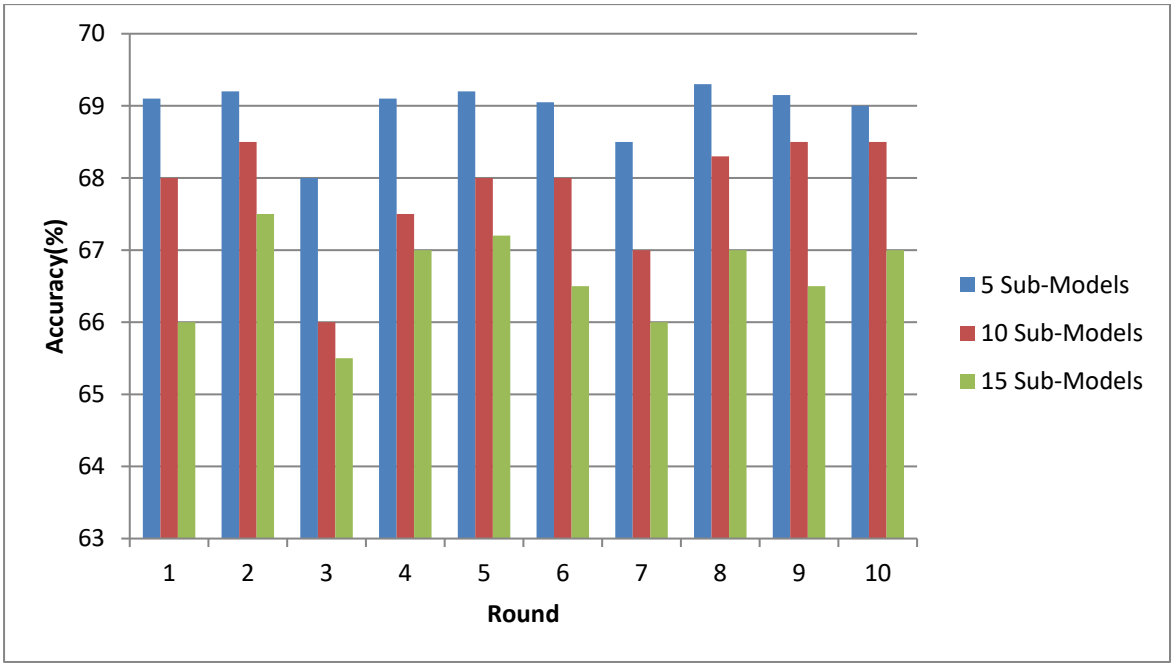


**Figure 5:** Shows classifier accuracy of ensemble method with different no. of sub models

As shown in Figure 5, it is evident that the details of different rounds are shown in horizontal axis while the vertical axis shows classifier accuracy (%). The results are made as per the ensemble method. The number of sub models has its influence on the classifier accuracy.

| Round | Accuracy | | |
|---|---|---|---|
| | Hadoop | Sampling | Sub models-ensembles |
| 1 | 67.9 | 67 | 69.1 |
| 2 | 68.7 | 67.1 | 69.2 |
| 3 | 67.5 | 66.9 | 68.8 |
| 4 | 68.5 | 67 | 69.1 |
| 5 | 68.3 | 67 | 69.2 |
| 6 | 68.8 | 67.1 | 69.1 |
| 7 | 67.9 | 67 | 68.9 |
| 8 | 68.5 | 67 | 69.2 |
| 9 | 68.6 | 66.9 | 69.1 |
| 10 | 68.4 | 67 | 69 |

**Table 3:** Comparison of classification accuracy of the three models

As presented in Table 3, it is evident that the classifier accuracy is compared among the three approaches. Out of all the approaches, the ensemble method is found to show highest accuracy in all the rounds observed.

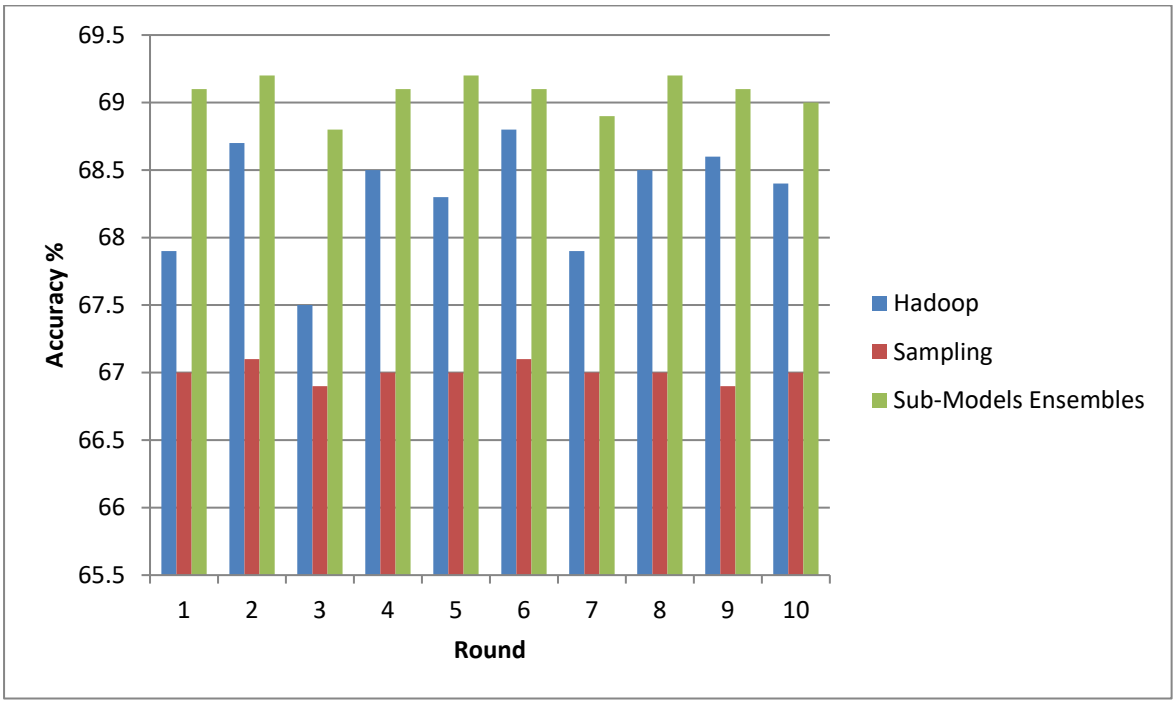


**Figure 6:** Shows classifier accuracy comparison among the methods

As shown in Figure 6, it is evident that the details of different rounds are shown in horizontal axis while the vertical axis shows classifier accuracy (%) of the three methods. The results revealed that the ensemble method with different sub-models showed better performance. The classifier accuracy of Hadoop is better than that of sampling data mining method.

## 7. CONCLUSIONS AND FUTURE WORK

Processing voluminous data or big data demands huge amount of storage and computing resources. Cloud computing has emerged to solve the problem of big data. In other words, bit data is possible to be stored and mines with shared pool of computing resources provided on-demand by cloud in pay per use fashion. In this paper we investigated data mining approaches on big data with different approaches such as ensemble approach and MapReduce approach. MapReduce is the new programming paradigm that can exploit the parallel

processing power of cloud infrastructure. It has associated file distributed file systems to have massive amount of data stored and processed. Ensemble approach and sampling approach are evaluated in terms of classification accuracy. A prototype application is built to demonstrate proof of the concept.

**References**

[1] F. Chang, J. Dean, S. Ghemawat, W. C. Hsieh, D. A. Wallach, M. Burrows, T. Chandra, A. Fikes, and R. E. Gruber. Bigtable: a distributed storage system for structured data. In OSDI '06: Proceedings of the 7th symposium on Operating systems design and implementation, pages 205–218, Berkeley, CA, USA, 2006. USENIX Association.

[2] H. Cao, D. Jiang, J. Pei, Q. He, Z. Liao, E. Chen, and H. Li. Context-aware query suggestion by mining click-through and session data. In KDD '08: Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining, pages 875–883, New York, NY, USA, 2008. ACM.

[3] C. T. Chu, S. K. Kim, Y. A. Lin, Y. Yu, G. R. Bradski, A. Y. Ng, and K. Olukotun. Map-reduce for machine learning on multicore. pages 281–288. MIT Press, 2006.

[4] N. V. Chawla, L. O. Hall, K. W. Bowyer, and W. P. Kegelmeyer. Learning ensembles from bites: A scalable and accurate approach. J. Mach. Learn. Res., 5:421–451, 2004.

[5] J. Bacardit and X. Llor`a. Large scale data mining using genetics-based machine learning. In GECCO '09: Proceedings of the 11th Annual Conference Companion on Genetic and Evolutionary Computation Conference, pages 3381–3412, New York, NY, USA, 2009. ACM.

[6] E. Y. Chang, H. Bai, and K. Zhu. Parallel algorithms for mining large-scale rich-media data. In MM '09: Proceedings of the seventeen ACM international conference on Multimedia, pages 917–918, New York, NY, USA, 2009. ACM.

[7] C. Domingo, R. Gavald`a, and O. Watanabe. Adaptive sampling methods for scaling up knowledge discovery algorithms. Data Min. Knowl. Discov., 6(2):131–152, 2002.

[8] Y. Gu and R. Grossman. Sector and sphere: The design and implementation of a high performance data cloud. Theme Issue of the Philosophical Transactions of the Royal Society, E-Science and Global E-Infrastructure, 367:2429–2455, 2009.

[9] R. Grossman and Y. Gu. Data mining using high performance data clouds: experimental studies using sector and sphere. In KDD '08: Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining, pages 920–927, New York, NY, USA, 2008. ACM.

[10] J. Dean and S. Ghemawat. Mapreduce: simplified data processing on large clusters. In OSDI'04: Proceedings of the 6th conference on Symposium on Opearting Systems Design & Implementation, pages 10–10, Berkeley, CA, USA, 2004. USENIX Association.

[11] S. Ghemawat, H. Gobioff, and S. Leung. The google file system. SIGOPS Oper. Syst. Rev., 37(5):29–43, 2003.

[12] D. Gillick, A. Faria, and J. DeNero. Mapreduce: Distributed computing for machine learning.

[13] A. S. Das, M. Datar, A. Garg, and S. Rajaram. Google news personalization: scalable online collaborative filtering. In WWW '07: Proceedings of the 16th international conference on World Wide Web, pages 271–280, New York, NY, USA, 2007. ACM.

[14] B. GU, F. HU, and H. LIU. Sampling and its application in data mining: A survey. 2000.