Advances in Computer Science and Information Technology (ACSIT)
p-ISSN: 2393-9907; e-ISSN: 2393-9915; Volume 3, Issue 2; January-March, 2016, pp. 90-93

http://www.krishisanskriti.org/Publication.html

Survey: Integrated use of Data Mining on Cloud Computing

Deepali Dhama[1], Kavita Jaiswal[2] and Manju Bala (Mentor)[3]

[1,2,3]Department of Computer Science, Indraprastha College for Women
E-mail: [1]deepalidhama95@gmail.com, [2]kavitajaiswal811@gmail.com, [3]manjugpm@gmail.com

Abstract—*Cloud computing is a method of computing where data is managed, stored as well as processed on a remote network of remote servers but can be accessed on personal computers instantly on demand. On the other hand data mining is used for extracting useful information from raw data. Various data -mining techniques are being used in our day- to- day activities nowadays. Data mining techniques help businesses to find their targeted customers. This paper emphasizes on integrated use of data mining techniques on cloud storage areas such as open source software Hadoop and MapReduce framework on which Hadoop is implemented, along with Apache Mahout which is Java machine learning library and a data mining framework used to manage big data. Nowadays various Mahout Machine learning algorithms are implemented for the Collaborative filtering, Clustering, Classification and Frequent item-set mining of the exponential amount of data around us. The combined use of data mining and cloud computing will be time reducing, highly reliable, high availability, cost effective and easy to manage.*

Keywords*: datamining, Cloud Computing, Hadoop, Mahout, MapReduce.*

1. INTRODUCTION

In this era of internet, data is growing on an exponential rate. This trend leads to the advancement in data collection and storage techniques. Data mining can facilitate the process of finding relationships and patterns in raw data and the results can be utilized further, while Cloud computing is a new business model containing pool of data and resources consisting of large number of computers [1]. It divides the computation task to its pool of data so that applications can retrieve variety of software services as per demand. Cloud computing provides unlimited data storage and computing power which help us in mine mass amount of data [6].

Cloud computing hosts an increasing number of applications in private and public clouds. The cloud services that are categorized on the basis of type of service they provide: Software as a Service (SAAS), Platform as a Service (PAAS), Infrastructure as a Service (IAAS) and also Business as a Service (BAAS) [2]. Amazon, Microsoft, Google are some of the major cloud service providers. An example of paas provided by Google which allows web application hosting is Google App Engine (GAE).

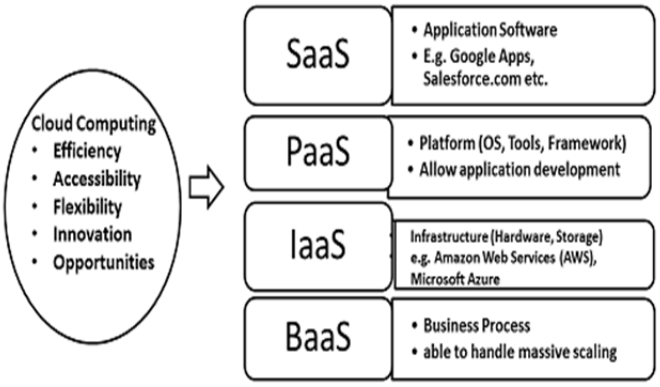

Fig. 1

Some of the services offered by Microsoft providing processing and storage capabilities for large data sets are Windows Azure, SQL Azure. Some cloud services provided by the Amazon are Amazon Web Services (AWS) including Simple Storage Service (S3), SQS and EC2 [4].

Big data mining has become essential for many industries to extract related and useful information from huge and data sets to support core operations and decision-making processes in the business. In order to manage the high demand of big data analytics, the main concern is the parallel data processing infrastructure [11]. Since it is essential in order to deal fault-tolerance and scalability, Hadoop has become the de facto standard in this direction. Based on Hadoop, Mahout is the open source library of data mining algorithms for various tasks such as clustering, classification, and recommendation [10].

This paper conducts an in-depth analysis and discussion on the big data components like Hadoop and Mahout. And also discuss why HADOOP on the cloud makes sense.

The rest of the paper is organized as follows: Section II briefly described related works to Hadoop, Mahout and explained MapReducing framework. Section III explained the use of Hadoop framework over the cloud and its advantages. Section IV concludes the paper also highlighting future works.

2. RELATED WORKS

2.1 Apache Mahout

We are in a world where huge information is available. The information has scaled to this height that, even our mailbox becomes so difficult to manage. Now is common for normal known websites to receive exponential amounts of information in bulk, and then to manage this big data, Mahout is a java machine learning library, and a data mining framework it runs with the Hadoop framework and manage exponential amount of data as a backend [23]. Apache Mahout was started by Isabel Drost, Grant Ingersoll and Karl Wettin. It was a part of the Lucene project but then it went on and become a top level project in April of 2010 [25]. Mahout implement four machine learning techniques [26].

1) *Collaborative filtering*, it mines user behavior and make product recommendations.Distributed Item-based Collaborative Filtering is an example of this. It estimates a user's preference for similar items by focusing at their preferences. it is done by Parallel Matrix Factorization algorithm. This algorithm predicts which items the user might prefer among a matrix of items that a user yet not seen [21].

2) *Clustering, it* takes items, group them in particular classes and organizes them into simple occurring groups [24] some algorithms that are used for clustering are Canopy, Dirichlet Process, K-Means Clustering, Fuzzy K-Means and Hierarchical Clustering. K-Means Clustering, partition n-observations into k-clusters in which each observation belongs to the cluster and Hierarchical Clustering builds a hierarchy of clusters using either an agglomeration "bottom up" or divisive "top down" approach, after that Canopy algorithm pre-processing data, where Dirichlet Process algorithm Performs Bayesian mixture modeling and Fuzzy K-Means algorithm is use to discovers soft clusters where a particular point can belong to more than one cluster and at last Latent Dirichlet Allocation automatically and jointly cluster words into "topics" and documents into mixtures of topics.[21].

3) *Classification*, it learns from existing distributions and then assign unclassified items to the different categories. Bayesian and Random Forests are the algorithms used in classification. Bayesian algorithm is used to classify objects into binary categories and Random Forests is an ensemble learning method for classification. It is achieve by building a multitude of decision trees [21].

4) *Frequent item set mining*, this machine learning technique analyzes the items and group them, then identifies which items consistently appear together [24]. It uses Parallel FP Growth Algorithm to group and analyze. [21].

2.2 Apache Hadoop

For the distributed processing of large data groups across clusters of computers, we use Apache Hadoop software library that is a framework in java. [20]. Doug Cutting and Mike Cafarella started this Open Source Project called HADOOP in 2005 and Daug named it after his son's toy elephant [19]. Now it is a registered trademark of the Apache Software Foundation. Apache Hadoop runs applications using the mapreduce algorithm [2].

HDFS is a distributed file system running on a cluster of common components, and the main storage system of the HADOOP framework [31]. For research and production almost all the type of companies and organizations use Hadoop. The latest version of Hadoop-Apache is Hadoop 2.7.1, that was released on 6 July 2015 [19].

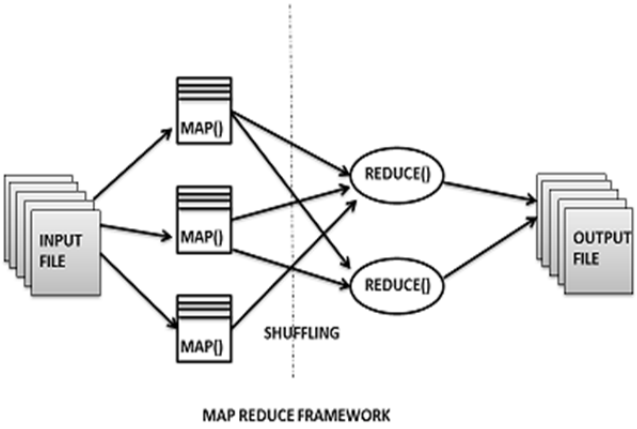

Fig. 2

2.3 Hadoop MapReduce

Hadoop MapReduce is a software framework, it process on large clusters of huge data in-parallel algorithm which is reliable and fault-tolerant. [29] MapReduce database commands and operations are performed with the help of mongodb [28]. In this map-reduce operation, mongodb handles the map phase to each input file and the Key-values pairs are emitted by map function and f or multiple values, mongodb implements the reduce phase, where the aggregated data collects and condenses then the results is stored in a collection by mongodb. And if needed the output of the resultant function go through to the final function to further condense or to process the results of the aggregated data. [28].

3. HADOOP AND CLOUD COMPUTING

Hadoop is the software framework for writing applications that rapidly process large amount of data in parallel on large clusters of compute nodes [4]. It provides a distributed file system and a framework for the analysis. It transform a huge data sets with the help of the MapReduce paradigm. The

data-sets are volume of data that is generated by the application is exponential. So, it needs an efficient processing for these large data-sets [11]. Hadoop framework helps to do this. It is capable enough to develop applications that runs on clusters of computers or large data-sets that perform statistical analysis for these exponential amount of data [34].

Hadoop framework includes four modules: Hadoop Common, Hadoop Distributed File System (HDFS™), Hadoop YARN and Hadoop MapReduce. The Hadoop modules supported by common utilities. High-throughput access to application data are provided by HDFS™. Hadoop YARN is a framework for job scheduling and cluster resource management and Hadoop MapReduce is a YARN-based system for parallel processing of large data sets [19].

As it has become indisputable that Hadoop in the cloud is a leading topic. These are some reason why this association makes sense:

Hadoop in the cloud requires no upfront investment in the on-site hardware or its support. Unlike static, on premise clusters Hadoop clusters in the cloud scale up or down as per data processing requirement. Nodes can be removed or added as per requirement which makes the combined use of cloud computing and Hadoop highly reliable and easy to manage. The cloud, with its pay as you use, model, is more efficient to handle batch workloads. Companies can schedule cloud based clusters to be available only at the period of time, during the day when the data needs to be crunched [30].

Hadoop can be used to implement MapReduce directly on the cloud storage. Hadoop on cloud platform also provide connectors that enable direct access to cloud storage [32]. For example *Google Cloud Storage Connector* for Hadoop provide direct data access with no need to transfer it into HDFS first [33].

The integrated use of data mining and cloud computing will be time reducing, higher availability and cost effective because data mining can be done directly on cloud storages and there is no need of fetching data again and again from HDFS.

4. CONCLUSION

Data mining is essential due to enormous increase in data. Speed of data mining is also an important aspect. This paper focuses on concept of using Hadoop framework on Cloud. This is beneficial since it will speed up data processing, lower the investment cost, provide business flexibility and save time and storage space also. Therefore Hadoop on cloud has a long term future scope.

Applications for processing data on cloud at various organization can be developed in future to make work faster, easier and more flexible.

REFERENCES

[1] Tingting Hu , Haishan Chen , Xiaodan Zhu and Lu Huang,,"A Survey of Mass Data Mining Based on Cloud Computing"

[2] Leila Ismail, Mohammad M. Masud and Latifur Khan, "FSBD: A Framework for Scheduling of Big Data Mining in Cloud Computing".

[3] Tao Chen, Jidong Chen and Baoyao Zhou, "A System for Parallel data Mining Service on Cloud".

[4] Deepti Mittal, Damandeep Kaur and Ashish Aggarwal, "Security data mining in Cloud using Homomorphic Encryption".

[5] Viki Patil and Prof. V. B. Nikam, "Study of Data Mining Algorithm in Cloud Computing using MapReduce Framework".

[6] Jianzong Wang, Jiguang Wan, Zhuo Liu and Peng Wang, "Data Mining of Mass Storage based on Cloud Computing".

[7] Hamid Malmir , Fardad Farokhi , Reza Sabbaghi-Nadooshan, "Optimization of Data Mining with Evolutionary Algorithms for Cloud Computing Application".

[8] LI Xiao-Feng, WANG Jian-Hua and GAO Wei-Wei, "Examination System in the Cloud Computing Platform based on Data Mining".

[9] Hongbo Yu, Yihua Lan, Xingang Zhang2, Zhidu Liu, Chao Yin and Changlin Long, "Research Of Data Mining In Cloud Environment".

[10] Inderjit Kaur and Deep Mann, "International Journal of Advanced Research in Computer Science and Software Engineering".

[11] Ruxandra-Ştefania PETRE, "Data mining in Cloud Computing".

[12] Mrs. Mishra Monika R. and Naskar Ankita "Using

[13] Cloud Computing to Provide Data Mining Services".Robert Vrbić "Data Mining and Cloud Computing".

[14] Xia Geng and Zhi Yang, "Data Mining in Cloud Computing".

[15] Anuja R. Yeole and Poonam Borkar, "Survey paper on Data Mining in Cloud Computing".

[16] Juan Li, Pallavi Roy, Samee U. Khan, Lizhe Wang and Yan Bai, "Data Mining Using Clouds: An Experimental Implementation of Apriori over MapReduce".

[17] International Journal of Computer Science and Information Technologies, Vol. 5 (2) , 2014, 2091-2094 www.ijcsit.com

[18] CH.Sekhar and S Reshma Anjum, "Cloud Data Mining based on Association Rule".

[19] Apache Hadoop .http://hadoop.apache.org/. [Accessed Sept 22nd, 2015]

[20] Hadoop introduction .http://www.tutorialspoint.com/hadoop/hadoop_introd uction.ht. [Accessed Sept 22nd, 2015]

[21] Mahout Algorithms. http://hortonworks.com/hadoop/mahout/#section_2.[Accessed Sept 21st, 2015]

[22] Machine Learning .http://www.tutorialspoint.com/mahout/mahout_machi ne_learning.htm.[Accessed Sept 21st, 2015]

[23] Apache Mahout: scalable machine learning and data mining .https://mahout.apache.org/general/downloads.html [Accessed Sept 22nd, 2015]

[24] Apache Mahout.http://hortonworks.com/hadoop/mahout/[Accessed Sept 21st, 2015]

[25] Mahout data mining.http://www.tutorialspoint.com/mahout/mahout_introduction.htm.[Accessed Sept 22nd, 2015]

[26] Recommendation.http://www.slideshare.net/Cataldo/apache-mahout-tutorial-recommendation-20132014.[Accessed Sept 19th, 2015]

[27] Cloud computing history.http://www.slideshare.net/beeniew/history-and-evolution-of-cloud-computing-safaricom-cloud?related=1[Accessed Sept,19th,2015].

[28] Map-Reduce. http://docs.mongodb.org/manual/core/map-reduce/[Accessed Sept 22nd, 2015]

[29] Map-Reduce overview. http://hadoop.apache.org/docs/current/hadoop-mapreduce-client/hadoop-mapreduce-client-core /MapReduceTutorial.html [Accessed Sept 21st, 2015].

[30] Big Data Analytics Use Cases. http:/www.qubole.com/resources/solution/best-use-cases-for-big-data-analytics/ [Accessed Sept 22nd, 2015]

[31] HDFS, http://fullforms.com/HDFS[Accessed Sept 18st, 2015] Hadoop on Google Cloud Platform.

[32] https://cloud.google.com/hadoop/what-is-hadoop?hl=en [Accessed Sept 23rd, 2105]

[33] Google Cloud Storage Connector.https://cloud.google.com/hadoop/google- Cloud-storage-connector. [Accessed Sept 23rd, 2105]

[34] Hadoop on cloud .http://www.techrepublic.com/article/hadoop-and-cloud-computing-collision-course-or-happy-symbiosis/ [Accessed Sept 23rd,2105]

[35] Virtual Hadoop. http://www.qubole.com/hadoop-as-a-service/ [Accessed Sept 23rd,2105]