International Journal of Technical Research and Applications e-ISSN: 2320-8163,
www.ijtra.com Volume 4, Issue 3 (May-June, 2016), PP. 276-278

COMPARISON OF ALGORITHMS FOR CHOOSING THE BEST ONE FOR THE FREQUENT ITEM SET MINING TECHNIQUES

Manu Mohan. P[1], Prof. Rasheeda Z Khan[2]
[1]MTECH in Computer Science and Engineering
[2]HOD , Department of Information Science and Engg
Shree Devi Institute Of Technology, Kenjar, Mangalore, Karnataka.
[1]manumohan234@gmail.com
[2]rasheeda_nr@yahoo.co.in

Abstract— **Massive amount of data is stored and transferred from the tremendous number of sources like sensor devices, mobile devices, social media networks, network operators, internet applications etc and those data are called as Big data. This big data is a set of structured and unstructured data as it is coming from various kinds of sources and will include text files, audio files, video clips, images, and even graphs and charts. So the management of big data is an important stage in the development of all kind of business fields. Big data management is not done using conventional tools and software techniques. The big data management is essential as it needs efficient techniques and the result will provide better insights about the stored data. There are many algorithms used for big data analysis. But the traditional methods need the entire data to be in main memory. But it is not possible to get all the data to be in main memory. Association rules and frequent itemset mining are the common techniques used for the big data management. To handle this drawback new Hadoop Mapreduce framework is used which has scalability and robustness features to manage big data sets. A new algorithm called clustBigFIM algorithm which is a modified bigFIM algorithm which makes use of Apriori algorithm and éclat algorithm for finding extensions had been implemented in HadoopMapReduce paradigm. The problem with hadoop mapreduce is that it stores the intermediate results in local discs. So it will become necessary to retrieve these data from the intermediate discs for further use and hence it will take time to access. This will lead to high latency problem. Spark gives a sequential execution model which leads to an in memory computational mechanism and querying data will be much faster than the disc based methods like MapReduce. So the paper mainly points out the advantages of spark framework to use clustBigFIM algorithm to enhance the speed of process and get better efficiency.**

Index terms- **Association Rule Mining, Big Data, Clustering, Frequent Itemset Mining, Hadoop, MapReduce. ,**

I. INTRODUCTION

World is now lead by massive amount of data that have collected and stored from tremendous types of devices and sources. These data will be stored and should be used for future works. The storage of such a big amount of data is not an easy task. It needs a well-controlled process to retrieve and process on it. The data fetched from mobile operator providers, sensing networks, business organizations, social media networks etc are collected and stored in a huge database. These data are structured as well as unstructured data and their processing is difficult using conventional tools and software techniques. Processing of big data is very important for many business firms, enterprise or organizations as it contains many hidden values and behaviors to make strategic planning for future.

Data mining is an important process of extracting the hidden values and special patterns of big data using some special techniques and algorithms. It provides the data a meaningful relationship with other information. Data mining can give the user and the administrator of a business firm, what precisely going on over the current trends of the world. So it is very helpful for the technical and non-technical users to understand about the business and get better answers. This allows their companies to make billions of dollars. Data Mining is a concept that is taking off in the commercial sector as a means of finding useful information out of gigabytes of data.

II. PROBLEM DESCRIPTION

The important technics od data set mining are frequent item set mining and association rule mining. Processing large data sets is challenging as it requires intensive calculation. So using traditional tools and techniques is not practical. So the new techniques could provide a better performance on big data mining. But these techniques need all the data inside the main memory, but it is not possible to store the entire data in main memory. Hadoop MapReduce overcomes the stated problem but it has latency problem. A new framework called Apache Spark overcomes the latency problem and provides a better platform for big data management using sequential query optimization techniques. When it explains about the business field the important term is profit improvement and depends on customer satisfaction. This depends mainly on understanding the behavior of customers and understanding their taste of item purchase. So that the items can be shelved in such a way that customers will find it easy for them to understand the item scheduling and do transaction. The online shopping websites

use the theoretical approach to improve their business. To know about the frequency of items purchased together, they use association rule and the minimum support values are used to find the transactional behavior of most sold items and item sets. The set of items purchased more together are shelved together for the easiness of customers and it is a successful term used practically in the business fields. So better and faster algorithms to find out the item sets are preferred by business firms. The comparison of algorithms are depicted in this paper.

III. METHODOLOGY

Here the existing system is given with a set of retail data as the input. The Apriori algorithm is allowed to compile first, then the packages for the algorithm have been created, and it is made to run. Now the process will start its execution within no time and the algorithm might check each and every data inputs to produce the output. The processing of this algorithm is interesting as it will be done in a three step processing. In the first step the data input is scanned from top to bottom and the input is ready to be worked with the data mining process. In the second step, all the values are started extracting from its structure whether it is structured or unstructured. The third step is very important as it is the step where the scanning is completed and a tidlist is generated.

But in the case of FP-Growth algorithm it is based on parallel execution that is the input data will be stored in parallel commodity clusters and the process will start running on them parallelly and simultaneously. The number of clusters connected in parallel can be increased or reduced according to the size of the input file. The performance of Apriori and FP-Growth are evaluated and compared using the time they have taken to complete the execution. Also the execution time has been measured for various numbers of instances and confidences the input retail data set.

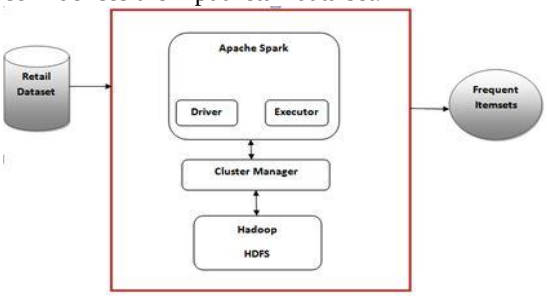
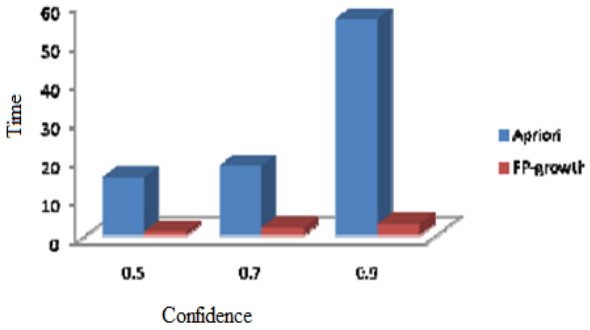


Figure1: System architecture

IV. PERFORMANCE EVALUATION

Confidences	Time of execution (in seconds)	
	Apriori	FPGrowth
0.3	17	0.3
0.6	19	0.6
0.9	60	0.9

The system is compared by taking different confidence level values of the system performance. The confidence levels are 0.3, 0.6 and 0.9. the performance is compared for both Apriori and FPGrowth algorithms. The Apriori showed slow execution timings than FPGrowth algorithm. So it is clear tghat FPGrowth is much good in the performance efficiency for big data management. Also a graph is plotted against the values obtained in the performanvce analysis. The values are plotted as a graph for confidence level values against time of execution.



It is clear that for any confidence level of any test case values, FPGrowth algorithm is efficient for the big data management process.

V. CONCLUSION

The big data management tools are very complex and difficult to understand and process. Since they play an important role in many fields of business industry their performance is of infinite importance. The enhancements in performance is characterized by the speed with which the huge datasets are processed and the result is stored in the intermediate discs. The clustBigFIM algorithm which is a new version of BigFIM algorithm shows better performance in a new sequential query optimization platform called Apache Spark by running tha FPGrowth algorithm on it. The characteristics of Apache Spark is giving efficiency improvements in the big data management and the project mainly focused on using the combination of Apache Spark framework and the FPGrowth algorithm for a better performance on massive data to get the in memory computed input data.

.

VI. FUTURE ENHANCEMENTS

The project will work with the big data management techniques and algorithms as it includes the comparison of two techniques. Algorithms which take too much space and time are replaced by new algorithms with fast in memory computation. The new methods are capable of processing even large amount of data with much efficient speed of tools and accurate outputs. Data are retrievable with all the algorithms but the hidden values and patterns are retrievable only through the new techniques which will make a big impact on the business field. It will drastically effect the improvements in the business level assuring better increase in profit.

REFERENCES

[1] J. Dean and S. Ghemawat. MapReduce: Simplified data processing on large clusters. In Proc. OSDI. USENIX Association, 2004.

[2] Moens, S.; Aksehirli, E.; Goethals, B., "Frequent Itemset Mining for Big Data," Big Data, 2013 IEEE International Conference on , vol., no., pp.111,118, 6-9 Oct. 2013 doi: 10.1109/BigData.2013.6691742

[3] Weizhong Zhao, Huifang Ma, and Qing He. 2009. Parallel K-Means Clustering Based on MapReduce. In Proceedings of the 1st International Conference on Cloud Computing (CloudCom '09), Springer-Verlag, Berlin, Heidelberg, 674-679.

[4] R. Agrawal and R. Srikant. Fast algorithms for mining association rules in large databases. In Proc. VLDB, pages 487–499, 1994.

[5] M. J. Zaki and K. Gouda. Fast vertical mining using diffsets. In Proc. ACM SIGKDD, pages 326–335, 2003.

[6]Zahra Farzanyar and Nick Cercone. 2013. Efficient mining of frequent itemsets in social network data based on MapReduce framework. In Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM '13).

[7] J Manyika, M Chui, B Brown, J Bughin, R Dobbs, C Roxburgh, AH Byers. Big data: The next frontier for innovation, competition, and productivity. McKinsey Global Institute, 1-137, 2011.

[8] Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth. 1996. The KDD process for extracting useful knowledge from volumes of data. Commun. ACM 39, 11 (November 1996), 27-34.