ISSN: 2350-0328

International Journal of Advanced Research in Science, Engineering and Technology

Vol. 4, Issue 3 , March 2017

Frequent Item Sets Mining Using MapReduce Technique

Megana S S , Uma N

P.G. Student, Department of Computer Science, New Horizon College Of Engineering, Bangalore, Karnataka, India

Sr.Assistant Professor, Department of Computer Science, New Horizon College Of Engineering, Bangalore,India

ABSTRACT: Parallel mining algorithms lacks some of the features like parallelization of sequential code, balancing load over cluster of computers, and distribution of data over computers in a cluster. To overcome these problems we have introduced the MapReduce technique for parallel frequent item sets mining. It uses the ultra-metric tree pattern for storage of data, it does not use the FP (frequent pattern) trees like existing ones uses it. Mining task is completed with the help of MapReduce technique. This technique incorporates three MapReduce job to mine the large amount of data conventionally and economically. In first MapReduce job, all frequent item sets are discovered. Secondly, it removes infrequent item sets. In Third step, it constructs small ultra-metric trees which are helpful to mine the frequent data conventionally and economically. It is implemented in Hadoop cluster.

I.INTRODUCTION

Currently there is an increase in volume of data in both Scientific and commercial domains. Knowledge Discovery & Data mining (KDD) both have been an important inference process. Knowledge discovery process provides a lot of useful information in all the fields of business, science, medicine etc. For example, in field of retail business, identifying customer purchasing patterns and customer groups provide wealth of information to improve the organization. When considering the availability of massive volume of data, analyzing and decision making is still a major issue.

Data mining is a techniqueof discovering the patterns from the huge amount of data. The technique of association rules discovering is one of the most known and the most explored techniques of data mining. Rules discovering is one of the popular and most explored technique in the field of data mining. Rules Discovering has two main phase. The first phase is the most expensive given the large number of accesses to transaction database and large number of candidate item sets. Since databases are generally very large, a solution to avoid the repetitive and costly accesses is to represent them by compact structures. In this paper, we are proposing a parallel binary approach for frequent item sets extracting, to deal with the great number of candidates and to take advantage of multicore architectures. We have implemented using compact data structure technique based on signature tree for the representations of the database to access it only once.
There are many data mining techniques like clustering, classification and association rule. The most popular one is the association rule that is divided into two parts i) generating the frequent itemset ii) generating association rule from all itemsets Therefore, most research is now directed towards the designing of more intelligent and autonomous information retrieval systems, known as Recommendation Systems.

II.IMPLEMENTATION

Number of Components: After careful analysis, the system has been identified to have the following components:

A. Installation of Hadoop using Horton Sandbox

B. Putty

C. Win SCP

D. MapReduce

ISSN: 2350-0328

International Journal of Advanced Research in Science, Engineering and Technology

Vol. 4, Issue 3 , March 2017

A. Installation of Hadoop

Hadoop is an open source project from Apache that has evolved rapidly into a major technology movement. It has emerged as the best way to handle massive amounts of data, including not only structured data but also complex, unstructured data as well. The Hadoop platform consists of two key services namely a reliable, distributed file system called Hadoop Distributed File System (HDFS)and the high-performance parallel data processing engine called Hadoop MapReduce.Hadoop Clusters a special type of computational cluster designed for storing and analyzing vast amount of unstructured data in a distributed computing environment. These clusters run on low cost commodity computers.

Hortonworks Sandbox is a personal, portable Apache Hadoop and its ecosystem environment comes with dozens of interactive tutorials and the most exciting developments from the Apache community. Numerous varieties of open source projects have been integrated, tested and combined as part of the Hortonworks Data Platform (HDP). The platform components comprising the Hortonworks Data Platform (HDP) are released under the Apache 2.0 License. HDP is also commonly used with 3rd-Party Components (ex. Oracle's JDK – Java Platform) and Optional Add-Ons (ex. Hive ODBC Driver).HDP enables enterprises to deploy, integrate and work with unprecedented volumes of structured and unstructured data.

B.PuTTY

PuTTY is a client program for the SSH, Telnet and Rlogin network protocols. These protocols are all used to run a remote session on a computer, over a network.

C.WinSCP

WinSCP is an open source free FTP and SFTP client for Windows. Legacy SCP protocol is also supported. Its main function is safe copying of files between a local and a remote computer.

D. MapReduce

MapReduce is a processing technique and a program model for distributed computing based on java. The MapReduce algorithm contains two important tasks, namely Map and Reduce. MapReduce algorithms has two important tasks, namely Map and Reduce. Map takes a set of data and converts it into another set of data, where individual elements are broken down into tuples (key/value pairs). Secondly, reduce task, which takes the output from a map as an input and combines those data tuples into a smaller set of tuples. As the sequence of the name MapReduce implies, the reduce task is always performed after the map job.

The major advantage of MapReduce is that it is easy to scale data processing over multiple computing nodes. Scale data processing over multiple computing nodes is the major advantage of Map Reduce.

Under the MapReduce model, the data processing primitives are called mappers and reducers. Mappers and reducers are the two data processing primitives in MapReduce model.

Decomposing a data processing application into mappers and reducers is sometimes nontrivial. But, once we write an application in the MapReduce form, scaling the application to run over hundreds, thousands, or even tens of thousands of machines in a cluster is merely a configuration change. This simple scalability is one of the main advantage of MapReduce which has attracted many programmers to use the MapReduce model.

Map stage: The map or mapper's job is to process the input data. Generally the input data is in the form of file or directory and is stored in the Hadoop file system (HDFS). The input file is passed to the mapper function line by line. The mapper processes the data and creates several small chunks of data.

ISSN: 2350-0328
International Journal of Advanced Research in Science, Engineering and Technology
Vol. 4, Issue 3 , March 2017

Reduce stage: This stage is the combination of the Shuffle stage and the Reduce stage. The Reducer's job is to process the data that comes from the mapper. After processing, it produces a new set of output, which will be stored in HDFS.

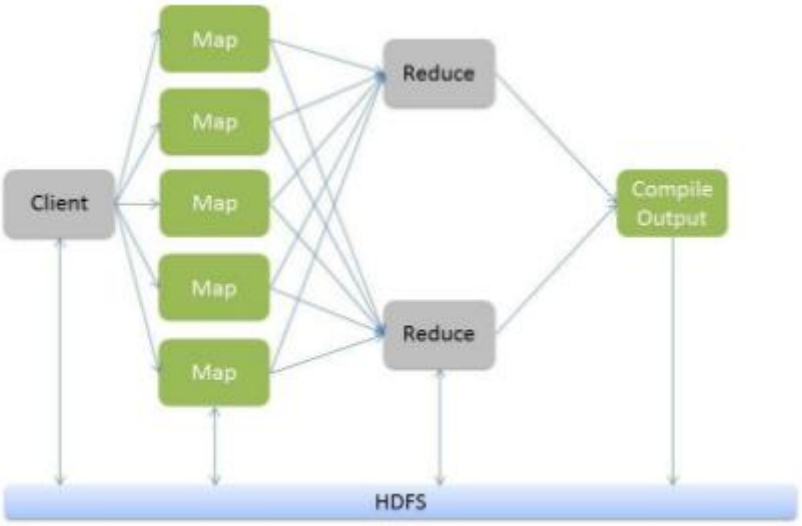

Fig 1: System Architecture

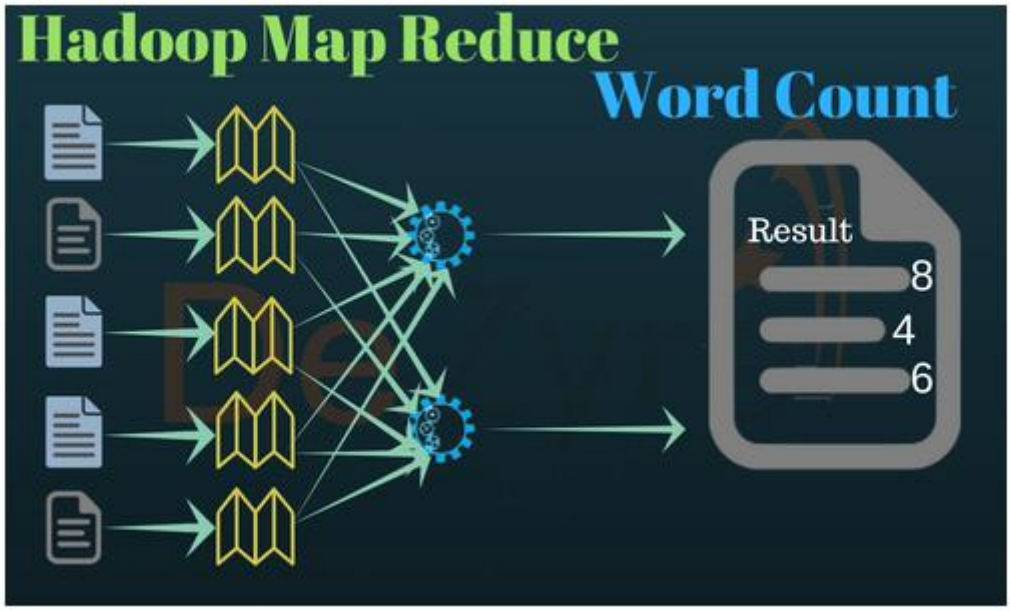

Fig 2:Map Reduce Technique to output the word count.

ISSN: 2350-0328

International Journal of Advanced Research in Science, Engineering and Technology

Vol. 4, Issue 3 , March 2017

III.CONCLUSION

Mapreduce programming model is applied for existing parallel mining algorithm for mining frequent itemsets from database and solves the load balancing and scalability. This paper gives the overview designed for parallel mining of frequent itemsets.The Apriori and FP tree algorithm were used for mining frequent item sets. Main drawback of Apriori algorithm is that the database has to be scanned many number of times and huge candidate keys needs to be exchanged between the processor. I/O and synchronization are the other problems in the Apriori algorithm. The disadvantage of FP-growth, however, lies within the impracticableness to construct in-memory FP trees to accommodate large-scale databases. This drawback becomes a lot of pronounced once it comes to huge and two-dimensional databases. To overcome these problems,this technique employs parallel frequent itemset mining. It incorporates the ultra metrictree rather than Apriori or FP-growth algorithm.

.

REFERENCES

[1] “Parallel Mining of Association rule.” Rakesh AgarwalJohn C Safer.
[2] H. D. K. Moonesinghe, Moon-Jung Chung, Pang-Ning Tan. Fast Parallel Mining of Frequent Item sets.
[3] R. Agrawal and J.C. Shafer, "Parallel Mining of Association Rules,"IEEE Tran. Knowledge and Data Eng.”.
[4] MapReduce Tutorial http://pages.cs.wisc.edu/~gibson/mapReduceTutorial.html.
[5] Han Jiawei, KamberMiceline. Fan Ming, MengXiaofeng translation, “Data mining concepts and technologies”. Beijing: Machinery Industry Press. 2001..
[6] Y. Ye and C. Chiang, “A parallel apriori algorithm for itemsets mining,” in Proc. 4th International Conference on Software Engineering Research, Management and Applications, Washington, DC, USA, 2006, pp. 87-94.
[7] J. Han, J. Pei, and Y. Yin, “Mining frequent patterns without candidate generation,” in Proc. of ACM SIGMOD International Conference on Management of Data, Dallas, USA, 2000, pp. 1-12.
[8] “FiDoop: Parallel Mining of Frequent Itemsets Using MapReduce” Yaling Xun, Jifu Zhang, and Xiao Qin, Senior Member, IEEE
[9] “Implementation of Parallel Apriori Algorithm On Hadoop Cluster” A. Ezhilvathani1, Dr. K. Raja. International Journal of Computer Science and Mobile Computing.
[10] “Frequent Itemset Mining for Big Data Sandy Moens, Emin Aksehirli and Bart Goethals UniversitiesAntwerpian, Belgium.