www.ijecs.in

International Journal Of Engineering And Computer Science ISSN:2319-7242
Volume 4 Issue 6 June 2015, Page No. 12539-12543

Study of Enhancing ARM Algorithms with Map Reduce Framework

***K.Mangayarkkarasi* [1] *, M.Chidambaram*[2]**

[1]Research Scholar,Research and Development Centre,

Bharathiar University,Coimbatore

[2]Assistant Professor,Rajah Serfoji College,

Bharathidasan University,Tanjavur

Tamil Nadu –India

[1]kanthimangai@gmail.com

[2]chidsuba@gmail.com

Abstract: ***This paper studies the requirements to add efficiency in data mining. As the volume of data increases in tera bytes and that too remains distributed in nature , the data mining algorithms need to be adoptive to this trend. Scaling is one essential issue to be considered. Cloud gives a virtual platform for parallel and distributed processing. This paper analyses the scope for adding efficiency to association rule mining algorithms with cloud technology.In the present scenario as data are distributed in nature to mine them efficiently algorithms with distributed and parallel in nature are needed. Cloud environment supports distributed and parallel platform. Data privacy is still a very big problem to be addressed in cloud environment. This paper throws light on many encouraging aspects of cloud for data mining. Also it gives a study on adaptation of cloud environment for association rule mining.***

Keywords—Association Rule Mining, Distributed data, Parallel processing, Scaling, Hadoop, Map Reduce

1. INTRODUCTION

Data mining is the process of analysing vast amount of data to find useful relations or patterns. Today organizations have huge amount of data store with them. Data volume is getting increased in terms of tera bytes. Data dimension is increasing and also the nature of data is dynamic. But there are no efficient and feasible solution to analyse and get valuable information hidden inside. Even though data are stored in volumes unless processed, effective and useful information or knowledge can not be extracted from them .Many number of data mining algorithms have been developed and implemented on various platforms. Current developments and advances in many growing areas in engineering ,science ,business, etc., are producing tremendous amount of data every day which causes high requirement of storage. Basically big data mining systems were rare and expensive because scaling a system to process large datasets is very difficult. It has been limited to the processing power that can be built into a single computer.

Fundamentally there are two approaches to scaling a system as the size of the data increases, and are referred to as scale-up and scale-out. In most enterprises, data processing is performed on large computers. As the size of the data grows, they need to move to larger server or storage array. The advantage of simple scale-up is that the architecture does not change through the growth. They need to use larger components but basic relationship stays the same. The software handles the complexities. There are practical limits on how big a single host can be, so at some stage scale-up can not be extended further. Instead of growing a system onto larger and larger hardware ,the scale-out approach spreads the processing onto more and more machines. If the data set doubles use two servers instead of single double-sized one. If it doubles again, move to four hosts. The benefit of this approach is that purchase costs remain much lower than for scale-up. Server hardware costs tend to increase sharply when one seeks to purchase larger machine. The other side is that strategies are to be developed for splitting the data processing across a fleet of servers.

Hence deploying a scale-out solution needs significant engineering effort. And the system developer needs to design mechanisms for data partitioning and reassembly. Again scheduling the work across the cluster and managing machine failures are great challenges.

The extraction of useful information from huge storage of data by applying datamining techniques is influenced by two factors. One is how much data is accessed and processed and the second factor is the data mining algorithm used to extract information. If the algorithm is applied on partial data the result of the mining will not be complete and meaningful. Hence if the huge amount of data is segmented and assigned to different cloud providers the data privacy gets improved. As each provider possesses partial data even if data mining techniques are applied over the data the information extracted will not reveal any accurate information. The idea is to categorize user data, split them into chunks and provide these chunks to the various cloud providers. For this categorization,fragmentation and distribution of data are to be done. According to the need of privacy level required data are categorized . Data which demand high level of security are assigned to trust worthy providers. As they are very sensitive to the organization they should not be left to be misused. Fragmentation and distribution of data helps to minimize the amount of data among individual providers. This ensures higher level of security. The data chunks are distributed to cloud providers according to their reliability level. The reputation of a cloud provider stands as the measure for their reliability. Due to distribution of data the attackers are prevented from access to vast data. Hence by adopting distributed architecture data mining based security risks can be solved to a great extent.

Cloud computing applies a virtualized platform with elastic resources on demand by provisioning hardware, software and datasets dynamically. The idea is to move desktop computing to a service-oriented platform using server clusters and huge databases at data centers. Cloud computing leverages its low cost and simplicity to benefit both users and providers.. Cloud computing intends to satisfy many user applications simultaneously. A cloud is a pool of virtualized computer resources. A cloud can host a variety of different workloads, including batch-style backend jobs and interactive and user-facing applications. Cloud providers use data mining to provide clients a better service. If clients are unaware of the information being collected, privacy problem arises. Misuse of data by the providers is the serious issue. Client privacy is a tentative issue .Clients can have various levels of privacy demands and according to that they may assign their data to trust worthy providers.

The cloud environment consists of some physical or virtual platforms. Virtual platforms have unique capabilities to provide isolated environments for different applications and users. With large datasets cloud data storage services provide large disk capacity and the service interfaces that allow users to put and get data. The distributed file system offers massive data storage service. It can provide similar interfaces as local file systems. With the underlying concepts of virtualization, clustering and parallel and distributed processing, cloud environments enhance efficiency for data mining.

This paper discusses on Association Rule Mining with cloud Technology. It throws light on the promising aspects, for enhancing efficiency of association rule mining algorithm with the adaptation of cloud technology. This paper is organized as follows. Section 2 gives the background work on ARM algorithms Section 3 gives the programming models in cloud environment section4 gives overview of adaptation of data mining algorithm in cloud environment and Section 5 concludes the paper.

2.DATA MINING ALGORITHMS

2.1Background work

Number of algorithms and techniques are developed for data mining. One such popular and important technique is association rule mining, that is used to find out the relationship or association between various items. This problem of finding relation between items is known as market basket analysis. Here the presence of items within basket is identified so as to assess the buying pattern of customers. This technique is used in inventory management, product layout, etc., [1]

The efficiency of association rule mining is majorly influenced by finding the frequent itemsets. This requires multiple passes through the database. In general the ARM algorithms aim at reducing the number of passess by generating a candidate set which should turn out to be frequent sets. Various algorithms are designed to find out the association rules. The algorithms differ on the basis of how they handle candidate sets and how they reduce number of scans on the database.

There are two important factors related to association rules: Support and Confidence. The support of an item or the set of items is the percentage of transactions in which that item occurs. The confidence measures the strength of the rule and is defined as the ratio of the number of transactions that contain X or Y to the number of transactions that contain X .[2]. The two thresholds namely minimal support and minimal confidence is set to find out reasonable support and confidence.

2.2 Sequential Vs Parallel AR mining.

The association rule mining is usually carried out in two steps. In the first step those items from the database are found out which exceed the predefined threshold. Such items are stated as frequent items or big items . In the second step the association rules are generated out of frequent items found in the first step.

The Apriori algorithm is the most general and widely used algorithm [3] .It uses an iterative method called layer search to generate (k + 1) item sets from k item sets. The concept of Apriori and AprioriTid was given by 1994 Agrawal et al. Then various algorithms Dynamic Hashing and Pruning (DHP) algorithm by Jong Soo Park and colleagues, Partition algorithm by Ashok Savasere and others, Sequential Efficient Association Rules algorithm by Andreas Mueller(SEAR), SPEAR with partition technique, Dynamic ItemsetCounting (DIC) algorithm by Sergey Brin and others ,and a completely different Equivalence class-based algorithm by Mohamed J.zaki and others,Eclat and MaxClique are developed with variations in data structure, data layout and data scaning .

With large amount of data and with the advent of parallel computing technology various association mining algorithms like count distribution algorithm, data distribution algorithm , and improved Apriori algorithms have been proposed[4].The parallelism is obtained in two ways. One by task parallelism and the other by data parallelism. To improve the performance of ARM alogorithm it is necessary to support scaling of massive data set and optimize response time. The message passing between various sites has to be made cost effective. That is communication cost is to be minimized. Also load balancing is one important factor for efficiency.

However most of the ARM algorithms are made for centralized systems, where there are no external communication [5] with the increase in the size of data, the computation time and the memory requirements increase to a great extent[6]. The available ARM algorithms can handle only few thousand dimensions or items. Current ARM algorithms are not suitable to handle terabytes of data. Large scale datasets are logically and physically distributed. Organizations that are geographically distributed need a decentralized approach to ARM. Most algorithms require low data skewness for good load balancing. But a high skew is needed to apply global pruning and cut down the candidate set. To overcome these problems many parallel and distributed algorithms are developed. To accomplish this the concept of dividing the database and then distributing it to different nodes is used.

3. PARALLEL AND DISTRIBUTED PROGRAMMING MODELS

Many models have been proposed for distributed computing with expected scalable performance and application flexibility. Among these models, there are MPI, MapReduce and Hadoop.

3.1 Message-passing Interface

The most popular programming model for message-passing systems is MPI the message passing Interface. It is a library of subprograms that can be called from C or FORTRAN to write parallel programs running on distributed computer systems. Main features specify synchronous or asynchronous point-to-point and collective communication commands and I/O operations in user programs for message passing execution. The idea is to embody clusters, grid systems and P2P systems with upgraded web services and utility computing applications.

3.2 MapReduce

This is a web programming model for scalable data processing on large clusters over large data sets. This is the framework developed by borrowing the “map’ and “Reduce” functions from functional programming paradigm. The model is applied mainly in web-scale search and cloud computing application. The user specifies a Map function to generate a set of intermediate key/value pairs. Then the user applies a Reduce function to merge all intermediate values with the same intermediate key. MapReduce is highly scalable to explore high degrees of parallelism at different job levels. A typical MapReduce computation process can handle terabytes of data on tens of thousands or more client machines. Hundreds of MapReduce programs can be executed simultaneously; in fact, thousands of MapReduce jobs are executed on Google’s clusters every day.

This software framework abstracts the data flow of running a parallel program on a distributed computing system by providing users with two interfaces in the form of two functions: Map and Reduce. Users can override these two functions to interact with and manipulate the data flow of running their programs.

3.3 Hadoop

The Google File System (GFS) and MapReduce are two frameworks developed by Google. Then came the open source of the same two fundamental works as Hadoop. The two components of Hadoop are Hadoop Distributed File System (HDFS) and Map Reduce. These are direct implementation of Google’s own GFS and Map Reduce. HDFS stores files in blocks typically at least 64 MB. It is optimized for throughput over latency, it is very efficient at streaming read requests for many small ones. It is optimized for workloads that are generally of the write-once and read many type. Each storage node runs a process called a DataNode that manages the blocks on that host and these are co-ordinated by a master NameNode running on a separate host. Instead of handling disk failures by having physical redundancies in disk arrays or similar strategies, HDFS uses replication. Each of the blocks comprising a file is stored on multiple nodes within the cluster, and the HDFS Name Node constantly monitors reports sent by each Data Node to ensure that failures have not dropped any block below the desired replication factor. If this does happen, it schedules the addition of another copy within the cluster.

Hadoop provides a standard specification for the map and reduce functions and implementation of these are refered to as mappers and reducers. A typical Map Reduce job will comprise of a number of mappers and reducers. The developer focuses on expressing the transformation between source and result data sets and the Hadoop framework manages all aspects of job execution, parallelism and coordination.

Hadoop uses the Writable interface based classes as the data types for the map reduce computations. These data types are used throughout the map reduce computational flow, starting from reading the input data, transferring intermediate data between Map and Reduce tasks, and finally, when writing the output data. Choosing the appropriate writable data types for input, intermediate and output data can have a large effect on the performance and programmability of the Map Reduce programs. Hadoop’s Writable-based serialization framework provides a more efficient and customized serialization and representation of the data for Map Reduce programs than using the general purpose java’s native serialization framework. Hadoop’s writable framework does not write the type name with each object expecting all the clients of the serialized data to be aware of the types used in the serialized data. Omitting the type names make the serialization process faster and results in compact, random accessible serialized data formats that can be easily interpreted by non-java clients. Hadoop’s writable-based serialization also has the ability to reduce the object-creation overhead by reducing the Writable objects, which is not possible with java’s native serialization framework.

4. ADOPTATION OF ARM ALGORITHM WITH MAP REDUCE IN CLOUD

Various sequential association algorithms are developed using number of data structures and layout. They all attempted to reduce the number of scans of the data base. But the parallel algorithms developed based on data parallelism and processing parallelism proved remarkable improvement in performance. The frequent pattern Tree algorithm and

Count Distribution algorithms are popular parallel association rule mining algorithms.

Map Reduce is a programming model that supports large datasets. The map reduce frame work is based on the functional programming .MapReduce libraries have been written in many languages. This framework is suitable for processing parallelizable problems which need to process data from various sources of data that are huge in size and distributed in nature. Partitioning and reassembling of huge data are handled automatically by this system itself.

Parallel association rule mining algorithms can very well be adopted for cloud environment. Input and output for hadoop needs to be in key,value pairs. Key is the transaction ID or offset of every line and values will be the comma separated list of items in that transaction. A simple two phase and loosely coupled architecture can be used for the implementation of wider range of data mining problem. For each phase in Map Reduce the data should be in the form of key and value pairs so the intermediate key value structure is to be decided. These intermediate key value pairs are passed to the reduce phase at the end of the map phase to extract the frequency count of the itemsets.

The data flow of the Map-Reduce framework will be like fig.i

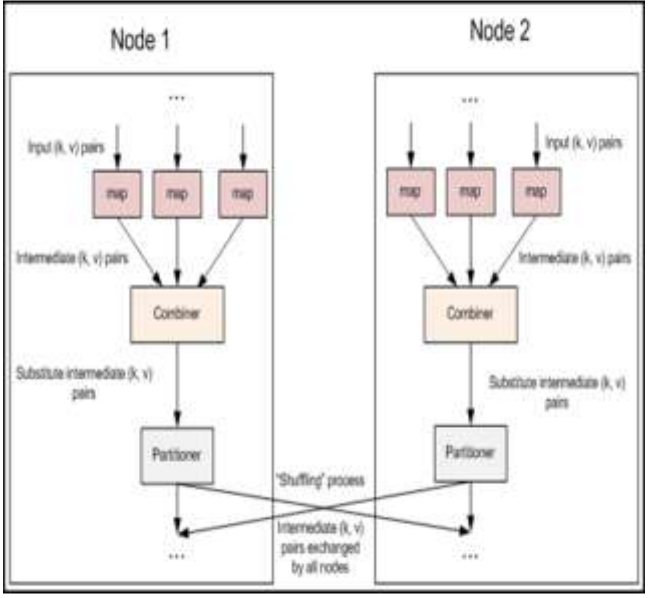
Node 1
Node 2
...
Input (k, v) pairs
map
map
map
Intermediate (k, v) pairs
Combiner
Substitute intermediate (k, v) pairs
Partitioner
'Shuffling' process
Intermediate (k, v) pairs exchanged by all nodes
...

Fig.i Data flow in mapreduce framework

To run a MapReduce job, users should furnish a map function ,a reduce function ,input data and an output data location. When executed Hadoop carries out the following steps.

i. Hadoop breaks the input data into multiple data items by new lines and runs the 'map' function once for each data item, giving the item as the input for the function. When executed, the map function outputs one or more key-value pairs.
ii. Hadoop collects all the key-value pairs generated from the 'map' function, sorts them by key and groups together the values with the same key.
iii. For each distinct key, Hadoop runs the reduce function once while passing the key and list of values for that key as input.
iv. The reduce function may output one or more key-value pairs and Hadoop writes them to a file as the final result.

5.CONCLUSION

The main advantage of Hadoop MapReduce is that it allows the users to easily handle analytical risk over large datasets. Hadoop Distributed File System is a block structured , distributed file system that is designed to run on a low-cost commodity hardware. HDFS supports storing massive amounts of data and provides high throughput access to the data. HDFS stores file data across multiple nodes with redundancy to ensure fault-tolerance and high aggregate bandwidth.It gives satisfied performance in scaling large clusters. It supports distributed data and computation. Computation is performed local to data and tasks are independent hence it can easily handle partial failures. When the node fails it can automatically restart. It has the ability to process the large amount of data in parallel. HDFS has the capability for replicating the files which can easily handle situations like software and hardware failure. In HDFS data can be written only once and it can be read for many times. Map Reduce is a batch-based architecture which means it does not allow itself to use cases that needs real time data access. It is one promising environment for data mining tasks. Association Rule Mining algorithms implemented on cloud environment have high scope to add efficiency.

References.

[1] *Qureshi, Jaya Bansal, Sanjay Bansal, A survey on Association Rule Mining in cloud computing-International Journal of Emerjing Technology and Advanced Engineering, Volume 3, Issue 4, April 2013 P.No 318-321.*

[2] *Juan Li,Pallavi Roy, Samee U. Khan, Lizhe Wang, Yan Bai "Data Mining Using Clouds : An experimental implementation of Apriori over MapReduce"*

[3] *Ling Juan Li,min zhang, "The strategy of mining association rule based on cloud computing" International Conference on business computing and global Information,2011.*

[4] *Lu,Lin Pan Rongsheng Xu,Wenbao Jiang," An improved Apriori based Algorithm for Association Rules Mining" sixth International conference on Fuzzy systems and knowledge discovery,2009 FSKD '09.*

[5] *Mafruz Zamam Ashrafi, david taiar, Kale smith "ODAM :An optimized Distributed Association Rule Mining Algorithm " ,IEEE DISTRIBUTED SYSTEMS ONLINE 1541-4922 @ 2004 Published by the IEEE computer Society, Vol 5,No 3 ;March 2004.*

[6] *Lin.K.W. "A fast parallel algorithm for discovering frequent patterns ", IEEE International conference on Granular Computing, 2009 ,GRC'09.*

[7] *Jiawei Han Micheline Kamber, Data Mining concepts and Techniques,2nd Edition.*

[8] *Mohammed J.Zaki: " Parallel and Distributed Association Mining : A survey" IEEE Concurrency October-December 1999.*

[9] *Viki Patil,Prof.V.B.Nikam "Study of Data Mining Algorithm in cloud computing using MapReduce Framework" JEC & AS ,Volume 2,No 7, July 2013.*

[10] *Srinath Perera,Thilina Gunarathne,"Hadoop MapReduce Cookbook" PACKT Publishing.*

[11] *Garry Turkington,"Hadoop Beginner's Guide" Pckt Publishing.*

[12] *R.Agrawal and R.Srikant. Fast algorithm for mining association rules. In Proc.1994International Conference on very Large DataBases Pages 48-499, Santigo,Chile,September 1994.*

[13] *D.Cheung et al. "A fast Distributed Algorithm for Mining Association Rules,"Proc.4th Int'l Conf.Parallel and Distributed Information Systems,IEEE Computer Soc PressAlamitos,Calif.,1996,pp.31-42.*