A Map Reduce Based Parallel Algorithm

D.O.I - 10.51201/12486

https://doi.org/10.51201/12486

Priyanka Gupta
Computer Engineering
Dwarkadas J Sanghvi College Of Engineering
Mumbai,Marashtra,India

Dr. Vinaya Sawant
Head Of Information Technology
Dwarkadas J Sanghvi College Of Engineering
Mumbai,Marashtra,India

Abstract*—In Today's world, Big Data became a vital role in our way of life. Growing rapidly and logical algorithms that will handle large datasets becoming a challenging task. This paper aim is to explain the Hadoop- Map Reduce Framework with Association rule Mining. One of the famous algorithm of ARM is that the Apriori algorithm, which is implemented in MapReduce artificial language, which is executed on the Hadoop framework. These rules are tradition to observe facts that always occur along with datasets. Using association rule, an issue arises when data becomes enormous. To beat this situation they used Hadoop. This paper provides an outline of the Hadoop – MapReduce framework. Our big challenge of this paper are to resolve the matter of scalability, apart from the other problems are execution time and communication overhead. This paper will solve the matter of reliable, scalable and distributed computing.*

Keywords—Data Mining, Association Rule Mining, Apriori Algorithm, Big Data, Hadoop, MapReduce

I. INTRODUCTION

Data Mining and Association Rule Mining (ARM)

In this era, the flexibility to gather this massive data has increased thanks to advances in software platforms. For instance, Wal-Mart alone can handle large data, customer transaction each hour, and imports those into databases. With this extremely huge data set, it should exist difficult for one device to process and association pattern rules between data set. Data processing is defined because of the process of finding a hidden pattern in an exceeding database. The most focus of information mining is to power the information into knowledge. Association rule mining could be quite a data processing process.

Apache Hadoop Framework

Hadoop is a large-scale distributed execution framework for data processing of massive amounts of data on a group of computers. Hadoop makes the use of the Hadoop Distributed File System (HDFS). Hadoop responsibility for assigning the data over multiple nodes and the number of duplications. It provides the map and reduce functions via implementations of interfaces, abstract classes. Hadoop had three modes of operation: Standalone mode Pseudo-Distributed mode and Fully Distributed mode. Because the HDFS is employed, benefits could also be gained from it.

Map Reduce Programming Model

Map Reduce model is a parallel programming model. MapReduce model is used for data processing for the huge size of data by splitting the tasks into independent tasks over the number of items. They used two idioms: map and reduce. It also includes two functions such as mapper and reducer. Both functions are in form of (key, value) pairs. The mapper function work as it takes input (k1, v1) pairs and creates the lists of intermediary (k2, v2) pairs. The output pairs are sorted and grouped on the identical key and computing to combiner to form a sum. The output which produces the intermediary pairs is shuffled and exchanged between the number of items to all the pairs with the same key to one reducer. The communication step is handle by the Hadoop MapReduce platform. The reducer function work as it takes (k2, v2) value as input to make a sum of values to produce new pairs (k3, v3).

II. LITERATURE SURVEY

Srinivasa R, and V Sucharitha [1] analyze the present Apriori algorithm, but it gives a several issues. So that they improve the Apriori algorithm which is employed for locating the 1- itemset which is frequent in the dataset. Also, they used the advance Apriori calculations by the handling of MapReduce Programming. Their productivity provides efficient time and memory by using this technique. They examine the present system of productive approaches to find the itemsets which decrease the time and communication costs through the implementation of parallelism.

Sandhya W, Sanchitta S, Shweta K, and Karishma M [2] implemented an Apriori algorithm, to gather the itemset that happens frequently which relies on the MapReduce programming model. They need to implement and execute the Apriori- MapReduce algorithm. During this paper, they implemented three MapReduce tasks to absolute the mining tasks. All the transactions of databases are stored within the HDFS system. The primary job is to search out all frequent itemset. The second job is to again scan the things to come up with k-items by pruning the infrequent items. The work is to search out the frequency of frequent items.

Marwa B, Ines B and Amel T [3] proposed a distributed manner for mining the common basis of association rule which relies on the MapReduce framework. The algorithm proposed is iterative and it takes the identical as three jobs of MapReduce programming. In this, the input is defined in key-value pairs(k,v). The goal of this paper is to test if the present transactions hold the present candidate and also mine the sole frequent items which reject the pairs, whose value is less than the minimum threshold value.

Seema T, and Bharti V [4] explores an algorithm called Enhanced Apriori Algorithm (EA) which is employed to unravel the matter of the Frequent Itemset Ultrametric Tree(FIUT) algorithm. The most use of this algorithm is to cut back the time that's required for scanning the transactions. This algorithm also uses new data partitioning technique to balance the load which one the disadvantage of the proposed system. During this, they used a market basket dataset for performing the algorithm.

Sonal L. and Varsha D [5] describe the information partitioning technique to balance the load among the cluster nodes is developed. Using this they meet the necessity for prime dimensionality of information processing. The author has developed the principles for providing the worth for minimum support and minimum confidence. They proposed an approach called PARMA i.e parallel mining approach which is the advantage of randomization for removing the frequent itemsets from several datasets. They explained in such a manner with the two steps, the first step is to gather all the information samples and the second step uses the parallel algorithm to increase mining speed.

III. ALGORITHM USED

A. Parallel Apriori Algorithm

Apriori Algorithm is used for locating the frequent itemsets carried candidate generation. Candidate containing all the frequent itemsets. The Apriori Algorithm relies on a nonempty subset of the frequent itemsets that should also be frequent. The main step of the algorithm is that the candidate of the k-itemset C_k from the frequent (k-1) - itemsets L_{k-1} and it will perform the join and prunes method. In the Join step, $L_{k-1} \bowtie L_{k-1}$ is allocated to the candidates set Ck. In the prune step, it reduces the dimensions of C_k using Apriori property.

Support= (# of transactions involving A and B)/ (total no. of transactions)

= P (A and B)

Confidence= (# of a transaction involving A and B)/ (total no. of transactions that have A)

=P (B if A)

=P (B/A)

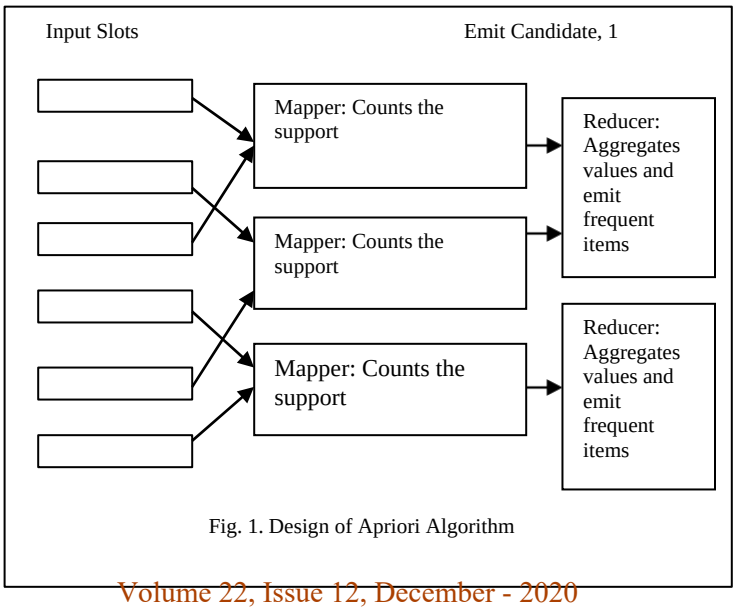

Fig. 1. Design of Apriori Algorithm

The typical Apriori algorithm is built on association rule. The main aim of this algorithm is to find frequent items or sets of items from the transactional database.

Step 1. The User gives the value of minimum support and minimum confidence.

Step 2. At first instance, it creates the itemsets having a single candidate, and then it removes the itemsets which there support value is lower than selected minimum support.

Step 3. It combines the itemsets that are specified in step 2 with each other to create the itemsets having two candidates, to create the frequent itemsets having two candidates.

Step 4. Repeat the steps likewise step 3 until no more itemsets.

B. Apriori Algorithm on Hadoop MapReduce

MapReduce framework was introduced by Google in 2004 for supporting distributed computing. The main feature of MapReduce was modified the distributed computing. This includes the main two points such as Map and Reduce. Both the function is to change the datasets within the (key, value) pairs. Mapper and Reducer function, both are executed in a parallel manner but the output is acquired only after the completion of reducer tasks. To execute the multiple stages of the MapReduce framework need to desire the outcome when the algorithm works recursively.

MapReduce 1

The First MapReduce Task is to organize for generating all frequent itemset.

Input- Each mapper sequentially scans each transaction from its specific input split, where each transaction is used in the format of pair (k, v) pairs.

Output- Globally frequent one-item sets are generated.

Map Reduce 2

The second task again scans the database to generate item set by pruning infrequent items in each transaction by using the output of the first Map Reduce.

Input-Output of the first Map Reduce is given as input to the second Map Reduce.

Map Reduce 3

The third Map Reduce task is to find out association rules using frequent item sets.

Input- The output of the second Map Reduce

Output- Find all the frequent item sets using the support threshold value.

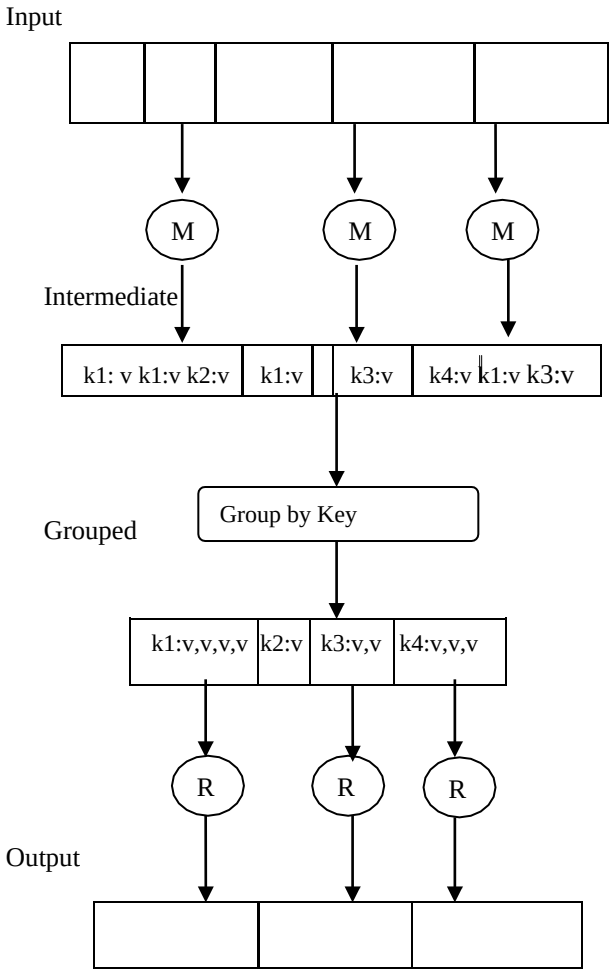

Fig. 2. Working of Apriori on MapReduce

IV. IMPLEMENTATION

By using the parallel apriori algorithm, reduce the consumption of network bandwidth and it's scalable. And also reduce the execution time and communication overhead. The programs execute the Mapreduce apriori algorithm on Hadoop. To judge the efficiency of our approach, we performed on two real datasets i.e Online Retail I and II and Wholesale Customer Datasets. To implement our algorithm we have used the Java Programming language with JDK.

Fig shows the execution of the project which we performed using the datasets.

Fig. 3. Input is taken for the Online Retail I datasets

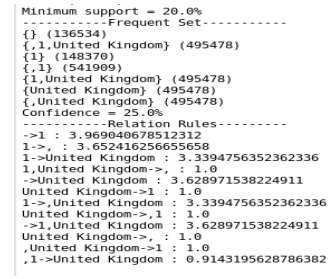

Fig. 4. Output producing the frequent itemsets

Fig. 5. Input is taken for Online Retail II datasets

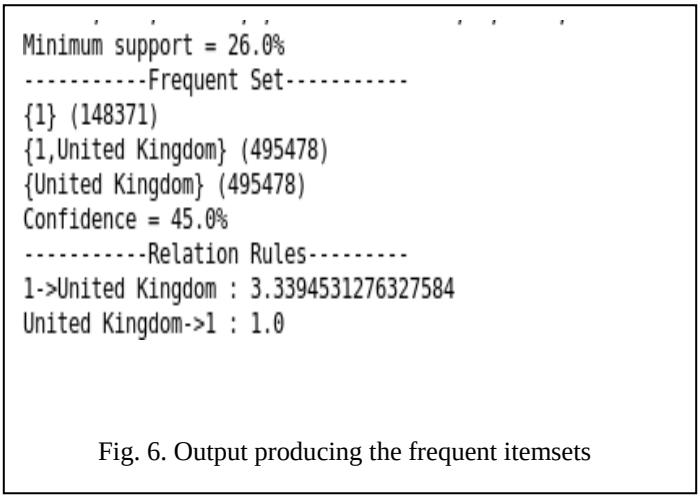

Fig. 6. Output producing the frequent itemsets

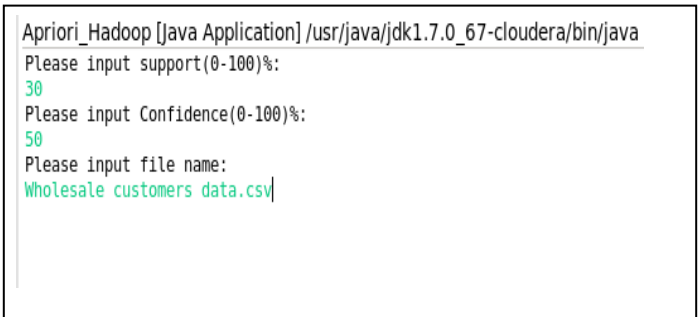

Fig. 7. Input is taken for Wholesale Customer datasets

```
Minimum support = 30.0%
-----------Frequent Set-----------
{3} (325)
{1,2} (424)
{2} (189)
{1} (375)
{2,3} (417)
{1,2,3} (414)
{1,3} (427)
Confidence = 50.0%
-----------Relation Rules---------
3->1,2 : 1.2738461538461539
3->2 : 1.283076923076923
3->1 : 1.3138461538461539
1,2->3 : 0.9764150943396226
2,3->1 : 0.9928057553956835
1,3->2 : 0.9695550351288056
1->2,3 : 1.104
2->3 : 2.2063492063492065
2->1 : 2.2433862433862433
1->2 : 1.1306666666666667
1->3 : 1.1386666666666667
2->1,3 : 2.1904761904761907
```

Fig. 8. Output producing the frequent itemsets

Table -1 Experiment Result

Datasets	Support	Confidence	Frequent Set	Relation Rules
Online Retail I	**20 %**	**25 %**	**136534**	**0.91431**
Online Retail II	**26 %**	**45 %**	**148371**	**1.0**
Wholesale Customer	**30 %**	**50 %**	**325**	**2.19**

Table 1 shows the input which we have taken and as correspondence, the output is produced. The performance of the output is given by the frequent set and the relation rules. It also is shown by the images of the result.

CONCLUSION AND FUTURE SCOPE

In this paper we discuss to conclude is the planning, performance, and analysis of using the Hadoop framework for ARM. This demonstrates that employing a parallelized style of the Apriori algorithm will give you efficient and simple to use on the Hadoop platform and also the MapReduce model. The matter of finding association rules need plenty of computations cost and memory. During this paper, a parallel Apriori algorithm is predicted in the MapReduce model. This algorithm also extracts the frequent patterns itemsets or among sets of things within the transaction database using the Apriori algorithm in the device. Using the Hadoop platform, a prediction example is implemented. Data are often stored on the Hadoop platform cost-effectively and it is retrieved easily when needed. MapReduce is gainful for a parallel Apriori Algorithm on big data on large datasets. Our main goal wants to optimize the Apriori algorithm but to check whether it is often implemented on Hadoop with a satisfactory result. It is often parallelized and simple to implement.

ACKNOWLEDGMENT

The authors would like to express their gratitude towards their guide, Dr. Vinaya Sawant, The Head of Information Technology Department, and the Principal, Dr. Hari Vasudevan for the opportunity provided to them to carry out implementation and research on this topic.

REFERENCES

[1] Srinivasa R, and V Sucharitha, “Efficient Algorithm using Big Data for Frequent Itemsets Mining”, IJITEE, pp. 394-396, 2019.

[2] Sandhya W, Sanchita S, Shweta K, Karishma M, “Apriori Algorithm Using MapReduce”, IJRR, Vol. 5, Issue. 5, pp. 129-132, 2018.

[3] Marwa Bouraoui, Ines Bouzouita, Amel Grissa Touzi, “Hadoop Based Mining Of Distributed Association Rules from Big Data”, IEEE, pp. 185-190, 2017.

[4] Seema Tribhuvan, Bharti Vasgi,” Parallel Frequent Itemset Mining for Big Datasets using Hadoop-MapReduce Paradigm”, IJRCCE, Vol. 6, Issue 6, pp. 188-192,2017.

[5] Sonal Londhe, Varsha Dange, “Mining Frequent Itemset using Hadoop Framework”, IJIRSET, Vol. 6, Issue 6, pp. 11369-11376,2017.

[6] Kavitha Mohan, Talit George, “Mining of Frequent Itemset in Hadoop”, IJRASET, Vol. 5, Issue 6, pp. 1068-1071, 2017.

[7] Jadhav Kalyani, Tamhana Manisha, Surwase Sonali, “A New Approach for Frequent Itemset Data Mining In Hadoop Environment”, ICRTESM, pp. 971-976,2017.

[8] Shivani Deshpande, Harshita Pawar, Amruta Chandras, Amol Langhe, “Data Partitioning in Frequent Itemset Mining on Hadoop Clusters”, IRJET, Vol. 3, Issue. 11, pp. 1231-1234,2016.

[9] Maedeh Afzali, Nishant Singh, Suresh Kumar, “Hadoop - MapReduce: A Platform For Mining Large Datasets”, IEEE, pp. 1856-1860, 2016.

[10] Mehdi Zitouni, Reza Akbarinia, Sadok Ben Yahia, Florent Masseglia, “ A Prime Number Based Approach For Closed Frequent Itemset Mining in Big Data”, ICDESA, pp. 1-10, 2015.

[11] P. Stanisic, S. Tomovic,” Apriori Multiple Algorithm For Mining Association Rules”, ITAC, Vol.37 No. 4, pp. 311-320,2015.

[12] S.O. AbdulSalam, K.S. Adewole, A.G. Akintola. “Data Mining in Market Basket Transaction: An Association Rule Mining Approach”, IJAIS, Vol.7 No.10, pp. 15-20, 2014.

[13] Sudhakar Singh, Rakhi Garg, P K Mishra,” Review of Apriori Based Algorithms on MapReduce Framework”, ICC, pp. 1-12, 2014.

[14] Phani Prasad, Murlidher Mourya,” A Study on Market Basket Analysis Using a Data Mining Algorithm”, IJETAE, Vol. 3, Issue. 6, pp. 361-363,2013.