HYBRID ALGORITHM WITH MAP REDUCE FRAMEWORK TO MINE DISTRIBUTED ASSOCIATION RULES FROM BIG DATA

[1]Jaini Doshi, [2]Prof. Anjuman Ranawadia [3]Prof. Rakesh Shah
[1]Student, [2]Asst. Professor, [3]PG Coordinator
[1] Computer Engineering,
[1]GMFE, Himmatnagar, India

Abstract : When its big data, its extremely large data sets. These datasets are analyzed computationally to reveal patterns, trends, and associations. That could be related to human behavior and interactions and product reordering ratio or understanding the symptoms or related signs of a disease. Many data analysis techniques are already defined and used by researchers. Results are still showing the scope of improvement. Based on the volume, variety, and velocity of data, the techniques are needed to be used or improved. Association rule mining is one of the technique to solve issues of accuracy in retrieved results. They are used to detect changes in customer behavior, buying trends and reasons that affect such process. Researches till date has proven the results are better than the earlier one. Though several methods have been suggested for the extraction of association rules, problems arise when data is in growing pattern with large volume. To overcome such issue, we propose, in this paper, a hybrid approach based on ARM techniques with Map Reduce framework, modified for processing large volumes of data in an increasing manner. Furthermore, because real life databases lead to a huge number of rules' including many redundant rules, our algorithm proposes to mine a compact set of rules with no loss of information. The results of experiments tested on large real world datasets highlight the relevance of mined data. Additionally in this research, the experiments are performed in continuous growing data which still yields comparative results.

IndexTerms **– Map – Reduce framework, Fp-growth, Hadoop, Association rules mining. Big data**

I. INTRODUCTION

Data Mining is a set of method that applies to large and complex databases. This is to eliminate the randomness and discover the hidden pattern. In other words, we can say that Data mining is the process of extraction of information from large databases and it is a powerful new technology having a great potential to help researchers as well as companies on the most important information in their data warehouses. Data mining tools are used to predict the future trends and behaviors thus allowing businesses to make knowledge-driven decisions.[2]

II. ASSOCIATION RULE

Association rule mining is a methodology that is used to discover unknown relationships hidden in big data. Rules refer to a set of identified frequent itemsets that represent the uncovered relationships in the dataset. In association rule mining (ARM), an often-used data mining task, provides a strategic resource for decision support by extracting the most important frequent patterns that simultaneously occur in a large transaction database. A typical application of ARM is the famous market basket analysis.[4]

Fig: 1 Market basket Analysis

III. BIG DATA WITH HADOOP

Big data is a collection of large datasets that cannot be processed using traditional computing techniques. It is not a single technique or a tool, rather it has become a complete subject, which involves various tools, techniques and frameworks. Big data is referred to extremely massive data sets that might be used to analyze computationally to uncover associations, patterns, and trends, particularly identifying with human conduct and collaborations.[6] The data in it will be of three types.[11]

1.Structured data − Relational data.
2.Semi Structured data − XML data.
3.Unstructured data − Word, PDF, Text, Media Logs.

IV. MAPREDUCE FRAMEWORK

MapReduce is a parallel programming paradigm for handling data with a very large volume[1]. The MapReduce algorithm contains two important tasks, specifically 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 progression of the name MapReduce implies, the reduce task is always performed after the map job.

Inputs and Outputs

The MapReduce framework works on <key, value> pairs[8], that is, the framework views the input to the job as a set of <key, value> pairs and produces a set of <key, value> pairs as the output of the job, conceivably of different types. The key and the value classes should be in sequential manner by the framework and hence, need to implement the Writable interface. furthermore, the key classes have to implement the Writable-Comparable interface to facilitate sorting by the framework. Input and Output types of a MapReduce job: (Input) <k1, v1> -> map -> <k2, v2>-> reduce -> <k3, v3>(Output).

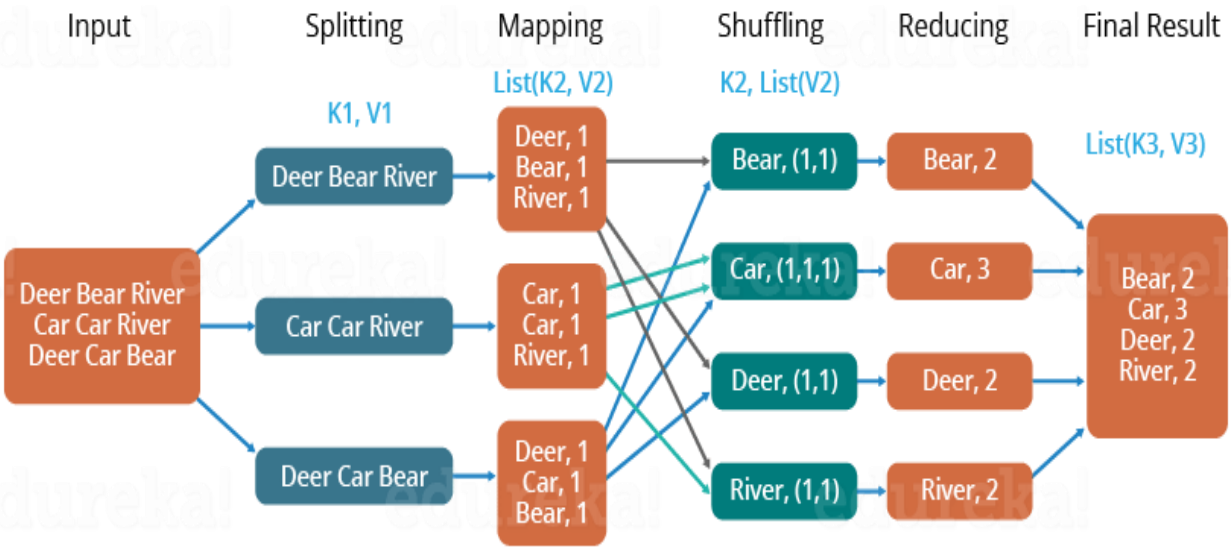


Fig: 2 Map Reduce Example

3.1 Problem Statement

Distributed data extraction for association rule is not much efficient when the data in large volume . Real life databases have a wide variety of data thus it leads to complex association rule and complex form of data . Selecting of data for different types of item sets will lead to in efficient result .

3.2 Research Gap

Association rule mining has been modified and hybridized in each separate research with big data, Hadoop and Map reduce has used it for grouping of value. Improvement scope in such data with novel formation of proven ARM algorithm will lead to better result.

3.3 Software Requirement:

Technology: JAVA-10, JDK-11
Plateform :UBUNTU OS-14.04 LTS 64-bit or windows platform
Framework : HADOOP – 2.6.5
Dev. Tool-IDE : ECLIPSE IDE

I. PROPOSED METHODOLOGY

3.1 Flow Chart of Proposed Algorithm

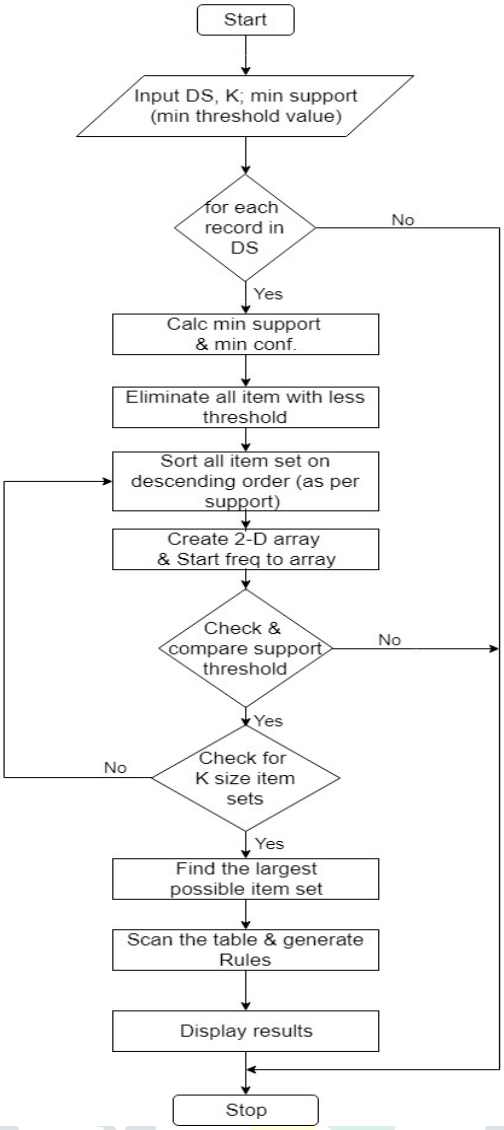


Fig. 3: flow chart of proposed work

3.2 Algorithm Steps

Input :

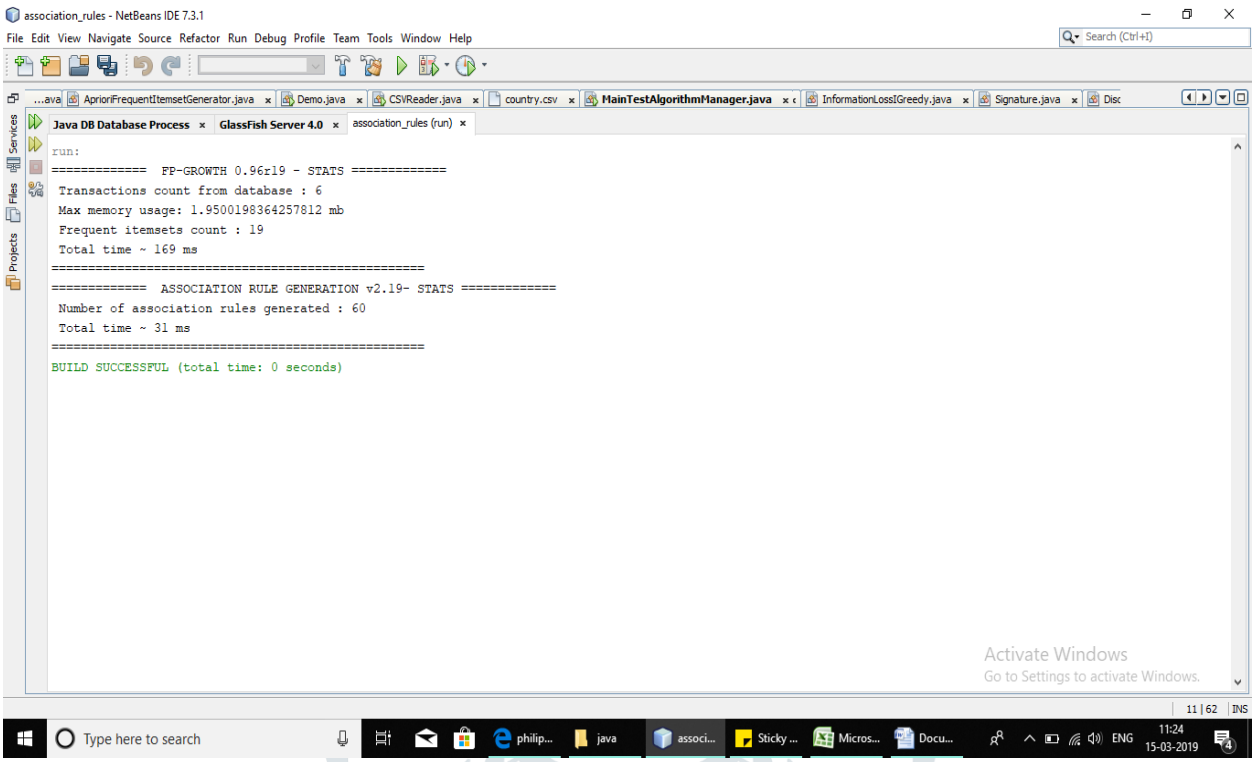
L1=large Itemset;C1=D=Dataset;K = Size of itemset;C = Single record;T= Temporary dataset;

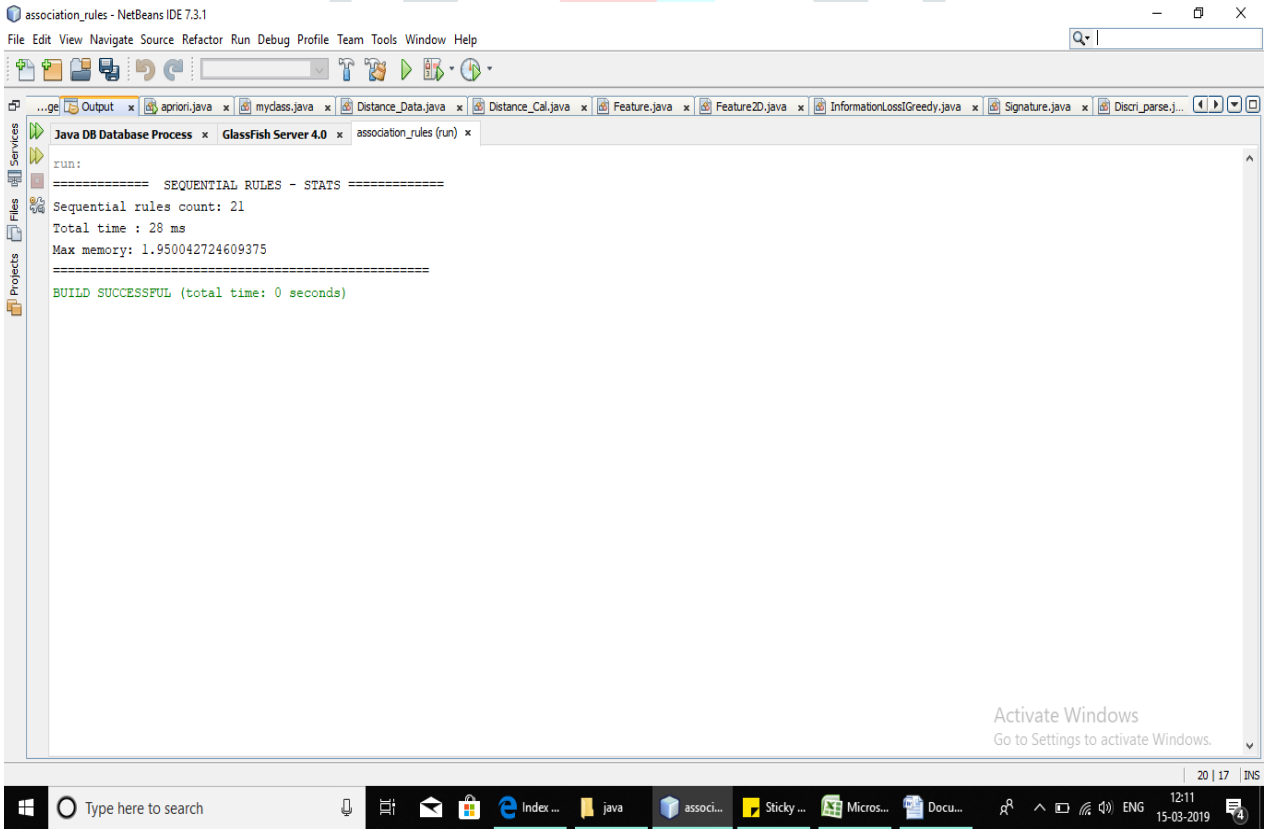
Algorithm :

Step 1 : L1 = {large 1-itemsets};
Step 2 : C1 = database D;
Step 3 : for (k = 2; Lk-1 $\neq \emptyset$; k++) do begin
Step 4 : Ck = FP-tree(Lk-1); // New candidates
Step 5 : Ck = $\emptyset$;
Step 6 : For all entries t $\in$ Ck-1 do begin
// determine candidate item sets in Ck contained in the transaction with identifier t.TID
Step 7 : Ct = {c $\in$ Ck | (c - c[k]) $\in$ t.set-of-item sets A (c - c[k - 1) $\in$ t.set-of-item sets};
Step 8 : For all candidates c $\in$ Ct do
Step 9 : c.count++;
Step 10 : if (Ct $\neq \emptyset$) then ck += < t.TID, Ct ;
Step 11 : for endStep 12 : Lk = {c $\in$ ck | c.count >= minsup}
Step 13:For End
Step 14 : Answer = Uk Lk;

V. IMPLEMENTATION RESULT

Processing for min support check with criteria





4.1 Number of rules generated and time taken for it

Table 4.1: Connect dataset

Generated rules		Time taken	
Base paper	**Proposed**	**Base paper**	**proposed**
80	85	25	20
60	64	42	34
30	36	75	60

Table 4.2: Webdocs dataset

	Generated rules		Time taken	
	The below row data is for number of rules			
Rules	50	40	30	20
Base paper	5	5	13	22
Proposed	3	4	11	18

V. CONCLUSION

In this proposed approach, Performed over big data based on massive small file processing strategies. We utilize the Sequence Files method to integrate all the small files which contain a large number of transaction datasets stored in HDFS into a large transaction data file as transaction database. The complexity of this algorithm is O(nlogn) which is far better than the existing algorithm's complexity. The scalability is high as it requires minimum communication cost and comparison costs.

VI. ACKNOWLEDGMENT

I forward my sincere thanks to Prof. Anjuman Ranavadiya and Prof. Rakesh Shah for there valuable help during the report design of Research Skills. There suggestions were always there whenever I needed it. As supervisor they sparred there valuable time for the in depth discussion on the topics. Also I forward my hearty thanks to other Faculty Members Department of Computer Engineering for their support.

REFERENCES

[1] Marwa Bouraoui, Ines Bouzouita, Amel Grissa Touzi, "Hadoop based Mining of Distributed Association Rules from Big Datas",IEEE, December 21-23, 2017.
[2] Divya.M.G, Nandini.K, Priyanka.K.T, Vandana.B," Weighted Itemset Mining from Bigdata using Hadoop", ICICN16, CSE, RRCE.
[3] Tushar M. Chaure, Kavita R. Singh," Frequent Itemset Mining Techniques - A Technical Review", WCFTR 2016.
[4] Yaling Xun, Jifu Zhang, Xiao Qin, Senior Member," FiDoop-DP: Data Partitioning in Frequent ItemsetMining on Hadoop Clusters", IEEE 2016.
[5] Ashwini A. Pandagale, Anil R," Hadoop-HBase for Finding Association Rules using Apriori MapReduce Algorithm",IEEE, May 20-21, 2016,
[6] Vijay M Bande, Ganesh K Pakle, "CSRS : Customized Service Recommendations System for Big Data Analysis using Map Reduce" , 2016 International Conference on Inventive Computation Technologies (ICICT)
[7] Xin Yue Yang, Zhen Liu, Yan Fu, ''Map Reduce as a Programming Model for Association Rules Algorithm on Hadoop'', IEEE - august-2010
[8] Sudhakar Singh, Rakhi Garg, P K Mishra, "Review of Apriori Based Algorithms on Mapreduce Framework", ICC – 2014
[9] https://www.tutorialspoint.com/hadoop/hadoop_big_data_solutions.htm
[10] https://searchdatamanagement.techtarget.com/definition/Hadoop
[11] https://www.tutorialspoint.com/hadoop/
[12] https://www.slideshare.net/sravya/hadoop-technology-ppt-57921409
[13] https://archive.ics.uci.edu/ml/datasets/online+retail#
[14] https://archive.ics.uci.edu/ml/machine-learning-databases/connect-4/connect-4.names
[15] https://www.bogotobogo.com/Hadoop/BigData_hadoop_Install_on_ubuntu_single_node_cluster.php