# Implement Mapreduce Apriori Algorithm to Generate Frequent Itemsets

Jyoti Yadav
Research Scholar
SDBCT, Indore

Neha Sehta
Asst Prof., Depaertment of IT
SDBCT, Indore

## ABSTRACT
For Mass data cloud computing is used as solution for storing and analyzing .Cloud is suitable for big data computation for processing large parallel data sets. Hadoop is used for distributed computing that would be required to enable big data. MapReduce is a component of Hadoop used for parallel processing. In this paper strategy of mining association rules is discussed with apriori algorithm. Modified MapReduce apriori algorithm with TopDown approach is implemented on datasets .The results show that strategy designed has higher efficiency and takes less time for execution for calculating frequent itemsets.

## General Terms
Modified Algorithm,Frequent Item sets,Apriori Algorithm,Cloud computing.

## Keywords
Hadoop,Map-Reduce, Apriori algorithm , Data mining Support Association rules,Top Down Approach.

## 1. INTRODUCTION
Data mining and knowledge discovery are used to extract useful ,hidden and unknown patterns and knowledge from large database.[1]Apart from basic dissolution steps, it involves database and data handling aspects ,data captivating and data resolution[2].Big data is the popular term used to express exponential growth of data. It is difficult to store ,collect, maintain, analyze and visualize. It has three characteristics volume ,velocity, variety. Cloud Computing is the development of distributed computing, parallel processing and grid computing which represents an emerging business computing model. We can use cloud computing techniques with data mining to reach high capacity and high efficiency. Hadoop [3] is an open source version of the MapReduce from Apache. It is the software framework for writing and running distributed applications that rapidly process large amounts of data on large clusters of computer nodes.MapReduce Apriori Algorithm is modified using top down Approach and implemented on datasets. The remainder of this paper is organized as follows. In Section 2 Background is discussed. Further in section 3,some related work is discussed .In section 4 Proposed methodology is given . In section 5results are shown.Finally conclusion is written un section 6..

## 2. HADOOP
Hadoop is open source framework provided by Apache to process very huge volume of data.It is written in Java.Google ,Facebook,LinkedIN,Yahoo,Witter used Hadoop as it allows distributed processing on cluster of commodity.Hadoop consist of Hadoop Coomon,Hadoop distributed File Sysyem(HDFS). Yet Another Resource Negotiator(YARN), MapReduce . Common utilities are used for other Hadoop Modules. HDFS provides high throughput accesss to application data. YARN is framework for job scheduling and cluster resource management. MapReduce is YARN based system for parallel processing of large datasets.Figure 1 shows working of MapReduce Model.It consist of two different task,Map Task and Reduce Task.Input data is splits in parallel by different machines by automatically partitioning.In Key Value Pair Map Node data is stored and output from Map Task is taken as Input in Reducer Node and after merging final output is send to Reducer Node.Hadoop Distributrd File System is designed to run on commodity hardware and suitable for applications having large datasets.

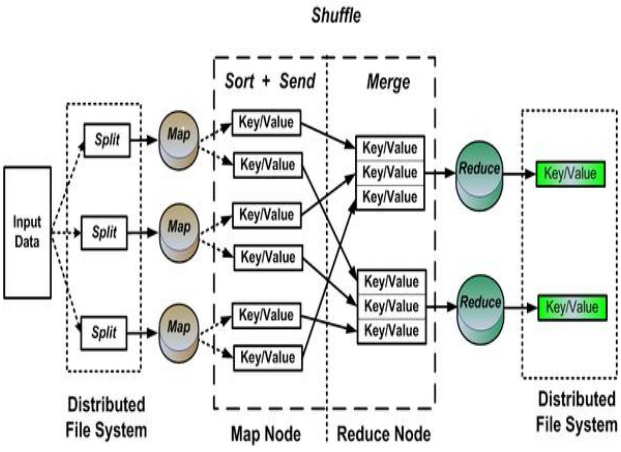


**Figure 1.Map Reduce Model**

MapReduce framework takes care of scheduling tasks ,monitoring them and re-executes the failed tasks.It consist of a single master JobTracker and one slave TaskTracker per cluster node.Using MapReduce In data mining Market Basket Analysis is the impottant term regarding discovery of association rule.It explained the frequently bought items together by the customer.Apriori Algorithm is also used for generating association rules on the basis of user defined support.

## 3. RELATED WORK
Author [1] gives the overview of parallel apriori algorithm implemented using mapreduce framework. Map and Reduce functions are classified further as passes. Generating frequent itemsets is same as MapReduce computing model.Triangular matrix data structure is used for counting of support in frequent 1,2 itemsets in one step. Author[2] in Parallel Implementation of modified Apriori algorithm on multicore system.By using Hadoop customer shopping behavior in Supermarket was analyzed.HDFS is used with Apriori algorithm to generate frequent itemsets outcome is used for generating association rules Author [4 ] in A Comprehensive Evaluation System of Association Rules on Multi-Index.Association rules were generated using multi angle and multi dimensions.On giving different weighs for each index association rules can befound..Support,Lift,Confidence are

parameters for rules evaluation and the comprehensive coefficient of each rule are calculated. Author [5]in Hadoop – MapReduce :A Platform for mining large datasets .Hadoop framework and its components are explained.Using MapReduce Platform apriori algorithm is modified and converted into parallel computation.Software is created and by using modified Hadoop Mapareduce Algorithm and frequent itemsets are generated. Author [6] in Hadoop – Hbase for Finding Association Rules using Apriori MapReduce Algorithm. As the frequent 2-itemsets offered from both time and space complexity .Apache . HBase is used as solution for access random data of low latency. Pruning is done on itemsets which are unnecessary using defined minimum support .It is shown by experimental results that it takes less time when HBase is used on different number of nodes.Author[7] in Novel Method of Apriori Algorithm using Top Down Approach. Classifical apriori algorithm was redesigned using Topdown approach.As apriori algorithm uses bottomup approach and suffrered from large number of database scan and perform better if itemsets are short.New Proposed Top down Apriori algorithm overcomes the deficiency by reducing number of database scan also useful for large amount of data base scan.In[8]Implementation of Apriori Algorithm based on MR Apriori algorithms.Three algorithms are implemented MR Apriori and existing algorithms(one phase and k-phase)based on hadoop mapreduce programming model.In[9]customer shopping behavior was analysed using Hadoop as a platform.Modifies Apriori algorithm is used to generate frequent itemsets.In[10]Parallelization in apriori and Improved Apriori algorithm is explained.On combining Improved Apriori algorithm with MapReduce MR_Apriori Algorithm is proposed and experiments were done on hadoop platform.

## MapReduce Apriori Algorithm

i. Scan the dataset to calculate support S of each item**.**

ii. min_supp=number/total number of items.

iii. If support S is greater than min_sup then add an item to frequent 1-itemset.

iv. Compute frequent item set for each map node using min_sup and collect all together in reduce phase.

v. Remove items that do not meet the min_sup.

vi. Again find frequent k – itemsets,calculate frequent itemset with an additional item by joining each map node.

vii. Collect the frequent item set at the reduce node and count item frequencies compared with min_sup.

viii. Remove the items that do not meet the min_sup in Reduce Node using prune().

## 4. PROPOSED METHODOLOGY

i. Scan the dataset to get the support count of each item.

ii. Remove item with support<minimum support .Let the item remain=n.

iii. setk=n-1.

iv. Compute ${}^{n}c_{k}$ subsets and count support..

v. Repeat step 4 for the itemset with maximum support count and k=k-1 till the itemset has value equal to minimum support. .

On the basis of above steps an example is solved. Given a set of transactions having minimum support=60% and minimum confidence=80%.

**Table 1. Input Transactions**

| TID | Itemset |
|---|---|
| 1000 | M,O,N,K,E,Y |
| 1001 | D,O,N,K,E,Y |
| 1002 | M,A,K,E |
| 1003 | M,V,C,K,Y |
| 1004 | C,O,O,K,E |

**Table 2. support count of each Itemset**

| Itemset | Support Count |
|---|---|
| M | 3 |
| O | 4 |
| N | 2 |
| K | 5 |
| E | 4 |
| Y | 3 |
| D | 1 |
| A | 1 |
| V | 1 |
| C | 2 |

Support count of each item is calcaulated and compare with min_sup.min_sup=60/100*number of transactions.Remove the items which are less than min_sup.

Min_sup=60/100*5=3.

**Table 3. Frequent-1 Itemsets.**

| Itemset | Support Count |
|---|---|
| M | 3 |
| O | 4 |
| K | 5 |
| E | 4 |
| Y | 3 |

MAP 1 HAVE ITEMSET OF TID 1000 AND TID 1001,MAP 2 HAVE ITEMSET OF TID 1002,1003,1004.

**Table 4.Mapper Input**

| MAP 1 | MAP2 |
|---|---|
| MOKEY=1 | MOKEY=0 |
| MOEY=1 | MOEY=0 |
| MOEK=1 | MOEK=0 |
| MOKY=1 | MOKY=0 |
| OKEY=2 | OKEY=0 |

**Reducer Output-OKEY=2**

**Table 5.Mapper Input**

| MAP 1 | MAP2 |
|---|---|
| OKEY=2 | OKEY=2 |
| OEY=2 | OEY=0 |
| EYK=2 | EYK=0 |
| OYK=2 | OYK=0 |
| OEK=2 | OEK=1 |

## 5. RESULTS

Datasets are of WHO region of different countries based on the national smoking ban in different organizations.size of dataset is 188kb .

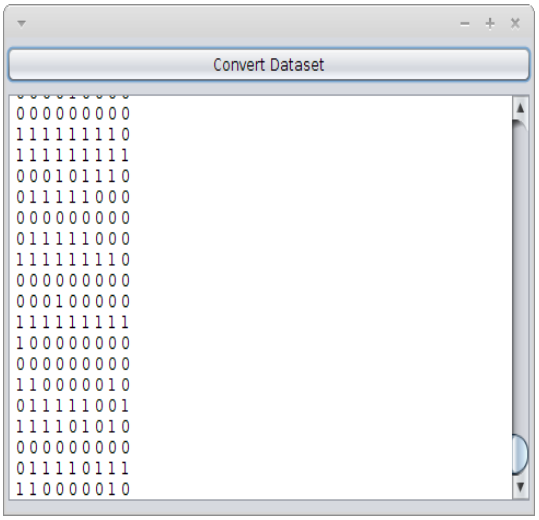


**Figure 2 Binary format of dataset**

This section provides the details of the user interface design .Below figure shows the initial project screen.On datasets implementation is done for evaluating the results with hadoop mapreduce.

**Reducer Output O,E,K=3**On the basis of this output

different associationRules are generated

**Table 6 Generation of association rules**

| Association Rule | Support | Confidence |
|---|---|---|
| O^K=>E | 3 | 3/3=1=100% |
| O^E=>K | 3 | 3/3=1=100% |
| K^E=>O | 3 | ¾=0.75=75% |
| E=>O^K | 3 | ¾=0.75=75% |
| K=>O^E | 3 | 3/5=0.6=60% |
| O=>K^E | 3 | ¾=0.75=75% |

Final association rules are O^K=>E and O^E=>K.

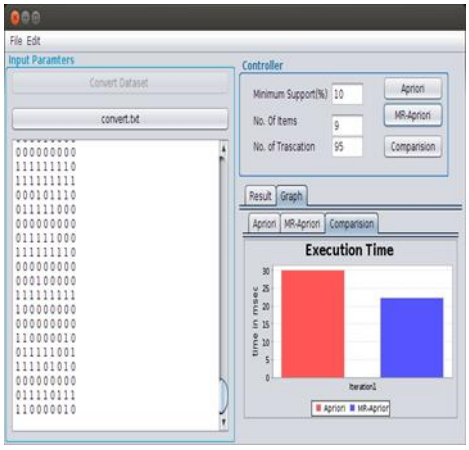


**Figure 3.Execution time comparison between Apriori and MR Apriori**

MR-Apriori takes less time than Apriori algorithm because it takes all itemset at once in map which satisfies the min_sup criteria.Input is divided in Mapon the basis of number of transactions.If a transactions has 5values in MAP1- 1,2 Transactions are send and in MAP2- 3,4,5transactions are taken.Parallel execution of input occurs in MapReduce which takes less time as compare to apriori.Itemset which has highest value is taken.and further n-1 set are calculated.This steps were repeated till the itemset having value equal to or greater than min_sup is obtained.


**Figure 4.Comparison between Apriori and MR Apriori for multiple iterations**

**Table 6.Evaluation Metric for Multiple Iterations**

| Iteration | Apriori | MR-Apriori |
|---|---|---|
| 1 | 25 Milli Seconds | 14 Milli Seconds |
| 2 | 19 Milli Seconds | 15 milli Seconds |
| 3 | 20 milli Seconds | 16 milli Seconds |

Apriori takes more time because first the frequent-1 itemset are calculated.After this those items which are below min_sup is removed. Then with remaining items pairs formed for frequent-2 itemsets then after items below min_sup criteria is removed ,with remaining items triplets were formed for frequent-3 itemsets here also items below min_sup were removed the item which fulfill min_sup criteria is takeforgenerating association rules.Each time dataset is scanned for generating frequent itemwsets and candidate generation.

## 6. CONCLUSION

In this paper an algorithm is implemented which is hybrid approach of modified apriori with hadoop map reduce for generating frequent itemsets. It will overcome the deficiency of classical Apriori algorithm hence reduces the number database scan and it is useful for large amount of database scans. Comparison of apriori and MR-Apriori is done.MR-Apriori is modified proposed algorithm which takes less time than classical apriori algorithm. On the basis of results association rules are generated. In future it can be enhanced for variety of datasets directly for feasibility

## 7. REFERENCES

[1] Sudhakar Singh ,Review of Apriori Based Algorithms on Map Reduce Framework.ICC-2014,pp 593-604

[2] Sruthi M,Parallel Implementation of modified Apriori Algorithms on Multicore Systems,Proceeding of IMCIC-ICSIT 2016.

[3] Apache hadoop.http:/hadoop.apache.org.

[4] Shunli Ding,A comprehensive Evaluation System of Association Rules on Multi-Index,(978-1-4673-6593-2/15)2015 IEEE

[5] M.Afzali,Hadoop-MapReduce:A Platform for finding Large datasets,978-9-3805-4421-2/16/$31.00 c 2016 IEEE".

[6] Anil R.Surve,Hadoop-HBase for Finding Association Rulesusing MapReduce Apriori Algorithm,2016 IEEE.

[7] S.Maheshwari,Novel Method of Apriori Algorithm using Top Down Approach,International Journal of Computer Applications(0975-8887)volume 77-N

[8] Othman Yaha,An Efficient Implementation of Apriori Algorithm based on Hadoop-MapReduce Model. International Journal of Reviews in Computing 31st December 2012. Vol. 12 © 2009 - 2012 IJRIC & LLS.

[9] Sadhana Shetty, Implementation of Modified APRIORI Algorithm Using HADOOP ,Journal of Data Mining and Knowledge Engineering Volume 1 Issue 2Page 1-11© MANTECH PUBLICATIONS

[10] Xueyan Lin.MR-Apriori:Association Rules Algorithm Based on MapReduce.978-1-4799-3279-5/14/$31.00 ©