International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 04 Issue: 10 | Oct -2017 www.irjet.net p-ISSN: 2395-0072

Association Rule Mining using RHadoop

Anurag Agrahari[1], Prof D.T.V. Dharmaji Rao[2]

[1]M.tech Student, Deptt. Of Computer Sci & Engg, AITAM College, Tekkali, Srikakulam, Andhra Pradesh, India.
[2]Professor, Deptt of Computer Sci & Engg, AITAM College, Tekkali, Srikakulam, Andhra Pradesh, India.

Abstract - With the advancement of computer related technology, the world is generating several Exabyte of data per year. Data generated through many sources like business processes, transactions, social networking sites, web servers, and many other source are remains in structured as well as unstructured form increased by exponential rates. Processing of huge amount of data or extracting hidden information is always a challenging and tuff task. Many Data Mining and statistical techniques are being used to find out the hidden pattern hidden in it. Association Rule Mining is one of core Data mining techniques being widely used in Big Data Analytics. Many tools are used in the Big Data Analysis. R is statistical language, widely used in the analysis. Hadoop can be integrated with R known as RHadoop. In this paper, we will briefly discuss what is big data, Hadoop, integration with R and analysis of result on training data set using association Rule.

Key Words*: *Big Data, Data mining, Association Rule mining, RHadoop, Ambari.

1. INTRODUCTION

Today large datasets are a part of every field that somehow involves digital information. Companies, programs and network sensors produce huge amount of fresh data, around 2.5 EBs (Exabyte's) every single day. The total amount of data in the world was 4.4 zettabytes in 2013. According to prediction, this is set to rise steeply to 44 zettabytes by 2020[1].A lot of statistical techniques and data mining techniques are used to analysis the Big data set. Big Data analysis refers to data that exceeds the typical storage, processing, and computing capacity of conventional databases using data analysis techniques [17]. Most popular distributed computing paradigms nowadays uses Hadoop framework for handling the large scale data. Data mining, a technique to understand and find out the hidden knowledge from raw data and convert into useful information, is increasingly being used in a variety of fields like marketing, business intelligence, scientific discoveries, biotechnology, Internet searches, and multimedia. Data mining is an interdisciplinary field combining ideas from statistics, machine learning, and natural language processing. Data mining in such environments requires a utilization of the available resources. The application of Big data in expert system in agriculture for making of decision support system, will open the new research area [21][22].

2. RELETED WORKS

Association rule mining works was carried out by [2] where she applied association rule with Map Reduce concept in cloud. Cloud itself provides the platform to analysis the Big Data. Most of Organization is using cloud for their application. Furthermore a lot of work for Association Rule Mining (ARM) is one of important and popular technique of data mining which find interesting correlation or association between set of items or attributes and also frequent patterns. in large database [3].Association Rule mining is widely used in Many other area like business application, bi-informatics, medical diagnosis, text analysis and business Intelligence and consumer behavior. A lot of works carried out using Hadoop and Association Rule mining.

Furthermore, very few and limited works carried out using Rhadoop. The concept of Rhadoop is discussed by[4],where they conclude the concept of Rhadoop, how it work with the Hadoop Component, integration with the hadoop, ODBC and other component. Memory-Sensitive Heterogeneous Earliest Finish Time (MSHEFT) algorithms were implemented by [5] where he used the Hadoop and spark platform with R to analysis and performance evaluation of the algorithms. Being well known statistical language, open source software, and new package can be developed for computation of large and complex data, large genotype–phenotype datasets [6]. A new tool was developed by [6] named BlueSNP using R and Hadoop for large scale cluster. Linear regression analysis was developed by Bogdan Oancea using R and Hadoop[7]. A state of art research on sentiment analysis was done by [8][9] using Rhadoop of twitter data. R is also providing the API interface for extracting the twitter data. Furthermore, R can be integrated with other Big data other Hadoop technology such as Spark known SparkR[10].Comparative study between the Rapriori and spark was done by[34].

Furthermore, in Map Reduce model, some parallel Apriori algorithms are implemented using Hadoop-MapReduce model [23][24][25][26]. The name of the algorithm Apriori is based on the Apriori property which states that

all nonempty subsets of a frequent itemset must also be frequent [2]. Many sequential and parallel algorithms have been proposed to improve the performance of Apriori algorithm. Some popular sequential approaches are Hash-based technique, Transaction reduction, Partitioning, Sampling and Dynamic itemset counting (DIC) [26][27][28][29][30]. A algorithms proposed for three parallel version of Apriori algorithm,Count Distribution (CD), Data Distribution (DD) and Candidate Distribution[32].

3. ASSOCIATION RULE MINING

Association rule mining (ARM) is a popular method of data mining method for discovering interesting relations between terms in the dataset. The concept of strong rules was used by Agarwal et al [11] to find association rules in items. An association rule defines relation between two set of items for e.g.

{A, B} => {C}.

In a purchase relation this would indicate if a person buys A and B together, he/she is more likely to also buy C. Mining association rule consists of following two steps:

Finding the item set which are frequent in the data set: The frequent item sets are set of those items whose support (sup (item)) in the data set is greater than the minimum required support (min_sup). Considering the above example all three A, B and C belongs to frequent item set and sup {A, B} and sup {C} would be greater than the min_sup. The support of an item set is defined as proportion of transactions which contains the item set.

Generating association rule from frequent item set: Generating the interesting rules from the frequent item sets on the basis of confidence (conf). The confidence of the above rule will be sup {A, B} divided by sup{C}. If the confidence of the rule is greater than the required confidence, the rule can be considered as an interesting one.

The frequent item set required for generation of association rule can be generated using Apriori algorithm. If the confidence of the rule is greater than the required confidence, then the rule can be considered as an interesting one.

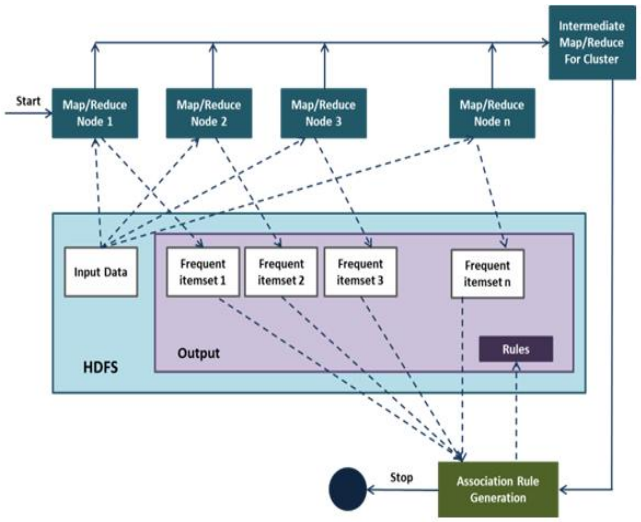

Fig 1. Map/reduce at n node showing n iterations [33].

The following Step in association mining will take place in Map reduce**.** As figure 1 show that data submitted to the process will have the multiple phase/step. Those step/phase are as below[23][24][25]. Hadoop paradigm have two function i.e. mapper and reducer.The input and output of these functions must be in form of *(key, value)* pairs. The Mapper takes the input (k1, v1) pairs from HDFS and produces a list of intermediate (k2, v2) pairs. An optional *Combiner* function is applied to reduce communication cost of transferring intermediate outputs of mappers to reducers.

Phase 1

```
Mapper (key, value)
        // key: TID
        // value: itemsets in
                transaction Ti
for each transaction Ti
        assigned to Mapper do
                for each itemset in Ck
                        do
                        if itemset ∈ Ti
                        output (itemset, 1);
                        end if
                end for
end for
```

Phase 2

```
Combiner (key, value)
        // key: itemset
        // value: list (1)
for each itemset do
        for each 1 in list (1) of
                corresponding itemset do
                itemset.local_sup + = 1;
```

```
        end for
        output (itemset,itemset.local_sup);
end for
```

Phase 3

```
Reducer (key, value)
        // key: itemset
        // value: list (local_sup)
for each itemset do
        for each local_sup in
                list (local_sup) of
                corresponding itemset do
                itemset.sup + = local_sup;
        end for
        if itemset.sup ≥ minimum
        support;
        output (itemset, itemset.sup);
end for
```

These algorithms can be classified in two categories: 1-phase of map-reduce and k-phase of mapreduce [2]. Some algorithms used all the three functions mapper, reducer and combiner while some used only mapper and reducer function.

4. RHADOOP

4.1 Integrating with R with Hadoop

R is a language and environment for statistical computing was developed at Bell Laboratories (formerly AT&T, now Lucent Technologies) by John Chambers and colleagues [12]. It is upgraded version of S. R was initially written by *Robert Gentleman* and *Ross Ihaka*—also known as "*R & R*" of the Statistics Department of the University of Auckland[13].

There are several approaches to integrate R and Hadoop: R and Streaming, Rhipe and RHadoop but in this paper we will present only RHadoop. RHadoop is an open source project developed by Revolution Analytics (http://www.revolutionanalytics.com/) that provides client-side integration between R and Hadoop. This allows running Map Reduce jobs within R and consists of a collection of several packages:

plyrmr - provides plyr like processing functions for structured data type, having capabilities of handling large data sets stored with Hadoop;
rmr - contains a collection of functions that provide Map Reduce model implementation in R;
rdfs - is an interface between R and HDFS, providing file management operations in R for data stored in HDFS;
rhbase - is an interface between R and Hbase, and provides management functions in R for Hbase databases;
ravro -A package that adds the ability to read and write avro files from local and HDFS file system and adds an avro input format for rmr2[16].

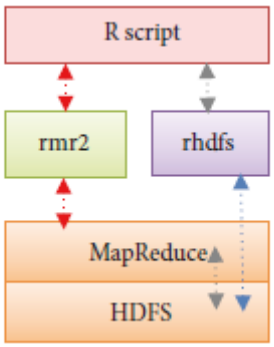

Fig 2. RHadoop Framework[14]

As figure 1 clearly show the working with different component of Mapreduce and HDFS. The package rhdfs is used to read and write function used for data in cluster. Same wise,rmr2 and rmr is used to do task of instruction with Mapreduce programming paradigm. The rhdfs package provides the mechanism to store and fetch the data from HDFS. It only installed at the each data node of the cluster. The rmr2 and rmr(for hadoop version 1) provide the facilities to do map reduce in cluster. R script or program only runs over the rmr and rhdfs package. Running the R script need some other package of R, will be discussed on next subsection.

4.2 R Programming With Rhadoop

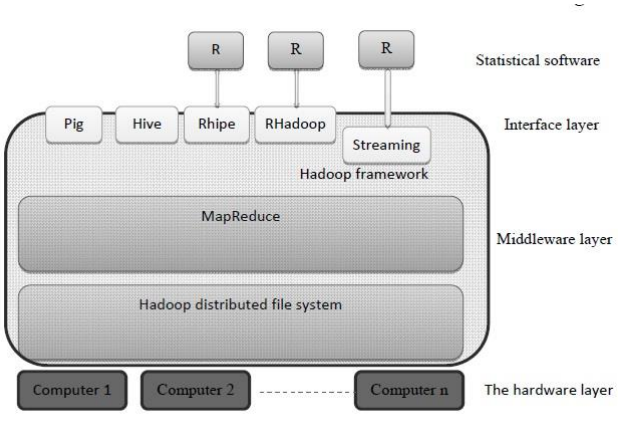

Fig 3. R interface RHadoop over Hadoop [15].

As Figure 3 show that the detailed Architecture of Hadoop and Rhadoop. The brief explanation was covered by author on [15]. The first layer known as hardware layer which include the commodity computer over the network or in cloud cluster. Middle layer manage the Map reduce and HDFS system component of Hadoop. The script

provided over the RHadoop via R is act through Rhadoop to Hadoop. The next layer is a layer that provides an interface for data analysis. Interface layer which having streaming component to other language like c,c++,Python and other programming language. R script communicates to hadoop component through the Rhadoop libraries which defined above section. We can use Rhipe or Rhadoop libraries that build an interface between Hadoop and R, allowing users to access data from the Hadoop file system and write their own scripts for implementing Map and Reduce jobs [15].R con not communicate to hadoop component, al through it need some more and necessary component of R core language detailed discussed on next section of Project setup.

5. PROJECT SETUP

For implementing the concept, we had created the environment of 8 node cluster AITAM Central Library. Ambari, open source tools for managing the cluster, monitor the cluster and installing the component of Hadoop component. For the implementing the concept the Horton work's HDP 2.2 ware used along with Ambari Version 1.7.1 over the Cent Os version 6.6 using the Virtual Platform.

Commodity Specification detail-

For implementing the algorithms, we had used different node with different processor based system, memory, RAM, cache Size, Different clock speed, and different core system. This is heterogeneous kind of system cluster. Along with this we have used the VMware for some system (n0.exmple.com, n1.example.com, n2.example.com). No.example.com machine have two installation i.e one for ambari server of RAM size 2 GB and rest for the Name node server based on the window 7 base operating system. The guest operating system was cent os 6.6.

Name	Core	Ip address	RAM
N0.example.com	8(i7)	172.16.6.181	7.67 GB
N1.example.com	8(i7)	172.16.6.182	3.73 GB
N2.example.com	8(i7)	172.16.6.183	3.73 GB
N3.example.com	4(i3)	172.16.6.184	3.73 GB
N4.example.com		172.16.6.185	
N5.example.com		172.16.6.186	
N6.example.com		172.16.6.187	
N7.example.com		172.16.6.188	
N8.example.com		172.16.6.190	

Table 1. Cluster commodity Hardware Specification

Ambari Installation A completely open source management platform for provisioning, managing, monitoring and securing Apache Hadoop clusters [18][17]. Ambari enables System Administrators to:

Provision a Hadoop Cluster Ambari provides a step-by-step wizard for installing Hadoop services across any number of hosts. It handles configuration of Hadoop services for the cluster.

Manage a Hadoop Cluster Ambari provides central management for starting, stopping, and reconfiguring Hadoop services across the entire cluster.

Monitor a Hadoop Cluster Ambari provides a dashboard for monitoring health and status of the Hadoop cluster. Ambari leverages Ambari Metrics System for metrics collection. Ambari leverages Ambari Alert Framework for system alerting and will notify you when your attention is needed (e.g., a node goes down, remaining disk space is low, etc).

Apart from this, Ambari is very useful in large scale cluster setup because it provide consistent, secure platform for operational control.

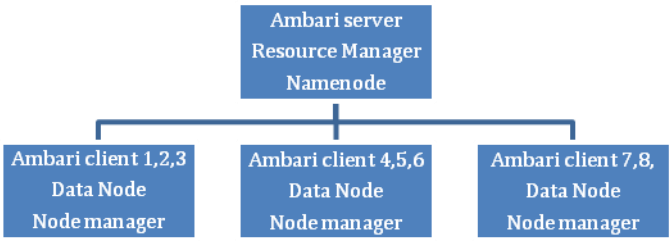

Fig 4. Conceptual Layout Of Ambari server installation

It is Simplified Installation and Configuration and Management, Centralized Security Setup, Full Visibility into Cluster Health and Highly Extensible and Customizable [17].

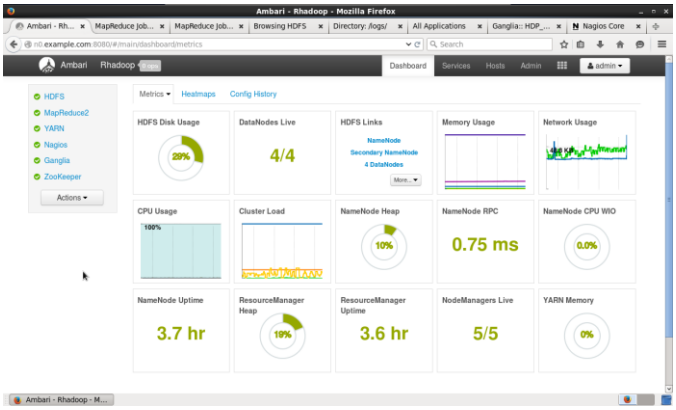

Fig 5. Screenshot of ambari server of 4 node cluster.

More node can be add and remove from ambari server dashboard. Service can start and stop from same dash board. For installing the Ambari we need to create repo for installation of hadoop service over node/client.

name	Address/base Url
HDP-2.2	http://n0.example.com/hdp/HDP/centos6/2.x/GA/2.2.0.0
HDP-UTILS-1.1.0.20	http://n0.example.com/hdp/HDP-TILS-1.1.0.20/repos/centos6

Table 2- Ambari repo base Url(Cent os6.6)

We have to provide the repositories' for hadoop. In our case we have used Horton work product HDP 2.2 and other hadoop utilities. The components of HDP 2.2 are summarized in Table 2.

SERVICE	VERSION	DESCRIPTION
Zookeeper	3.4.6.2.2.0.0	Centralized service which provides highly reliable distributed coordination
HDFS	2.6.0.2.2.0.0	Apache Hadoop Distributed File System
YARN + MapReduce2	2.6.0.2.2.0.0	Apache Hadoop NextGen MapReduce (YARN)
Ganglia	3.5.0	Ganglia Metrics Collection system (RRDTool will be installed too)
Nagios	3.5.0	Nagios Monitoring and Alerting system

Table 3- HDP 2.2 Cluster Stack Descriptions

Name	ip address	services
n1.example.com	172.16.6.181	App Timeline Server Ganglia Monitor HDFS Client History Server NameNode NodeManager ResourceManager
n2.example.com	172.16.6.182	DataNode Ganglia Monitor Ganglia Server HDFS Client MapReduce2 Client Nagios Server YARN Client ZooKeeper Client
n3.example.com	172.16.6.183	DataNode Ganglia Monitor HDFS Client MapReduce2 Client SNameNode YARN Client ZooKeeper Client ZooKeeper Server
n4.example.com n5.example.com n6.example.com n7.example.com n8.example.com	172.16.6.184 172.16.6.185 172.16.6.186 172.16.6.187 172.16.6.188 172.16.6.189	DataNode Ganglia Monitor HDFS Client MapReduce2 Client YARN Client ZooKeeper Client

Table 4- Description of service running at nodes

Rhadoop And R installation- R and Rhadoop installation is easy task. Before installing Rhadoop, The dependent package of Rhadoop should be installed in each node of cluster. Revolution R Open (RRO) is the enhanced distribution of R from Revolution Analytics. It is a complete open source platform for statistical analysis and data science. RRO 8.0.1 is based on (and 100% compatible with) version 3.1.2 of the R language, and includes additional capabilities for performance, reproducibility and platform support. Along with this Rstudio ware integrated with R. The concerning package of the Rhadoop and R software shown in table 3.

SERVICE/Package	Version	Description
rmr2	3.4.6.2.2.0.0	Master, client
rhdfs	1.0.8	Only at Master
plyrmr	0.6	Master, client
rHadoopClient	0.2	Master, client
Nagios	3.5.0	Master, client
Rcurl	1.95-4.5	Master, client
rJava	0.9-6	Master, client
RJSONIO	1.3-0	Master, client
iterators	1.0.7	Master, client
itertools	0.1-3	Master, client
digest	0.6.8	Master, client
Rcpp	0.11.4	Master, client
jsonlite	0.9.14	Master, client
mime	0.2	Master, client
R6	2.0.1	Master, client

functional	0.6	Master, client
httr	0.6.1	Master, client
stringr	0.6.2	Master, client
memoise	0.2.1	Master, client
whisker	0.3-2	Master, client
evaluate	0.5.5	Master, client
rstudioapi	0.2	Master, client
brew	1.0-6	Master, client
roxygen2	4.1.0	Master, client
reshape2	1.4.1	Master, client
plyr	1.8.1	Master, client
devtools	1.7.0	Master, client
caTools	1.17.1	Master, client
R.methodsS3	1.7.0	Master, client
RColorBrewer	1.1-2	Master, client
pryr	0.1	Master, client
proto	0.3-10	Master, client
dichromat	2.0-0	Master, client
gtable	0.1.2	Master, client
colorspace	1.2-4	Master, client
scales	0.2.4	Master, client
munsell	0.4.2	Master, client
labeling	0.3	Master, client
DBI	0.3.1	Master, client
ggplot2	1.0.0	Master, client
acepack	1.3-3.3	Master, client
latticeExtra	0.6-26	Master, client
lhs	0.10	Master, client
rmr2	3.3.1	Master, client
sp	1.0-17	Master, client
Hmisc	3.15-0	Master, client
hydroPSO	0.3-4	Master, client
plyrmr	0.6.0	Master, client

Table 5. R and Rhadoop Package Description in Node/master

6. RESULT AND ANALYSIS

Data Set was taken from well known source UCI data set repository and other source. The dataset named *chess*, *connect*, and *mushroom* ware downloaded from UCI data set repository [19][20]. The short description about the data set summarized in table 4.

Data set	Instance	attributes
Chess	3196	36
pumsb	49046	74
Connect	67557	42

Table 6. Data set And there Description [19][20].

The dataset provided by [19] are in integer form and space delimiter. Each row of the dataset represents the new data in the file.

For the conceptualizing the concept, we had fix the support value as **(.5)** and count value as **(.9)**. Table 5 show the result of training data set with same support and count value in without hadoop and 2 node, 4 node, 6 node and 8 node cluster. Running time count in second and process run 4 times each to find out the average running time of the cluster.

Dataset	Without hadoop	2 node	4 node	6 node	8 node
chess	31.157	24	12	10	7
pumsb	271.117	37	36	34	25
connect	82.150	42	40	25	18

Table 7. Result showing in different node

Result may have difference because of the network and one single system is running logically two system. The reduction of execution time is more evident for those datasets which have longer execution time. Moreover, we can see that when the size of the dataset increases, the performance on multi-node Hadoop also gets better. The result could get better if the ambari server is running on separate system and master node running separately in the cluster and having more RAM size. The value of running time is average mean value and round of value in second.

7. CONCLUSIONS

We implemented the improved Apriori Algorithm on MapReduce programming model on the Hadoop platform. The R language is widely used in the hadoop due to easy and comfort for the analyst. Moreover, this distributed algorithm can also cater to the distributed nature of the input data. Furthermore, details of the management of the distributed systems, such as data transferring among nodes and node failures are taken care by Hadoop, which adds a great deal of robustness and scalability to the system. The performance could be enhanced if we have the facilities to increase the RAM and high performing Machine.

ACKNOWLEDGEMENT

I really thankful to the Director, Shri Prof. V.V Nageswara Rao and Head of computer science department Shri Dr. G. S. N. Murty and my guide shri prof D.T.V. Dharmaji Rao, for there kind support and help.

REFERENCES

[1] D. Kornack and P. Rakic, "Cell Proliferation without Neurogenesis in Adult Primate Neocortex," Science, vol. 294, Dec. 2001, pp. 2127-2130, doi:10.1126/science.1065467.

[2] Pallavi roys(2012), MINING ASSOCIATION RULES IN CLOUD, MS thesis, North Dakota State University of Agriculture and Applied Science, Fargo, North Dakota August 2012Mr.

[3] Neeraj Raheja, Ravish Kumar(2012) , Optimization of Association Rule Learning in Distributed Database using Clustering Technique,International Journal of Scientific and Research Publications, Volume 2, Issue 12, December 2012 ,ISSN 2250-3153.

[4] HARISH D,ANUSHA M.S ,Dr. DAYA SAGAR K.V(2015).,"BIG DATA ANALYSIS USING RHADOOP", ISSN: 2349-2163,Issue 4, Volume 2 (April 2015) pp 180-85.

[5] Bao Rong Chang, Yun-Da Lee, and Po-Hao Liao, "Development of Multiple Big Data, Analytics Platforms with Rapid Response" https://doi.org/10.1155/2017/6972461.

[6] H.Huang et al., BlueSNP: R package for highly scalable genome-wide associationstudies using Hadoop clusters, BIOINFORMATICS APPLICATIONS NOTE , Vol. 29 no. 1 2013, pages 135–136 doi:10.1093/bioinformatics/bts647.

[7] Bogdan OANCEA, LINEAR REGRESSION WITH R AND HADOOP, Challenges of the Knowledge Society. IT in Social Sciences.

[8] Thasnim K M1, M Sudheep Elayidom, IMPLEMENTATION OF A NOVEL SENTIMENT ANALYSIS TECHNIQUE FOR TWITTER DATA USING RHADOOP, IJRET: International Journal of Research in Engineering and Technology, eISSN: 2319-1163 | pISSN: 2321-7308.

[9] Shubham S. Deshmukh et al(2017)., Twitter Data Analysis using R, International Journal of Science, Engineering and Technology Research (IJSETR) Volume 6, Issue 4, April 2017, ISSN: 2278 -7798.

[10] Shivaram Venkataraman et al(2016), SparkR: Scaling R Programs with Spark, SIGMOD '16, June 26–July 1, 2016, San Francisco, CA, USA., ISBN 978-1-4503-3531-7/16/06.

[11] R. Agrawal and R. Srikant(1994). Fast algorithms for mining association rules. In Proc. 1994 Int. Conf.Very Large Data Bases, pages 487-499, Santiago, Chile, September 1994.

[12] https://www.r-project.org/about.html access on 03/09/17.

[13] R project,Cran project https://www.r-project.org/contributors.html access on 05/09/17.

[14] Bao Rong Chang, Yun-Da Lee, and Po-Hao Liao(2017), Development of Multiple Big Data Analytics Platforms with Rapid Response, Hindawi Scientific Programming Volume 2017, Article ID 6972461, https://doi.org/10.1155/2017/6972461.

[15] Bogdan OANCEA, Raluca Mariana DRAGOESCU(2014), Integrating R and Hadoop for Big Data Analysis, Romanian Statistical Review nr. 2 / 2014.

[16] Rhadoop wiki https://github.com/RevolutionAnalytics/RHadoop/wiki access on 03/09/2017.

[17] Anurag Agrahari, Prof D.T.V. Dharmaji Rao(2017), "A Review paper on Big Data: Technologies, Tools and Trends," www.irjet.net,Volume: 04 Issue: 10,Oct -2017,e-ISSN: 2395-0056 pp 640-649.

[18] Ambari wiki,https://ambari.apache.org/ access on 2/08/2017.

[19] http://fimi.ua.ac.be/data/ access on 12-04-2015.

[20] Lichman, M. (2013). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

[21] Violeta nicoleta opriş & ciprian racuciu(2015), the expert systems analysis using the concept of big data and cloud computing services, Military Technical Academy, Faculty of Military Electronic and Information Systems, Bucharest, Naval Academy Scientific Bulletin, Volume XVIII – 2015 – Issue 2.

[22] Anurag Agrahari ,Dr. ShashiKant Tripathi(2012),A Theoretical Framework for Development of Decision Support System for Agriculture, RESEARCH INVENTY: International Journal of Engineering and Science ISSN: 2278-4721, Vol. 1, Issue 6 (October 2012), PP 50-55.

[23] Li L. & Zhang M. (2011). The Strategy of Mining Association Rule Based on Cloud Computing. Proceeding of the 2011 International Conference on Business Computing and Global Informatization (BCGIN '11). Washington, DC, USA, IEEE: 475- 478.

[24] Li N., Zeng L., He Q. & Shi Z. (2012). Parallel Implementation of Apriori Algorithm Based on MapReduce. Proc. of the 13th ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel & Distributed Computing (SNPD '12). Kyoto, IEEE: 236 – 241.

[25] Lin M., Lee P. & Hsueh S. (2012). Apriori-based Frequent Itemset Mining Algorithms on MapReduce. Proc. of the 16th International Conference on Ubiquitous Information Management and Communication (ICUIMC '12). New York, NY, USA, ACM: Article No. 76.

[26] Yang X.Y., Liu Z. & Fu Y. (2010). MapReduce as a Programming Model for Association Rules Algorithm on Hadoop. Proc. of the 3rd International Conference on Information Sciences and Interaction Sciences (ICIS '10). Chengdu, China, IEEE: 99 – 102

[27] S. Park, Ming-Syan and P. S. Yu, "Using a Hash-Based Method with Transaction Trimming for Mining Association Rules," in IEEE Transactions on Knowledge and Data Engineering, vol. 9, no. 5, pp. 813-825, 1997.

[28] J. S. Park, Ming-Syan and P. S. Yu, "An Effective Hash-Based Algorithm for Mining Association Rules," in SIGMOD ACM, vol. 24, no. 2, pp. 175-186,1995.

[29] M. J. Zaki, S. Parthasarathy, W. Li and M. Ogihara, "Evaluation of Sampling for Data Mining of Association Rules," in Proceedings IEEE 7th International Workshop on Research Issues in Data Engineering, 1997, pp. 42-50.

[30] A. Savasere, E. Omiecinski and S. Navathe, "An Efficient Algorithm for Mining Association Rules in Large Databases," in Proceedings 21st VLDB Conference,Switzerland, 1995, pp. 432-444.

[31] S. Brin, R. Motwani, J. D. Ullman and S. Tsur, "Dynamic Itemset Counting and Implication Rules for Market Basket Data," in ACM SIGMOD Record, vol. 26, no. 2, pp. 255-264, 1997.

[32] R. Agrawal and J.C. Shafer, "Parallel Mining of Association Rules," in IEEETransactions on Knowledge and Data Engineering, vol.8, no. 6, pp.962-969,1996.

[33] Poonam Modgi et al, / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (3) , 2014, pp 3093 – 3097.

[34] Sanjay Rathee, Manohar Kaul, Arti Kashyap, R-Apriori: An Efficient Apriori based Algorithm on Spark, http://www.iith.ac.in/~mkaul/papers/pikm09-rathee.pdf. access on 03/09/17.