ijournals International Journal of Software & Hardware Research in Engineering
ISSN No: 2347-4890 Volume 2 Issue 2, February 2014

Improving Apriori Algorithm to get better performance with Cloud Computing

Zeba Qureshi[1]; Sanjay Bansal[2]
Affiliation: A.I.T.R, RGPV, India[1], A.I.T.R, RGPV, India[2]

ABSTRACT

Cloud computing has become a big name in present era. It has proved to be a great solution for storing and processing huge amount of data. It promises on demand, scalable, pay-as-you-go compute and storage capacity. Data mining techniques implemented with cloud computing paradigm can be useful to analyze big data on clouds. In our project we have used association rule mining as a data mining technique. In particular we have used Apriori algorithm for association rule mining. It has been observed that the original Apriori algorithm was designed for sequential computation so directly using it for parallel computation doesn't seems a good idea. So we have improved the Apriori algorithm so as to suit it for parallel computation platform. We have used Amazon's web services namely EC2, S3 and EMR for cloud computing.

General Terms

Association rule mining, Improved Apriori algorithm

Keywords

Data mining, cloud computing, data mining with cloud computing, association rule mining, association rule mining in clouds.

1. INTRODUCTION

The Internet is a vital tool in our professional and personal life and its users are becoming more numerous day by day. May business transactions are conducted over the Internet. One of the most revolutionary concepts of internet is Cloud Computing. In fact cloud computing is the evolution of internet computing. Cloud computing is an innovative technology that provides IT resources on demand [9]. Many definitions are proposed for cloud computing. The one definition by Vaquero is as follows. Cloud computing can best be described as a giant pool which comprises software, hardware and other services that can be retrieved through the "cloud". All these resources can be accessed when necessary. Generally there are cloud providers which provide these services on pay-per-use basis. This ensures flexibility and cost reduction because you only pay for what you use and extra resources are always available.

Cloud computing is a developing concept and technology for delivering IT resources on demand. Low cost, mobility and flexibility adds beauty to the cloud computing though security is a major concern in cloud computing [10]. An equally significant field in last few years is data mining that focuses on changing raw data into useful information [3]. Data mining is significantly and increasingly used in the fields of marketing, internet, business intelligence, biotechnology, scientific discoveries etc [5].

Data mining techniques and applications are significant in the cloud computing [6]. The data mining techniques when combined with Cloud computing can retrieve meaningful information from virtually integrated data warehouse and this in turn will reduce the cost of infrastructure and storage. The execution of data mining techniques along with cloud computing can benefit users to a great extent [7] [11].

Data mining is nothing but extracting useful or interesting patterns out of raw data. The raw data is analysed to extract useful information out of it. Association rule mining, clustering, classification etc are some of the important techniques of data mining [4]. Data mining is usually used in market basket analysis and can be applied in various fields.

2. BACKGROUND AND RELATED WORK

All material on each page should fit within a A4 size, centered on the page, beginning 2.54 cm (1") from the top of the page and ending with 2.54 cm (1") from the bottom. The right and left margins should also be 2.54 cm (1"). The text should be in two 8.45 cm (3.33") columns with a .83 cm (.33") gutter.

2.1 Data Mining

The process of extracting useful patterns or information from large amount of data is known as data mining. Most of the people think data mining as a synonym of knowledge discovery [1]. But actually data mining can be considered as a step of knowledge discovery in databases (KDD). KDD process includes data cleaning (to remove noise and inconsistent data), data integration (where multiple data sources may be combined), data selection (where data relevant to the analysis task are retrieved from the database), data transformation (where data are transformed or consolidated into forms appropriate for mining by performing summary or aggregation operations), data mining (an essential process where intelligent methods are applied in order to extract data patterns, pattern evaluation (to identify the truly interesting patterns representing knowledge based on some interestingness measures) and knowledge presentation (where visualization and knowledge representation techniques are used to present the mined knowledge to the user).Data mining has attracted great deal of attention in information industry as well as in business areas because of the need of turning large data into useful information. Data mining is useful in an explanatory scenario in which there are no predefined notions about what will constitute an interesting outcome. The database system industry has witnessed an evolutionary path in the development of the following functionalities like data collection and database creation, data management and advanced data analysis that includes data warehousing and data mining. Prediction and description are considered as two primary goals of data mining. Predictive data mining which produces the model of system described by the given data set and descriptive data mining, which produces new, non trivial information based on the available data set. The goal of prediction and description are achieved through data mining tasks such as classification, discovering association rules and clustering [11].

2.2 Cloud Computing

Cloud computing can be defined as the use of IT resources (such as softwares, platforms, storage etc) that are delivered as a service over a network. With traditional computing paradigms we run the software and store data on our computer system. These files could be shared in a network. The importance of cloud computing lies in the fact that the software are not run from our computer but rather stored on the server and accessed through internet. Even if a computer crashes, the software is still available for others to use. The concept of cloud computing has developed from clouds. A cloud can be considered as a large group of interconnected computers which can be personal computers or network servers; they can be public or private. The concept of cloud computing has spread rapidly through the information technology industry. The ability of organizations to tap into computer applications and other software via the cloud and thus free themselves from building and managing their own technology infrastructure seems potentially irresistible. In fact some companies providing cloud services have been growing at double digit rates despite the recent economic downturn.

2.3 Association rule mining

One of the well techniques of data mining is association rules which are used to find out the relationship or association between various items. The problem of finding relation between items is often termed as market basket analysis. In this problem the presence of items within baskets is identified so that the customers buying habits can be analysed [2]. The technique is used in inventory management, sales promotion etc. The discovery of association rules is primarily dependent on finding the frequent sets. This can require multiple passes through the database. The algorithms aims at reducing number of passes by generating a candidate set which should turn out to be frequent sets. Many different algorithms are designed to find out the association rules. The algorithm differs on the basis of how they handle candidate sets and how they reduce number of scans on the database. Some of the recent algorithms of association rule mining do not create candidate set. Practically the frequent sets generated are very large in number and this can be constrained by selecting only those items in which the user is interested. Let us consider a set of items and a transaction database which is again a set of transactions. The association rule takes the following form for a transaction database: X=>Y, where X and Y are the sets of items called item sets. Now there are two important terms related to association rules: support and confidence. The support of an item or the set of items is the percentage of transactions in 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. The two thresholds namely minimal support and minimal confidence is set to find out reasonable support and confidence.

The association rule mining is the method that finds out the association rules which satisfy the predefined minimum support and minimum confidence. The association rule mining is usually carried out in 2 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 first step. Various algorithms like Apriori algorithm, partition algorithm, pincer search algorithm, dynamic item set counting algorithm, FP tree growth algorithm, Éclat and dEclat etc have been developed to find out the frequent items from the transaction database [1].The Apriori algorithm is the most general and widely used association rule mining 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 Apriori Tid was given by 1994 Agrawal et al. Other algorithms like SETM and AIS were also used for association rule mining but the performance of Apriori and Apriori Tid was better than these algorithms. This is because SETM and AIS generated too many candidate sets which were later found out to be infrequent among data sets. With large amount of data and with the advent of parallel computing technology various association mining algorithms like count distribution algorithm, data distribution algorithm, candidate distribution algorithm and improved Apriori algorithms have been proposed [8]. These algorithms can be used under cloud computing environment. Reducing time for generating frequent item sets can boost the performance of an association rule mining algorithm. Keeping this in mind various other algorithms were developed later. The concept of hashing can be used for pruning (removing infrequent item sets) which reduces time to generate frequent item sets. The process of association rule mining can be fastened by removing the infrequent item sets as quickly as possible though pruning can be problematic sometimes. The frequent patterns algorithm without candidate generation eliminates the costly candidate generation. It also avoids scanning the database again and again. So, we can use Frequent Pattern (FP) Growth ARM algorithm that is more efficient structure to mine patterns when database grows. FP tree growth algorithm is also used for mining and it does not create the candidate set. It rather creates a tree like structure to find the frequent sets.

2.4 Hadoop and Mapreduce

Hadoop is a programming framework that supports the processing of large data in a distributed computing environment. Hadoop is a free and java based programming framework [14]. Hadoop was created by Doug Cutting and Mike Cafarella in 2005. Hadop can make use of a single or thousands of machines for computation and storage. It creates a cluster of machines and coordinates work. Hadoop consists of two modules: (1) Hadoop Distributed File System (HDFS) for reliable storage and (2) Map/Reduce for high-performance parallel data processing.

HDFS splits and distributes the user data across servers in a cluster and replicates data to avoid data loss in case of multiple nodes failure. Hadoop also implements MapReduce, which is a distributed parallel processing system. MapReduce takes advantage of the distribution and replication of data in HDFS to spread execution of any job across many nodes in a cluster.

MapReduce is a programming framework which is used to simplify data processing across massive data sets. As people rapidly increase their online activity, organizations are finding it vital to quickly analyse the huge amounts of data their customers and audiences generate to better understand and serve them. MapReduce is the tool that is helping those organizations.

3. METHODOLOGY

Various data mining techniques have been implemented in cloud computing [12]. The Apriori algorithm is a famous algorithm for association rule mining. But the existing implementation used the original Apriori for cloud computing paradigm. Some have tried parallelism but have failed to reduce number of steps in Apriori algorithm. Using original Apriori for cloud paradigm doesn't make a good choice because the original Apriori algorithm was designed for the sequential computing. The current implementations have the drawbacks that it does not scale linearly as number of records increase. Secondly, the execution time increases when a higher value of k-itemsets is required.

So in this project we have optimized the Apriori algorithm that is to be used on the cloud platform. We have tried to remove the above limitations and our improved Apriori algorithm has the following features:

1. It will scale linearly as number of records increase.
2. Time taken will be agnostic to the value k. That is whatever k-itemsets run happens, it will take same time for a given number of records.

The execution time of existing Apriori algorithm increases exponentially with decrease in support count. The proposed algorithm below will reduce the execution time for lower values of support count, which is often desired.

Following improvements are proposed:

- Instead of running n more passes till we get frequent-n itemsets, we propose to have first pass generating frequent-1 itemsets and just one more pass generating all other frequent(2,n) itemsets. Here mapper will emit all the possible subsets of a transaction in the first pass itself rather than going for 2-itemset creation first and then go for 3-itemset, 4-itemset etc. as is the case with the classic implementation of Apriori for Hadoop. It would bring down the total execution time considerably for a very low support count.
- As done in existing implementations why even emit value as 1 to show existence of an element and then count that in the reducers, just existence of a value would be enough to count occurrences of a given item Id. Hence we can just emit an empty string there by reducing the amount of data written in the mapper phases.
- Also we shall not emit anything for the case when there is just one element in the transaction; it is not going to help generate any association rule. Hence any such transaction having just a single item is not going to be helpful.
- The custom key format is used which improves the performance by saving on the checks for equality of associative sets. In reducers the identity of itemsets have to be established which needs considerable computation mileage, for example following itemsets all represent the same set:
  {1, 2, 3}
  {2, 1, 3}
  {3, 2, 1}
- Hence, in order to minify the string comparisons required and perhaps one of the hindrances in earlier implementations not attempting to get it done in 2 steps, we will be now implementing a custom key format which would take a set itself as a key rather than a text / string. This would be achieved using the Java's collection library.

4. IMPLEMENTATION

The following Cloud Services are used in the project

4.1 Amazon Elastic Compute Cloud (EC2) Service

EC2 is one of the most famous services that come under Amazon web services. This web service allows cloud users to deploy instances with minimal or little configuration. Different types of instances are available here. These instances are actually virtual machines that are accessible over internet via Amazon EC2 service.

4.2 Amazon Simple Storage Service (S3)

Amazon S3 is widely used AWS. S3 is used to store and retrieve large amounts of data, at anytime, anywhere on the web. With S3 one can store any amount of data and this data can be easily retrieved from anywhere. The data is actually uploaded into S3 buckets whose capacity varies according to the user's need.

4.3 Amazon Elastic MapReduce (EMR) Service

Amazon Elastic MapReduce (Amazon EMR) is a web service process vast amounts of data .the advantages of EMR is that this service reduces the cost to process huge amount of data and also saves time.

The EMR service makes use of Hadoop, which is an open source framework to distribute user's data across a resizable cluster of Amazon EC2 instances. Amazon EMR is widely used in web indexing, data warehousing, log analysis, financial analysis etc. In this research work, EMR has been used to launch clusters and execute Hadoop jobs.

The following steps are followed while using Amazon EMR:

- Develop the data processing application.
- Upload application and data to Amazon S3.
- Configure and launch the cluster.
- Monitor the cluster (Optional).
- Retrieve the output.

We have implemented improved Apriori algorithm on Amazon EC2 to evaluate the performance. The data input files and application were saved on S3 which is a data storage service. Data transference between Amazon S3 and Amazon EC2 is free which makes S3 attractive for Hadoop EC2 users. The output data is also written back in S3 buckets at the end. The temporary data is written in the HDFS files. Amazon Elastic MapReduce takes care of provisioning a Hadoop cluster, running the job flow, terminating the job flow, moving the data between Amazon EC2 and Amazon S3, and optimizing Hadoop. Amazon EMR removes most of the difficulties associated with the Hadoop configuration, such as setting up the hardware and networking required by the Hadoop cluster including monitoring the setup, configuring Hadoop, and executing the job flow.

Hadoop job flow using cloud service EMR, EC2 and S3 cloud is described here. To start the job, a request is sent from the host to the EMR. Then after the Hadoop cluster with master and slave instances is created. This Hadoop cluster does all processing on the job. The temporary files created during the execution of the job and the output files are stored on S3. Once the job is completed, a message is sent to the user indicating the completion of the job [12].

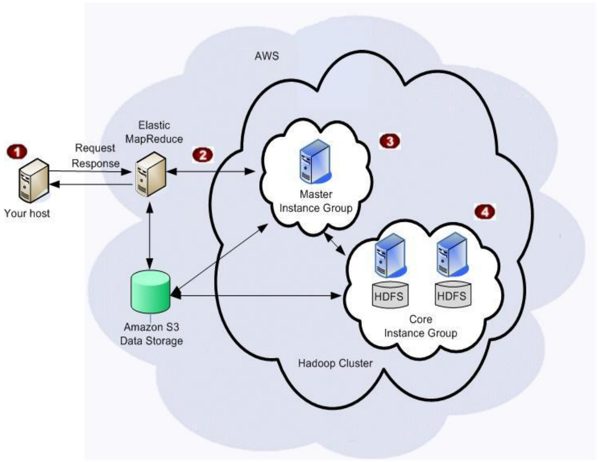


Fig 1: Job flow in the EC2 cloud for Hadoop [13]

5. RESULTS

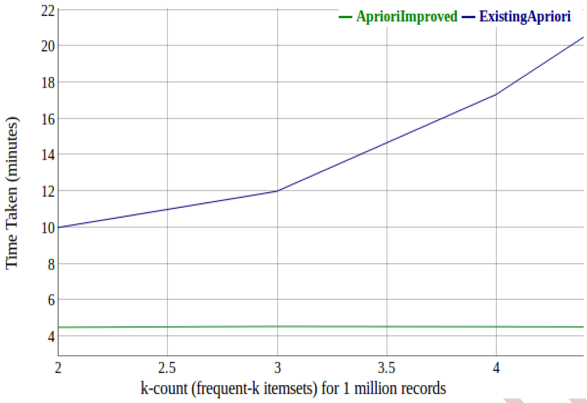


Fig 2: A graph showing comparison of time taken by existing Apriori and Apriori improved for frequent k-itemsets

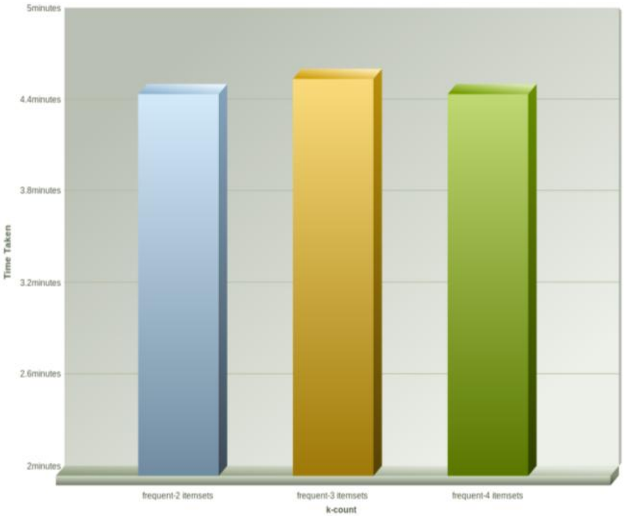


Fig 3: A graph showing comparison of time taken by Apriori improved for frequent-2 itemsets, frequent-3 itemsets and frequent-4 itemsets

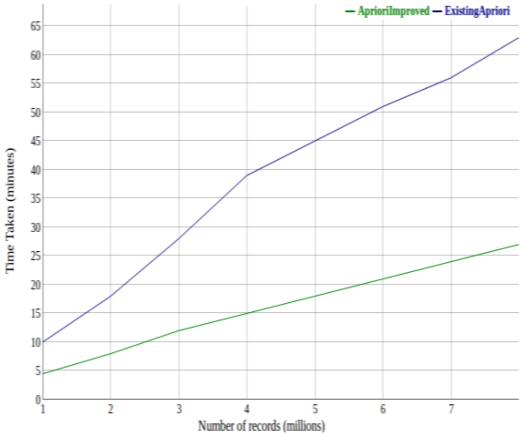


Fig 4: A graph showing comparison of time taken by existing Apriori and Apriori improved

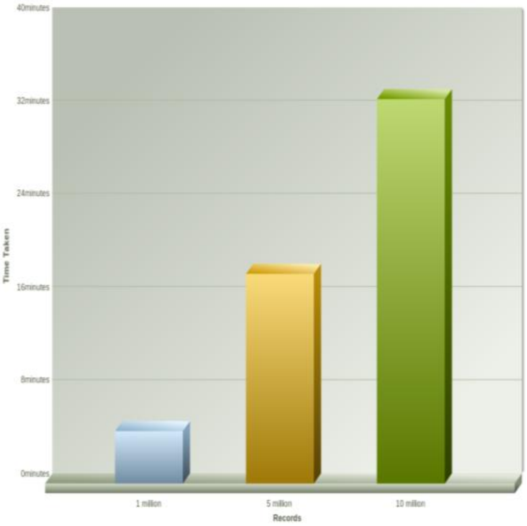
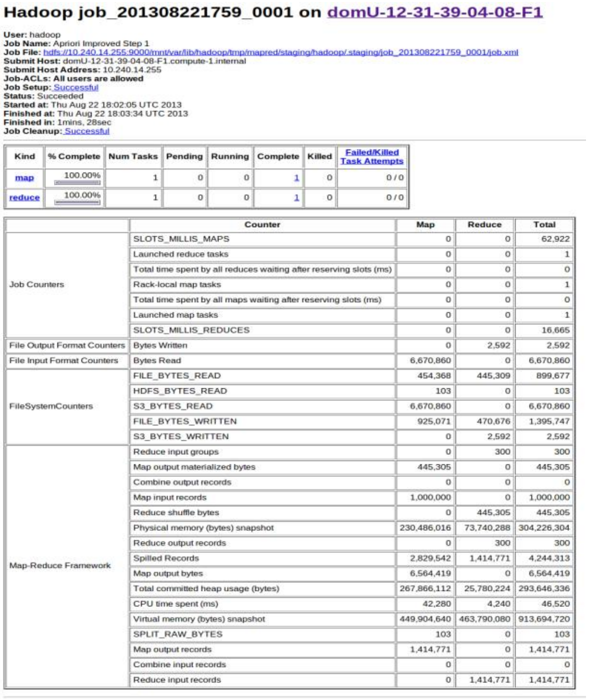


Fig 5: A graph showing comparison of time taken by Apriori improved for different no. of records



Hadoop job_201308221759_0001 on domU-12-31-39-04-08-F1

User: hadoop
Job Name: Apriori Improved Step 1
Job File: hdfs://10.240.14.255:9000/mnt/var/lib/hadoop/tmp/mapred/staging/hadoop/.staging/job_201308221759_0001/job.xml
Submit Host: domU-12-31-39-04-08-F1.compute-1.internal
Submit Host Address: 10.240.14.255
Job-ACLs: All users are allowed
Job Setup: Successful
Status: Succeeded
Started at: Thu Aug 22 18:02:05 UTC 2013
Finished at: Thu Aug 22 18:03:34 UTC 2013
Finished in: 1mins, 28sec
Job Cleanup: Successful

Kind	% Complete	Num Tasks	Pending	Running	Complete	Killed	Failed/Killed Task Attempts
map	100.00%	1	0	0	1	0	0 / 0
reduce	100.00%	1	0	0	1	0	0 / 0

	Counter	Map	Reduce	Total
Job Counters	SLOTS_MILLIS_MAPS	0	0	62,922
	Launched reduce tasks	0	0	1
	Total time spent by all reduces waiting after reserving slots (ms)	0	0	0
	Rack-local map tasks	0	0	1
	Total time spent by all maps waiting after reserving slots (ms)	0	0	0
	Launched map tasks	0	0	1
	SLOTS_MILLIS_REDUCES	0	0	16,665
File Output Format Counters	Bytes Written	0	2,592	2,592
File Input Format Counters	Bytes Read	6,670,860	0	6,670,860
FileSystemCounters	FILE_BYTES_READ	454,368	445,309	899,677
	HDFS_BYTES_READ	103	0	103
	S3_BYTES_READ	6,670,860	0	6,670,860
	FILE_BYTES_WRITTEN	925,071	470,676	1,395,747
	S3_BYTES_WRITTEN	0	2,592	2,592
Map-Reduce Framework	Reduce input groups	0	300	300
	Map output materialized bytes	445,305	0	445,305
	Combine output records	0	0	0
	Map input records	1,000,000	0	1,000,000
	Reduce shuffle bytes	0	445,305	445,305
	Physical memory (bytes) snapshot	230,486,016	73,740,288	304,226,304
	Reduce output records	0	300	300
	Spilled Records	2,829,542	1,414,771	4,244,313
	Map output bytes	6,564,419	0	6,564,419
	Total committed heap usage (bytes)	267,866,112	25,780,224	293,646,336
	CPU time spent (ms)	42,280	4,240	46,520
	Virtual memory (bytes) snapshot	449,904,640	463,790,080	913,694,720
	SPLIT_RAW_BYTES	103	0	103
	Map output records	1,414,771	0	1,414,771
	Combine input records	0	0	0
	Reduce input records	0	1,414,771	1,414,771

Fig 6: Result of Amazon's EMR

6. CONCLUSION

Cloud computing is the next evolution of internet computing which provides cost effective solutions for storing and analysing huge amount of data. Data mining on cloud computing paradigm can benefit us to a great extent. That is why we have implemented data mining technique on cloud platform. Out of many data mining techniques we have studied association rule mining technique in this paper. More specifically we have association rule mining in cloud computing environment. For association rule mining we have used improved Apriori algorithm that has comparatively better performance than the original Apriori on cloud paradigm.

7. REFERENCES

[1] Jiawei Han Micheline Kamber, Data Mining concepts and techniques, 2nd Ed.

[2] Ling Juan Li, min Zhang, "The strategy of mining association rule based on cloud computing", International conference on business computing and global information, 2011.

[3] Ven Katadri .M, Dr. Lokaanaathaa C.Reddy, "A review on data mining from past to future", International Journal of Computer Applications (0975 – 8887) Volume 15– No.7, February 2011.

[4] Musa J. Jafar, "A Tools-Based Approach to Teaching Data Mining Methods", Journal of Information Technology Education, Volume 9, 2010.

[5] http://en.wikipedia.org/wiki/Data_mining/

[6] T.V.Mahendra, N.Deepika, N.Keasava Rao, "Data Mining for High Performance Data Cloud using Association Rule Mining", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 2, Issue 1, January 2012.

[7] Joanna Gordon, Chiemi Hayashi, "Exploring the Future of Cloud Computing," World Economic Forum, 2010.

[8] 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.

[9] Wang, L., Tao, J., Kunze, M., Castellanos, A. C., Kramer, D., and Karl, W. 2008. "Scientific cloud computing: early definition and experience". In Proceeding of the 10th IEEE International Conference on High Performance Computing and Communications.

[10] Anthony T.Velte, Cloud Computing: A Practical Approach, TATA McGraw Hill Edition.

[11] Bhagyashree Ambulkar, Vaishali Borkar, "Data Mining in Cloud Computing", Proceedings published by International Journal of Computer Applications® (IJCA) ISSN: 0975 – 8887.

[12] Juan Li, Pallavi Roy, Samee U. Khan, Lizhe Wang, Yan Bai, "Data Mining Using Clouds: An Experimental Implementation of Apriori over MapReduce".

[13] http://docs.amazonwebservices.com/ElasticMapReduce/latest/Develo perGuide/Introduct ion_EMRArch.html.

[14] http://citeseerx.ist.psu.edu/viewdoc/summary10.1.1.111.9204