IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-6, Issue-5, May 2018 E-ISSN: 2347-2693

Finding Most Efficient Algorithm for Segregating and Saving Data: A Comparative Analysis

Heli P. Vyas[1], Sanjay M. Shah[2]

[1] Department, Chaudhari Technical Institute, Gandhinagar, India
[2]Narsinhbhai institute of Computer Studies and Management, Kadi, India

Corresponding Author: helivyas@yahoo.co.in, Tel.: 9979431730

Available online at: www.ijcseonline.org

Accepted: 24/May/2018, Published: 31/May/2018

***Abstract* -** Data mining is the process of making patterns for large data sets. There are many data mining algorithms proposed in recent years. All algorithms work on different data type so one algorithm may not be used for all applications. On the basis of requirement and compatibility of dataset, an algorithm will be applied. Apriori is a classical algorithm which is use for all dataset. But there are some shortfalls of Apriori algorithm. Our, proposed Data segregation algorithm is introduced to address the effect of these short comings of apriori. The comparison of two algorithm studied on the basis of some criteria discussed in the paper.

Keywords: comparative analysis, data segregation model, cloud, data mining, methodology, Apriori algorithm.

I. INTRODUCTION

History of cloud computing:

Just a few decades ago, people used a pen and paper to write letters, notes or even books of accounts or big and small enterprises. These books where then kept in a safe place and preserved for future references. Picture were to be printed on a photo paper and stored in albums for years, and videos were only seen if they were recorded in large book sized cassettes. In short every means of data storage consumed a lot of space and again it was unreliable. Then came the era of desktop computers, where anyone could generate data digitally and store it in the form of small or large files, depending on their type, on their personal computers. And suddenly life was so easy for all. Larger and small business enterprises used such computers to save their day to day data locally. But with advancement of technology and new software and advanced hardware become cheaper and easily affordable by everyone. As a result data generation capacity of people and organisations greatly increased and data was generated in millions of Gigabytes daily. Soon storage capacity of our local computers and servers started filling fast and people had to invest heavily in buying more data storage space. Also too much reliability on local servers and computers raised questions of reliability and security of the sensitive and important data.

Just when we all started facing the heat of all of these, Cloud computing came as a one stop solution to all our problems. “Cloud” in cloud computing can be defined as the set of hardware, networks, storage, services, and interfaces that combine to deliver aspects of computing as a service [1]. Cloud services include the delivery of software, infrastructure, and storage over the Internet based on user demand. Cloud computing opened doors to limit less opportunities and to trillions and trillions of terabytes of storages space at rates affordable by one and all. Not only does Cloud computing offers unlimited storage capacity but also provides advantages like lower investments in computer hardware especially the cost of hardware upgrades due to advancements in technology, improvement in work efficiency, increased data reliability etc [2]. There are hundreds of cloud storage providers on the Web like Google Drive, Dropbox Yahoo Mail, Picassa, You Tube etc. With all these advantages came a few disadvantages like data redundancy, data hording, data security and portability [3].

Today we have a choice whether to keep a piece of data on our personal computers or local servers or to push it on the cloud servers. Large organisations are always in a dilemma about where it would be safe and secure to store sensitive data that can be accessed only by a few people and common data or reports that needs to be accessed by hundreds of employees. In large businesses where hundreds of files are generated, modified and accessed daily by several people together, data processing, storage and fetching of stored data becomes extremely important as it will affect the performance and efficiency of expensive hardware that is being is used to manage this humongous data [4].

In our research we have made an attempt to find a solution to this concern by designing a hybrid model. This model will make the best use of both the local servers and the cloud and independently decide on the storage location of a data.

II. METHODOLOGY:

We propose a data segregation model to overcome these disadvantages. In our model there are seven components viz. temporary memory, data storage history, data access history, pattern detection manager, storage strategy selector, data access predictor, request generator and data propeller. Each component has a different role to play in the data segregation process. These seven component can collectively be called as Data Segregation Server. Based on a predefined parameters, this Data storage server first arranges the oncoming data in a pattern and then decides whether it should save a particular data or a set of data on the local server or should the data be pushed on to the cloud [5].

Since the data will be arranged in a particular pattern or order and saved in location decided by our by our proposed model, it will make data fetching faster and more efficient, so much so that the need for prefetching of data will be almost negligible. To make this possible we have to make use of a data mining algorithm and hence we propose our Data Segregation Algorithm which will be the heart of our Data Segregation Server. There are several data mining algorithms that arrange the data in a particular pattern and help in selection of a particular type of data based on predefined criteria. All though each algorithm has its own benefits, but each algorithm has a few shortfalls as well.

Over the course of our research we compared 4 of the most popular data mining algorithms viz. KNN, Apriori, K-mean and EM and below is a table which briefly describes this comparison.

Table 1. Comparison of Algorithms

Parameters	**Apriori**	**k-means**	**kNN**	**EM**
Large datasets	Yes	Yes	No	Yes
Easy implementation	Yes	Yes	Yes	Yes
Frequent item set	Yes	No	No	No
Works on any data set	Yes	No	Yes	Yes

One of the most popular and robust algorithms is Apriori. As it can be seen from the comparison chart that Apriori is better than other algorithms in more than one ways no wonder, it is also one of the most widely used algorithms in data mining. But unfortunately, Apriori algorithm loses efficiency when the database is huge. [6]

III. INTRODUCTION TO APRIORI ALGORITHM:

Apriori is a seminal algorithm for finding frequent item sets using candidate generation. It is characterized as a level-wise complete search algorithm using item sets. Apriori uses a "bottom up" approach, where frequent subsets are extended one item at a time (a step known as candidate generation), and groups of candidates are tested against the data. The algorithm terminates when no further successful extensions are found. The pattern of this algorithm is as such that it scans the database at every level or step and rearranges it after discarding the unselected data. Since there are multiple scans involved in every process of data segregation it consumes more time if dataset is large. [7]

Steps of Apriori algorithm:

1. Find all large 1-itemsets. (scan the DB, and count the number of times that item appears in a basket)
2. Initialize count of candidate with zero
3. Find out frequency of each candidate.
4. Discard candidates whose frequency count < minimum support.
5. Repeat process until large-item set is empty.

IV. DATA SEGREGATION ALGORITHM:

We propose a Data Segregation Algorithm (DSA) which addresses this shortfall of Apriori Algorithm and is more efficient. It will allow a user to set one or more criteria. Based on this criteria DSA will first arrange the data in a particular pattern and then in the next step, discard the data which does not meet the user defined criteria. The process will terminate when all the entire dataset is filtered and the desired data is identified. Since now our Data Segregation Server knows which data to push it will in turn go ahead and save the data to either local server or cloud, as decided by the DSA. For the scope of our research we have kept frequency as the criteria of data segregation. So a user can set a minimum frequency of the number of times a particular type of data is accessed by one or more users. For example a monthly sales report of a company like Hindustan Liver, can be accessed by several hundred sales staff of the company and several times at the end of each month, as opposed to an yearly balance sheet which might be seen only by a handful of people in the top management and accounts department. So once the algorithm finds the data that meets the minimum threshold frequency it will save that data on the location either set by the user or it will automatically decide by the Data Segregation Server based on the past records available with it. Our algorithm is designed in such a way that unlike Apriori, the database will not be scanned on each step and so time taken to execute this algorithm comes out to be lesser. Due its efficient and faster processing times this algorithm can work accurately on large dataset

Steps to generate pattern in DSA:

Arrange candidates of each transaction in either alphabetical or ascending order of their occurrence.
Find out frequency of each candidate.
Discard candidates whose frequency count < minimum support.
Rearrange candidates in ascending order (candidates whose frequency count is maximum will come first) in each transaction.
Find frequency of 2-3 combination of candidates.
Discard combination candidates whose frequency count < minimum support. [8].

V. COMPARISON OF DSA AND APRIORI ALGORITHM

To compare DSA with Apriori we have done the following experiment wherein we have taken 7 sets of varied number of transactions. And then we have run both the algorithms on the entire database one by one to check their performance.

Below is an experiment which we conducted to compare the two algorithms, expressed in the form of tables and a graph.

Table 2. Set of Transactions

Transactions	Dataset
T1	15 transactions
T2	30 transactions
T3	40 transactions
T4	70 transactions
T5	100 transactions
T6	150 transactions
T7	200 transactions

Table 3. Time taken to run two algorithms

Transactions	Data Segregation algorithm (in sec)	Apriori Algorithm (in sec)
T1	31	33
T2	47	49
T3	63	67
T4	125	131
T5	62	65
T6	94	98
T7	150	163

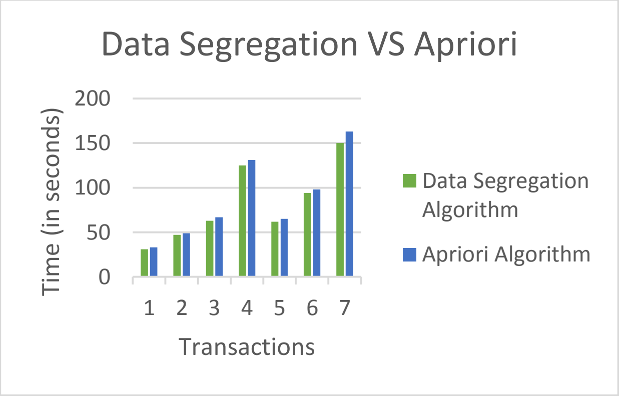

Figure. 1 Column chart representing Dara segregation vs. Apriori algorithm

VI. RESULT AND DISCUSSION:

We can see in the first comparison chart i.e. Table-1 that Apriori is a better algorithm amongst other algorithm. Apriori algorithm can work on frequent item set and work on any type of dataset. But due to the reasons mentioned above there are some shortcomings of Apriori which makes it slower on large databases.

Figure3 shows a comparison between DSA and Apriori of the time taken by each algorithm to perform Data segregation on the given dataset. Figure1 shoes the graphical representation of comparison between Apriori and DSA. With the help of statistics we can conclude that our proposed Algorithm is faster than Apriori in performing data segregation on a given database. And when we extrapolate the database to millions of entries we can see a major difference in the performance of the two algorithms.

Table 4. Summary of Comparison of DSA and Apriori

Parameters	**Apriori Algorithm**	**Data segregation Algorithm (DSA)**

Efficient in large datasets	No	Yes
Easy implementation	Yes	Yes
Frequent item set	Yes	Yes
Works on any data set	Yes	Yes
Database scan at every level	Yes	No
Time consuming process	Yes	No

VII. CONCLUSION:

In this paper there is comparative analysis of Apriori and Data Segregation algorithm. The comparison table shows that the Apriori algorithm outdoes other algorithms in cases of closed item sets whereas Data Segregation algorithm displayed better performance in all the cases.

REFERENCES

[1] Rajleen Kaur, Amanpreet Kaur, "A Review Paper on Evolution of Cloud Computing, its Approaches and Comparison with Grid Computing", (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (5) , 2014.

[2] Abdulaziz Aljabre ," Cloud Computing for Increased Business Value", International Journal of Business and Social Science Vol. 3; January 2012.

[3] T. Dillon, C. Wu, and E. Chang, "Cloud Computing: Issues and Challenges," 2010 24th IEEE International Conference on Advanced Information Networking and Applications(AINA), pp. 27-33, DOI= 20-23 April 2010

[4] Santosh Kumar , R. H. Goudar, "Cloud Computing – Research Issues, Challenges, Architecture, Platforms and Applications: A Survey", International Journal of Future Computer and Communication, Vol. 1, No. 4, December 2012.

[5] Heli P. Vyas, Sanjay M. Shah," A New Perception in Cloud Computing: Hybrid Model", International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169 Volume: 3 Issue: 9 5702 – 5704.

[6] Ahilandeeswari.G.,Dr. R. ManickaChezian," A comparative study of Frequent pattern mining Algorithms: Apriori and FP Growth on Apache Hadoop", International Journal of Innovations & Advancement in Computer Science, ISSN 2347 – 8616, March 2015.

[7] Heli P. Vyas, Sanjay M. Shah," A New Approach: Data Segregation Model", International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169 Volume: 4 Issue: 7, p 38 – 40.

[8] Heli P. Vyas, Sanjay M. Shah," A Comparative Analysis of Frequent Pattern Mining Algorithms", International Journal for Research in Applied Science & Engineering Technology (IJRASET) ISSN: 2321-9653; IC Value: 45.98; Volume 5 Issue XI November 2017

[9] R. V. Dharmadhikari, S. S. Turambekar , S. C. Dolli , P K Akulwar, " Cloud Computing: Data Storage Protocols and Security Techniques", International Journal of Scientific Research in Computer Science and Engineering Vol.6, Issue.2, pp.113-118, April (2018), E-ISSN: 2320-7639.

[10] Rajesh Piplode , Pradeep Sharma and Umesh Kumar Singh, "Study of Threats, Risk and Challenges in Cloud Computing", International Journal of Scientific Research in Computer Science and Engineering, Volume-1, Issue-1, Jan-Feb-2013

Authors Profile

Ms. Heli P. Vyas is currently pursuing Ph.D. from CHARUSAT University, India since 2012 and currently working as Assistant Professor in Masters of Computer Application Department in Chaudhari Technical Institute since 2008. Her main research work focuses on Cloud Computing Data Segregation. She has 10 years of teaching experience and 6 years of research Experience.

Dr. Sanjay M. Shah is currently working as a director in Narsinhbhai institute of Computer Studies and Management, Kadi, India. His main research area is Data Mining and Artificial Intelligence. He has 30 years of teaching experience and 11 years of research experience.