ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 10, October 2015

A Review on Data Mining in Cloud Computing Environment

R.Kabilan, Dr.N.Jayaveeran

Research Scholar, P.G & Research Dept. of Computer Science, Khadir Mohideen College, Adirampattinam, Thanjavur District. India.

Associate Professor and Head, P.G & Research Dept. of Computer Science, Khadir Mohideen College, Adirampattinam. Thanjavur – District. India.

ABSTRACT: Data mining is considered as an important process as it is used for finding new, valid, useful and understandable forms of data. Cloud computing is a resourceful technology that can support a wide range of applications. Data mining techniques and applications can be effectively used in cloud computing paradigm. Data Mining and Cloud Computing are considered as major technologies that can support proper resource sharing. The data mining tasks in cloud computing provides a flexible and scalable architecture which can reduce the cost of infrastructure and storage and used for efficient mining of huge amount of data from virtually integrated data sources with the goal of producing useful information which is helpful in decision making to predict the future trends and behavior. But it has the risk of privacy of data user and security. Recently cloud computing has been facing lots of security issues regarding privacy of data. This paper reviews the concepts of data mining in cloud computing environment.

KEYWORDS: Data Mining, Cloud Computing, Association Rule Mining, Apriori Algorithm, Classification Algorithm, Clustering.

I. INTRODUCTION

A. *Relation between Data Mining and Cloud Computing:*

Data mining is the task of discovering interesting patterns from large amounts of data, where the data can be stored in databases, data warehouses or other information repositories. Various data Mining tasks are: Classification. Clustering, Association Rule mining, Regression, Summarization, Sequence Discovery and Time Series etc. Cloud means network or Internet which is present at remote location. Cloud computing is an internet-based computing in which shared resource, software and information are supplied to computers and other devices on demand It provides services available on Wide Area Network and Local Area Network. Data Mining using cloud computing has become an area because it now covers almost all business and scientific computing. The data mining in Cloud Computing allows organizations to centralize the management of software and data storage, with assurance of efficient, reliable and secure services for their user.[1] The cloud provider provides a tool for data mining for better service. On the other hand outside attackers can use data mining tasks to unauthorized access private data by interacting it. Interaction of data can involve two factors (i) Appropriate quantity of data (ii) Suitable mining algorithms. There are number of mining algorithms which can be used to interact private data and hence threat to data privacy.

II. BACKGROUND

B. *Data Mining in the cloud:*

Using data mining through Cloud Computing reduces the barriers that keep small companies from benefiting of the data mining instruments. The implementation of data mining techniques through Cloud Computing will allow the

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 10, October 2015

users to retrieve meaningful information from virtually integrated data warehouse that reduces the cost of infrastructure and storage.

Data mining in the Cloud "DMCloud" offers tremendous potential for analyzing and extracting the (useful) information in various fields of human activities: business, economics, health care medicines, heredity, biology, pharmacy, advertising, etc. The use of this technology should allow that with just a few clicks of the mouse one can reach the preferred information about clients, behavior, wellbeing, purchasing power, and regularity of purchases of certain items, spot and so on.

DMCloud is, from technical point of view, a very tedious process that requires a special infrastructure based on application of new storage technologies, handling and processing.

The Microsoft suite of cloud-based services includes a new technological preview of DMCloud. It permits to perform some fundamental data mining tasks leveraging a cloud-based Analysis Services connection. It is valuable capability for IWs that would like to begin considering SQL Server Data Mining without the added burden of needing a technology professional to first install Analysis Services. Additionally, IWs can use the DMCloud services no matter where they may physically be located as long as they have an Internet connection.[2] The data mining tasks we can perform with DMCloud are the same Table Analysis Tools found in the traditional Excel Data Mining add-in.

These data mining tasks include:

i) Examine Key Influencers
ii) Notice Categories
iii) Fill From Example
iv) Forecast
v) Highpoint Exceptions
vi) Scenario Analysis
vii) Prediction Calculator
viii) Shopping Basket Analysis.

III. SCOPE OF RESEARCH

A. *Merging data mining and cloud Computing:*

The following figures illustrates the overlap between cloud computing service models and the data mining techniques in the merging data mining and cloud computing environment.

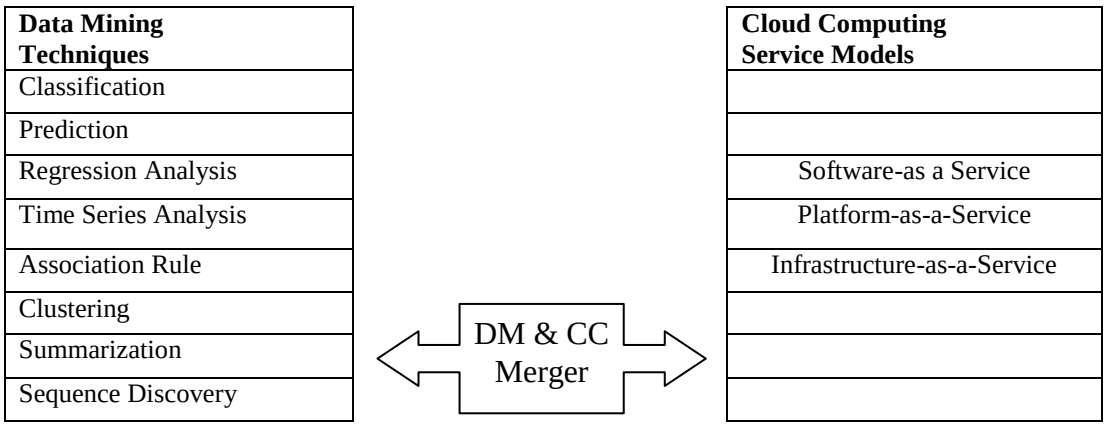

Fig.1. Data Mining and Cloud Computing incorporation

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 10, October 2015

Data mining methods and application are very essential in cloud computing field. It is explored that how the data mining techniques like classification, clustering, association rule analysis etc, are used in cloud computing models such as SaaS, PaaS and IaaS to extract the information.

Cloud provides a benefit for small sized companies to have an opportunity to rent a cloud service for efficient analysis of all the data in the organization which was earlier reserved only for big companies.

Data Mining is preferably used for a large amount of data and related algorithms often require large data sets to create quality models. Cloud providers use data mining to provide clients better services such as data extraction, infrastructure.

B. *Advantages of Data Mining in Cloud Computing:*

Data mining in the cloud computing environment can be considered as the future of data mining because of the advantages of cloud computing paradigm. Cloud Computing provides greater capabilities in data mining and data analytics. The major concern about data mining is that the space required by the operations and item sets is very large.

The following are the advantages [7] of the integrated data mining and cloud computing environment.

- Virtual computers that can be started with short notice.
- No query structured data.
- Message queue for communication.
- The customer only pays for the data mining tools that he needs.
- The customer doesn't have to maintain a hardware infrastructure as he can apply data mining through a browser.
- Redundant robust storage.

C. *Disadvantages of Data Mining in Cloud Computing:*

There are certain issues linked with data mining in the cloud computing. The major issue of data mining with cloud computing is security as the cloud provider has complete control on the underlying computing infrastructure. The attackers can use cheap and raw computing for hacking the data base in the storage of the cloud so the data can be loss in server[14]. Special care has to be taken so as to ensure the security of data under cloud computing environment.

IV. REVIEW OF DATA MANAGEMENT IN CLOUD COMPUTING

A. *Data Management in cloud:*

Data Centers fundamentally are now reliant on cloud computing to deliver the flexibility, scalability, efficiency and speed to the data. Completion of the data in modern ways helps to explore businesses unparalleled opportunities to understand and respond to the needs of a rapidly moving and changing market. Business analysts need access to a wide variety of data, in real time, for diverse uses such as systematic and focused reporting, online analytical processing (OLAP), advanced analytics such as data mining and data marts, and staging for detailed or real-time data sources. The strategic requirement for real-time data warehousing also adds online transaction processing (OLTP) workloads to the mix, increasing the strain on existing infrastructures. In today's world now we see a large amount of data is being used and therefore all accommodation possibilities have to be made for this huge volume of data. So a number of data centers are now making use of the cloud just to have a pool of resources. For data processing and analysis, having a shared, standardized, and consolidated database architecture for all Data Warehouse and online transaction processing workloads is an effective strategy.[12]

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 10, October 2015

V. REVIEW OF DATA MINING ALGORITHMS IN CLOUD COMPUTING WITH RESULTS

Cloud computing provides innovative ways for data mining algorithms. Combining data mining algorithms with cloud computing technology will become the future research trend. Increasingly algorithms to be parallelized and transplant to the cloud computing platform. Theory and experiments show that data mining algorithm play a vital role mining the valuable data under cloud computing environment. It produces theory and cost-effective significance for academic research and commercial operations. Distributed and parallel data mining algorithms can be used for sharing of resources in Cloud Computing.

A . *Data Stream Mining Algorithms:*

Data Stream Mining is the method of extracting knowledge structures from continuous, rapid data records. A data stream is an ordered sequence of instances. In many applications, it can be read only once or minimum number of times using limited computing and storage capabilities. It is a sequence consisted of data which arrives by the time sequence. It is continuous, dynamic data. Unlike traditional static data, its collection process and data mining process are carried out at the same time. Many researches improving the existing data mining algorithms to make it suitable for cloud computing environment. There are different algorithms namely parallel association rule mining algorithm, parallel k-means clustering algorithm and MapReduce parallels Naive Bayes text classification algorithm on Hadoop platform.[2]

B. *Association Rule Mining Algorithms:*

The main purpose is to discover rules linked with regularly co-occurring items, used for market basket analysis, root-cause analysis and cross-sell. The reason is to precious information which describes links between data items from large volume of data.[2] Most commonly used algorithm is Apriori. It finds all the frequent item sets by scanning the database time after time and it will consume a lot of time and memory space when scanning the database with mass data which will become the bottleneck of Apriori algorithm to realize the parallelization.

C. *Classification Algorithms:*

The most commonly used technique for predicting a specific outcome such as high/medium/low value customer, response or no response, likely to buy or not buy. Its purpose is to propose a classification function/model which can map the data item in the database to one of the given categories.[2] Decision tree algorithms originally intended for classification

D. *Clustering:*

The main purpose is to discover group of objects such that the objects in a group will be similar to one another and different from the objects in other groups. The K-mean clustering algorithm can be used to cluster the huge dataset into smaller cluster. There are different clustering algorithms namely K-means and its variants, Hierarchical clustering and density based clustering.[2]

Authors	Year	Algorithm	Results
Lijuan Zhou Xiang Wang	2014	FP-Growth Algorithm Based on Cloud Environments	This algorithm improves the shortcomings of the traditional FP-Growth algorithm.
Harneet Khurana, Kailash Bahl	2014	An Approach to Mine Frequent Itemsets in Cloud Using Apriori and FP-tree Approach	Suitable for parallel computation platform.
Sanjeev Rao, Priyanka Gupta	2012	Improved Algorithm Over APRIORI Data	New scheme for extracting association rules that considers

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 10, October 2015

		Mining Association Rule Algorithm	the time, number of database scans, memory consumption, and the interestingness of the rules.
Bo He	2012	Fast Mining Algorithm of Association Rules Base on Cloud Computing	The algorithm is fast Mining and effective.
A. Mahendiran, N. Saravanan, N. Venkata Subramanian and N. Sairam	2012	Implementation of K-Means Clustering in Cloud Computing Environment	Storing large database with less cost

VI. CONCLUSION

In Cloud computing, the data is being shifted from one server to another server in a peer to peer transaction. Data mining technologies provided through Cloud Computing is an completely essential feature for today's businesses to make practical, knowledge driven decisions as it helps them have future trends and behaviors predicted. Data mining based on Cloud computing is an important characteristic for today's infrastructure to make efficient and better knowledge driven decisions. This paper provides review of data mining concepts in cloud computing and different types of algorithms that can be used for sharing of resources using Data mining and Cloud computing. Both Data Mining techniques and Cloud Computing helps the business organizations to achieve exploited profit and cut costs in dissimilar possible ways.[5]

REFERENCES

1. http://en.wikipedia.org/wiki/Cloud_storage
2. Ms. Aishwarya S. Patil , Ms. Ankita S. Patil A review on Data Mining based Cloud Computing International Journal of Research In Science & Engineering e-ISSN: 2394-8299 Volume: 1 Special Issue: 1 p-ISSN: 2394-8280
3. D. Bhu Lakshmi1 and S. Arundathi, "Providing Privacy and Security for Cloud Data Using Data Mining," International Journal of Innovation and Scientific Research (Vol. 11 No. 2, Nov 2014) ISSN 2351-8014., pp. 264-272.
4. Anuja R. Yeole, Poonam Borkar, Survey Paper on Data Mining in Cloud Computing ,International Journal of Science and Research (IJSR) ISSN (Online): 2319-7064 Index Copernicus Value (2013): 6.14 | Impact Factor (2013): 4.438
5. Mahendiran, N. Saravanan, N. Venkata Subramanian and N. Sairam Implementation of K-Means Clustering in Cloud Computing Environment, Research Journal of Applied Sciences, Engineering and Technology 4(10): 1391-1394, 2012 ISSN: 2040-7467
6. CH.Sekhar. S Reshma Anjum "Cloud Data Mining based on Association Rule"(IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (2) , 2014, 2091-2094
7. Kamala, "A Study On Integrated Approach Of Data Mining And Cloud Mining", International Journal of Advances in Computer Science and Cloud Computing (IJACSCC), Volume- 1,Issue-2,pp 35-38 ,2013.
8. Vanishree K Prakruthi S N Pratiba D " A Study on Association Rules and Clustering Methods" International Journal of Computer Science and Information Technologies, Vol. 5 (1) , 2014, 233-235.
9. Fei Long,Research on algorithms of data mining under cloud computing environment Journal of Chemical and Pharmaceutical Research, 2014, 6(7):1152-1157.
10. Chetna Kaushal, Aashima Arya, Shikha Pathania "Integration of Data Mining in Cloud Computing" Advances in Computer Science and Information Technology (ACSIT) Volume 2, Number 7; April – June, 2015 pp 48 – 52.
11. Praveen Pappula, Ramesh Javvaji, Rama B Experimental Survey on Data Mining Techniques for Association rule mining International Journal of Advanced Research in Computer Science and Software Engineering, Volume 4, Issue 2, February 2014.
12. "Oracle In The Cloud", Oracle Datasheet, www.oracle.com
13. Shweta Singhal , Ankita Sharma "SUBSTANTIAL DATA IN CLOUD COMPUTING" IJICCT–JUL 2013;Vol 1,Issue 1;ISSN 2347-7202
14. Mr.A.Srinivas, M. Kalyan Srinivas, A.V.R.K.Harsha Vardhan Varma"A Study On Cloud Computing Data Mining" International Journal of Innovative Research in Computer and Communication Engineering Vol. 1, Issue 5, July 2013.
15. Sanjeev Rao, Priyanka Gupta "Implementing Improved Algorithm Over APRIORI Data Mining Association Rule Algorithm", IJCST Vol. 3, Iss ue 1, Jan. - March 2012
16. Bo He "Fast Mining Algorithm of Association Rules Base on Cloud Computing" 2nd International Conference on Electronic & Mechanical Engineering and Information Technology (EMEIT-2012)
17. Lijuan Zhou "Research of the FP-Growth Algorithm Based on Cloud Environments" JOURNAL OF SOFTWARE, VOL. 9, NO. 3, MARCH 2014.
18. Harneet Khurana, Kailash Bah "An Approach to Mine Frequent Itemsets in Cloud Using Apriori and FP-tree Approach" International Journal of Computing and Technology, Volume 1, Issue 7, August 2014