ISSN:0975-9646

Cloud Data Mining based on Association Rule

CH.Sekhar[1,] S Reshma Anjum[2]

[1,2] *Sr.Asst Prof Dept of CSE Vignan's Institute Of Information Technology, Duvvada, Visakhapatnam, Andhra Pradesh*

Abstract: This paper describes how data mining is used in cloud computing. Data Mining is used for extracting potentially useful information from raw data. The integration of data mining techniques into normal day-to-day activities has become common place. Every day people are confronted with targeted advertising, and data mining techniques help businesses to become more efficient by reducing costs.

Cloud computing provides a powerful, scalable and flexible infrastructure into which one can integrate, previously known, techniques and methods of Data Mining. This paper also discusses such technology- the technology of big data mining, known as cloud data mining (CDM). Data security and access control are the most challenging in cloud computing because users send their sensitive data to the cloud service providers. The service providers must have a suitable way to protect their client's sensitive data.

Association rules are dependency rules which predict occurrence of an item based on occurrences of other items. Apriori is the best-known algorithm to mine association rules.

Keywords: Cloud computing, data mining, Cloud Data Mining, Association, Apriori.

1. INTRODUCTION:

Every day business activities specially, in recent years, as a consequence of using on-line services, enormous amounts of data are being created. All this accumulated data is potentially hiding in (useful) information, such as the buying preferences, financial situation, interests, political views etc. of users or clients, which can significantly improve the decision-making. Cloud infrastructure can be effectively used for intensive and demanding operations with data that is typical for processes of data mining to get the potentially hidden useful information. Although cloud computing is a powerful means of achieving high storage and computing services at a low cost, it has not lived up to its reputation. Many potential users and companies yet lack interest in cloud based services. One of the main reasons behind this lack of interest involves security issues. As the data owners store their data on external servers, there have been reportedly increasing demands and concerns for data confidentiality, authentication and access control. Earlier to the expansion of the concept of cloud computing, crucial industrial data used to be stored internally on the storage media, protected by security measures including firewalls, to avoid external access to the data and including organizational policy to ban unauthorized internal access.

The use of Cloud Computing is gaining popularity due to its mobility, huge availability and low cost. At an equally significant extent in recent years, data mining techniques have evolved and become more used and discovering knowledge in databases becoming increasingly vital in various fields: business, medicine, science and engineering, spatial data etc.

2. CLOUD COMPUTING:

The Cloud, as it is often referred to, involves using computing resources – hardware and software – that are delivered as a service over the Internet. Cloud computing represents both the software and the hardware delivered as services over the Internet. Cloud Computing is a new concept that defines the use of computing as a utility, that has recently attracted significant attention. NIST defines cloud computing as a model that provides ubiquitous, simple and on demand network access to a shared set of resources (e.g. network resources, servers, data storage, applications and services) that can be readily available for use or if necessary shut down and **all with minimal intervention of service providers.** This cloud model is composed of five essential characteristics, three service models, and four deployment models." The essential characteristics of cloud computing are on-demand self-service, broad network access, resource pooling, rapid elasticity and measured service.

The service models that compose cloud computing are:

Software as a Service (Saas) –
A technology platform that allows access to applications via the internet in the form of services.

Platform as a Service (Paas) - this model allows the user to build his own applications that run on the provider's infrastructure.

Infrastructure as a Service (IaaS) – providers provide the **ability** to use computer infrastructure. Users do not buy the servers, software, data storage or network equipment.

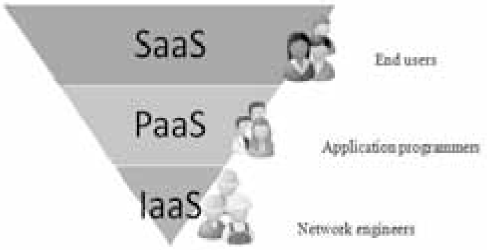


Fig2.1 Cloud Services

The deployment models of cloud computing are:

Private Cloud - Cloud computing infrastructure is accessible to only one organization. It can be managed by the organization itself or by someone

Public Cloud - Platform available and open to the public, regardless of whether they are individuals or organizations and hybrid cloud.

Community Cloud - Model of implementation that provides the ability for more organizations to share the same Cloud computing structure.

Hybrid cloud – It is a model which consists of two or more previously discussed types of the cloud computing structure.
Thus, cloud computing represents all possible resources on the internet, offering infinite computing power.

3. DATA MINING:

Data mining is carried out over large volumes of data in order to pull “new information out of them that will be the basis for making (better) business decisions.

DM is highly multidisciplinary field, which has its roots in statistics, mathematics, information theory, artificial intelligence, machine learning theory, data bases and in the whole series of other related fields. DM involves activities of searching large databases and data warehouses with the aim to find the hidden, so far unknown facts, regularities or patterns. Data mining represents finding useful patterns or trends through large amounts of data. “Data mining, the extraction of hidden predictive information from large databases, is a powerful new technology with great potential to help companies focus on the most important information in their data warehouses. Data mining tools predict future trends and behaviors, allowing businesses to make proactive, knowledge-driven decisions.

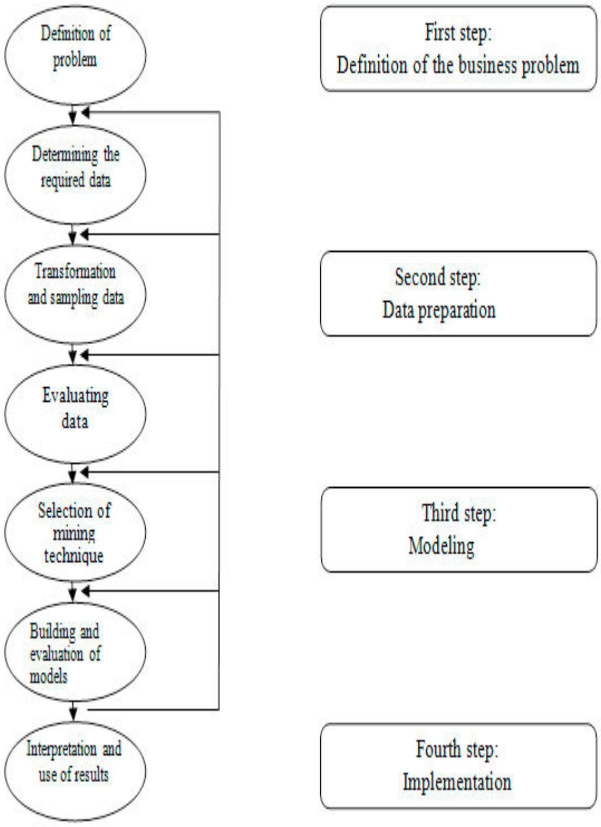

Definition of problem
Determining the required data
Transformation and sampling data
Evaluating data
Selection of mining technique
Building and evaluation of models
Interpretation and use of results
First step: Definition of the business problem
Second step: Data preparation
Third step: Modeling
Fourth step: Implementation

Fig3.1Data Mining Process steps

3.1Data Mining Techniques:

Data Mining Name	*Key Features*
Clustering	Useful for exploring data and finding natural groupings. Members of a cluster are more like each other than they are like members of a different cluster. Common examples include finding new customer segments and life sciences discovery.
Classification	Most commonly used technique for predicting a specific outcome such as response/no-response, high/medium/low value customer, likely to buy/not buy.
Association	Find rules associated with frequently co-occurring items, used for market basket analysis, cross-sell, root cause analysis. Useful for product bundling, in-store placement, and defect analysis.
Regression	Technique for predicting a continuous numerical outcome such as customer lifetime value, house value, process yield rates.
Attribute Importance	Ranks the attributes according to strength of relationship with target attribute. Use cases include finding factors most associated with customers who respond to an offer, factors most associated with healthy patients.
Anomaly Detection	Identifies unusual or suspicious cases based on deviation from the norm. Common examples include health care fraud, expense report fraud, and tax

4. CLOUD DATA MINING:

Data mining techniques and applications are very much needed in the cloud computing paradigm. As cloud computing is penetrating more and more in all ranges of business and scientific computing, it becomes a great area to be focused by data mining. Data mining in cloud computing is the process of extracting structured information from unstructured or semi-structured web data sources.
The data mining in Cloud Computing allows organizations to centralize the management of software and data storage. Using data mining through Cloud Computing reduces the barriers that keep small companies from benefiting of the data mining instruments. 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 users. The implementation of data mining techniques through Cloud Computing will allow the users to retrieve meaningful information from virtually integrated data warehouse that reduces the cost of infrastructure and storage.
CDM (Cloud Data Mining) offers tremendous potential for analyzing and extracting the (useful) information in various fields of human activities: finance, banking, medicine, genetics, biology, pharmacy, marketing, etc. The application of this technology should enable that with just a few clicks of the mouse one can reach the desired information about customers, their habits, interests, purchasing power, frequency of purchases of certain items, location and so on.

Cloud provides technology that can "handle" huge amounts of data, which cannot be processed efficiently and at reasonable cost using standard technologies and techniques. Data mining in Cloud (CDM) is, from a technical point of view, a very tedious process that requires a special infrastructure based on application of new storage technologies, handling and processing. Big Data/Hadoop is the latest type in the field of data processing.

5. INTEGRATED DATA MINING AND CLOUD COMPUTING:

Data mining in Cloud Computing allow the organizations to centralize the management of software and data storage with assurance of efficient, reliable and secure services for their users. It provides technology that can handle large amounts of data which cannot be processed efficiently at reasonable cost using standard technologies and techniques. It also allows the users to retrieve meaningful information from virtually integrated data warehouse that reduces the cost of infrastructure and storage. We can provide new ways and means to effectively solve the distributed storage of massive data mining and efficient computing through Cloud Computing, mass data storage and distribution of computing, massive data mining environment for cloud computing. Extension of Cloud Computing will drive the Internet and technological achievements in the public service to promote the depth of information resources sharing and sustainable use of new methods and new ways of traditional data mining. 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 users.

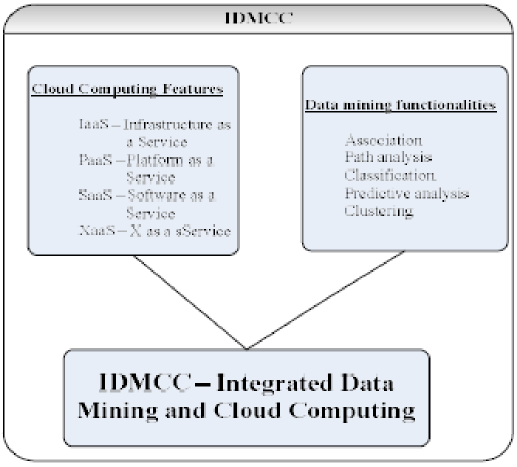

IDMCC
Cloud Computing Features
IaaS – Infrastructure as a Service
PaaS – Platform as a Service
SaaS – Software as a Service
XaaS – X as a sService
Data mining functionalities
Association
Path analysis
Classification
Predictive analysis
Clustering
IDMCC – Integrated Data Mining and Cloud Computing

Fig 5.1 IDMCC Integration

5.1 Advantages of IDMCC Integration:

The following are the advantages of the Integrated Data Mining and Cloud Computing Environment.

- Virtual computers that can be started with short notice
- Redundant robust storage
- 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

5.2 Advantages Of Using Data Mining With Cloud Computing:

Cloud Computing combined with data mining can provide powerful capacities of management. Due to the explosive data growth and amount of computation involved in data mining, an efficient and high performance computing is an excellent resource necessary for a successful data mining application. 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.

5.3 Disadvantages Of Using Data Mining With Cloud Computing:

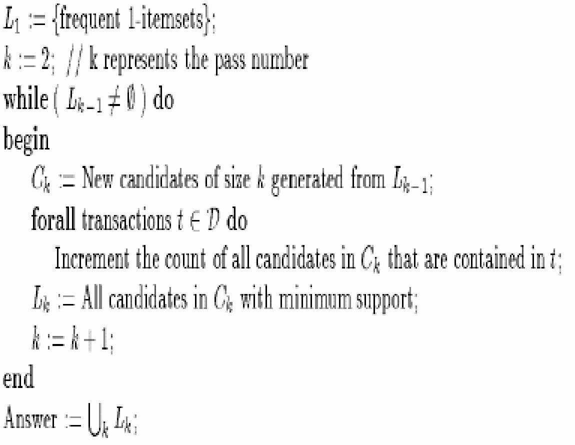
There are certain issues associated with data mining in the cloud computing. The major issue of data mining with cloud computing is security as the cloud provider has complete control on the underlying computing infrastructure. Special care has to be taken so as to ensure the security of data under cloud computing environment.

6. ASSOCIATION RULES:

Association rule is very popular and well researched method for discovering interesting relations between variables in large databases. Given a set of transactions, where each transaction is a set of items, an association rule is an expression X => Y, where X and Y are sets of items. The intuitive meaning of such a rule is that transactions in the database which contain the items in X tend to also contain the items in Y. An example of such a rule might be that 98% of customers who purchase tires and auto accessories also buy some automotive services; here 98% is called the confidence of the rule. The support of the rule X => Y is the percentage of transactions that contain both X and Y. The problem of mining association rules is to find all rules that satisfy a user-specified minimum support and minimum confidence. Applications include cross marketing, attached mailing, catalog design, loss-leader analysis, add-on sales, store layout, and customer segmentation based on buying patterns. The problem of mining association rules can be decomposed into two sub problems: 1. Find all sets of items (item sets) whose support is greater than the user-specified minimum support. Item sets with minimum support are called frequent item sets. 2. Use the frequent item sets to generate the desired rules. The general idea is that if, say, ABCD and AB are frequent item sets, then we can determine if the rule AB => CD holds by computing the ratio conf = support(ABCD)/support(AB). If conf >= minimum confidence, then the rule holds. That is, the rule will have minimum support because ABCD is frequent. Much of the research has been focused on the first sub problem as the database is accessed in this part of the computation and several algorithms have been proposed. In association rule mining algorithm, most of the algorithms are based on Apriori algorithm to calculate and in the mining process they can produce amount of option set which reduce the efficiency of association rule mining.

6.1 APRIORI ALGORITHEM:

Apriori is designed to operate on databases containing transactions. Apriori uses a "bottom up" approach, where frequent subsets are extended one item at a time and groups of candidates are tested against the data. The algorithm terminates when no further successful extensions are found. Apriori uses breadth-first-search and a tree structure to count candidate item sets efficiently. It generates candidate item sets of length k from item sets of length k - 1. Then it prunes the candidate which has an infrequent sub pattern. According to the downward closure lemma, the candidate set contains all frequent k-length item sets. After that, it scans the transaction database to determine frequent item sets among the candidates. The above code gives an overview of the Apriori algorithm. The first pass of the algorithm simply counts item occurrences to determine the frequent item sets. A subsequent pass, say pass k, consists of two phases. First, the frequent item sets Lk-1 found in the (k-1) the pass are used to generate the candidate item sets Ck, using the Apriori candidate generation procedure. Next, the database is scanned and the support of candidates in Ck is counted. For fast counting, we need to efficiently determine the candidates in Ck contained in a given transaction t. A hash-tree data structure is used for this purpose.



```
$L_1$ := {frequent 1-itemsets};
$k$ := 2; // k represents the pass number
while ( $L_{k-1} \neq \emptyset$ ) do
begin
    $C_k$ := New candidates of size $k$ generated from $L_{k-1}$;
    forall transactions $t \in \mathcal{D}$ do
        Increment the count of all candidates in $C_k$ that are contained in $t$;
    $L_k$ := All candidates in $C_k$ with minimum support;
    $k$ := $k + 1$;
end
Answer := $\bigcup_k L_k$;
```

7. CONCLUSION:

Cloud Computing provides storage of data in a server by protecting data by using data mining concept. Actually, we are discussing the cloud computing data mining for the advance use of security in data loss purpose. 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 absolutely necessary characteristic for today's businesses to make proactive, knowledge driven decisions as it helps them have future trends and behaviors predicted.

REFERENCES:

[1] Mell, P. and Grance, T. (2011). The NIST Definition of Cloud Computing (Draft). Recommendations of the National Institute of Standards and Technology, NIST.

[2] ORACLE, "Oracle Data Mining Techniques and Algorithms"

[3] Pejić Bach, M. (2005). Data mining in the banking industry. Proceedings of Faculty of Economics, University in Zagreb.

[4] http://samisa-abeysinghe.blogspot.com /2011/07/cloud-computing-explained.html (Accessed: July 2012)

[5] Janardhan. N, T. Sree Pravallika, Sowjanya Gorantla, "An efficient approach for integrating data mining into cloud computing", International Journal of Computer Trends and Technology (IJCTT) - volume4 Issue5–May 2013

[6] Rahul Sharma, Rohit Sharma, "Excavate the Cloud via Autonomous agents & Data mining", International Journal of Computer, Electronics & Electrical Engineering, Volume3 – Issue 1

[7] Ruxandra-Ştefania PETRE, "Data mining in Cloud Computing", Database Systems Journal vol. III, no. 3/2012

[8] Ch.Sekhar, U Nanaji- "Secure Cloud by IT Auditing" International Journal of Modern Engineering Research (0975 – 888), Volume 2, Issue Sep 2011

[9] Jing Ding, Shanlin Yang, Classification Rules Mining Model with Genetic Algorithm in Cloud Computing‖, International Journal of Computer Applications (0975 – 888), Volume 48– No.18, June 2012.

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

AUTHOR'S

Mr.CH. Sekhar received the B.Tech Degree from JNTU, Hyderabad in 2005 and M. Tech Degree in Computer Science Engineering JNTU, Kakinada in 2011.He is working as a Sr. Asst Prof in CSE Dept, Vignan's IIT, Vizag and also worked as Associate Prof in Avanthi Groups of College from Sep 2005 to May 2012. Research interests include Data Mining, Cloud Computing and Computer Networks.

Mrs. Shaik Reshma Anjum received the MCA P.G from Shadan College, Osmania University in 2004 and M. Tech degree in Computer Science Engineering, GITAM University, Visakhapatnam in 2012. She is working as Assistant Professor in CSE Dept, Vignan's IIT,Vizag and also worked as Sr.Lecturer in CSE Dept, R G Kedia P G College from 2004 to 2007.She is an Oracle Certified Associate. Research interests include Data Mining, Cloud Computing and Data bases.