IJISET - International Journal of Innovative Science, Engineering & Technology, Vol. 2 Issue 11, November 2015.
www.ijiset.com
ISSN 2348 – 7968

An Integrated Cloud Data Mining Using APRIORI

Sumalatha Potteti[1], Ravi Kumar G[2]
[1]Assistant Professor, Department of CSE,BRECW,Hyderabad,India,sumalatha.po@gmail.com
[2]Assistant Professor, Department of CSE,BRECW,Hyderabad,India,ravisslm@gmail.com

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.

Key words: Cloud Computing, Data Mining.

1. INTRODUCTION:

The increasing ability to generate vast quantities of data brings potentials to discover and utilize valuable knowledge from data. Data mining has been an effective tool to analyze data from different angles and getting useful information from data. It can also help in predicting trends or values, classification of data, categorization of data, and to find correlations, patterns from the dataset. On the other hand, utilizing the vast amount of data presents technical challenges as data storage and transfer approaches needs to deal with prohibitive amounts of data. The management of data resources and data flow between the storage and compute resources is becoming the main bottleneck. Analyzing, visualizing, and disseminating these large data sets has become a major challenge and data intensive computing is now considered as the "fourth paradigm" in scientific discovery after theoretical, experimental, and computational science [15].

To shift to this fourth paradigm, researchers have to plan strategically to perform data analysis. Cloud computing, a business model containing a pool of resources, provides an effective paradigm for this purpose. In this context, cloud computing is a distributed computing paradigm that enables large datasets to be sliced and assigned to available computer nodes where the data can be processed locally, avoiding network-transfer delays. This makes it possible for people to understand and utilize the trillions of rows of information in a data center. We strongly believe that cloud computing can be an effective platform for data mining. To gain more experience in cloud-assisted data mining, in this paper, we use association rule based algorithm, Apriori [5], as an example to study how data mining algorithms can be adjusted to fit the increasing demand for parallel computing environment of cloud.

2. CLOUD COMPUTING:

Cloud computing is an interactive communication model that is constituted in more than one place synchronously, easy to use, can be accessed whenever user needs, consist of configurable computing resources and needs minimal effort to achieve maintainability [5]. Nowadays cloud computing users are using services that they need from providers' computing resources and charged as they profited [6]. Cloud computing has many definitions in different resources, which are similar to each other. As a summary, cloud computing can be defined as today's computing technology that as time and location independent services, shaped with user's needs, has a minimum effort to maintain and charged as the service usage.

2.1 Structure of Cloud Computing Architecture:

Cloud computing architecture contains some types of actors, which can be either an individual or an organizational unit who attend cloud services/tasks. NIST defines five main actors [7]:

Consumer uses the cloud computing service and can be either an individual or an organizational unit. A consumer chooses the most appropriate service or services, which are provided by the cloud provider. Besides, the services are charged against to the agreement that is signed between the consumer and the provider.

Provider is an entity that is responsible for developing resources and services, which are used by individuals, organizations or consumers. Provider manages software, platform or infrastructure that is needed by consumers, and it builds obligatory technical infrastructure and provides specified service levels (mostly trust and security levels).

Auditor inspects whole information technology processes, performance and security issues independently within predefined criteria. Auditor must be a third party and can be either an individual or an organizational unit.

Broker manageability of cloud systems is very complicated because of its nature. Consumers can use cloud services not only get in contact with the provider

ISSN 2348 – 7968

directly but also broker. Broker organizes the connection between provider and consumer, and also manages performance and availability of the system.

Carrier realizes connection, communication and transfers between provider and consumer, and also it enables consumers can access to the services over communication infrastructure and other devices such as desktops, laptops or any mobile devices. Distribution of the cloud services can be realized via network and communication infrastructure or communication agents, which have high storage capacity opportunities.

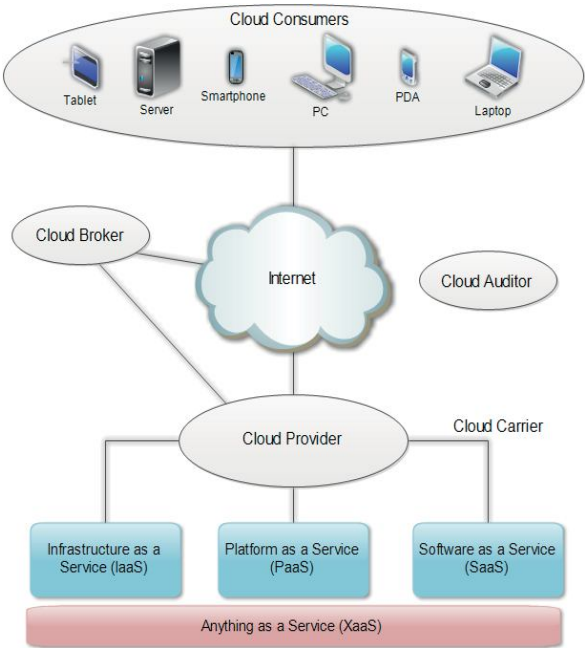

Fig 1: Service models and actors in cloud computing.

2.2 Service Models in Cloud Computing:

Software as a Service (SaaS): These services are applications over Internet. Normally the user can run these applications using a web-browser. User abstract totally about the hardware and software that is using and simply access to a interface with a web browser and from there he have access to some information and functionalities. It‘s dedicated to current users; an example to this kind of services may be Google Docs.

Platform as a Service (PaaS): These services are focused on the deployment of applications or services online letting to the developer manage the hardware or software necessary, including also a solution stack. This service includes all the life-cycle of the deployment of application/ service such as design, implementation, testing, deployment, integrity with databases, etc.

Infrastructure as a Service (IaaS): These services are focused to offer a computer infrastructure. All the servers, connections, software and other resources are offered by the providers. And the users see it like an entire infrastructure hosted in the same organization.

3. CLOUD COMPUTING DEPLOYMENT MODELS

Cloud computing architects give following basic service models:

• Public cloud
• Private cloud
• Hybrid cloud
• Community Cloud

Conveying distributed computing can vary relying upon necessities, and the accompanying four arrangement models have been recognized, each with particular attributes that help the requirements of the administrations and clients of the mists specifically ways. There exist four separate sorts of mists on the groundwork of who claims and utilization them:

1) Public Clouds: A public cloud encompasses the traditional concept of cloud computing, having the opportunity to use computing resources from anywhere in the world. Public clouds are frequently hosted away from customer site, and they provide flexible infrastructure to cut down customer risk and cost.

2) Private Clouds: Private clouds are assembled for utilization of one customer solely, furnishing the most extreme control over information, security, and nature of administration. Here, the organization claims the foundation and has control over how requisitions are circulated on it. Private mists could be manufactured and deliver the goods by an organization's IT association or by a cloud supplier. In this model, an organization can introduce, arrange, and work the base to help a private cloud inside an organization's undertaking data center.

3) Hybrid Clouds: Hybrid clouds join the aspects of both open and private cloud models. They can help to give on-interest, remotely provisioned scale. The capacity to incorporate a private cloud with the assets of an open cloud might be utilized to administer administration levels. A half breed cloud likewise could be utilized to handle arranged workload spikes. Half and half mists present the intricacy of figuring out how to disperse provisions crosswise over both an open and private cloud. The cloud foundation comprises of various billows of any sort, yet the mists have the capability through their interfaces to permit information or provisions to be moved starting with one cloud then onto the next.

4) Community Clouds: In Community Cloud the cloud base is imparted by numerous associations that have imparted contemplations. It is ought to be overseen by the associations or a third gathering and might as well exist on-premises or off-premises.

ISSN 2348 – 7968

4. 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.

Data Mining Techniques and its 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: it 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.

5. 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.

6. 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.

ISSN 2348 – 7968

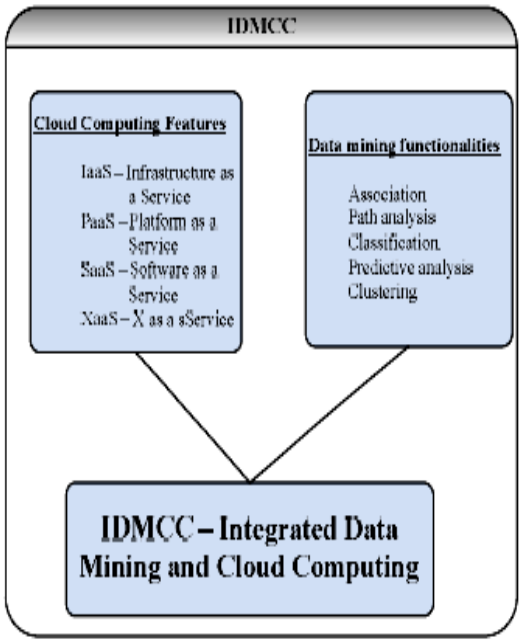

Fig 2 : Integrated Data Mining and Cloud Computing

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.

6.1 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
• The customer doesn't have to maintain a hardware infrastructure as he can apply data mining through a browser

6.2 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.

6.3 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.

7. 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

ISSN 2348 – 7968

amount of option set which reduce the efficiency of association rule mining.

8. APRIORI ALGORITHM:

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.

$L_1 := \{\text{frequent 1-itemsets}\};$
$k := 2;$ // 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 $:= \bigcup_k L_k;$

CONCLUSION:

Cloud Computing has given rise to a new services paradigm to the information technology. 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.

ACKNOWLEDGMENT:

It is not only customary but necessary for a researcher to mention his/her indebtedness to those who had helped in carrying out and enhance the research work.
I pay my deep regards to God, my Parents and my loving Friends for their support and wishes which made this tedious work easy and successful.
Finally, I would like to extend my thanks to all those who have contributed, directly or indirectly to make this project successful.

REFERENCES:

[1] Kleber Vieira, Alexander Schulter,CarlosBecker Westphall, and Carla Merkle Westphall "Intrusion Detection for Grid and Cloud Computing" (IT Professionals, Vol. 12, no. 4, 2010 pp 38-43)

[2] Hisham A. Kholidy, Fabrizio Baiardi, Salim Hariri, Esraa M. Elhariri, Ahmed M. Yousof and Sahar A. Shehata. 2012. A Hierarchical Intrusion Detection System For Cloud: Design and Evaluation. International Journal on Cloud Computing Services and Architecture (IJCSA).

[3] Ahmed Patel, Mona Taghavi, Kaveh Bakhtiyari, Joaquim Celestino Junior. 2013. An intrusion detection and prevention system in cloud computing: A systematic review. Journal of Network and Computer Applications.

[4] S.V. Narwane, S.L. Vaikol. 2012. Intrusion Detection System in Cloud Computing Envirnment .InInternational Conference on Advances in Communication and Computing Technologies (ICACACT).

[5] Ms. Parag K. Shelke, Ms. Sneha Sontakke, Dr. A. D. Gawande. 2012. Intrusion Detection System for Cloud Computing. International Journal of Scientific & Technology Research Volume 1.

[6] Hassen Mohammed Alsafi, Wafaa Mustafa Abduallah and Al-Sakib khan Pathan. 2012.IPS: An Integrated Intrusion Handling Model for Cloud Computing Environment, International Journal of Computing and Information Technology (IJCIT).

[7] Deris Stiawan, Abdul Hanan Abdullah, Mohd. Yazid Idris. 2011. Characterizing Network Intrusion Prevention System International Journal of Computer Applicationas.

[8] V. Jyothsna, V.V. Rama Prasad, K. Munivara Prasad, 2011, A Review of Anomaly Based Intrusion Detection System, International Journal of Computer Applications.

ISSN 2348 – 7968

[9] R. Base and P. Mell.2001. NIST Special Publication on Intrusion Detection Systems. National Institute of Standard and Technology. Dinesh Sequeira. 2002. Intrusion Prevention Systems- Security Silver Bullet?. SANS Institute.

[10] Kazi Zunnurhain, Susan Vrbsky “Security Attacks and Solutions in Cloud”

[11] M. Kuzhalisai and G. Gayathri, "Enhanced Security in Cloud with Multi-Level Intrusion Detection System", IJCCT, Vol. 3, Issue 3, 2012.

[12] Ajeet Kumar Gautam, Dr. Vidushi Sharma, Shiv Prakash and Manak Gupta, "Improved Hybrid Intrusion Detection System (HIDS): Mitigating False Alarm in Cloud Computing", JCT, 2012.

[13] Irfan Gul and M. Hussain, "Distributed Cloud Intrusion Detection Model", IJAST, Vol. 34, September, 2011.

[14] Pradeep Kumar Tiwari and Dr. Bharat Mishra ," Cloud Computing Security Issues, Challenges and Solution " ,IJETAE, Volume 2, Issue 8, August 2012.