ISSN : 2347 - 8446 (Online)
ISSN : 2347 - 9817 (Print)
Vol. 2, Issue 2, Ver. 1 (April - June 2014)
International Journal of Advanced Research in Computer Science & Technology (IJARCST 2014)

Survey on Secure Data mining in Cloud Computing

[I]Uppunuthula Venkateshwarlu, [II]Puppala Priyanka
[I,II]M.Tech,Computer Science and Engineering, JNTUH, Hyderabad, AP, India

Abstract
Data mining techniques are very important in the cloud computing paradigm. The integration of data mining techniques with Cloud computing allows the users to extract useful information from a data warehouse that reduces the costs of infrastructure and storage. Security and privacy of user's data is a big concern when data mining is used with cloud computing. Cloud computing is an emerging computing paradigm in which resources of the computing infrastructure are provided as services of the internet. An important security concern is privacy attacks based on data mining involving analyzing data over a long period to extract valuable information. In this dissertation our main objective is to provide information with the help of which we can make data secure from unauthorized users. As uses of data come on front we have to face concept of data mining. Data Mining is a field where accuracy matters a lot. Data mining techniques and applications are very much needed in the cloud computing paradigm. 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 costs of infrastructure and storage.

Keywords
Cloud Computing, Data mining, privacy preserving.

I. Introduction

The Internet is becoming a surprisingly vital tool in our daily life, both professional and personal, as its users are becoming more numerous. The Cloud, as it is often referred to, involves using computing resources – hardware and software – that are delivered as a service over the Internet. At an equally significant extent in recent years, data mining techniques have evolved and became more used, discovering knowledge in databases becoming increasingly vital in various fields: business, medicine, science and engineering, spatial data etc. It is not surprising that business is increasingly conducted over the Internet. Perhaps one of the most revolutionary concepts of recent years is 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. The Cloud, as it is often referred to, involves using computing resources – hardware and software – that are delivered as a service over the Internet. . The use of Cloud Computing is gaining popularity due to its mobility, huge availability and low cost. Cloud computing represents both the software and the hardware delivered as services over the Internet.

The emerging Cloud Computing trends provides for its users the unique benefit of unprecedented access to valuable data that can be turned into valuable insight that can help them achieve their business objectives. Cloud Computing is a new concept that defines the use of computing as a utility, that has recently attracted significant attention. The use of Cloud Computing is gaining popularity due to its mobility, huge availability and low cost. On the other hand it brings more threats to the security of the company's data and information. Computing is a new concept that defines the use of computing as a utility, that has recently attracted significant attention.

The deployment models of cloud computing are private Cloud, community cloud, public cloud and hybrid cloud. The deployment models of cloud computing are private Cloud, community cloud, public cloud and hybrid cloud. Many companies are choosing as an alternative to building their own IT infrastructure to host databases or software, having a third party to host them on its large servers, so the company would have access to its data and software over the Internet Cloud

Some perspectives regarding cloud mining- Data mining:

Cloud mining represents finding useful patterns or trends through large amounts of data. Data mining is defined as a "type of database analysis that attempts to discover useful patterns or relationships in a group of data. 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. The analysis uses advanced statistical methods, such as cluster analysis, and sometimes employs artificial intelligence or neural network techniques. A major goal of cloud mining is to discover previously unknown relationships among the data, especially when the data come from different databases."

The service models that compose cloud computing are Software as a Service (SaaS), Platform as a Service (PaaS) and Infrastructure as a Service (IaaS). The deployment models of cloud computing are private cloud, community cloud, public cloud and hybrid c loud. Cloud computing represents all possible resources on the Internet, offering infinite computing power. Considering the varied data mining techniques and the great need for discovering patterns and trends in data that would lead to knowledge that could not be obtained otherwise, it's no wonder that data mining is used in the most varies field of activity. As it is defined by the National Institute of Standards and Technology, "Cloud computing is a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction. This cloud model is composed of five essential characteristics, three service models, and four deployment models".

ISSN : 2347 - 8446 (Online)
ISSN : 2347 - 9817 (Print)

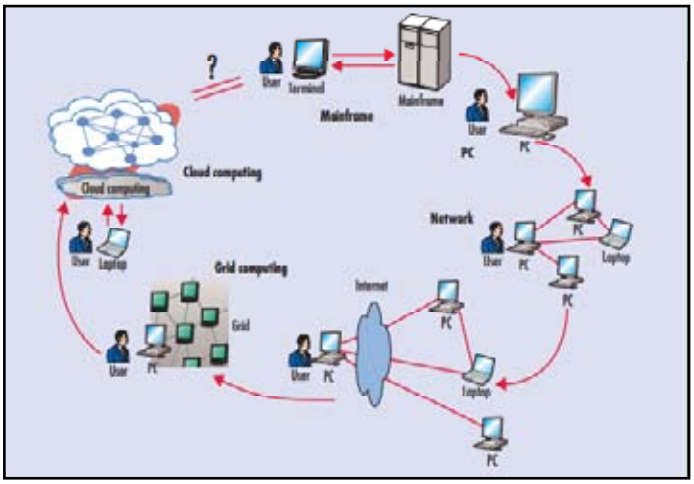

Fig.1: Computing paradigm

In Figure 1 below it is illustrated the computing paradigm shift on the last half century through six distinct phases:

- Phase 1: people used terminals to connect to powerful mainframes shared by many users.
- Phase 2: stand-alone personal computers became powerful enough to satisfy users' daily work.
- Phase 3: computer networks allowed multiple computers to connect to each other.
- Phase 4: local networks could connect to other local networks to establish a more global network.
- Phase 5: the electronic grid facilitated shared computing power and storage resources.
- Phase 6: Cloud Computing allows the exploitation of all available resources on the Internet in a scalable and simple way.

II. Areas For Secure Cloud- Data Mining Application

- Governments can discern illegal or embargoed activities done by individuals, associations or other governments with the implementation of the data mining techniques.
- Businesses can make predictions about how well a product will sell or develop new advertising campaigns by using these new relationships reflected by the data mining algorithms.
- The medical sector benefits from the data mining techniques, as well as the geographical data being better analyzed by using data mining.
- In short, data mining has developed uses in the majority of field of activity.

Cloud has been able to generate and collect large amount of information. The Internet has a great importance in the society providing information exchange and communication environment in trade relations and social interactions. The increasing use of the Internet and the fast advance of new technologies have motivated the development of the Future Internet.

Cloud computing needs to address three main security issues: confidentiality, integrity and availability. Cloud computing has transfigured the way computing and software services are delivered on demand to the clients. Due to the greater level of flexibility, the cloud has become the proliferating ground of a new generation of products and services. However, the flexibility of services of cloud imposes the risk of the security and privacy of users' data. Thus, users of cloud are more concerned about the security of their data and this is becoming a major barrier to the widespread growth of cloud computing. One of the security concerns of cloud is data mining based privacy attacks that involve analyzing data over a long period to extract valuable information. In particular, in current cloud architecture a client entrusts a single cloud provider with his data. It gives the provider and outside attackers having unauthorized access to cloud, an opportunity of analyzing client data over a long period to extract sensitive information that privacy violation of clients. This is a big concern for many clients of cloud.

Therefore the data mining based privacy risks on cloud data must be identified and solution such as a distributed architecture should be used to eliminate the risks. Most of the business organizations try to analyze their data to discover new patterns. Usually, analyzing such amount of data requires huge computational power and storage facilities that may not be available to these organizations. Cloud computing offers the best way to solve this problem. Storing the private data of different in the same cloud server enhances the mining process, but at the same time, raises privacy concerns.

It is, therefore, highly recommended to support privacy preserving data mining algorithms in the cloud environment. To assure privacy preserving data mining in the cloud a solution is needed providing an efficient and accurate cryptography-based scheme for mining the cloud data in a secure way without loss of accuracy. Many companies are opting for cloud storage; hence it is important to use an efficient and effective data mining strategy to mine the cloud storage to extract interesting patterns and relationship between variables in large databases.

These data patterns may be forecasting or predictions that can be used by the companies in near future to increase their sales. The generated predictions which are result of mining should be very well secured from interception. To provide a solution for this privacy concern a Secure Cloud Mining (SCM) architecture that will generate a Secure Forecasting Report (SFR) for the company can be used.

III. Data Mining On Cloud

Data mining is used by cloud providers to provide clients a better service. If the clients are not aware of the information collected through mining then the privacy of the client may be violated. If the cloud providers in any way or means misuse this information, the client privacy is endangered. Again attackers outside cloud providers may have a prohibited access to the cloud, and gain the opportunity to mine cloud data. Attackers can use computing power provided by cloud computing to mine data getting access to useful information from data. Cloud being a massive source of centralized data, data mining gives attackers a great advantage in extracting valuable information and thus violating clients' data privacy.

The implementation of data mining techniques through Cloud computing will provide the users an opportunity to retrieve meaningful information from integrated data warehouse that reduces the costs of infrastructure and storage. The important effect of data mining based cloud computing is that the customer needs to pay only for the data mining tool that he needs. Further the customer need not maintain an infrastructure of as he can use data mining through a browser.

IV. Some Issues of Cloud Computing- Based Data Mining

There are some problems of data mining based on cloud including-

- The design and selection of data mining algorithms.
- Using appropriate algorithms and adopting appropriate parallel strategy can assist in increasing efficiency.
- Setting appropriate parameters is also very important.
- Privacy protection is a very important issue.

A. Client privacy and its importance:

Companies dealing with financial, educational, health or legal issues of people are prominent targets and leaking information of such companies can do significant harm to their customers. Information in this context refers to the financial condition of a customer, the likelihood of an individual getting a terminal illness, the likelihood of an individual being involved in a crime etc.. Sometimes leaking information regarding a particular company leads to a national misfortune.

Data Mining as a threat to client privacy Some mining algorithms allow to extract information up to the limit that violates client privacy. For example, multivariate analysis identifies the relationship among variables and this technique can be used to determine the financial condition of an individual from his buy-sell records, clustering algorithms can be used to categorize people or entities and are suitable for finding behavioural patterns, association rule mining can be used to discover association relationships among large number of business transaction records etc. Thus analysis of data can reveal private information about a user and leaking this sort of information may do significant harm.

Thus, data mining is becoming more powerful and possessing more threat to cloud users. In upcoming days, data mining based privacy attack can be a more regular weapon to be used against cloud users. 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.

"Cloud computing denotes the new trend in Internet services that rely on clouds of servers to handle tasks. 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, with assurance of efficient, reliable and secure services for their users." As Cloud computing refers to software and hardware delivered as services over the Internet, in Cloud computing data mining software is also provided in this way.

B. The main effects of data mining tools being delivered by the Cloud are:

1. The customer only pays for the data mining tools that he needs – that reduces his costs since he doesn't have to pay for complex data mining suites that he is not using exhaustive;
2. The customer doesn't have to maintain a hardware infrastructure, as he can apply data mining through a browser – this means that he has to pay only the costs that are generated by using Cloud computing.

Using data mining through Cloud computing reduces the barriers that keep small companies from benefiting of the data mining instruments. "Cloud Computing denotes the new trend in Internet services that rely on clouds of servers to handle tasks. 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, 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 costs of infrastructure and storage.

C. Cloud mining techniques:

a. 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.
b. Association: Find rules associated with frequently co-occurring items, used for market basket analysis, cross-sell, and root cause analysis. Useful for product bundling, in store placement, and defect analysis.
c. Regression: Technique for predicting a continuous numerical outcome such a customer lifetime value, house value, process yield rates.
d. Attribute Importance: Ranks 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.
e. Feature Extraction: Produces new attributes as linear combination of existing attributes. Applicable for text data, latent semantic analysis, data compression, data decomposition and projection, and pattern recognition.
f. Data mining in Cloud Computing: Data mining techniques and applications are very much needed in the cloud computing paradigm.

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, with assurance of efficient, reliable and secure services for their users." As Cloud computing refers to software and hardware delivered as services over the Internet, in Cloud computing data mining software is also provided in this way.

V. Secure Mining in Cloud

A Protection from Data Mining Based Attack using distributed cloud architecture Data mining can be a potential threat to cloud security because of the fact that entire data belonging to a particular user is stored in a single cloud provider. The provider gets an opportunity due to a single storage provider approach to use powerful mining algorithms or tools that can extract private information of the user. Mining algorithms require a reasonable amount of data as a result of which the single provider architecture suits the purpose of the attackers.

The job of attackers is also eased because of single cloud storage provider approach. These attackers have unauthorized access to the cloud and use data mining to extract information. In this approach data is distributed multiple cloud providers so that data mining becomes a difficult job to the attackers. The key idea of this approach is to categorize user data, split data into chunks and provide these chunks to the proper cloud providers. This approach consists of categorization, fragmentation and distribution of data. The categorization of data is done according to mining sensitivity.

ISSN : 2347 - 8446 (Online)
ISSN : 2347 - 9817 (Print)

Mining sensitivity in this context refers to the significance of information that can be leaked by mining.
A cloud provider is given a particular data chunk only if the provider is reliable enough to store chunks of such sensitivity. Distribution restricts an attacker from having access to a sufficient number of chunks of data and thus prevents successful extraction of valuable information via mining. Even if an attacker manages to access required chunks, mining data from distributed sources remains a challenging job. This distributed approach provides two major benefits First, it improves privacy by making the attacker's job complicated by increasing the number of targets and decreasing amount of data available at each target. Second, it ensures the greater availability of data.

This system consists of two major components:
Cloud Data Distributor and Cloud Providers. The Cloud Data Distributor receives data in the form of files from clients, splits each file into chunks and distributes these chunks among cloud providers. Cloud Providers store chunks and responds to chunk requests by providing the chunks.

i) Cloud Data Distributor
Cloud Data Distributor receives data (files) from clients, performs fragmentation of data (splits files into chunks) and distributes these fragments (chunks) among Cloud Providers. It also participates in data retrieving procedure by receiving chunk requests from clients and forwarding them to Cloud Providers. Clients do not interact with Cloud Providers directly rather via Cloud Data Distributor. To perform distribution and retrieval of data (chunks), the Cloud Data Distributor needs to maintain information regarding providers, clients and chunks. Hence, it maintains three types of tables describing the providers, the clients and the chunks.

ii) Cloud Providers
The important tasks of Cloud Providers are storing chunks of data, responding to a query by providing the desired data, and removing chunks when asked. Providers receive chunks from the distributor and store them. Each provider is considered as a separate disk storing clients' data. Certain factors such as distribution of chunks, maintaining privacy levels, reducing chunk size, addition of misleading data contributes to the effectiveness of the system.

VI. Secure Cloud Mining Architecture

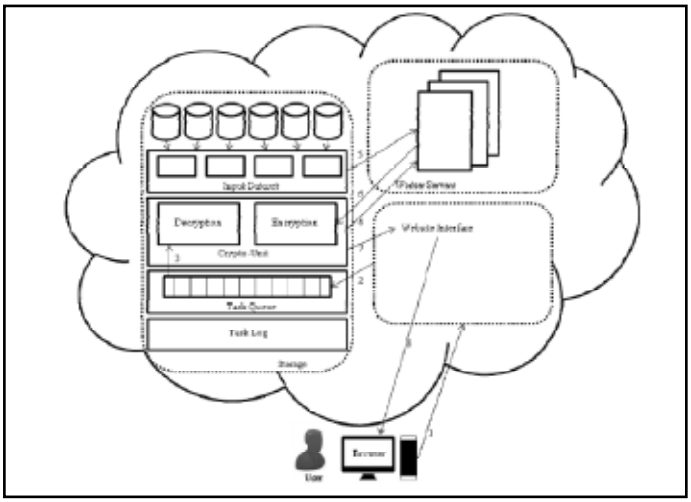

Fig. 3: Secure Cloud Mining Architecture

This system consists of a User who needs the mined information for his business. He makes the use of the Web Browser to interact with the Website Interface that will accept the users query and give the forecast report to the user through the browser. The website interfaces forward the users request to the storage and the Worker servers get the results for the request. The functional blocks of this architecture consists of the User, Web Browser, Website Interface, Worker Servers and the Storage.

VII. Comparative Analysis
The comparison of the three techniques is done on the following parameters:

- ***Data mining algorithm:*** The algorithm used to implement data mining in the cloud.
- ***Method:*** The method used to implement the respective technique
- ***Cloud Architecture:*** Whether single or distributed architecture used.
- ***Accuracy:*** Accuracy of the data mining algorithm.
- ***Effectiveness:*** Effectiveness is about the system as to how the system does the right task as expected, completing activities and achieving goals.
- ***Application area:*** The type of application scenario in which the technique can be used.
- ***Advantages:*** The benefits of each technique.
- ***Disadvantages:*** The drawbacks of each technique.

VIII. Conclusions
This paper provides an overview of the necessity and utility of data mining in cloud computing, as cloud service providers as well as other third parties use different data mining techniques to acquire valuable information. As the need for data mining tools is growing every day, the ability of integrating them in cloud computing becomes more and more stringent. Finally, we also discussed a technique to secure or protect the privacy of the data in cloud. The ontology is used to build up the role hierarchy for a specific domain. Ontology transformation operations algorithms are provided to compare the similarity of different ontology. The proposed idea can ease the design of security system in cloud and reduce the complexity of in the field of mining.

IX. Acknowledgment
We would like to thank everyone who has motivated and supported us for preparing this manuscript.

References
Xia Geng ,Zhi Yang , School of Computer Science and Telecommunication Engineering , Jiangsu University, Jiangsu Zhenjiang, P.R. China: Data Mining in Cloud Computing- © 2013. The authors - Published by Atlantis Press.

[1]. Ruxandra-Ştefania PETRE ,Bucharest Academy of Economic Studies: Data mining in Cloud Computing- Database Systems Journal vol. III, no. 3/2012

[2]. Preeti Aggarwal, CS/IT, KIIT College of Engineering, M. M. Chaturvedi ,SET, Ansal University: Application of Data Mining Techniques for Information Security in a Cloud: A Survey - International Journal of Computer Applications (0975 – 8887) Volume 80 – No 13, October 2013

[3]. Masooda M.Aslam Modak, Dr. Vijayalakshmi M. Information Technology Department Vivekanand Education Society's Institute Of Technology,: Privacy Preserving Data Mining

Techniques In The Cloud :A Comparative Analysis- VESIT , International Technological Conference-2014 (I-TechCON), Jan. 03 – 04

[4]. Naskar Ankita, Mrs. Mishra Monika R., Information Technology Dept Smt. Kashibai Navale College of Engineering Pune, India: USING CLOUD COMPUTING TO PROVIDE DATA MINING SERVICES- International Journal Of Engineering And Computer Science ISSN:2319-7242 Volume 2 Issue 3 March 2013 Page No. 545-550*

[5]. Himel Dev, Tanmoy Sen, Madhusudan Basak and Mohammed Eunus Ali Department of Computer Science and Engineering (CSE) , Bangladesh University of Engineering and Technology (BUET) : An Approach to Protect the Privacy of Cloud Data from Data Mining Based Attacks.

[6]. Zeba Qureshi1, Jaya Bansal2, Sanjay Bansal3 Mtech, Professor & Head, department of CSE, A.I.T.R: 318 A Survey on Association Rule Mining in Cloud Computing- International Journal of Emerging Technology and Advanced Engineering Website: www.ijetae.com (ISSN 2250-2459, ISO 9001:2008 Certified Journal, Volume 3, Issue 4, April 2013)

[7]. Xiaoni Wang ,School of Applied Science, Beijing Information Science and Technology University, Beijing, 100192 : Research on the Distributed Data Mining Platform Based on Cloud Computing Security - 2013 International Conference on Innovation, Management and Technology Research (ICIMTR 2013)

[8]. T.V.Mahendra ,N.Deepika ,N.Keasava Rao Professor & HOD, IT, Narayana Engg. College, Nellore,: 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 ISSN: 2277 128X

[9]. Ms. Sunita, Prachi CSE, SES, BPSMV India : Efficient Cloud Mining Using RBAC (Role Based Access Control) Concept - International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 7, July 2013 ISSN: 2277 128X.*

Authors

Uppunuthula Venkateshwarlu pursuing Post Graduate in Master of Technology with specialization of Computer Science & Engg. at RVR Inst. of Engg.& Tech, Hyderabad, AP, India. Her interested research area is Data warehousing & Data Mining, Cloud Computing and Data Structures.

Puppala Priyanka 2 is pursuing her Post Graduate in Master of Technology with specialization of Computer Science & Engg. at AVN Inst. of Engg.& Tech, Hyderabad, AP, India. Her interested research area is Data warehousing & Data Mining, and Network Security.