Secure Mining Methodology for Cloud-based Computing Services

B V V Sony Sowjanya
PG Scholar, Department of Computer Science and Engineering,
BABA Institute of Technology and Sciences , Vizag

K S N Murthy
Asst.Prof, Department of Computer Science and Engineering,
BABA Institute of Technology and Sciences , Vizag

Abstract- This paper is of how data mining is utilized in cloud computing. Data Mining is a procedure of separating potentially helpful data from huge data. How SaaS is extremely valuable in cloud computing. The reconciliation of data mining systems into ordinary everyday exercises has turned out to be regular place. We are gone up against day by day with focused promoting, and organizations have turned out to be progressively effective using data mining exercises to lessen costs. Data mining applications can determine much statistic data concerning clients that was already not known or covered up in the data. We have as of late observed an expansion in data mining strategies focused to such applications as fraud detection, recognizing criminal suspects, and prediction of potential terrorists. All things considered, data mining frameworks that have been produced to data for clusters, disseminated clusters and lattices have accepted that the processors are the rare asset, and subsequently shared. The data mining procedures like grouping, order, neural system, hereditary calculations help in finding the concealed and beforehand obscure data from the database. Cloud Computing is an electronic innovation whereby the assets are given as shared services. The extensive volume of business data can be put away in Cloud Data focuses with minimal effort. The two Data Mining methods and Cloud Computing causes the business associations to accomplish expanded benefit and cut expenses in various conceivable ways. The fundamental point of the work is to actualize data mining system in cloud computing utilizing Google App Engine and Cloud SQL.
Keywords Data Mining, cloud Computing, how data mining are used in cloud computing, key Characteristics.

I. Introduction

Data Mining over Big Data and Cloud Computing are considered as significant innovations that can bolster reasonable asset sharing. Data mining is considered as an imperative strategy as it is utilized for finding new, reasonable, significant and clear types of data. Huge Data is another term used to perceive the datasets that because of their gigantic size and intricacy, we can't oversee them with our present data mining devices. Data Mining over Big Data is the potential of removing helpful data from these expansive datasets or floods of data, that because of its amount, irregularity and speed, it was impractical before to do it. Cloud computing is a clever innovation that can bolster an extensive variety of utilizations. Data mining undertakings and applications can be viably utilized in cloud computing model. The data mining assignments in cloud computing gives a flexible and adaptable basic plan which can lessen the expense of framework and storage and utilized for effective mining of tremendous measure of data from for all intents and purposes consolidated data sources with the point of creating valuable data which is steady in basic leadership to anticipate the future patterns and conduct. However, it has the danger of protection of data client and security. Today, computing turns out to be relentlessly progressively essential and increasingly utilized. The measure of data traded over the system or put away in a PC is always expanding. In this manner, the handling of this expanding mass of data requires more PC gear to meet the different needs of associations [2]. Cloud computing is an unavoidable pattern later on computing advancement of innovation. Its basic significance lies in its capacity to furnish every one of the clients with superior and steady estimation. Cloud computing is the development of disseminated

computing, network computing, and numerous different strategies. In cloud computing data is moving from work area framework to data focuses. By methods for virtualization innovation, one physical host can be virtualized into different virtual has and utilize these hosts as an essential computing unit. [3]. In this scientific classification we have endeavored to expand cloud computing engineering alongside its quality, shortcoming, difficulties and applications in current situation dependent on the present advances from the scholarly world.

II. Related work

By cloud we can state that it is a foundation that comprises of services conveyed through offer data focuses and showing up as a solitary purpose of access to shoppers computing needs and furthermore gives requested assets or services over the web. The idea of cloud computing does not give offices to the learning revelation and data recovery. Besides, it is required that the purported information revelation ought to be in agreement with the structure, composition and engineering of that learning. The rising learning cloud is viewed as deficient to recover data successfully and consequently, Chang, Yang and Luo (2011) attempted an examination to a gigantic number of ways just in light of the fact that the things are common to the point that they can't resist the opportunity to seem together. This is known as the uncommon thing issue. It implies that utilizing the Apriori calculation, we are probably not going to produce decides that may show uncommon occasions of potentially emotional propose "a cosmology based operator age system for data recovery in an adaptable, straightforward and simple path on cloud condition". The perfect beginning stage is a data distribution center containing a blend of inside data following all client contact combined with outside market data about contender movement. Foundation data on potential clients likewise gives a phenomenal premise to prospecting. This distribution center can be executed in an assortment of social database frameworks: Sybase, Oracle, Redbrick, etc, and ought to be enhanced for adaptable and quick data get to. An OLAP (On-Line Analytical Processing) server empowers a progressively advanced end-client plan of action to be connected while exploring the data stockroom. The multidimensional structures enable the client to examine the data as they need to see their business – condensing by product offering, locale, and other key points of view of their business. The Data Mining Server must be coordinated with the data distribution center and the OLAP server to implant ROI-centered business investigation straightforwardly into this foundation. A progressed, to zero, RSAA sets aside a comparative measure of opportunity to that taken by Apriori produce low-bolster runs in among the high backings rules. Processcentric metadata layout characterizes the data mining destinations for explicit business issues like crusade the board, prospecting, and advancement streamlining. Joining with the data distribution center empowers operational choices to be straightforwardly actualized and followed. As the distribution center develops with new choices and results, the association can ceaselessly mine the prescribed procedures and apply them to future choices. This structure speaks to an essential move from customary choice emotionally supportive networks. As opposed to just conveying data to the end client through inquiry and announcing programming, the Advanced Analysis Server applies users‟ plans of action straightforwardly to the distribution center and returns a proactive examination of the most pertinent data. These outcomes upgrade the metadata in the OLAP Server by giving a dynamic metadata layer that speaks to a refined perspective of the data. Revealing, perception, and different examination devices would then be able to be connected to design future activities and affirm the effect of those plans.

III. The Responsibility of Data Mining

In Cloud 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. The main effects of data mining tools being delivered by the Cloud are 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. 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. These data mining tasks include: Analyze Key Influencers, Detect Categories, Fill From example, Forecast, Highlight Exceptions, Scenario Analysis, Prediction Calculator, and Shopping Basket Analysis. The implementation of data mining techniques through Cloud computing will allow the users to retrieve carrying great weight in order to virtually integrated data warehouse that reduces the costs of infrastructure and storage.

IV. Current Research Challenges in Cloud

Countless obtainable issues have not been entirely addressed, while new challenges keep emerging from various applications. This section describes some of the challenging research issues in cloud computing for researchers who are much interested in this field.

Automated service provisioning

One of the primary key features of cloud computing is the capability of acquiring and releasing resources on demand. The objective of a service provider in this case is to allocate and de-allocate resources from the cloud to satisfy its service level objectives while minimizing its operational cost.

Virtual machine migration

Virtualization can provide significant benefits in cloud computing by enabling virtual machine migration to balance load across the data center. In addition, virtual machine migration enables robust and highly responsive provisioning in data centers. Now, detecting workload hotspots and initiating a migration lacks the agility to respond to rapid workload changes.

Server consolidation

Server consolidation is an effective innovative approach to take full advantage of resource utilization while minimizing energy consumption in a cloud computing environment. Existing virtual machine migration technology residing on multiple servers onto a single server at that time the remaining servers can be set to an energy-saving state. The problem of optimally consolidating servers in a data center is often formulated as a variant of the vector bin-packing problem which is an NP-hard optimization problem.

Energy management

Improving energy efficiency is another major issue in cloud computing. Infrastructure providers are under enormous pressure to reduce energy consumption. Designing energy-efficient data centers has recently received considerable attention. This problem can be from several directions such as Energy efficient hardware architecture, Energy-aware job scheduling and server consolidation. In this admiration, the minority researchers have recently started to investigate coordinated solutions for performance and power management in an active cloud environment.

Traffic management and analysis

Examination of data traffic is significant for today's data centers in cloud environment. Network operators also need to know how traffic flows through the network in order to make many of the management and planning decisions. Currently, there is not much work on measurement and analysis of data center traffic.

Data security

Data security is another important research topic in cloud computing. Since service providers typically do not have access to the physical security system of data centers, they must rely on the infrastructure provider to achieve full data security. The hardware layer must be trusted using hardware TPM. Secondly, the virtualization platform must be trusted using secure virtual machine monitors. VM migration should only be allowed if both source and destination servers are trusted. Recent work has been devoted to designing efficient protocols for trust establishment and management.

Storage technologies and data management

These file systems are different from traditional distributed file systems in their storage structure, access pattern and application programming interface. In particular, they do not implement the standard POSIX

interface, and therefore introduce compatibility issues with legacy file systems and applications. Several research efforts have studied this problem but not yet optimum.

Novel cloud architectures

Now, most of the commercial clouds are implemented in large data centers and operated in a centralized fashion. Although this design achieves economy-of-scale and high manageability, it also comes with its limitations such high energy expense and high initial investment for constructing data centers. Recent researchers suggests that small size data centers can be more advantageous than big data centers in many cases: a small data center does not consume so much power, hence it does not require a powerful and yet expensive cooling system; small data centers are cheaper to build and better geographically distributed than large data centers. The researchers think to build Nano-Data centers with full pledged manner. The Data mining technologies provided throughout Cloud computing is an extremely essential trait for today's businesses to make proactive, knowledge driven decisions, as it helps them have future trends and behaviors predicted. This chapter provides an overview of the necessity and utility of data mining in cloud computing. As the need for data mining tools is growing every day, the aptitude of integrating them in cloud computing becomes progressively stringent. The current technologies are not matured enough to realize its full potential. In future the researchers highly deliberate the data mining technologies to implement in fog computing.

V. Data Mining In Cloud Computing

"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." Cloud computing refers to the delivery of computing resources over the Internet. Instead of keeping data on your own hard drive or updating applications for your needs, you use a service over the Internet, at another location, to store your information or use its applications. Doing so may give rise to certain privacy implications. The Microsoft suite of cloud-based administrations presents another specialized sneak peak of Data Mining in the Cloud as "DMCloud".

The data mining tasks include:

- Analyze Key Influencers
- Detect Categories
- Fill From Example
- Forecast

V. Some Issues of Cloud Computing- Based Data Mining

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

www.ijarcst.com

- 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 cooccurring 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.

VI. 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.

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.

VII. Proposed 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.

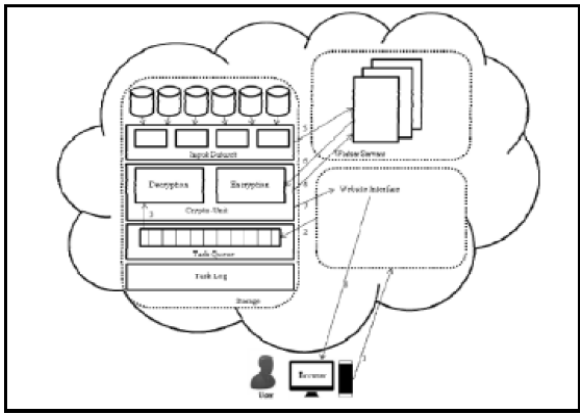

Fig. Proposed Architecture diagram

VIII. Conclusion

Data mining is used in various applications such as Health care, Student management, mathematics, Science, in various website. 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. Here we explore the how the data mining tools like SAS, PAS and IaaS are used in cloud computing to extract the information. A cloud provider for a data mining and natural language processing system. Leading cloud computing providers Amazon Web Services, Windows Azure, OpenStack. People use this feature to build information listing, get information about different topics by searching in forums etc. Company's use this service to see what kind of information is floating in the world wide web for their products or services and take actions based on the data presented The information retrieval practical model through the multiagent system with data mining in a cloud computing environment has been proposed. It is however, recommended that users should ensure that the request made to the IaaS is within the scope of integrated data warehouse and is clear and simple. Thus, making the work for the multi-agent system easier through application of the data mining algorithms to retrieve meaningful information from the data warehouse. Cloud computing allows the users to retrieve meaningful information from virtually integrated data warehouse that reduces the costs of infrastructure and storage.

References

[1] Assessing Invariant Mining Techniques for Cloud-based Utility Computing Systems, Antonio Pecchia, Member, IEEE, Stefano Russo, Senior Member, IEEE,
and Santonu Sarkar, Member, IEEE
[2] A. Khiyaita, M. Zbakh, H. El Bakkali and D. El Kettani, Load balancing cloud computing: State of art, Proc. of 2012 National Days of Network Security and Systems (JNS2), 20-21 April 2012, 106-109.
[3] Kai Zhu, Huaguang Song, Liu Lijing, Gao Jinzhu and Guojian Cheng, Hybrid Genetic Algorithm for Cloud Computing Applications, Proc. of The 2011 IEEE Asia-Pacific Services Computing Conference (APSCC), 12-15 December 2011, 182-187,
[4] Yuqi Zhang, Jun Wu, Yan Ma, Xiaohong Huang and Mingkun Xu, Dynamic load-balanced multicast based on the Eucalyptus opensource cloud-computing system, Proc. of 2011 4th IEEE International Conference on Broadband Network and Multimedia Technology (ICBNMT 2011), 28-30 October 2011, 456-460.
[5] R. Jeyarani, N. Nagaveni and R. Ram Vasanth, Design and Implementation of an efficient Two-level Scheduler for Cloud Computing Environment, Proc. of The 10th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing, 17-20 May 2010, 585-586.
[6] IBM, Fundamentals of Cloud Computing, Instructor Guide ERC 1.0, (November 2010) 17-230.
[7] Shyam Patidar, Dheeraj Rane and Pritesh Jain, A Survey Paper on Cloud Computing, Proc. of 2012 second international conference on Advanced Computing and Communication Technologies (ACCT), 7-8 January 2012, 394-398.
[8] Cloud Security Alliance, Security guidance for critical areas of focus in cloud computing v2. 1, (December 2009) 13-19.
[9] Deyan Chen and Hong Zhao, Data Security and Privacy Protection Issues in Cloud Computing, Proc. of International Conference on Computer Science and Electronics Engineering (ICCSEE) 2012, 23- 25 March 2012, 647-651.
[10] Kui Ren, Cong Wang and Qian Wang, Security Challenges for the Public Cloud, IEEE Internet