www.ijecs.in
International Journal Of Engineering And Computer Science ISSN:2319-7242
Volume 2 Issue 3 March 2013 Page No. 545-550

USING CLOUD COMPUTING TO PROVIDE DATA MINING SERVICES

Naskar Ankita*, Mrs. Mishra Monika R.
*Information Technology Dept
Smt. Kashibai Navale College of Engineering
Pune, India
ankmit24@gmail.com

Information Technology Dept
Smt. Kashibai Navale College of Engineering
Pune, India
monika.r.mishra@gmail.com

Date of Submission: 2nd March, 2013

Abstract— Data security and access control are the most challenging research work going on, at present, in cloud computing. This is because of the users sending their sensitive data to the cloud providers for acquiring their services. In cloud computing, the data is going to be stored in storage area provided by the service providers. The service providers must have a suitable way to protect their client's sensitive data, especially to protect the data from unauthorized access. A common method of information privacy protection is to store the client's data in encrypted form. If the cloud system is responsible for both storage and encryption/decryption of the data, the system administrators may simultaneously obtain encrypted data and the decryption keys. This will allow them to access the information of the client without any authorization. This leads to the risk of sensitive information leak and the method involved of storage and encryption/decryption is costly too. Hence, to overcome these problems, a model (cloud server) has been proposed in this paper which accepts only those data which are required in an encoded form, performs the service opted by the client and sends the result in the encoded format to be understood by the respective client.

***Keywords*-cloud computing, data mining, apriori algorithm, k-means algorithm,cloud server,XML, web services, GlassFish, SaaS.**

I. INTRODUCTION

The major fields which are being talked about in this paper are: Cloud Computing and Data Mining.

In the recent years, cloud computing has become a hot topic in the global technology industry. Cloud computing also faces the data security challenges like any other communication model. 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. In addition to confidentiality and privacy breaks, the external servers could also use part or whole of the data for their financial gain. Therefore, ruining the data owners market or even bringing economic loss to the data owners. These concerns start off from the fact that cloud servers are usually operated by commercial providers which are probably from outside of the trusted domain of users.[1]

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.

In the cloud computing environment, storage service providers should have prepared data security practices to make sure that their clients' data is safe from unauthorized access and disclosure. Essentially, the rules and measures for preventing privileged users such as system administrators from unauthorized access must be strictly established and implemented. Service providers track specific policies and practices to protect their users' data, and these policies are usually fixed in the service contract. For example, a Gmail user should read the service contract online and show his/her consent to the service contract before he/she can use the webmail service. The substance of the service contract covers definitions of service items, service scope, service change notification, scope of privacy protection, regulations on user data collection, use, sharing and release, and statements regarding user responsibilities.

In a cloud computing environment, the service content presented by service providers can be accustomed according to the needs of the user. For example, the candidate can ask for different amounts of storage, transmission speeds, levels of data encryption and other services. Adding together to defining the service items, the agreement usually also states the time, quality and performance requirements provided with the service. Generally, these service agreements are referred to as Service Level Agreements (SLA). By signing an SLA, the user shows that he/she has understood and approved to the contents of the application service, and are in accord with the provider's data privacy and protection policies.

A general way to protect user data is that user data is encrypted before it is stored. In a cloud computing environment, a user's data can also be stored after performing additional encryption, but if the storage and encryption of a given user's data is done by the same service provider, the service provider's internal staff (e.g., system administrators, authorized staff, etc) can use their decryption keys and internal access privileges to access the user data. From the user's point of view, this can put his/her stored data at risk of unauthorized leak. Creation of user's trust by the protection of the user's data is the key to the extensive approval of the cloud computing.

Data mining is the process of extracting useful patterns or knowledge from large databases. [2] Though, data mining also poses a risk to privacy and information protection if not done or used properly. For example, association rule analysis is an accepted tool for discovering useful associations from huge amount of data and some valuable hidden information could be simply discovered using this sort of tool. Hence, the security of sensitive hidden information has become a significant issue to be resolved. The aim of privacy preserving data mining is to hide certain information so that they cannot be exposed through data mining techniques such as association rule analysis. There have been two significant approaches for privacy preserving data mining are: output and input privacy.

The output privacy approach is to modify the data before delivery to the data miner so that real data is hidden and mining result will not reveal certain privacy. For example, blocking, merging, swapping and sampling are some methods that have been proposed for this type of output privacy . The input privacy approach, on the other hand, is to change the data using data distribution methods. In this approach, mining result is not affected or minimally affected. For example, reconstruction based and cryptography based are some techniques that have been proposed for this type of input privacy.

Data mining has also emerged as a way for identifying patterns and trends from large quantities of data. For example, shopping centres found out that male customers who buy diaper usually buy beers by analyzing consuming lists. This forms the relation between diaper and beer through rearranging these goods. This improvement of goods arrangement after analysis yields more sale. This kind of analysis can be used in many fields such as Credit Cards, Banking sectors, etc. Hence, techniques of data mining without leaking the private information are needed. Research on privacy preserving data mining is developed for this purpose. [3]

The privacy preserving data mining and knowledge discovery should be developed aiming at these problems. In order to secure an openly available system , it must be ensured that not only that private sensitive data should be trimmed out, but also to make

sure that certain inference channels should also be blocked as well. Under privacy constraints, the association rule mining problem was extensively researched. Many efficient methods for privacy preserving association rule mining were found. However, most of these methods resulted in information loss and side-effects to some extent, such as non-sensitive rules falsely hidden and spurious rules falsely generated, may be formed in the sensitive rule hiding process.

Sequential pattern mining can be defined as finding the complete set of frequent subsequences in a set of sequences. Sequential pattern mining can be used for discovering meaningful sequential patterns among a large quantity of data. For example, let us see the sales database of a bookstore. The revealed sequential pattern could be “70% of people who bought Twilight also bought Harry Potter at a later time”. The bookstore can make use of this information for shelf placement, promotions, etc.

II. Literature Survey

A. *Problem Statement*

The basic idea behind this paper can be proposed as "Developing a Cloud Server for providing Data Mining Services" where in data mining services can be offered over cloud for any one to use.

B. *Origin and Definition of Cloud Computing*

The Internet rapidly began to grow up in the 1990s and, the progressively more complicated network infrastructure and enlarged bandwidth developed in the recent years have considerably improved the strength of various application services available to users through the Internet, hence, marking the beginning of cloud computing network services. Cloud computing services use the Internet as a communication medium and convert information technology resources into services for end-users, including software services, computing platform services, development platform services, and basic infrastructure leasing.

Primary significance of Cloud computing lies in allowing the end users to access computation resources through the Internet. The unusual features of cloud computing include the storage of user data in the cloud and the lack of any need for software installation on the client side. Provided that the user is able to connect to the Internet, all of the hardware resources in the cloud can be used as client-side infrastructure. Normally, cloud computing applications are demand-driven, providing various services according to user requirements, and service providers charge by metered time, instances of use, or defined period.

Cloud computing can be defined as “a type of parallel and distributed system which consists of a collection of interconnected and virtualized computers that are dynamically provisioned and presented as one or more unified computing resources based on service-level agreements established through negotiation between the service provider and consumers”.[1]

The cloud computing concept can be understood in a more better way by following the below given figure:[4]

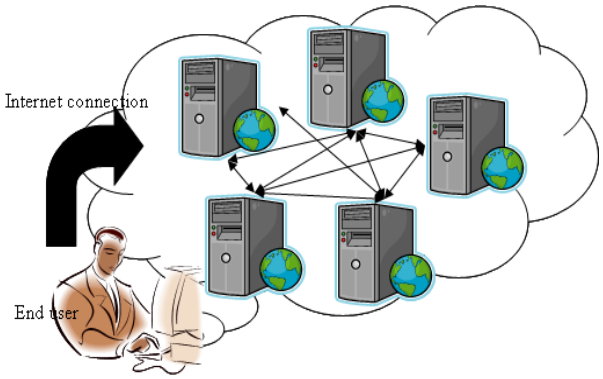

Figure 1: Figure to represent cloud computing concept map

The architecture of cloud services can be divided into three levels: infrastructure, platform, and application software. Application software builds the user interface and shows the application system’s functions. To build a cloud computing application as a service requires infrastructure, platform and application software which can be obtained from a single provider or from different service providers. If the income for cloud services mainly comes from charging for infrastructure, this business model can be referred to as Infrastructure as a Service (IaaS). If income comes mainly from charging for the platform,

the business model can be referred to as Platform as a Service (PaaS). If income mainly comes from charging for applications or an operating system, the business model can be referred to as Software as a Service (SaaS). The model being proposed in this paper uses SaaS concept.

C. *Origin and Definition of Data Mining*

Data mining is to find knowledge, and knowledge is represented through certain patterns. Association rule is the most often used method in data mining, which finds out the association between data and various objects by finding the potential dependence among data. Classification and clustering can be used to sort out things by characterizing the common significance among different things. The disadvantage of data mining in centralized database, generally have the several following points: network traffic is considered less, mining efficiency is low and the degree of spatial complexity is high. The most classic classification data mining are classification methods based on distance, classification methods based on decision tree, Bayesian classification and so on.

Data mining techniques have been extensively used in various applications. However, the mistreat of these techniques may lead to the discovery of sensitive information. Researchers have recently made efforts at hiding sensitive association rules. However, undesired side effects, e.g., non-sensitive rules falsely hidden and spurious rules falsely generated, may be formed in the rule hiding process. [5]

Privacy has become an significant issue in Data Mining. Many methods have been brought out to solve this problem. The basic aspect which we are concerned about in this paper is of association rule mining which preserves the confidentiality of each database. In order to find the association rule, each participant has to share their own data. Thus, a lot of privacy information may be put out or been illegally used. [6] Data mining can be defined as "the process that attempts to discover patterns in large data sets". The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use.

The following figure explains the different steps which comprise the overall data mining process:

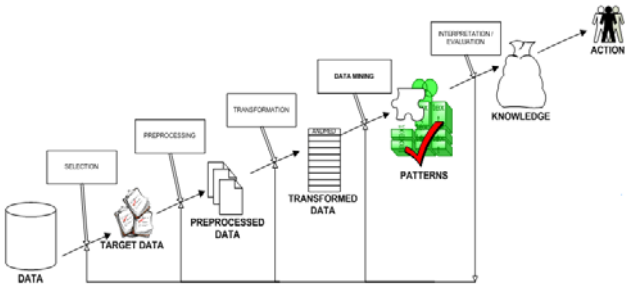

Figure 2: Figure to represent steps in data mining process

III. PROPOSED SCHEME

In this paper, a Cloud Server model is proposed. Using this model, the clients or users can avail different data mining services which the cloud providers promise to provide. As stated in the first part of the paper, the need to develop this kind of model is the problem of sensitive information leak and costly services which are come across when a general client-server model is used or when the client's sensitive data is sent to the cloud while availing its servics. Using this model, these problems will not be faced any more in the future.

When a general client-server model is being used while communicating between a server and a client, both the client and the server need to share a common shared library so that they use the same constructs and formats to communicate. This created problem because all the clients needed to be made aware of this library if they are not. This turned out to be costly, more time taking and wasting of resources. Another reason why this model is being proposed is that when a client wants to avail some of the services provided by the cloud, he/she needs to send their whole database to the cloud. This is done because for using the cloud's services the database of client is considered as input to the service routine. This leads to the sensitive information leak. Even if the database is encrypted in the cloud, then also the system administors and other officials can manage the decrytion key. Hence, this procedure also turned out to be insecure and did not work.

As a result of the above stated problems, this new model is proposed where we need to develop a cloud server which does not need any shared library. Figure

3 shows the proposed model. The figure shows that the cloud server at first intakes the database from the client. This database is not transfered as a whole, but those parts of it which are essential for the cloud server for providing the service asked by the client, in an indexed format. In this manner, the client database is transferred and accessed in a secure way. This cloud server only employs SaaS. The proposed model provides data mining services using web services and it also uses GlassFish to implement the model.There can be 'n' number of clients using services from this cloud server at a time.The various elements used in the model are explained in detailed manner as below:

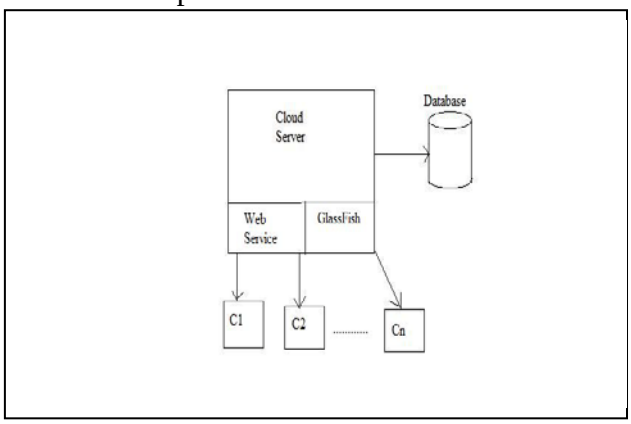

Figure 3: Figure to represent proposed model

Web service is a method of communication over a network between two electronic devices. These were intended to solve three main problems such as Firewall Traversal, Complexity, and Interoperability.

The **database** represents the client's database which consists of the details about the transactions on the client's side.

Software as a Service (SaaS) can be defined as a software distribution model in which applications are hosted by a vendor or service provider and made available to customers over a network, normally the Internet. It is becoming a gradually more widespread delivery model as underlying technologies that support Web services and service-oriented architecture (SOA) mature and new developmental approaches, such as Ajax, become popular. SaaS is closely related to the ASP (application service provider) and on demand computing software delivery models. Benefits of the SaaS model include: easier administration, automatic updates and patch management, compatibility, easier collaboration, and global accessibility. The term "software as a service" (SaaS) is considered to be part of the classification of cloud computing, together with infrastructure as a service (IaaS) and platform as a service(PaaS). Most of SaaS solutions are based on a multi-tenant architecture. Using this model, a single version of the application, with a single configuration (hardware, network, operating system), can be used for all customers, i.e, tenants. To support scalability, the application is installed on multiple machines. Saas has an exception that some of its solutions do not use multi-tenancy, or use other mechanisms ,such as virtualization, to cost-effectively manage a large number of customers in place of multi-tenancy.

GlassFish is a project started by Sun Microsystems which is an open-source application server for the Java EE platform. It is now sponsored by Oracle Corporation. The supported version of GlassFish is known as Oracle GlassFish Server. It is a free software. It is the reference implementation of Java EE and also supports Enterprise JavaBeans, JPA, JavaServer Faces, JMS, RMI, JavaServer Pages, servlets, etc. This also allow developers to develop enterprise applications that are portable and scalable, and that combine with legacy technologies.

XML (Extended Markup Language) is used to set the rules for exchange of information in the proposed model. It is a markup language that is used to define set of rules for encoding documents in a format that is both understood by human and as well as machines. The design goals of XML include simplicity, generality, and usability over the Internet. It can also be defined as a textual data format with strong support through Unicode for the languages of the world. Even though the design of XML focuses on documents, it is generally used for the representation of arbitrary data structures, for example in web services.

The proposed model can be used to provide any data mining service, but in this paper, I am concerned about only 2 services. The first service provides information about only those entities which occur frequently in the database and the second service is

used to group together those entities from the client database, which have common characteristics between them. The first service uses Apriori algorithm to find out the frequent itemsets and the second service uses K-means algorithm.

Apriori [7] is an influential algorithm proposed by R. Agrawal and R. Srikant in 1994 for the purpose of mining frequent itemsets for Boolean association rules. The name of this algorithm has been derived by the fact that the algorithm uses prior knowledge of frequent itemset properties. The algorithm uses iterative approach which is called level-wise search. In this algorithm, k-itemsets are used to find out (k+1)-itemsets. At first, the set of frequent 1-itemsets is obtained by scanning the database to accumulate the count for each item, and collecting the items that satisfy minimum support count for each item, and accumulating only those items that that have minimum support count.

K-means [8]algorithm takes 'k' as the input parameter and divides a set of 'n' objects into 'k' clusters. This will result into high similarity in the intracluster whereas low similarity in interclusters. Cluster similarity can be measured by the mean value of the objects in a cluster.

IV. BENEFIT ANALYSIS

This paper presented the proposed model of cloud server which eliminates the sensitive data leak and cost issues which came across when using the normal scenario. Using this model, the client database will be secured and unauthorized access will be denied because the cloud only uses those parts of the data which are required and in an indexed format. As a result, no encryption/decryption will be performed on the cloud and no storage area is required on the cloud. This removes the cost issue. The proposed model uses only the essential parts of the client database to provide the service requested by the client. After the execution of the service routine, which includes execution of one or both of the algorithms, the result of the service is send to the client. The client can then infer from the result the information which he/she required. Future extensions will include adding up of more data mining services to be provided by the cloud server.

V. FUTURE SCOPE

In the future, I plan to add more data mining algorithms to this approach other than the ones implemented in this paper. Right now, the main category of data mining algorithms covered here are Clustering and Association Rule Mining (Apriori Algorithm).

REFERENCES

[1] Sunil Sanka, Chittaranjan Hota, Muttukrishnan Rajarajan, "Secure Data Access in Cloud Computing," in IMSAA '10, 2010, p. 1-6.

[2] S. M. Mahajan and A. K. Reshamwala, "Data Mining Ethics in Privacy Preservation - A Survey " in International Journal of Computer Theory and Engineering, Vol. 3, No. 4, August 2011.

[3] Manoj Gupta and R. C. Joshi, "Privacy Preserving Fuzzy Association Rules Hiding in Quantitative Data" in International Journal of Computer Theory and Engineering, Vol. 1, No. 4, October, 2009.

[4] Jing-Jang Hwang and Hung-Kai Chuang, Yi-Chang Hsu and Chien-Hsing Wu, "A Business Model for Cloud Computing Based on a Separate Encryption and Decryption Service," in ICISA '11, 2011, p. 1-7.

[5] Yi-Hung Wu, Chia-Ming Chiang, and Arbee L.P. Chen, "Hiding Sensitive Association Rules with Limited Side Effects," in IEEE Transactions on Knowledge and Data engineering, Vol. 19, No. 1,pp. 29-42, January 2007.

[6] Tinghuai Ma, Sainan Wang, ZhongLiu, "Privacy Preserving Based on Association Rule Mining," in Advanced Computer Theory and Engineering (ICACTE), Vol. 1, pp. 637-640, August 2010.

[7] Jiawei Han, Micheline Kambe. *Data Mining, Concepts and Techniques* , 2nd Ed. CA: Morgan Kaufmann Publishers, 2006,pp. 234-239.

[8] Jiawei Han, Micheline Kambe. *Data Mining, Concepts and Techniques* , 2nd Ed. CA: Morgan Kaufmann Pub