Realisation of Resourceful Data Mining Services Using Cloud Computing

Ankita Naskar
Information Technology Dept
Smt. Kashibai Navale College of Engineering
Pune, India
ankmit24@gmail.com

Prof. Mrs. Monika R. Mishra
Information Technology Dept
Smt. Kashibai Navale College of Engineering
Pune, India
monika.r.mishra@gmail.com

***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 general method of information privacy protection is to store the client's data in encrypted form. If the cloud system is in charge for both storage and encryption/decryption of the data, the system administrators may concurrently obtain encrypted data and the decryption keys. This will let them to use the information of the client devoid of any permission. This will lead to the risk of vulnerable information disclose and the method concerned with storage and encryption/decryption is expensive 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. This paper also contains the implementation results and the benefit analysis of the proposed system.

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

I. INTRODUCTION

The most important fields which are being talked about in this paper are: Cloud Computing and Data Mining.In the latest years, cloud computing has grown to be a hot topic in the universal technology industry. Cloud computing also faces the data protection challenges like any other communication model. As the data owners stock up their data on peripheral servers, there have been allegedly increasing hassle and concerns for data confidentiality, authentication and access control. Along with confidentiality and privacy breaks, the external servers might also use part or entire of the data for their economic gain. Therefore, tarnishing the data owners market or even getting economic loss to the data owners. These concerns begin from the fact that cloud servers are generally operated by commercial providers which are most likely from outside of the trusted domain of users as specified in Reference[1].

Prior to the growth of the thought of cloud computing, critical industrial data used to be stored within the storage media, protected by safety measures including firewalls, to stay away from external access to the data and counting organizational policy to disallow unauthorized internal access. In the cloud computing atmosphere, storage service providers should have organized data security practices to ensure that their clients' data is protected from unauthorized access and leak. Effectively, the policy and measures for preventing privileged users such as system administrators from illegal right to use must be firmly recognized and implemented. Service providers follow precise policies and practices to guard their users' data, and these policies are typically set in the service agreement. For example, a Gmail user should study the service contract online and confirm his/her approval to the service contract before he/she can utilize the webmail service. The essence of the service contract include definitions of service items, service scope, service change notification, scope of privacy security, policy on user data gathering, use, distribution and discharge, and statements concerning user responsibilities.

In a cloud computing atmosphere, the service substance offered by service providers can be adapted according to the requirements of the user. For instance, the contender can ask for diverse amounts of storage, broadcast speeds, levels of data encryption and other services. Adding up to defining the service substance, the contract usually also gives the time, quality and performance necessities given with the service. Usually, these service

contracts are referred to as Service Level Agreements (SLA). By signing an SLA, the user clears that he/she has understood and allowed to the contents of the application service, and are in concurrence with the provider's data isolation and security policies.

A common method to guard user data is that user data is encrypted prior to it is stored. In a cloud computing atmosphere, a user's data can also be kept after performing extra encryption, but if the storage and encryption of a given user's data is prepared by the same service provider, the service provider's in-house team (e.g., system administrators, authorized staff, etc) can use their decryption keys and internal access rights to use the user data. From the user's point of view, this can put his/her stored data in threat of unofficial disclose. Formation of user's faith by the safeguard of the user's data is the solution to the widespread sanction of the cloud computing.

Data mining is the course of extracting useful patterns or knowledge from large databases. Reference [2] was that despite the fact that, data mining also poses a danger to isolation and information security if not completed or used properly. For example, association rule analysis is an acknowledged means for finding valuable relations from vast amount of data and some precious unseen information could be mainly revealed by means of this type of tool. Thus, the protection of sensitive secreted information has become a major issue to be determined. The intend of privacy preserving data mining is to conceal certain information so that they cannot be out in the open through data mining techniques such as association rule analysis. There have been two major approaches for privacy preserving data mining are: output and input privacy. The output privacy method is to alter the data before release to the data miner so that actual data is unknown and mining result will not disclose certain privacy. For example, jamming, integration, exchange and sampling are some methods that have been planned for this type of output privacy. The input privacy method, on the other hand, is to change the data with the help of data distribution methods. In this approach, mining result is unaffected or minimally affected. For example, reconstruction based and cryptography based are some techniques that can be used for this type of input privacy.

Data mining has also emerged as a means for identifying patterns and trends from large bulk of data. For example, shopping centres found that male customers who buy diaper usually purchase beers by studying consumers lists. This forms the link between diaper and beer through rearranging these goods. This development of goods arrangement after studying gives more sale. This kind of examination can be used in many fields such as Credit Cards, Banking sectors, etc. Hence, techniques of data mining without leaking the delicate information are needed. Research on privacy preserving data mining is developed for this intention. [3] Sequential pattern mining can be defined as discovery of complete set of recurrent subsequences in a set of sequences. It can be used for finding significant sequential patterns among a large magnitude of data. For example, let us see the sales database of a bookstore. The shown 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 swiftly began to raise up in the 1990s and, the gradually more complex network infrastructure and enlarged bandwidth developed in the latest years have by far enhanced the power of various application services accessible to users using the Internet, hence, marking the start of cloud computing network services. Cloud computing services make use of the Internet as a communication medium and alter information technology resources into services for end-users, together with software services, computing platform services, development platform services, and basic infrastructure leasing.

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" as specified in Reference [1].

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

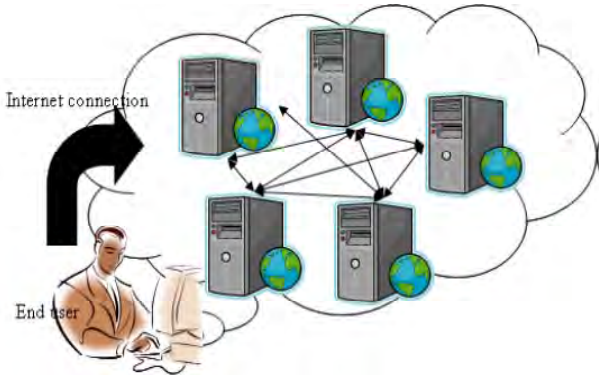

Figure 1: Cloud Computing concept map

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 the process to uncover 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 are widely used in many areas. However, the misuse of these techniques may direct to the detection of susceptible information. Researchers have lately taken pains at hiding susceptible association rules. Though, unwanted side effects, e.g., non-sensitive rules wrongly secreted and bogus rules wrongly generated, might form the rule hiding process as specified in Reference [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. Reference [6] was that "Data mining can be defined as "the process that attempts to discover patterns in huge data sets". The whole goal of the data mining process is to extract information from a data set and transform it into an understandable format for future use."

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

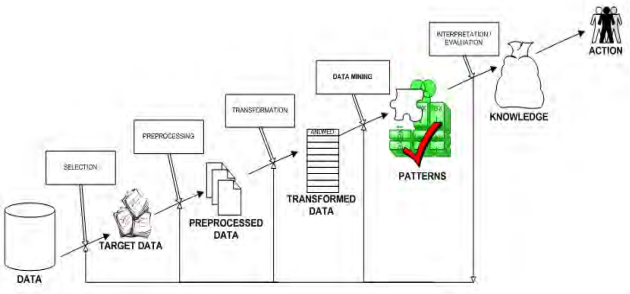

Figure 2: Steps in Data Mining process

III. PROPOSED SYSTEM

After the 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:

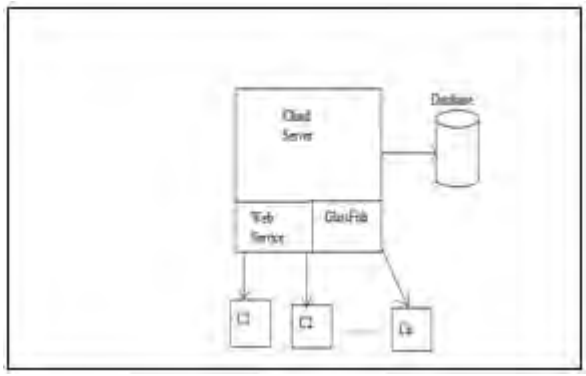

Figure 3: Figure to represent proposed model

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) is defined as a software sharing mould where applications are hosted by a trader or service supplier and offered to clients over a network, usually the Internet. It is now a more extensive delivery mould as basic technologies that support Web services and service-oriented architecture (SOA) become older and new developmental methods, such as Ajax, become well-liked. SaaS is closely linked to the ASP (application service provider) and on-demand computing software delivery moulds. Advantages of the SaaS mould include: easier management, regular updates and space management, compatibility, easy teamwork, and universal ease of access. The term "software as a service" (SaaS) is considered to be part of the categorization of cloud computing, along with Infrastructure as a Service (IaaS) and Platform as a Service(PaaS).

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 main goals of XML include simplicity, generalization, and usability over the Internet. It can be defined as a textual data design with huge hold up through Unicode for

the languages of the world. Even though the aim of XML focuses on documents, it is usually used for the depiction of random 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.

The mathematicatical model of the proposed approach is as shown below:

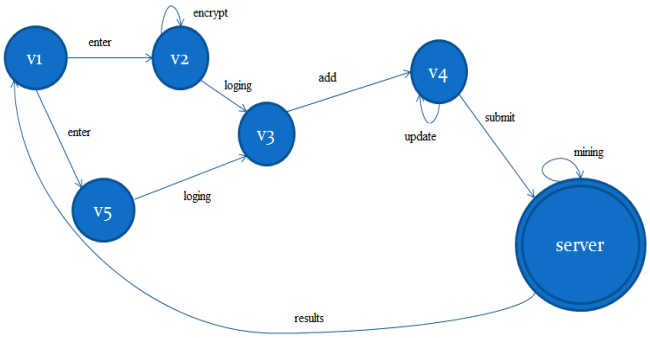

Figure 4: Mathematical model of proposed approach

In the figure above, the colored circles represent the states and the arrows represent actions. v1 is the start state and server is the end state. The user enters user Id and password in v1. The user ID goes to v5 state from where it is send to v3 state for loging. Similarly, the password gets encrypted in v2 state and the encrypted information is send to v3 state for loging. Both the details are then send to v4 state for adding it in the database and the database gets updated in v4 state. From v4, the user's details are send to the server where mining takes place and the results are send to the v1 state.

The different components required to understand the mathematical model:

- Set = { users, database, services, password, uid}
- u1, u2,u3…….un Є users
- d1, d2, d3…….dn Є database
- s1, s2 Є services
- p1, p2, p3…….pn Є password
- i1, i2, i3……..in Є uid
- Mining_results1= Apriori(database)
- Mining_results2 = K-means(database)

IV. APRIORI ALGORITHM

Reference [7] was that "**Apriori** 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.

V. K-MEANS ALGORITHM

Reference [8] was that "**K-means** 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." At first, the 'k' of the objects is selected by a random method which initially represent a cluster mean or center. For each of the remaining objects, an object is assigned to the cluster to which it is most similar, on the basis of the distance between the object and the cluster mean. The mean of each cluster is then computed and the process then iterates till no more mean can be calculated.

VI. IMPLEMENTATION AND RESULTS

We have successfully implemented the proposed system choosing Java as the coding language. MySQL is required for the storage purpose of the client side database. The different entities specific to the client side

operations become the different fields of the database in MySQL. Netbeans IDE is used as platform for the implementation. GlassFish is used to create the cloud server. JDBC is used to enable interaction between the application and the database. To connect with individual databases, JDBC (the Java Database Connectivity API) requires drivers for each database. The JDBC driver gives out the connection to the database and implements the protocol for transferring the query and result between client and database.

As a result, we get the following screenshots:

Figure 5: Login page for Users

Above figure shows the form which is used to accept the user ID and password of the users. If any new user wants to access the system, he/she can add their details and set their user ID and password by clicking the New User option.

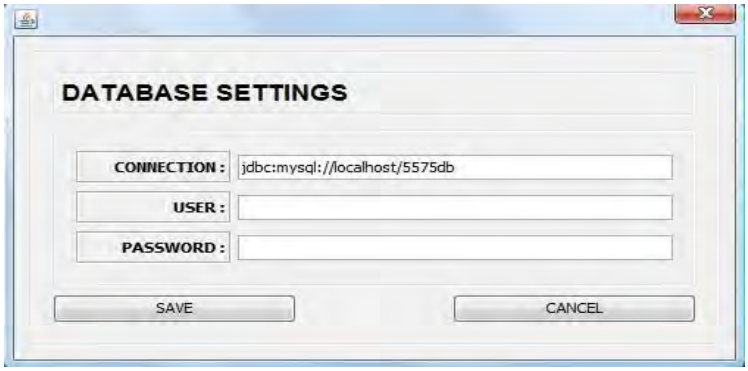

Figure 6: Creating connection to the database

Above figure shows the form through which the application is connected to the system's database. Here, the user enters the database's user ID and password.

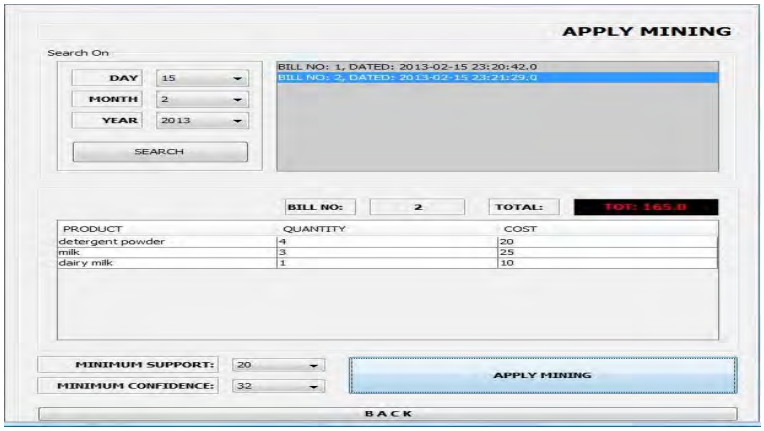

Figure 7: Applying Apriori algorithm to single bill

Above figure shows how a particular invoice bill is selected to apply Apriori rule on that bill. Here, the user needs to select the minimum support and minimum confidence which are required in the apriori algorithm.

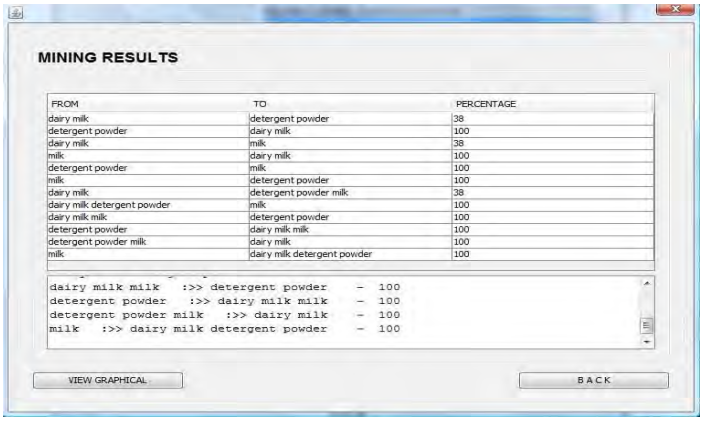

Figure 8: Frequent itemsets obtained

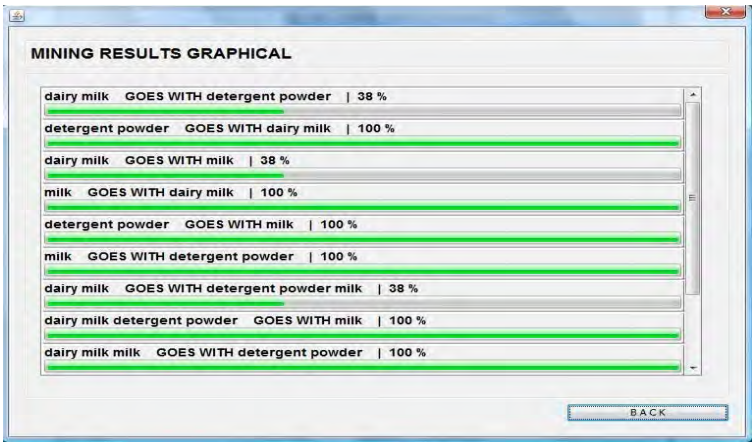

Figure 9: Apriori result in graphical form

Figure 8 & figure 9 shows the results obtained after applying Apriori algorithm on the invoice bill. Figure 8 shows the results in tabular form and figure 9 shows the results in graphical form.

The figure below shows the result obtained after applying K-means algorithm to the dataset. The dataset can be generated in a random fashion by using the generate random option shown in figure, or any file containing data values with .csv extension can be imported to use the values stored in it.

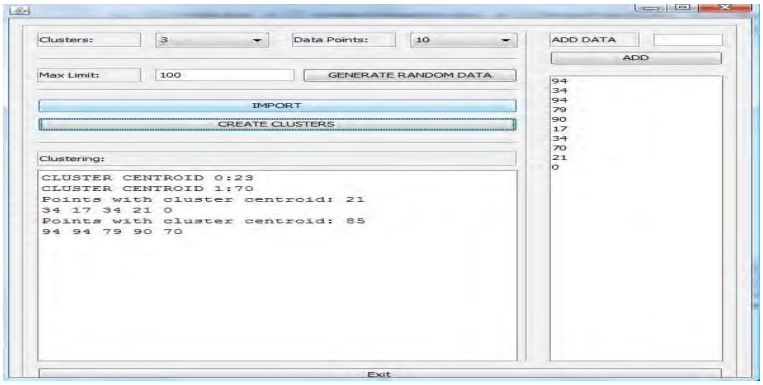

Figure 10: Creating clusters using K-means algorithm

VII. 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.

VIII. FUTURE SCOPE

In the future, we 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 Publishers, 2006, pp398-402.

[9] Ankita Naskar, Monika R. Mishra, "Using Cloud Computing to Provide Data Mining Services," in 978-93-81583-85-2 ICRITO '13, 2013, p. 435-439.

AUTHORS PROFILE

1. Name of first author: Ankita Naskar
   Dept: Information Technology Dept
   College Name: Smt. Kashibai Navale College of Engineering, Pune, India
   Email id: ankmit24@gmail.com
2. Name of second author: Prof. Mrs. Monika R. Mishra
   Dept: Information Technology Dept
   College Name: Smt. Kashibai Navale College of Engineering, Pune, India
   Email id:monika.r.mishra@gmail.com