Data Classifier for Encryption in Cloud

Ms. Arnika
Asst. Professor, Department of CSE, SRM University
NCR Campus Modinagar Ghaziabad Uttar Pradesh
jain.arnika2009@gmail.com

Vivek Verma
Research Scholar, Department of CSE, SRM University
NCR Campus Modinagar Ghaziabad Uttar Pradesh
vivek11235813@gmail.com

***Abstract*—** The aim of this project is to provide the cloud security infrastructure with a security policy such that the data encryption takes place as per need of data and provide suitable security to data based on the security needs of data with resource optimization in mind, as most of the data stored in cloud database would not be a classier document. Some of, or a high percentage of, data must be no encryption required-data. A model is needed which would not only classify the data but also handle the security concerns of data according to class which could be based on any security parameter, security criteria, or a set of conditions which are to be decided in accordance with the cloud service provider .On the basis of these predefined security criteria classes are formed and eventually various types of policies can be applied to the classes for which the cloud service provider agrees and is capable of. The data have different values and characteristics that must be identified before sending to cloud severs. As per the literature review so far the model present is a combination of K-Nearest Neighbor (K-NN) classifier and the Rivest, Shamir and Adelman (RSA) algorithm ,where data is classified into sensitive and non sensitive data ,only the sensitive data is encrypted using RSA algorithm. In a cloud server, the data are stored in two ways. First encrypt the received data and store on cloud servers. Second store data on the cloud servers without encryption. As the literature states, after implementing this model it is found that the confidentiality level of data is increased and this model is proved to be more cost and memory friendly for the users as well as for the cloud services providers. This classification technique uses binary number of class. The sensitivity of data can vary from more than just two levels. The data can grow two large to handle by K-NN and better technique exists to handle large data sets. For encryption any outsource algorithm would do more appropriate for cloud environment. This project proposes a model with more number of target classes which can handle the sensitivity of data more precisely and satisfy the need of security architecture. The model uses decision tree induction algorithm for classifier and the encryption algorithm can be used in varieties with a single class with no encryption. The classified data is then loaded into the servers into different data centres. The model aims at classification of data on basis of various security parameter or cloud policies.

Keywords: *Cloud Computing, Data Encryption, Classification before encryption, Decision Tree Classifier.*

*****

I. INTRODUCTION

Cloud Computing is an internet based distributed virtual environment. All computational operations are performed on cloud through the Internet [1]. The cost of the resource management is more than the actual cost of the resources. So, it is often better to get the required resources by renting despite purchasing one's own resources. Basically, the cloud computing provides all IT resources for rent. The simple definition of cloud computing is: "A distributed virtual environment provides virtualization based IT-as Services on rent". Beside all of the services like Software-as-a-Service (SaaS), Platform-as-a-Service (PaaS) and Infrastructure-as-a-Service (IaaS), cloud also provides storage as a service, in which distributed database servers are available for rent to consumers. These services are available for all users without any business or data bias. Consumers nowadays are using cloud services to avoid IT infrastructure purchasing and maintenance cost. A large amount of data can be stored on cloud. Cloud computing poses a number of challenging threats in the distributed storage model [2]. The data security is always the main challenging threat for quality of services and also stops the users to adopt cloud services [3].Data is a vital asset for any organization. Data could be in any forms, i.e. numbers, words, images etc. Data privacy and security is a crucial issue for any organization. Data deals with various properties such as accuracy, validity, reliability etc. Basic security issues of data include confidentiality, integrity and availability [4]. Data confidentiality deals with privacy of data which includes authentic and authorized access of sensitive data. Data integrity deals with the data content. Consistency and accuracy of data is required to achieve integrity. Data availability issues pertaining to foolproof storage, storage type, provisions for disaster recovery and backup plan.

In cloud storage, all kinds of data are stored on servers and they are stored through two storage methods. The first method is to encrypt received data and store on cloud servers. The second method is to store data on servers without encryption. These data storage methods can face data confidentiality issue. It is known that data are often not of the same type and have different properties and characteristics. In a cloud environment, a consumer's data are stored on remote servers that are not physically known by the consumer and there is high chance of confidentiality leakage. This study focuses on the confidentiality threat in the cloud environment. When a dataset is being transferred to cloud, it passes through a security mechanism, such as data encryption (without understanding the features of data) or directly being stored on servers without encryption. All data have different kinds of sensitivity levels. So, it would be non-technical to just send data into a cloud without understanding its security requirements. To address the security requirements of data, we have proposed a data classification model in the cloud environment to classify data according to its sensitivity level. Data is first classified using classification algorithm based on security parameters and then separate encryption for classified data is done. Data Classification is the process of defining various data levels and deciding a level of sensitivity to it. It is an essential activity at various stages as it is being created, modified, stored, or transmitted. The classifications of the data determine the extent to which the data needs to be secured and its value in terms of Business Assets. Data classification is done based on the various aspects. Some classify the data according to the risk associated with the disclosure. They are public, internal, confidential (or highly confidential), restricted, regulatory, or top secret. Some classify the data based on the way it is created, user personal data, their usage patterns etc. The classification mechanism we propose is decision tree induction classifier. As the tree classifier works a tree is induced out of sample data which is provided by cloud service provider, on the basis of which classification mechanism is defined. Once the tree is induced the test data can be classified into different sections or classes. Different encryption mechanisms can be applied to the classes. The accuracy of the classifier also depends upon the sample data provided.

II. RELATED WORK:

Data privacy and security concerns in cloud computing (or in any similar distributed information system) is always an issue. Storage and access mechanism proposed by various researchers and experimentations shows that in spite of having provisions for data security, various attacks and data leakage problems and still part of the cloud ecosystem. In such environment encryption mechanisms always exist. Data characteristics are analyzed with respect to online social networks by authors in [5]. They have identified the information and their leakage at the time of sharing data in the network. A crypto co-processor was suggested in [6] to solve the data security threats in cloud. The crypto co-processor is a tool which provides security-as-a-service on demand controlled by a third party. Crypto co-processor allows the users to select the encryption technique to encrypt the data and divide data into different fixed chunks. This is to make hacker not knowing the starting and ending points of the data. But the limitation with this study is that the end user may not be technically savvy enough to select powerful technique for data encryption and the additional overhead of a co-processor makes it unsuitable for many environments. IBM proposed a new concept of inner-cloud in 2010 which is completely different in terms of working and providing a secure cloud model. The inner-cloud is the clouds of a cloud. The inner-cloud storage model is more reliable and trustworthy as compared to a single cloud storage model. In the inner-cloud model, the hash function and digital signature are hybridized to provide data authentication and integrity in a cloud. Whereas the data security key is divided and shared in multiple clouds; but this process of sharing of keys leads to a key issue when one cloud is not available. Data security is studied as a part of survey by authors in [7]. Various other security issues are also analyzed and a trust based solution for the same is proposed.

The techniques and methods discussed above make use of encryption policies most of which encrypt the data unaware of the security sensitivity of the data. To encrypt complete data, it is very expensive in the context of time and memory. It would be better to separate the sensitive data from the public data first and encrypt only the sensitive data. The data classification is fundamental to risk assessment and useful for security control in organization [8]. Without understanding the importance of data, it is impossible to secure the business operations on data .We consider that data must be classified before any security mechanism can be applied to it. This should be done both to save the resources wasted in encryption of unnecessary data and to provide data with different classes on the basis of sensitivity of data. The technique discussed by authors in [9] is the most closest to this idea and provides us with a new paradigm in security “How to classify data on the basis of confidentiality in cloud environment?” In this model author used the K-NN machine learning technique in the cloud computing environment to solve the data confidentiality problem. To separate sensitive and non-sensitive data, the K-NN classifier is used in a designed simulation environment. The value of k is maintained to 1 for accuracy. The data is classified into two classes, confidential and public (non-confidential) data. The classification of the data depends on the attributes of the data. After finding the sensitive and non-sensitive data, the sensitive data is further transferred to the RSA encryption algorithm for data encryption to protect sensitive data from unauthorized users. Therefore, the non sensitive data is directly allocated a Virtual Machine (VM) without encryption. The VM will process the data and communicate with storage servers for the data storage on the cloud servers. This model do not superimpose any encryption technique on whole of the data but classify the data on the basis of security sensitivity of data and then applies public key encryption wherever necessary ,thus saving the data from under-security or over-security problem. The data classification helps determine what baseline security requirements/controls are appropriate for safeguarding that data. When talking about the shortcomings of author[9],the main drawback is the binary classification of data which is doubtable sufficient for data in cloud .Data in the cloud can vary in the level of security sensitivity .There is for sure more than two type of data stored in cloud ,considering the security concerns. Moreover, [9] treats all data as confidential except public data, suggesting no security requirements for public data. Secondly, model assumes that the user is the one who verifies the results of the data classification that the model has classified data accurately.

PROPOSED MODEL:

The proposed model uses the basic idea of data classification before encryption and enhances the method directed in [9], overcoming its potholes of less number of service levels and no encryption for public data. The method uses a decision tree induction based classifier trained on the basis of sample data provided ,hence replacing the K-NN classifier .The accuracy of the classifier will now depend on the sample data provided and manual quality check is not required. Things to be kept in mind when designing a sample data set is that if , there is noise in learning data sets, or the number of training examples is too small to produce a representative sample of the true target function, ID3 can lead to inaccurate decision making

.Data is classified into three classes ,first one for the data which do not need any security concerns ,second the private data of cloud encrypted with the symmetric encryption algorithm, third public data which has security requirement encrypted with public key encryption mechanism. The data can be handled separately classwise.Fig.1 shows the working of the proposed model assuming that the sample data provided for the classifier training is accurate enough to design a decision tree.

Our model uses the ID3 variant of decision tree induction algorithms. A decision tree is a tree in which each branch node represents a choice between a number of alternatives, and each leaf node represents a decision. Decision tree are commonly used for gaining information for the purpose of decision -making. Decision tree starts with a root node on which it is for users to take actions. From this node, users split each node recursively according to decision tree learning algorithm. The final result is a decision tree in which each branch represents a possible scenario of decision and its outcome. Decision tree learning is a method for approximating discrete-valued target functions, in which the learned function is represented by a decision tree

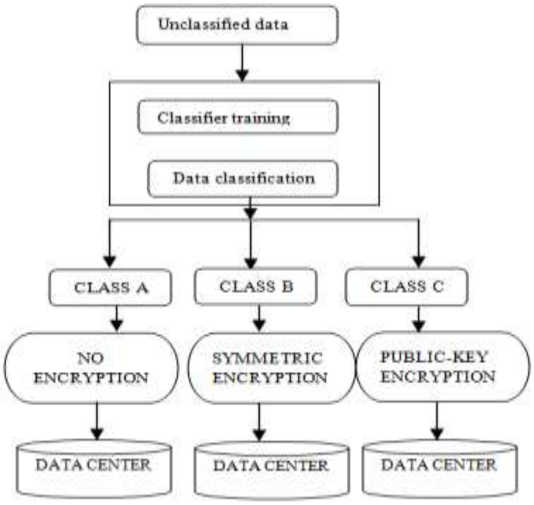

Fig.1 Proposed Model

ID3 is a simple decision tree learning algorithm developed by Ross Quinlan (1983). The basic idea of ID3 algorithm is to construct the decision tree by employing a top-down, greedy search through the given sets to test each attribute at every tree node. In order to select the attribute that is most useful for classifying a given sets, we use a metric---information gain. Information gain is calculated for each attribute and it works as the splitting criteria while constructing decision tree. Entropy can be easily generalized for number of classes' n > 2 using following formulae.

$$E(S)=-p_1\log p_1-p_2\log p_2...-p_n\log p_n=-\sum_{i=1}^{n} p_i\log p_i$$

Where p_i is the proportion of examples in S that belong to the i^{th} class.

For implementing the security based classification the sample data provided must be such that the tree build out of it can classify the data as per the need. While providing the sample data for classifier training the entropy gain has to be kept in mind because it is the main factor on which accuracy of the classifier depends.

The classifier divides the data into three classes according to the security policy decided by the cloud service provider which are expressed through the sample data. The number of classes are subjected to change according to the need of the cloud environment. In this model we assume three type of data .Class A contains the public data which has no security concerns and don't require any encryption mechanism .This classes data is directly stored in data centre. Class B contain data which require adequate level of concern and is of high confidentiality. This data can be internal data of cloud which is to be kept safe of the outer world or the public data which has trust on the service provider .This data is encrypted using symmetric key encryption. Symmetric-key algorithms are algorithms for cryptography that use the same cryptographic keys for both encryption of plain text and decryption of cipher text. The keys may be identical or there may be a simple transformation to go between the two keys. Any block cipher like AES can be used for this purpose. Class C contain public data which require safety measures. Keeping in mind the condition of not trusting the cloud service provider, a public key encryption can be used with more involvement of the client in the security procedure. Outsourcing algorithm can also be used.

Data can be stored on various data centre or in the same according to the design of the cloud .For the security purpose data can be spread across geographically dispersed data centre. And on the contrary can be located on the same data centre in separate relations .This totally depends on the architecture of the cloud.

III. CONCLUSION AND FUTURE WORK:

Data privacy and security is one of the major issues while dealing with the data storage in cloud. And encryption of the confidential data is the viable solution to the data leakage or any other attack .this study aims at finding the solution for keeping the data safe with minimal resource use .Classification of data before encryption doesn't only provides cloud infrastructure with various levels of security on which data is classified but also saves the resources otherwise wasted in case of over-security. In this way, it is easy to know which data need what security and which data do not need any security. Without data classification, the consumer may over-secure or under secure his/her data. The decision tree classifier provides policy builders to built the classifier on the basis of security policies of their choice .With certain level of designing the sample data provided can lead to a tree which is accurate enough to reflect the security policies in terms of classification. Analysis of variants of decision tree induction algorithm is proposed .The splitting criteria can be linked with the security-importance of the attribute. In future, work can be done in direction of a classifier which can be dynamic in sense of number of classes and its policies of classification which can be controlled on the basis of security requirements.

IV. REFERENCES:

[1] Rawat, P.S., G.P. Saroha and V. Barthwal, "Quality of service evaluation of Saas modeler (Cloudlet) running on virtual cloud computing environment using CloudSim", Int. J. Comput. Appl.2012, 53(13): 35-38.

[2] Deepanchakaravarthi P. and S. Abburu, 2012."An approach for data storage security in cloud computing", IJCSI Int. J. Comput. Sci. Issues, 9(2),: 1694-0814.

[3] Rittinghouse, J.W. and J.F. Ransome, 2009. "Cloud Computing Implementation, Management, Security", CRC Press by Taylor and Francis Group, LLC.

[4] Rizwana Shaikh and M. Sasikumar .Data "Classification for achieving Security in cloud computing", Procedia Computer Science ,45 (2015) 493 – 498.

[5] Balachander Krishnamurthy and Craig E. Wills, "Characterizing Privacy in Online Social Networks", Proceedings of the first workshop on Online social networks, WOSN '08, Pages 37-42, ACM New York, 2008.

[6] Ram, C.P. and G. Sreenivaasan, 2010. "Security as a service (SasS): Securing user data by coprocessor and distributing the data",. Proceeding of Trendz in Information Sciences and Computing (TISC, 2010), pp: 152-155.

[7] Rizwana Shaikh , "Security Issues in Cloud Computing: A survey". International Journal of Computer Applications 44,(19):4-10, April 2012

[8] Etges, R. and K. McNeil, 2006. "Understanding data classification based on business and security requirements". J. Online, 5: 1-8.

[9] Munwar Ali Zardari, Low Tang Jung and Mohamed Nordin B. Zakaria, ''Data Classification Based on Confidentiality in Virtual Cloud Environment", Research Journal of Applied Sciences, Engineering and Technology 8(13): 1498-1509, 2014.