THINK INDIA (Quarterly Journal)

ISSN:0971-1260
Vol-22-Issue-4-October-December-2019

Upgrade- Data Security in Cloud by Machine Learning and Cryptography Techniques

[1]Puli Srilakshmi, [2]K Lakshmi Viveka,[3]Y.Jnapika
[1]Asst.Professor, [2]Asst.Professor,[3]Asst.Professor
[1]Department of Computer Science and Engineering, [2]Department of Computer Science and Engineering,[3]Department of Computer Science and Engineering
[1]Pragati Engineering College, Surampalem, East Godavari, India,[2]Pragati Engineering College, Surampalem, East Godavari, India,[3]Pragati Engineering College, Surampalem, East Godavari, India,

[1]thotasrilakshmi1@gmail.com ,[2]viveka509@gmail.com, [3] jnapikay@gmail.com

ABSTRACT

Cloud computing is termed as the shared pool of configurable computer system resources and high level services that can be rapidly provisioned with a minimal management effort over the internet. Cloud computing has transformed the way organizations approach IT, enabling them to become more agile, introduce new business models, provide more services, and reduce IT costs. Cloud computing technologies can be implemented in a wide variety of architectures, under different service and deployment models, and can coexist with other technologies and software design approaches. The prevalent problem associated with cloud computing is data privacy, security, anonymity and reliability. Cloud computing provides the way to share distributed sources and services that belong to different organizations or sites. Since it shares distributed resources via network in open environment that makes it cause security issues. In this paper, the proposed work plan is to eliminate the concerns regarding to data privacy using encryption algorithms to enhance the security in the cloud. In this method some important security services including authentication, encryption and decryption, compression are provided in cloud computing system.

KEYWORDS

Keywords: Cloud computing, Cryptography, Data Classification, Data Security, Decryption, Encryption, HMAC function, KNN technique, Machine Learning, RSA Algorithm

1. INTRODUCTION

Cloud computing is an emerging virtual distributed environment that uses the ideas of sharing, power processing, storing, connectivity, and virtualization. Communicating through broad network i.e., Internet cloud facilitate a large pool of resources, storage media and sharing media that helps to supply on-demand services. This will help the end users in complying with the ideas of isolation, elasticity, security and distribution. Security issues are the foremost difficult problem in cloud domain and the most vital obstacle for aggrandize of IT based companies that provide users on-demand services. These security issues can be visualized at application phase, network phase, authentication phase, authorization phase and virtualization phase.
There are two reasons for the security concerns in the cloud computing:

ISSN:0971-1260
Vol-22-Issue-4-October-December-2019

1. Now a days everyone is storing their data on cloud database, so the main focus is that the user's data should be safe and the vital information shouldn't get tampered when travelling from one place to different across the network. It is necessary that Integrity, Confidentiality and Availability of user data must be ensured.
2. The unauthenticated user tries to access the authenticated user's data.

We can apply cryptographic algorithms in cloud servers to solve these threats. However when a user is revoked, using a single cryptography algorithm is not adequate to assure the confidentiality of data and managing the access control methods in cloud computing environment. For data security these techniques are applied on encryption. Encrypting complete data can turn to be very expensive in terms of memory as well as for time. So, to solve this problem it would be better if we first separate our sensitive data and then apply encryption algorithms. It would address reliable results if we classify the data according to its sensitivity level. In machine learning field, the data classification is a method of distinguishing the category of unclassified data sample set with the help of build classifier. The classifier is constructed by constructing a training set of familiar data samples. To create an appropriate classifier, a large range of justified training data samples are required. This advancement invites a new paradigm of services where data classification is offered by servers in a cloud to its various clients/users. Specifically, the server will process the data automatically and hence, categorise the clients data samples present on remote private servers. However un-trusted third party-servers can access the private data. Moreover, any vital detail or training data set specifications may not be disclosed by the servers even if it provides the classification services to its client. Thus, a mechanism that ensures the privacy of the server's training set and client data samples is required. Hence, a re-encryption model is essential requirement to forestall the revoked user from accessing the encrypted information as well as to generate reliable keys for valid users.

Objective of Our System:

1. To develop a system that will Provide Security and Privacy to Cloud Storage.
2. To Establish an Encryption Based System for protecting Sensitive data on the cloud and Structure how owner and storage Service Provider to operate on encrypted Data.
3. To Create a System where the user store its data on the cloud the data is sent and stored on the cloud in encrypted form As in normal cases in cloud computing when a user login to the cloud and they store data on cloud storage device the data stored on the server cloud is not such secure as it can be readable to anyone which have permission to access and Leaving data vulnerable.
4. To Develop a retrieval System in which the data is retrieved by the user in encrypted form and is decrypted by the user at its own site using a public and private key encryption both the keys working at the user level.

ISSN:0971-1260
Vol-22-Issue-4-October-December-2019

PROPOSED WORK

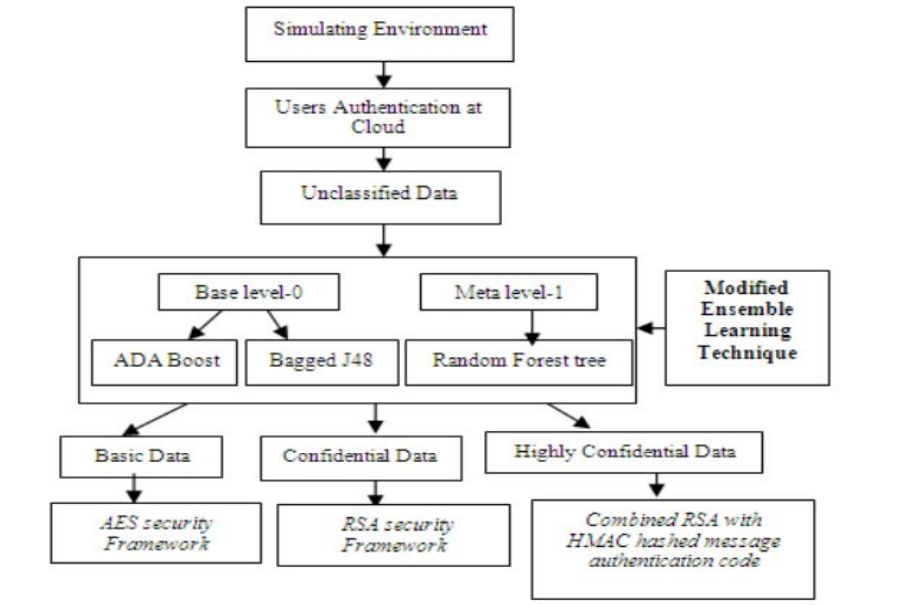


fig 1: Proposed Model

Data classification is the task of identifying data sets with respect to its data value. These values are based on usage consumption of data by its users and restrictions on access control methods. KNN (K-Nearest Neighbours) technique is used in machine learning artificial intelligence that helps to categorise the class of unclassified data with the help of build classifier. It is constructed by employing a training set of familiar data samples. In this proposed work, the modified Ensemble Learning Technique is used to enhance the execution of existing KNN technique. Ensemble learning method comprises a set of different models are group together to improve the prediction and stability power of any model. It has two levels: base level-0 and Meta level-1 as shown in fig: 2. At base level a no. of algorithms can run i.e., AdaBoost and bagging algorithm. At Meta level, also known as decision making algorithm, random forest tree is used. To utilize the training sets of data which is given by KNN model, the training set is computed with Euclidean distance function. To reduce the computational density, we improved the basic algorithm of ensemble learning.

Working of the Algorithm:

The classifier is evaluated using a cross validation (K-fold) technique. Each layer is trained as:

ISSN:0971-1260
Vol-22-Issue-4-October-December-2019

The dataset is split into two sets: training set and testing set.

- For each base level layers classifiers are generated using Training set.
- The predictions generated by the base level is assembled and forwarded to meta level. At meta level, validation process is performed and provides the final predicted results.
- Finally, by using complete layer's datasets (not the use of simplest out of k folds of the dataset) we can train each layers of the training set.
- Now to secure highly confidential data we combine RSA encryption technique with HMAC. The HMAC(hashed message authentication code) is a cryptographic checksum. It is stored at local machine as well as appended with the cipher text and sent to the cloud. The HMAC+ cipher text must be matched with the HMAC checksum i.e, already stored at the local machine of user. The hacker finds it difficult to guess the HMAC checksum while hacking the data. Thus, results in increasing the security of user's HC data by using RSA encryption technique.
- HMAC is calculated by following formula: HMAC(Key,msg)=HMAC(KeyXORout)‖H((Key XORin)‖msg)) Where Key is the original key, m is the message that needs to be authenticated.

Simulation on Cloud Computing:

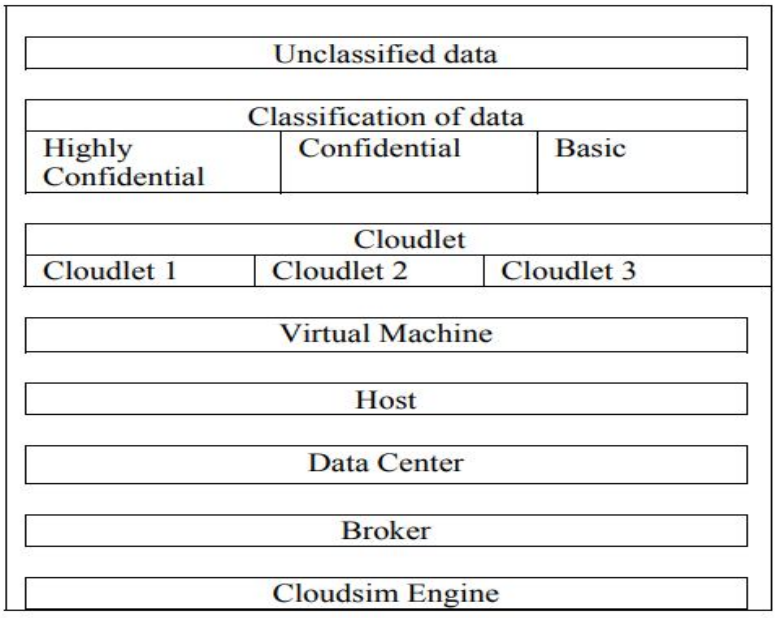


fig 2: Proposed Simulation Environment

The cloudsim as shown in fig-2 is used for the proposed model that solves the problem of confidentiality and classification of data. Firstly we run the simulation then data centres' are build. The virtual machine manager (VMM) is used to handle the VMs and for allocating

ISSN:0971-1260
Vol-22-Issue-4-October-December-2019

VMs to cloudlets i.e., the cloud task. Authentication process is performed so that only authenticated users can access the data. Then classification KNN is combined with modified Ensemble learning technique. This will improve the prediction capabilities and accuracy of existing KNN.

Algorithm for data classification:

Step 1: To select Beta and Meta layers

```
I/p: o.ds: DS, folds: Int
DS←o.ds
AoC as classifiers array that consist AdaBoost and Bagged
Decision Tree
For lr= 0 to 2 do:
For each fold in folds do:
If lr ≠2 do:
AOC←Trn_S.Lr (lr, t.s, folds)
In_New←Classify (test-set,AoC)
Adding In_New to ds [lr+1]Else
Trn_S.Lr (lr, ts, folds)
Lr = lr + 1
Trn-S.Lr (lr 0, o-ds)
```

Step 2: Single Layer Classification

```
Trn-S.Lr I/p: Lr
No: Int, ds: DS, folds: Int
O/p: Scr-Ds: DS
Scr DS ← emp Grp
 For each fold in folds do:
 B.C (Lr No, ts)
For each In. in ts
Produce probabilities-vector by applying In. on current
layer's classifiers.
build a new In. from probabilities-vector
Add the new In. to Scr-DS
Ret Scr-DS
```

Where: o.ds: Original dataset, DS: Dataset, Int: Integer, AoC: Array of classifier, Lr: Layer, Ts: Training set, In: Instances, Scr: Successor.

Properties and description of Cloud IaaS Model:

Before starting the simulation it is necessary to set the properties of IaaS(Infrastructure as a Service). The properties are as follows:

Table 1: The properties of IaaS are associated with the data centres. These data centres are assigned to VMs.

Data centre ID	Storage limit	RAM in MB	Architecture of data	Operating System	Bandwidth

ISSN:0971-1260
Vol-22-Issue-4-October-December-2019

0	100000000	4096	x84	Linux	10000
1	100000000	4096	x84	Linux	10000

Encryption Algorithm:

This algorithm emphasizes on improving classical encryption techniques by integrating substitution cipher and transposition cipher. Both substitution and transposition techniques haveused alphabet for cipher text. In the proposed algorithm, initially the plain text is converted into corresponding ASCII code value of each alphabet. In classical encryption technique, the key value ranges between 1 to 26 or key may be string (combination alphabets). But in proposed algorithm, key value range between 1 to 256. This algorithm is used in order to encrypt the data of the user in the clouds. Since the user has no control over the data after his session is logged out, the encryption key acts as the primary authentication for the user. The algorithm is described below.

Following are the steps involved in the encryption of data

Step 1: Count the No. of character (N) in the plain text without space.

Step 2: Convert the plain text into equivalent ASCII code. And form a square matrix(SXS>=N).

Step 3: Apply the converted ASCII code value from left to right in the matrix. Divide matrix into three part namely upper, diagonal and lower matrix.

Step 4: Read the value from right to left in each matrix.

Step 5: Each matrix use three different key K=K1, K2, K3 for encryption. Do the encryption.

Step 6: Apply the encrypted value into the matrix in the same order of upper, diagonal and lower.

Step 7: Read the message by column by column. Here the order in the columns read from the matrix is the key K4.

Step 8: Convert the ASCII code into character value.

Results:

The results of our system are the comparison of existing work with the proposed work has been shown. According to the following analysis, it has been observed that the system proposed in this research work is giving improved and reliable results. For higher degree of confidentiality and security different cryptographic techniques are required. The performance of the proposed system has been evaluated by following parameters i.e., Accuracy, Classification details, Encryption Time and Decryption Time, Error Rates.

1. Accuracy: It is the measure of correctly classified instances to the total number of correctly and incorrectly classified instances. TP: True Positive, TN: True Negative, FP: False Positive, FN: False Negative.

$$\text{Accuracy}=\frac{TP+TN}{TP+FP+TN+FN}$$

2. Classification Details:
   - Precision and recall: Both are used for evaluating execution capability and in data analysis field like retrieving data and data mining. Precision measures the exactness and recall measures the completeness of data.
     i. $\text{Precision}=\frac{TP}{TP+FP}$
     ii. $\text{Recall}=\frac{TP}{TP+FN}$
   - F-measure: It is the Harmonic mean of precision and recall.

$$\text{f-measure}=\frac{2*Recall*Precision}{Recall+Precision}$$

ISSN:0971-1260
Vol-22-Issue-4-October-December-2019

3. Encryption time: Encoding information or any message in such a way that only authorised party can be able to access it.

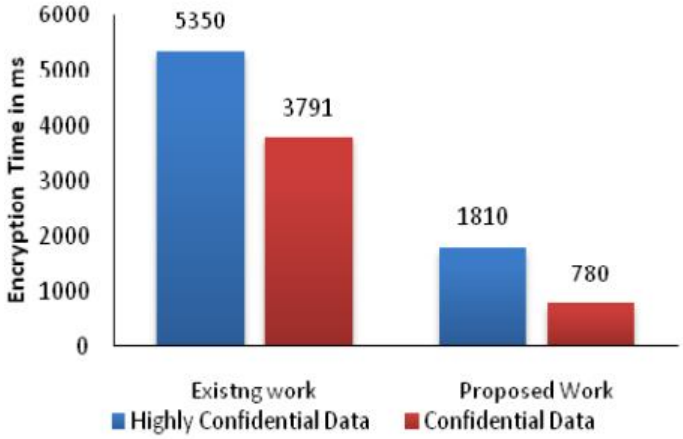


fig 3: Comparison of encryption time b/w the proposed and existing model

4. Decryption time: It is the process of reconverting the encrypted data or message into its original form.

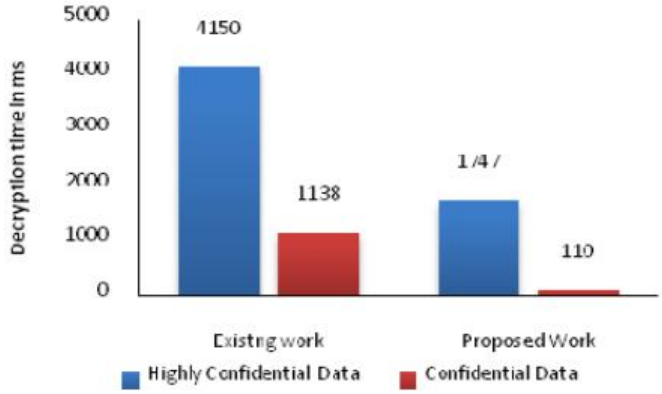
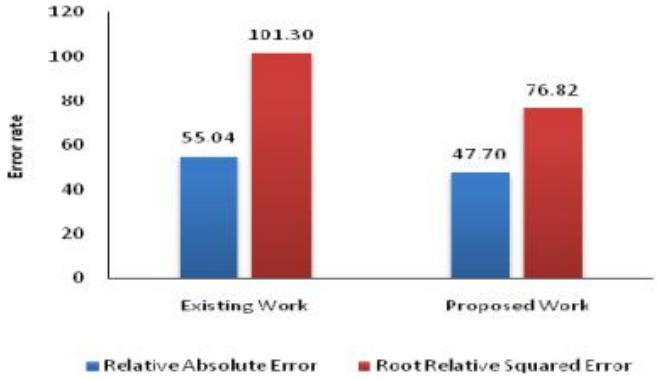


fig 4: Comparison of decryption time b/w the proposed and existing model

5. Error rate: The measurement of the effectiveness of a communications channel.



ISSN:0971-1260
Vol-22-Issue-4-October-December-2019

fig 5: Comparison of error rates b/w the proposed and existing model

$$\text{Error rate} = \frac{no\,of\,erroneous\,units\,of\,data}{total\,data}$$

Advanced techniques used in securing data in the cloud:

Below are the three main techniques that are involved in securing data in the cloud. These usually perform with the option of blocking of confidential data. They are:

1. Data Masking
2. Secure logic migration and execution technology
3. Data traceability technology
   - **Data Masking:** Data masking is a method of creating a structurally similar but inauthentic version of an organization's data that can be used for purposes such as software testing and user training. The purpose is to protect the actual data while having a functional substitute for occasions when the real data is not required.

   In data masking, the format of data remains the same; only the values are changed. The data may be altered in a number of ways, including encryption, character shuffling, and character or word substitution. Whatever method is chosen, the values must be changed in some way that makes detection or reverse engineering impossible.

   Vendors of data masking products include Compuware, dataguise, IBM, Informatica and Oracle.
   - **Secure logic migration and execution technology:** For confidential data that cannot be released outside of the company, even formed by concealing certain aspects of the data, by simply defining the security level of data, the information gateway can transfer the cloud-based application to the in-house sandbox (A protected program execution environment that prevents fraudulent tampering of data) for execution. The sandbox will block access to data or networks that lack pre-authorized access, so even applications transferred from the cloud can be safely executed. Moreover, because the execution status of applications is recorded, application providers are able to confirm if there is any inappropriate use of the data.
   - **Data traceability technology:** The information gateway tracks all information flowing into and out of the cloud, so these flows and their content can be checked. Data traceability technology uses the logs obtained on data traffic as well as the characteristics of the related text to make visible the data used in the cloud. For example, in a joint development project, one can check how textual data collected in the cloud have been used, including whether portions have been copied, thereby enabling any inappropriate usage to be identified.

ISSN:0971-1260
Vol-22-Issue-4-October-December-2019

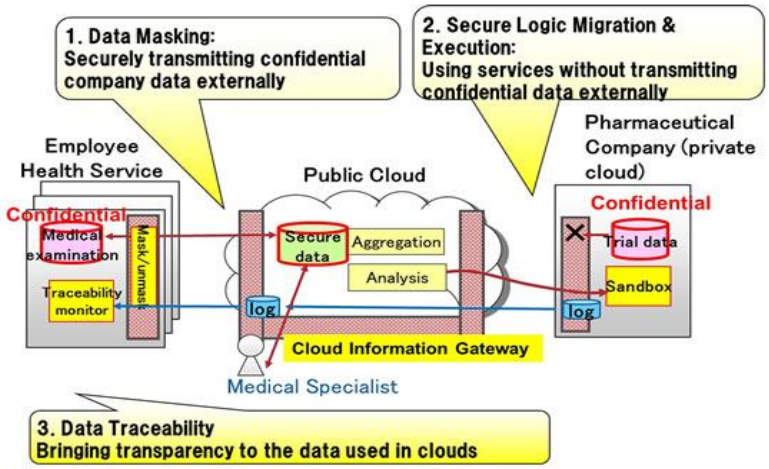


Fig 6: usage scenario for newly developed data gateway

Conclusion:

This paper aims at achieving data confidentiality, data access control management and provides an automatic classification of data in cloud. On the contrary, the existing system was lacking in providing automatic classification of data. In this paper the functionality and performance of existing KNN technique is improved with modified Ensemble Learning technique. The data will be automatically classified by the machine on the basis of data security parameters. Also to emphasis more focus on highly confidential data HMAC (Hash-based Message Authentication Code) function has been appended with the existing RSA encryption algorithm. The proposed system proves to be more accurate and economical. And also saves the user time for encrypting/decrypting different classes of data (basic, confidential and highly confidential).

Future Work:

This technology will now undergo verification in environments where multiple clouds are working in collaboration, with commercialization targeted for 2019.

References:

[1] http://en.wikipedia.org/wiki/Cloud_computing
[2] http://www.roughtype.com
[3] F. F. Moghaddam, M. Vala, M. Ahmadi, T. Khodadadi, and K. Madadipouya, “A reliable data protection model based on re-encryption concepts in cloud environments,” Proc. - 2015 6th IEEE Control Syst. Grad. Res. Colloquium, ICSGRC 2015, pp. 11–16, 2016.
[4] N. Surv, B. Wanve, R. Kamble, S. Patil, and J. Katti, “Framework for client side AES encryption technique in cloud computing,” Souvenir 2015 IEEE Int. Adv. Comput. Conf. IACC 2015, pp. 525–528, 2015.
[5] V. K. Pant, J. Prakash, and A. Asthana, “Three step data security model for cloud computing based on RSA and steganography,” 2015 Int. Conf. Green Comput.Internet Things, pp. 490–494, 2015.
[6] L. Tawalbeh, N. S. Darwazeh, R. S. Al-Qassas, and F. AlDosari, “A secure cloud computing model based on data classification,” ProcediaComput.Sci., vol. 52, no. 1, pp. 1153–1158, 2015.
[7] R. Shaikh and M. Sasikumar, “Data classification for achieving security in cloud computing,”

ISSN:0971-1260
Vol-22-Issue-4-October-December-2019

ProcediaComput.Sci., vol. 45, no.C, pp. 493–498, 2015.
[8] M. A. Zardari, L. T. Jung, and N. Zakaria, "K-NN classifier for data confidentiality in cloud computing," 2014 Int. Conf. Comput. Inf. Sci. ICCOINS 2014 - A Conf. World Eng. Sci. Technol. Congr. ESTCON 2014 - Proc., 2014.
[9] Dr. A. Padmapriya, P. Subhasri," Cloud Computing: Reverse Caesar Cipher Algorithm to Increase Data Security", International Journal of Engineering Trends and Technology (IJETT) – Volume 4, Issue 4, pp 1067-1071, 2013.
[10] Quist-AphetsiKester, "A Hybrid Cryptosystem Based on Vigenere Cipher and Columnar Transposition Cipher", International Journal of Advanced Technology & Engineering Research (IJATER), Volume 3, Issue 1, pp 141-147, 2013.
[11] Randolph Barr, QualysInc, "How to gain comfort in losing control to the cloud"