The Secure Migration of Data to Cloud using Data Sanitization and MAC address based AES

Hitesh Marwaha, Rajeshwar Singh,

Abstract: Cloud computing is the most emerging technology in current information technology. It is internet based technology where a user has to pay as per usage. However the major challenges for the wider acceptance of cloud are privacy and security of data in cloud. The mathematical model of data sanitization for giving false look to sensitive data before transferring data to cloud and mac address dependent AES technique for transferring non sensitive data and sanitized data is proposed in the paper.

Keywords: Cloud Computing, CIS, Data Sanitization, Encryption, AES, MAC address.

I. INTRODUCTION

Cloud computing is an internet based technology used to share hardware and software resources. Cloud computing model is based on pay as use model like other utilities such as electricity, water and gas we use in our day to day life. Similarly, in information technology only useful information can be delivered customer whenever required. As in day to day life all internal complexities of generation utilities like electricity, water etc. are hidden from the user, in the similar way computing, is considered to fully virtualized in cloud [1]. Buyya et al. [2] have defined it as “Cloud is a parallel and distributed computing system consisting 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 (SLA) established through negotiation between the service provider and consumers.”

1.1 Services of Cloud Computing

Cloud computing services are divided into three classes, according to the abstraction level of the capability provided and the service model of providers namely [3]:

- Infrastructure as a Service (IaaS)
- Platform as a Service (PaaS)
- Software as a Service (SaaS).

The abstraction level is viewed as the layered architecture in which services of higher level can be composed from services of underlying layer. [4] Infrastructure as a Service (IaaS): In IaaS the end users are offered direct access to computing resources empowering them to willingly exercise their resources [3]. Examples of IaaS include Amazon Elastic Compute Cloud (EC2), Rack space, and IBM Computing on Demand.

- Platform as a Service (PaaS): In this model, using tools and libraries hosted by CSP, end-user creates, test and upload applications. [5]. in this model, using tools and libraries hosted by CSP, end-user can create, test and deploy applications. [9]. Example of PaaS includes Google App engine which offers the end user a scalable environment to deploy and test web applications using python or java.

- Software as a Service (SaaS) is where users are shifting to online systems from locally installed computer programs to SaaS is where the user has not to pay for purchasing a required software. User pay as per usage as software is in the possessions of CSP [6].

- Data as a Service (DaaS): DaaS implies that data can be offered on the users demand throughout universe irrespective of demographic organizational divorce between provider and consumer [11]. Data can be stored using two methods vis-à-vis locally or on cloud. Data stored locally in hard drive is more susceptible to various threats like theft, breaking down through natural disasters like floods or fire. Although storing data in the local site is cost wise cheaper than data stored in cloud.

1.2 Cloud Delivery Models:

The cloud appeared to be a public as per services and uses provided at large, however based upon the physical location of cloud utilities it is further classified as public, private, community, or hybrid [3] According Armbrust et al. [12] The public cloud is “cloud made available in a pay-as-you-go manner to the general public” and the private cloud is “Internal data center of organization, not made available to the general public.”

- Private Clouds

Cloud’s services owned by organizations for their personal use are known as Private clouds. It requires ample amount of resources to purchase all essentials for the cloud and the crew of IT professionals.

- Public Clouds

When the clouds services are open to any customer for the computing services is known as public cloud. Cloud service provider must be viable and reliable as the whole business of organizations is dependent on it; hence it is a disadvantage. Public clouds are the most deployed cloud environments [13].

- Hybrid Clouds

Hybrid cloud conflates both public and private clouds. An organization exercising private cloud can use public cloud resources in addition to their own cloud.
The Secure Migration of Data to Cloud using Data Sanitization and MAC address based AES

- **Community Cloud**
  A community cloud is shared by several organizations that have similar concerns like missions, security requirements and policies. [3]

II. DATA SANITIZATION:

“The process of camouflaging sensitive information in a test and development of databases by overwriting it with realistic looking but fictitious data of a similar type is known as Data Sanitization” [14]. Here an attempt has been made to provide tools which will enhance the current capabilities of the cloud while camouflaging display of sensitive data, thus, benefiting the cloud user and the cloud provider through enhanced business intelligence. Moreover, sanitized data can also be used for activities related to development, testing, and training because actual data is not an essential requirement for the same. So the integrity of the database is maintained while providing opportunities for progress. Following are some data sanitization techniques used for giving the data false look. Every technique has its own pros and cons: [14]

- **Masking Data**
  Masking data means replacing some fields in the data set with a Mask character (such as an X). This technique efficiently disguises the data content while maintaining the same formatting on front end screens and reports.

- **Substitution**
  This technique haphazardly replaces the contents of a column of data with information that is completely distinct but looks similar to real data. Substitution is efficient technique as it preserves the real look and feel of data.

- **Shuffling Records**
  In Shuffling records technique data in the columns is arbitrarily moved between rows and while shuffling it should be kept in mind there should not be any kind of correlation with information in the row. In other words there must not be symmetrical shifting of record. Like substitution technique, it also preserves the originality of data in term of look and feel.

- **Number Variance**
  The Number Variance technique is applicable on numeric data. In this technique each value in the record is modified by some percentage of real value. This technique provides the realistic disguise for the numeric data.

- **Gibberish Generation**
  While sanitizing data, one must take due consideration to eliminate all imbedded references to the real data. This is especially true if the original record can be determined via a simple join on a unique key. Sanitizing “vague” non-explicit data, for example, letters, notices and notes is one of the hardest methods in Information sanitization. Generally these kinds of fields are simply substituted with an arbitrary amount of comparably estimated nonsense or irregular words.

- **Encryption/Decryption**
  This technique offers the alternative of leaving the data in place and unmistakable to those with the proper key while remaining adequately pointless to anyone without the key. The quality of the encryption is likewise an issue. Some encryption is more secure than others. The experts opinion about, most encryption frameworks is that it can be broken – it is simply an issue of time and exertion.

III. RELATED WORK

Ahmad, N. (2017) has reviewed the major security issues in cloud and concluded that security is the major challenge in almost all the technologies concerned with the implementation of cloud. Not only the implementing technologies like virtualization and web services are lacking in security but the cloud architecture, SLA, and deployment model also requires proper security mechanism to win the trust of customer [15]. Sun, X. (2018) a survey was accomplished to review all critical security aspects of cloud computing less than three different parts viz. Computer security, network security, and information security. Finally, it was concluded that major security issues in cloud computing derived from both insider and outsider threats. A mathematical encryption based security protocols FHE is proposed in the paper which encrypts the sensitive at cloud server side [16]. Al-Ahmad et. al. (2018) demonstrated by numerous challenges because it is internet based technology, thus it requires high security and have many technical complexities due to physical location of data and data transfer bottleneck. The cloud security issues identified in this paper the major hindrance to maintain the confidentiality and privacy of data in cloud is sharing of resources, insiders rogue employee, and internet vulnerabilities [17]. Shaikh & Modak emphasized that data security in cloud is always a major concern from the point of view of both consumer as well as CSP. The paper also addressed the major security challenges as per comprehensive analytic report published by CSA concerning various security issues in the cloud. A model comprising of various parameters for data security is proposed. Some numeric weights are assigned to these parameters and these parameters acts as a reference to customer for selecting CSP [18]. Lee, B. H. et. al. (2018) proposed data security in cloud computing using AES in reference to Hurok cloud (a PaaS cloud) is proposed. The performance evaluation of AES shows that it can be used to encrypt the data before migrating to cloud security. It was concluded that larger size of data leads to data delay time while encryption of data using AES [19]. Rajeswari & Kalaiselvi (2017) conducted a survey of the various techniques proposed by different authors for data security in cloud is conducted on the basis of parameters such as privacy confidentiality, integrity, access control and storage security. Finally it was concluded that data and storage security must be provided without storage and computational overhead. Authentication, authorization confidentiality and integrity must be guaranteed for data security.
Akhil, & Pushpa(2017) proposed AES based encryption model. Akhil, & Pushpa (2017) proposed AES based encryption model for encrypting data at client end and concluded that it lessens the possibility of entering of intruder into the network[21]. Bouchaala, M. et. al. concluded that Security is the major concerns that obstruct the adoption of Cloud Computing at large. From the viewpoint of security an end-to-end cloud architecture is proposed in the paper. In the proposed model Cloud environment is divided into four major parts: end user, network, third parties and Cloud provider. The security aspect of all these entities is considered in the proposed reference architecture. However, the proposed model is not validated by the authors [22]. Sharma, P. K et. al. (2017) pointed out that security and privacy are major hindrance in adoption of cloud computing at large. Although it is verified reality that cloud computing platform are cost effective solutions to share hardware and software resources yet before actually moving into a cloud infrastructure, it must be ensured that proper security measures have been taken in to account. The paper also throws light on various security risks associated with various service levels of cloud [23].

### IV. PROPOSED WORK

After extensive study of various challenges in cloud, security and privacy of data are two major constraints due to which user hesitates to migrate their sensitive data in cloud. Many security techniques are proposed by different researchers. Every technique has its own pros and cons. Out of the various techniques for securing data in cloud and winning the trust of users by cloud service providers the encryption technique, although the most traditional technique yet appreciated by many researchers.

#### 4.1.1 Encryption based data sanitization technique:

The aforementioned technique sanitizes the sensitive data before transferring to CSP. As discussed, there are various methods to sanitize data, but we propose encryption method to sanitize the sensitive data before handing over data to CSP. To avoid the unauthorized access of sensitive data by rogue employees at CSP, it is advisable to give the false look to the data before transferring it to CSP. Data at any organization can be categorized into two parts:

(i) Sensitive Data

(ii) Non Sensitive Data

#### 1.1.2 Sensitive data

Sensitive data may include secret business data that put organization in risk if leaked in the competitive market, trade secrets and financial information of business organization etc. While migrating data to cloud both kinds of data may be given different treatment. Like in real life while going out of home we just hand over the keys of home to near or dear but never even tell the location of the keys of the locker where all precious things are placed. In similar manner Sensitive data items must be inferred from non-sensitive data through some inference process based on need and knowledge of the owner of data [32]. An attempt has been made to give typical false look to sensitive data before migrating data to cloud.

Seven layered architecture is proposed for the same.

#### 1. Data Layer:

It is the first most layers at the client side which consists of actual data comprises of sensitive as well non sensitive data. Here, the owner of data takes the decision which data is to be migrated by giving false look. Suppose DATA is whole data set to be migrated to cloud. First, it is classified into two parts X (sensitive data) and Y (Non Sensitive data).

Suppose following is the data set to be transferred:

\[ Y = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\} \]

#### 2. Segmentation Layer:

In this layer sensitive data to be stored in the cloud is divided into several parts according to the Mutually Exclusive and Mutually Exhaustive criteria. According to aforementioned criteria, we divide the data set into several parts in such a way that the union of the domains of each part must be equal to the domain of the whole data set and the intersection of domains of each part must be null.

In general, suppose the whole data set say \( X \) having domain \( D \) is divided into \( n \) parts (say)

\[ A_1, A_2, \ldots, A_n \]

An and having domains \( D_1, D_2, \ldots, D_n \) respectively such that

\[ D_1 \cup D_2 \cup \ldots \cup D_n = D \]

\[ D_1 \cap D_2 \cap \ldots \cap D_n = \phi \]

Divide the above data set into three parts say

\[ A_1 = \{2, 3, 5, 7, 9\} \]
\[ A_2 = \{1, 4\} \]
\[ A_3 = \{6, 8, 10\} \]

#### 3. Selection Layer:

To give sensitive data a false look proposed model will have \( n \) number of different mathematical functions. In this layer user will randomly select the appropriate user defined mathematical functions according to the number of data sets. If the data sets are in small in number then equal number of user defined functions can be selected otherwise data sets will further be divided into groups using some criteria and similar user defined mathematical functions are selected for a pair of data sets. The complexity of the type of the proposed user generated mathematical function can be increased to secure data from unauthorized access.

Let us define functions \( f(x) \), \( g(x) \), \( h(x) \) on data sets \( A_1, A_2, A_3 \) respectively as:

\[ f(x) = (x^2 + 0.5)^2 + 5 \]
\[ g(x) = x^3 + 9 \]
\[ h(x) = x + \pi \]

#### 5. Operational Layer:

In this layer appropriate mathematical functions are applied on the data sets produced in segmentation layer. Dividing data into different data sets and applying different sanitization technique on these sets will lead to multi-level encryption and gives false look to the data.
6. **Assembly Layer**: The various data set produced in operational layer are reassembled to form a single data set again and is ready to migrate to CSP over internet.

   \[ Y = \{10, 11.75, 17.75, 73, 35.75, 9.14, 61.75, 11.14, 95.75, 13.14\} \]

7. **Application Layer**: The data is transferred to the cloud service provider using secure standard protocols to ensure integrity and confidentiality of data. Some standard encryption algorithm using the concept of public key will be applied to the false data. The process of transferring the data through standard protocol is explained in section 4.2.

8. **Physical Layer**: It includes the physical infrastructure of CSP which includes various types of servers, networking equipment and hardware and software components and various information security algorithms for securely storing the data at their end to win the trust of the client.

---

**Fig. I**

---

**Data Layer**

**Segmentation Layer**

**Selection Layer**

**Operational Layer**

**Assembly Layer**

**Application Layer**

**Physical layer**

---

**Fig. II** (The process of transferring data to CSP)

---

**4.2 Enhanced AES Algorithm**

The proposed algorithm is modification over primitive AES in two manners:

- **Secure MAC key Generation**
- **Key Dependent S-Box Generation**

**4.2.1 Secure MAC Key Generation**

Secure MAC Key Generation algorithm (SMKG): The proposed technique generates unpredictable key in using MAC address of the computer from which we want to transfer the data to CSP.

It generates non identical block keys from the user provided 256 bit secret key to encrypt every block of plain text. Secured hash algorithm SHA 256 is applied to convert MAC address into 256 bits. Exclusive OR operation is applied on the secret key and the MAC address of the computer. The detailed process is depicted in fig. III.

---

**4.1.3 Benefits of proposed technique:**

1. **Multilevel Encryption**: Dividing data into different data sets and applying different encryption criteria on these sets will lead to multilevel encryption. If in case, hacker got succeeded to reveal one of your functions then it would be impossible for him to reveal the next value by using the same key.

2. **Save from Insider attack**: A malicious employee at CSP can quickly leak client’s confidential data and valuable trade secrets. The proposed technique avoids the aforementioned risk by sanitizing sensitive data before handing over to CSP.
Algorithm – I (Generation of Key)

1. Let KEY = private key.
2. Apply SHA-256 to KEY and Set it to KEY_256bit.
3. Let MAC = MAC address of the computer from where data is to be migrated.
4. Apply SHA-256 to MAC and Set it to MAC_256.
5. NEWKEY = KEY_256bit ⊕ MAC_256 [Apply XOR on secret key and 256 bit MAC address.]
6. Return

Algorithm – II (Generation of Key dependent S – Box)

Let us suppose after application of key generation algorithm following key in hexadecimal form is generated

| 59 | AF | 5a | E3 | 87 | F6 | DE | 98 |
| 3C | E9 | 01 | C7 | B9 | 18 | F8 | F1 |

The procedure generates a 256-bit secure key using private key (KEY) and MAC address (MAC) of the computer.

1. Call MACKEY(KEY, MAC)
2. Apply XOR on each byte of KEY returned by Step I and Set the value so obtained to X
3. A= S-BOX
4. Shift every element of matrix A by X bits and Set revised matrix to DYN_SBOX

4.2.2 Key Dependent S-Box Generation (Dynamic S-Box)

Advanced Encryption standard uses static S-box for encryption generated by calculating multiple Inverse of each byte ranging between 00 to FF in hexadecimal form. While S-box used in decryption is inverse of that used in encryption. In proposed technique S box is dynamic and generated by applying algorithm (Generation of key dependent S-Box). The process of encryption starts with the generation of dynamic S box which include following three steps [24].

- Generate round key
- Apply XOR operation on every bite of round key.

Fig. III (Key Generation Process)

Let matrix A is 16 X 16 matrixes. This algorithm generates the key dependent S Box

1. Call MACKEY(KEY, MAC)
2. Apply XOR on each byte of KEY returned by Step I and Set the value so obtained to X
3. A= S-BOX
4. Shift every element of matrix A by X bits and Set revised matrix to DYN_SBOX
The Secure Migration of Data to Cloud using Data Sanitization and MAC address based AES

Table – I (Bank Account Details)

<table>
<thead>
<tr>
<th>Customer’s Name</th>
<th>Bank Account Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>Suraj Singh</td>
<td>15452897</td>
</tr>
<tr>
<td>Kishore Lal</td>
<td>15445689</td>
</tr>
<tr>
<td>Ram Prasad</td>
<td>15345687</td>
</tr>
<tr>
<td>Prakash Singh</td>
<td>14785692</td>
</tr>
</tbody>
</table>

Table – II (Health record)

<table>
<thead>
<tr>
<th>Patient’s Name</th>
<th>Purpose of Visitation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anshu Ralph</td>
<td>Regular Check-Up</td>
</tr>
<tr>
<td>Riya Malik</td>
<td>Treatment of Typhoid</td>
</tr>
<tr>
<td>Sunil Sharma</td>
<td>Blood Donation</td>
</tr>
<tr>
<td>Mahesh Tripathi</td>
<td>Regular Check-Up</td>
</tr>
<tr>
<td>Surender Singh</td>
<td>Regular Check-Up</td>
</tr>
<tr>
<td>Shiv Singh</td>
<td>Removal of Appendix</td>
</tr>
<tr>
<td>Manoj Singh</td>
<td>Fracture</td>
</tr>
</tbody>
</table>

Now, the tables are separately stored in a Word document file and needs to be transferred to the cloud. But before transferring it to the cloud, the Bank/Hospital decides to encrypt their data as they want to maintain the privacy of their customers. Now we can test both the encryption methods (Modified AES and General AES) about which yields the better results. We will judge the methods on the basis of the time taken by both the methods to complete the process. We can deduce from the graph above that based on the run time recorder on the MATLAB’s inbuilt ‘Run and Time’ feature, the timings noted for the above tables using modified AES were 7.80 (for T-1) and 7.92 (for T-2). And the timing using General AES was 7.92 (For T-1) and 7.99 (for T-2). The plot above shows that the Modified AES is acting linearly in this case. Although in the above tables the data is not significantly large. So, some more tests should be tried out on relatively large data base. In order to do so, we have used for small on-

system files ranging from 25 kb to 700 kb. The timings were again noted from the inbuilt MATLAB’s feature Run and Time, the timings using Modified AES were 7.9200, 9.8000, 10.9000, 11.2000, 11.4000, 12.1000, 13.4000, 14.5000, and 15.2000 for file sizes as 25, 50, 150, 250, 300, 350, 400, 550, and 700 respectively. And the timings using general AES were 7.9000, 9.6000, 10.9000, 12.2000, 12.5000, 12.9200, 13.4000, 14.7000 and 16.3000 for files size of 25, 50, 150, 250, 300, 350, 400, 550 and 700 respectively.

VI. CONCLUSION:

The Modified Advanced Encryption System will provide a better security to our data as we have illustrated in the examples above. Moreover our proposed data sanitization technique even treats the sensitive data more seriously and apply numerous mathematical function on the data to make sure that the data is properly encrypted and hashed so that it will not be easy for attacker to steal the information stored on the cloud, or an employee if goes rogue and takes the data to their benefit, they can only have the encrypted part and it will useless to them.
REFERENCES


7. D. Nurmi, R. Wolski, C. Grzegorczyk, G. Obertelli, S. Soman, L. Youseff, and D.


