Mitigating the Threat due to Data Deduplication Attacks in Cloud Migration using User Layer Authentication with Light Weight Cryptography

Aruna M G, Mohan K G

Abstract: The widespread adoption of multi-cloud in enterprises is one of the root causes of cost-effectiveness. Cloud service providers reduce storage costs through advanced data de-duplication, which also provides vulnerabilities for attackers. Traditional approaches to authentication and data security for a single cloud need to be upgraded to be best suitable for cloud-to-cloud data migration security in order to mitigate the impact of dictionary and template attacks on authentication and data integrity, respectively. This paper proposes a scheme of user layer authentication along with lightweight cryptography. The proposed simulates its mathematical model to analyze the behavioral pattern of time-complexity of data security along with user auth protection. The performance pattern validates the model for scalability and reliability against both authentication and data integrity.

Keywords: Cloud computing, Authentication, De-duplication, Data security, Cloud-to-Cloud data migration, Hashing, Cryptography

I. INTRODUCTION

Right from the cloud prospect to the cloud users, the security of the data is a prime concern. In 2008, a survey conducted by IDC enterprises revealed that security is the primary factor in deciding to adopt the cloud [1]. Whereas, a recent survey conducted in 2018 by IDG provides surprising statistics that today 73% organization has already adopted cloud and rest planning to adopt in the very near future [2]. Out of 73% of organizations, 42% use a multi-cloud strategy. This paradigm shift raises a research question, that whether there is no security risk involved with the cloud by now or the business requirement compelled the enterprises to adopt the cloud services. Off-course, there is a business compulsion for enterprises to adopt the cloud to meet the feasibility of the global economy. In contrast, the security requirements have shifted from one layer to another dimension with adopting the trend of multi-cloud tenancy that offers resilience to adapt technologies and essential service to the critical applications [3-4]. One of the primary reasons to adopt a multi-cloud strategy is to gain the cost-saving of the storage as well as an additional backup by keeping data on the more cloud storage models [5]. But the question arises that how the cost of SaaS is minimizing to attract the existing cloud users to migrate their data to the new platform. This question is answered by understanding the mechanism of the process adopted by the cloud service provider for the data de-duplication. In order to reduce the cost of data storage scalability, the de-duplication is widely adopted for data storage management [6]. The models adopted for the data de-duplication introduces vulnerabilities that provide a way to compromise the user’s data and its privacy information by means of various attacks [7-9]. This paper proposes an attack framework and a mitigation model for secure data migration from cloud to cloud. Figure 1 shows the architectural diagram of the system model.

A. Cloud to cloud migration

The term cloud migration initially gives a sense of on premise data and system moving to the cloud infrastructure. The situations whereas already organization system and data are on the cloud, they move or migrate their storage as a Service from one vendor to another cloud vendor that makes companies to adopt multi-cloud tenancy and requirement of cloud to cloud data migration services. The cloud to cloud(C2C) Data migration services require to handle many of the complexities involved mainly to avoid speed collisions, false starts, and finally, a seamless migration with the least cost, time, and disruption. The data movement process needs to ascertain the location of the new datacenter and the bandwidth capacity. The services to C2C migration must consider the capabilities of security, period and automatic updates, scalability as well as device independence.

B. Reason for C2C Migration

The organization adopting multi-cloud tenancy requires building logical strategy as per business factor to choose the cloud vendors.
The business factors many include security, reliability, or many other compatible features requirements. Whereas, the business process being very dynamic, the key factors itself changes accordingly over which the initial C2C migration decision is taken, and that continues a thrust of C2C migration. The few compelling requirements may include:

- A new designated cloud may offer a performance improvement guarantee with the new system.
- The availability of additional services requires user applications, as well as a large number of APIs and integration tools.
- Price Benefit.
- Better service-level agreements and uptime
- Larger scalability to accommodate your business growth
- Decrease your DevOps and deployment timelines

These above reasons partly or fully could be the motivating factors to migrate from one cloud to another cloud or in simple word to have communication among multi-cloud tenancy environment.

C. Challenges in C2C Migration

In order to reduce the complexities for moving from on-premise to cloud infrastructure, every cloud service provider adopts its proprietary state of art close architecture. These architectures are strengthened with auto-update and maintain along with the associated requirements of hardware, operating systems, software, and storage to reduce the cost, time, and efforts require. The close architecture of cloud vendors brings synchronization issues while c2c migration. The synchronization components include seamless compatibility among operating systems, network architecture, VM configuration, databases, and many more management tools. Therefore, migrations itself is a cumbersome problem as a potential re-architecting effort is involved.

D. Decision perspective for C2C

In the decision perspective, the factors including cost, risk assessment, and overall performance and business need benefits require to be evaluated as it takes place while on-premise migration to the cloud at first instance. In a nutshell, it shall be feasible enough in multi-facet aspects of the operation, finance, and technological advantages. Yet another critical sense of decision says that to maintain a better synergy with the vendors and partners in both operational and cost aspects to manage and maintain reliable and high-performance bridges, the early decision of migration reduces the increasing complexities and challenges. This paper proposes a framework model for secure authentication using a token-based system for data deduplication layers of cloud security. The paper is organized as section II described the related work as a part of a review of literature, section III describes the proposed system model, section IV discusses the results and analysis and finally section V explains the conclusion.

II. RELATED WORK

There are different dimensions of the data in the modern edge of computing, where the Internet of things has a leading role to generate the data which requires cloud as a storage platform. Due to cloud vulnerabilities, it is necessary to protect the data generated by the IoT stored in the cloud from various data attacks, thereby maintaining the reliability of the storage by maintaining its integrity. The existing method, including RSA, poses huge computational overheads, whereas the signature-based methods do not fit suitable in the massive sensor data storage. The work of Zhu et al. (2019) proposes an algorithm based on a short signature that claims minimization of computational overhead and efficiency against a message attack. However, its dependencies for auditing on the third party does not make its reliable for secure data migration from cloud to cloud as many of the IoT application require multi-cloud infrastructure[10]. The traditional software as a service is evolving to meet the customized requirement of the users considering requirements like security in a multitenancy environment Ali et al. (2019), [11]. In the healthcare sector, the usage of the cloud to cloud(C2C) data migration is of high requirement, where the patient data are shared by the intra-cloud users. In these situations, the security requirements are at high demand to protect private information. The approach of the encryption for the attributes provides a way to protect privacy Edemacu et al. (2019) [12]. Zhang et al. (2019) presents scheme to handle the overhead of authentication with the balance of control on the untrusted clouds using certificates and one-way authentication to achieve lightweight mechanism [13]. The use of modified RSA, signature, encryption, and other authentication having their own limits, a new trend of use of blockchain is witnessed in many of the security paradigms. One of the works of Zuo et al. (2019) focuses on the security aspect cloud-based of medical on-demand services by overcoming traditional attribute-based encryption (ABE) to meet the challenges of both down and up-scaling of subscriptions by using blockchain along with ABE and validated it against the collusion attack [14]. The open architecture of the device brings various kind of attacks due to vulnerabilities of the device and inject to the cloud that can be controlled by the suitable device authentication methods beforehand uploading the data to the cloud. The authentication method proposed by Yanambaka et al. (2019) uses a concept of physical uncolorability by avoidance of saving of the device-related data into the cloud servers with optimal usage of keys and achieves robustness and lightweight with scalability [15]. The combined efforts of authentication of healthcare-related data with IoT based control systems ensure a secure mechanism of preserving privacy and data security on the cloud by preventing malicious access [16]. Apart from this, the privacy protection of the wearable device of the IoT ecosystem is being studied by Zhang et al. (2019), while publishing the data using distance vector mechanism [17]. With the balance between security and storage costs, the growing scale of digital medical data is being addressed through improved deduplication and ABE and minimized computing overhead. Ma et al. (2019) take up the problem of revocation of attributes while deduplication by exploiting the nature of prime number and agent key with ciphertext, where the computational overhead is reduced by means of third-party decryption [18].
The design approach by Amaral et al. (2019) assures the integrity of the data in a scalable manner to handle the tradeoff between communication and storage balances. The validation of the model takes place with an adversary model. Although, this model can handle various file types but on the non-incremental data [19]. Most of the applications, including health care and IoT based systems, generate large data periodically where the use of certificate poses additional overhead. One of the works by He et al. (2019) claims a secure data integrity scheme using a certificate less method [20]. There is a paradigm shift into the architectural layers of cloud computing to support a highly distributed deployment. The new model includes synchronization of cloud resources with the edge and fog. This brings a challenge to security requirements due to complex heterogeneity. A collaborative approach of development, security, and operation is proposed by Díaz et al. (2019) as a cybersecurity watchdog in an autonomous mode [21].

III. SYSTEM MODEL

The system model of multi-cloud architecture for cost-effective data storage is based on the assumption that due to the strong de-duplication process, the cloud can offer cost-effective data storage services so that the cloud user adopts to migrate from one cloud to another cloud. A detailed process flow of the proposed architecture of the trust mechanism used during data migration is shown in Figure 2. The data(D) is applied to the Hashing algorithm (HA) at the cloud -A(CA) and on the other side the same hashing algorithm at the cloud side (HB) if both HA is found equal to HB then it ensures the verification of integrity is true. If there is an attack that occurred on the hash function, the user needs to trust the cloud of choice to evaluate the trust factor, which is explained in the attack scenario.

A. Attack Scenario

The simple authentication process to the cloud provides a vulnerability to perform the de-duplication attack in the cloud data migration scenario. The process of the data integrity and cloud trust identifies the point of attack, and accordingly, the user decides their trust level to migrate to the particular cloud of not.

The system maintains a reference unit for every data and respective cloud hash as \( R_i = \{ R_{A_i}, R_{B_i}, ..., R_{A_i}, R_{B_i}, ... R_n \} \)

Check for Integrity: \( H_A = H_B \) True
\( H_A \neq H_B \) False

Check for Trust: \( R_A = H_A \) and \( R_B = H_A \) Trust: Yes
\( R_A = H_A \) and \( R_B = H_A \) Trust: No

\[ \text{Fig. 2 Process mechanism of the security model due to the threats of deduplication in data migration from C2C} \]
B. Token Based Authentication

In the token-based authentication framework, the cloud user initial registration takes certain credentials as fn, ln, a, e, and p, as their data uploader administrator’s first-name, last-name, age, email id and the password as the first level of the authentication credentials. The combined user-authentication credential ‘eAuth’ vector, which is used during the data upload and deduplication process. The authenticator engine-(AE), notify the cloud user through secure-socket layers auto-generated messages for successful entry as a cloud user.

![Cloud User Authentication Process]

**Fig. 3 New cloud user registration validation**

The attacker tries to get the user’s credentials by some means of phishing; therefore, the framework validates the duplicate credentials and rejects the re-registration with the same user credentials. In case of dictionary attack for the user credentials and authentication token, the framework proposes a random process of matching credential of first layer Auth-code-(MCFL-AC). The algorithm for MCFL-AC is as described below:

Algorithm 1: Matching Credential of First Layer Auth-code-(MCFL-AC)

**Input:** \(e, e_{gp}, p\)

**Output:** User: rejection or move to AC2

**Process:**

**Start**
- \(e_{cred} \leftarrow e \times e_{gp} \) and p
- Initialize, \(d_{cred}\)
- if \((e_{cred}, p)\) then reject user credential
- else, move to: AuthCode-2 generator

**End**

The credentials \((e)\) is concatenated with the associated service provider for the message notification service \((e_{gp})\) that provides the e-credential of the new cloud user who has been notified by the message notification. In the same way, the cloud user personal credential key as ‘p’ provided during the initial registration process forms attributes for the duplicate registration as well as a first layer authentication. For the duplicate \(e_{cred}\) check a pointer: \(d_{cred}\) is flag to 0. The fig 4 shows the process flow of AuthCode-2 generation.

**Fig. 4 Authentication rejection or AuthCode2 notification**

The cred: \([e, e_{gp}, p]\) goes to the datastore (DS), where all the previously authenticated or registered cloud users credentials are stored, and performs the matching operation with its values-(val). If the \(cred\neq val\) then the user first layer authentication is rejected, otherwise it moves to the operation of AuthCode-2 generation (AC2G) the algorithm of AuthCode-2 generation is described below:

Algorithm 2: Second Layer Auth-code-(AC2G)

**Input:** \(S_i, S_h\)

**Output:** AuthCode-2

**Process:**

**Start**
- Initialize \(S_i, S_h\)
- Compute:
  - temp \(AC2 \leftarrow \sum(S_i, S_h) \times f(rng)\)
  - AuthCode-2 \(\leftarrow \sum(S_i, temp \ AC2)\)

**End**

In order to make the AuthCode-2 more chaotic, to seeds, namely low-value seed (SI) and high value seed (SH) are introduced, which can be changed periodically and randomly by the AuthCode-2 generator engine synchronized between two clouds to mitigate the effect of guessing attack. The AuthCode-2 is generated with \(S_i, S_h\) and a pseudo uniform random generator function \(f(rng)\) using equation (1)

\[
\text{AuthCode-2} = \sum \left( S_i \left( \sum \left( S_i, S_h \right) \times f(rng) \right) \right) \text{...eq.}(2)
\]

The cloud user gets authenticated at two layers and in every cycle of authentication, notified automatically as one more layer of security.

**Data Upload and De-duplication:** The data (D) and de-duplicated data (DeD) holds their respective timestamp in the format of a template shown in the table 1.

| Table 1 Data and De-duplicated data storage format |
|---------------------------------------------|------------------|------------------|
| Header-1 | Header-2 | Header-3 |
| Cred | Timestamp (\(T_x\)) | (D/DeD) |

The successive upload of ‘D’ is stored in the payload structure shown in the table-1. The de-duplication process checks for the unique records in the data stored and append only the additional record which is not found in the record. The process of matching and appending is shown in fig-5.

**Fig. 5 Matching and Appending: data template, new data, de-duplication**

As the data uploads increase from the authenticated client, and efficient de-duplication process matches each block in the data store between the data template in the cloud store and each block of new data. The unique or non-duplicated blocks found during the matching operation are appended to the existing data template in the cloud store that minimizes the cost of storage. Since the data migration process among cloud-to-cloud requires on-fly encryption, therefore the computational cost of the appending, the non-duplicate data tokens should be linear with increasing data size to get a proportionate time complexity, whereewith increasing the size of data token or data file exhibits proportionate values of time as tabulated in table 2.
Table 2 Observations of time-(ms) Vs file size

| Time-T(ms) | 1  | 2.83 | 5.20 | 8  | 11.18 |
| Size-(mb)  | 1  | 2    | 3    | 4  | 5     |
| Time-T(ms) | 14.70 | 18.52 | 22.63 | 27 | 31.62 |
| Size-(mb)  | 6  | 7    | 8    | 9  | 10    |
| Time-T(ms) | 36.48 | 41.57 | 46.87 | 52.38 | 58.10 |
| Size-(mb)  | 11 | 12   | 13   | 14 | 15    |

A filed primer (α) is initialized with a random value to evaluate the computational cost for appending non-duplicate token. For each value of the data token stack (N), the time (T) is computed using equation 2.

\[ T \leftarrow (N_r)^\alpha \ldots \text{Eq}(2) \]

The figure 6 shows the pattern of the graph between time and size for the computational cost of appending non-duplicate data token as a notion of data storage cost.

![Fig.6 Computational cost of non-duplicate data token upload with varying file/token size](image)

The section 4 describes the performance analysis of different cryptographic processes on fly in detail.

IV. RESULT AND ANALYSIS

The data store of the entire user appends into the secure bucket of the respective cloud, which is modeled for the light-weight encryption with a time factor threshold with a filed primer (α) and another primer (β) for time complexity computation for both encryption and decryption respectively. Figure 7, illustrates the performance behavior of time consistency for both encryption and decryption with varied file sizes.

![Fig.7 The computational cost of encryption and decryption: File size-(MB) vs running time (ms)](image)

The computation of time factor takes place using equation 3 and equation 4.

\[ \alpha \left[ t_\alpha \right] \leftarrow (u_{θ})^\alpha \ldots \text{equ}(3) \]
\[ \beta(t_\beta) \leftarrow (u_{θ})^\beta \ldots \text{equ}(4) \]

The trend of both encryption time (t_\alpha) and decryption (t_\beta) for respective data (u_{θ}) shows a proportionate and consistent incremental pattern that ensures the suitability of the cryptography system for large data sizes.

Figure 8, illustrates the computational cost of re-encryption with a key size of 256 bit for varying file sizes by computing respective computation time in milli-second.

![Fig.8 The computational cost of role re-encryption(rkey=256bit) for block size: 1024,256,64 and 16, File size-(MB) vs Running time(ms).](image)

The computation of time factor for varying file size for respective block size = \{1024, 256, 64, and 16\} takes place using equation 5, equation, equation 7 and equation 8.

\[ \alpha \left[ t_{α} \right] \leftarrow (u_{θ})^α \ldots \text{equ}(5) \]
\[ \beta(t_{θ}) \leftarrow (u_{θ})^β \ldots \text{equ}(6) \]
\[ γ(t_{θ}) \leftarrow (u_{θ})^γ \ldots \text{equ}(7) \]
\[ δ(t_{θ}) \leftarrow (u_{θ})^δ \ldots \text{equ}(8) \]

The trend of re-encryption for varied block size(1024, 256, 64 and 16) with a key size of 256 bit shows a proportionate and consistent incremental pattern over increasing file size, that ensures fitness of the cryptography system for large data sizes. Figure 9 demonstrates the computational cost of re-encryption with a key size of 512 bit for different file size by computing respective computational time in milli-second.

![Fig.9 The computational cost of role re-encryption (rkey=512bit), of block size: 1024, 256, 64 and 16, File size-(MB) vs Running time (ms).](image)

The computation of time factor for varying file size for respective block size = \{1024, 256, 64, and 16\} takes place using a similar function as demonstrated in equation 5, equation, equation 7 and equation 8. The trend of re-encryption for varied block size with a key size of 512 bit shows a proportionate and consistent incremental pattern that ensures the suitability of the cryptography system for large data sizes.

Published By:
Blue Eyes Intelligence Engineering & Sciences Publication

International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075, Volume-9 Issue-3, January 2020

Retrieval Number: C8463019320/2020©BEIESP
DOI: 10.35940/ijitee.C8463.019320

2543
the architecture of the eco-system of cloud-to-cloud migration, is designed. The issue of authentication is handled by providing a secure auth-token by hashing mechanism along to combat the effect of dictionary attack by the adversaries to gain authentication codes, which is one-layer advancement as compared to traditional cloud service provider authentication schemes. Another essential issue of data security against the threat of de-duplication template attack is handled by a scheme of encryption and decryption while first data incident storage and re-encryption and re-decryption for the additional unmatched data while de-duplication process. The various performance analysis for time complexity with respect to varying file size, block size, and re-keying size shows a consistent, proportionate pattern, that validates the model for scalability, interoperability, and time-complexity. The proposed system is suitable as it offers a robust and efficient mechanism and reliable to be adopted in real-time implementations.

REFERENCES


### BIOGRAPHIES OF AUTHORS

**Mrs Aruna M G** is a research scholar in Visvesvaraya Technological University, Belagavi, India. She has completed her B.E and M.Tech from Bangalore university and MGR university in 2001 and 2006 respectively. Currently, she is working as an associate professor in M S Engineering college, Bangalore, India. Her research interests is in security in cloud computing. She has overall 15 years of experience as an academician and 04 years of research experience.

**Dr. K. G. Mohan** has received PhD from Anna university in 2007 in the domain of computer architecture. His area of research interest includes low power architecture design, Cloud Computing, Wireless Sensor Networks, IoT, Network Security, etc. He has overall 30 years of experience as an academician and 14 years of experience in research. He has published many international journal and conferences of repute.