Secure Data Deduplication

M.Asif Ali Khan
*(Student/CSE, Kamaraj College of Engineering and Technology, Madurai
Email: asif4225@gmail.com)
K.K.Jijendran
**(Student/CSE, Kamaraj College of Engineering and Technology, Madurai
Email: jijendrankk1997@gmail.com)
S.Thanga Pandi
***(Student/CSE, Kamaraj College of Engineering and Technology, Madurai
Email: goldenpandi892@gmail.com)
P.Swathika
**** (Assistant Professor/CSE, Kamaraj College of Engineering and Technology, Madurai
Email: swathikacse@kamarajengg.edu.in)

Abstract:

Deduplication process that eliminates repeated copies of data and reduces storage spaces. Data deduplication is necessary for cloud storage providers because of the continuous increase in number of the users and the size of the data. Storage and data transfer costs will be reduced by their cloud providers by storing a unique copy of duplicate data. Cloud computing offer services various resources by the Internet. Important cloud service is data storage. The privacy of data holders are important, so in order to secure the privacy of the data holders, most of the data is stored in the cloud in encrypted form. In cloud data deduplication, encrypted data will introduce new challenges, which becomes challenge for data storage and processing. There are some of the traditional schemes for deduplication that is not work on encrypted data. Some of the existing solutions are present for encrypted data deduplication which suffers from security i.e. brute force attacks means they are not flexibly supporting data access control and revocation. In this paper, we are proposing a scheme to deduplicate the encrypted data which is stored in the cloud. And also we use Convergent Encryption Technique for encryption and SHA for hash code generation for the data deduplication.

Keywords — **Deduplication, Convergent Encryption, SHA.**

I. INTRODUCTION

Cloud providers offer potentially infinite storage space, where users may use huge spaces if they want and vendors searches for techniques which aimed to minimize redundant data (multiple copies) and to save maximize space. To minimize redundant data that is the elimination of multiple copies, we make use of deduplication technique. The data deduplication is to store Data only once without repetation.

Therefore, if an user wants to upload a data that is already stored in the cloud, the cloud provider will not show deduplication to that user. Deduplication reduces storage needs for backup applications and provide standard file systems. Users need protection of their data and confidentiality along with low costs and flexibility supports which guarantees by encryption. if deduplication and encryption are used then it will be two conflicting technologies. but the aim of deduplication is to avoid duplicate data and store them only once, but after encryption it makes two identical data indistinguishable. This means that if data is encrypted by users, the cloud storage providers cannot apply deduplication because two identical data will be different after encryption. On the other hand, confidentiality will not be guaranteed and data will not be protected against attackers in cloud storage if data is not encrypted by users.

Convergent encryption is a technique which is proposed for these two conflicting requirements where the encryption key is the result from the hash of the data. If we want to achieve the deduplication and confidentiality at the same time, convergent encryption seems to be a good candidate but unfortunately, it suffers from various well-known weaknesses. Here we mainly focus on deduplication and cloud storage.

By linking network resources together, cloud computing provides a big pool of resource. It has various properties, such as elasticity, scalability, and pay-per-use, fault-tolerance. Thus, it became a promising service platform. Data storage service is the most important and popular cloud. Cloud users have some personal or confidential data which will be uploaded to the data center of a Cloud Service Provider and allow these data to maintain. Since we cannot avoid intrusions and attacks to sensitive data at CSP, we should assume that CSP cannot be fully trusted by cloud users.

But the same or different users can upload duplicated data in an encrypted format to CSP. Here the practical issue is how to manage many encrypted data in cloud storage with deduplication in efficient way. However, existing industrial deduplication solutions could not handle encrypted data. Current solutions for deduplication suffers from some security weakness i.e. brute-force attacks.

They cannot support data access control and revocation at the same time.There are number of reasons to allow data holders to manage deduplication. First, they cloud cause storage delay because data holders can not be always online or available for such a management. Second, deduplication could become too complicated in terms of communications and computations to involve data holders into deduplication process. Third, in process of discovering duplicated data, it may intrude the privacy. Forth, a data holder have no idea about how to issue deduplication keys or data access rights to a user in some situations when it didn't know other data holders due to data distribution. so, CSP couldn't cooperate with data holders on the data storage deduplication in many situations. The results show the superior efficiency and effectiveness of the scheme for efficient practical deployment, especially data deduplication in cloud storage.

II. LITERATURE SURVEY

DupLESS: Server-Aided Encryption for Deduplicated Storage :

DupLess Server-Aided Encryption for Deduplicated Storage provide simple storage interface. Cloud Storage provider like Dropbox, Mozy, and other providers can use deduplication technology to reduce space by storing single copy of data. Message lock encryption Technique is used to resolve the clients problems encrypt their file however saving are lock. Dupless method is used to provide secure Deduplicated storage as well as storage resisting brute-force attacks. Clients encrypt under message-based keys obtained from a key-server via an PRF protocol in dupless server. It allows clients to store encrypted data

Drawback:

- Operations are time consuming

Secure Client Side Deduplication Scheme in Cloud Storage Environment :

Ensure better confidentiality towards unauthorized users by cryptographic usage of symmetric encryption used for enciphering data file and the asymmetric encryption for the meta data files. Data access are managed by the data owner by providing two level of access control Only authorized user can decipher encrypted files.

Fast and secure laptop backups with encrypted de- duplication :

Data that is common between users to increase the speed of backup storage and reduce the storage requirement. it supports client-end peruser encryption that is necessary for confidential personal data.This is used to decrease backup times and storage requirement.

Drawback :

- Network bandwidth is a bottle-neck.
- Backing up directly to cloud is very costly

Weak leakage–resilient in client side Deduplication of encrypted data:

.Propose secure client–side deduplication scheme. Addresses an important security concern in cross-user client–side deduplication. But Convergent encryption and custom encryption method is not semantically secure.

Deduplication Techniques over Encrypted Data :

Convergent encryption, is an encryption approach that support deduplication. With convergent encryption, encryption key is generated out of hash of plain text. Thus applying these technique identical plaintexts would produce same cipher text, and this helps in performing deduplication further.

Drawback :

- we cannot compromise on both security and duplication of data across storage areas.

Secure Deduplication on Encrypted Big Data in Cloud Computing Environment :

This scheme flexibly supports data update and sharing with deduplication when data holders are offline.Encrypted data is securely accessed because only the authorized data holders can obtain the symmetric keys which is used for data decryption.

Secure Authorized Deduplication Systems :

The hybrid cloud approach is presented for the security purpose which has system with a differential privileges to different users accordinglyIt uses the convergent encryption technique for encrypting the data with convergent key. It also provides Differential Authorized duplicate checking, so, only the authorized user with specified privileges can perform the duplicate check.

Drawbacks :

- An attacker maynot know the entire file, but the partners who have the file. The partners follows bounded retrieval model, so that they can help the attackers to obtain the file.

Deduplication of Data in Cloud :

It compresses data by removing the repeated copies of same data and it is extensively used in cloud storage to save bandwidth and reduce the storage space.To secure the confidentiality of sensitive data during deduplication, the convergent encryption technique is used to encrypt the data.

Drawbacks :

- Issue of data deduplication authorization.

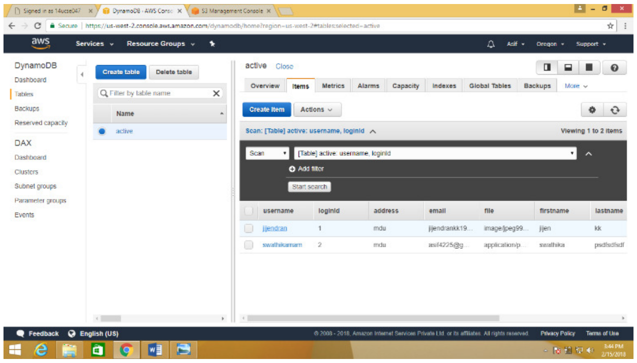
III. EXISTING SYSTEM :

Cloud storage service providers such as Google Drive, dropbox, Mozy, and others can perform deduplication to reduce space by only storing one copy of each file uploaded. However, if client encrypts their data storage savings by deduplication is totally lost because the encrypted data is saved as different contents by applying various encryption keys. Existing industrial solution fails in encrypted data deduplication. For example, Deduplication is an efficient system, but it cannot handle encrypted data. The existing system introduces one model for provable data possession(PDP) that allows client that to store data at an untrusted server to verify that the server possesses original data without retrieving it.

IV . PROPOSED SYSTEM :

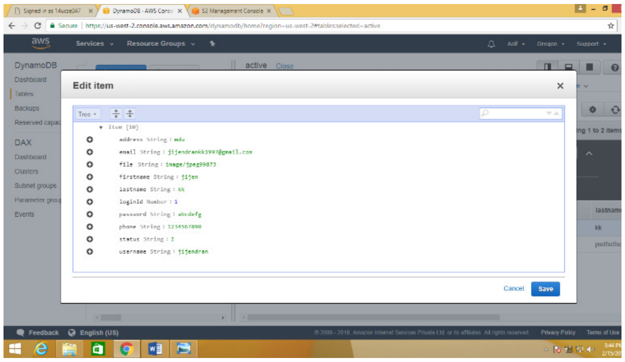
In this paper, we propose a scheme to deduplicate encrypted data that is stored in the cloud. A Hashing function can be used to return a unique key for a file, based only on the contents of the data.If two files are having the same data, so then the hashing function will return the same key for these two files.If this key is used as the index for storing file, then any attempt to store multiple copies of the same file will be detected immediately.

V . RESULTS :



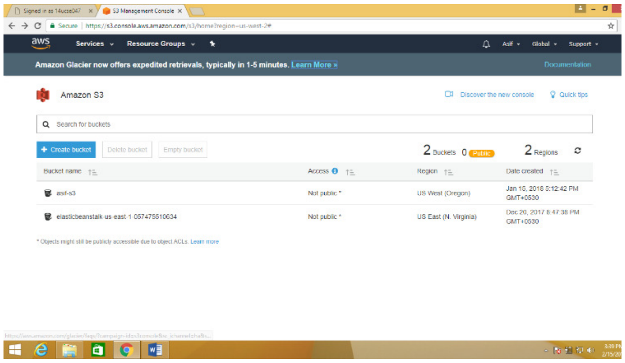
username	loginid	address	email	file	firstname	lastname
jijendran	1	mdu	jijendrankk19...	image/jpeg99...	jijen	kk
swathikamam	2	mdu	asif4225@g...	application/p...	swathika	psdfsdfsdf

Fig 1 : User Details and File Details within a single row in AWS Dynamo DB



```
▾ Item {10}
    address String : mdu
    email String : jijendrankk1997@gmail.com
    file String : image/jpeg99873
    firstname String : jijen
    lastname String : kk
    loginId Number : 1
    password String : abcdefg
    phone String : 1234567890
    status String : I
    username String : jijendran
```

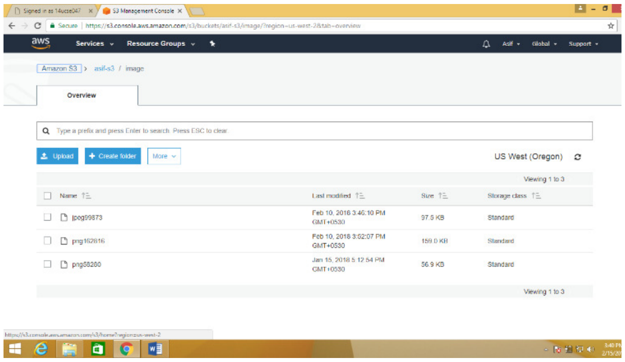
Fig 2: User Details and File Details in AWS Dynamo DB



Bucket name	Access	Region	Date created
asif-s3	Not public *	US West (Oregon)	Jan 15, 2018 5:12:42 PM GMT+0530
elasticbeanstalk-us-east-1-057475510634	Not public *	US East (N. Virginia)	Dec 20, 2017 8:47:38 PM GMT+0530

* Objects might still be publicly accessible due to object ACLs. Learn more

Fig 3: Bucket Created in AWS S3


Signed in as 14ucse047
S3 Management Console
Secure
https://s3.console.aws.amazon.com/s3/buckets/asif-s3/image/?region=us-west-2&tab=overview
aws
Services
Resource Groups
Asif
Global
Support
Amazon S3 > asif-s3 / image
Overview
Type a prefix and press Enter to search. Press ESC to clear.
Upload
Create folder
More
US West (Oregon)
Viewing 1 to 3

Name	Last modified	Size	Storage class
jpeg99873	Feb 10, 2018 3:46:10 PM GMT+0530	97.5 KB	Standard
png162816	Feb 10, 2018 3:52:07 PM GMT+0530	159.0 KB	Standard
png58280	Jan 15, 2018 5:12:54 PM GMT+0530	56.9 KB	Standard

Fig 4 : Deduplicated files in the folder image in AWS S3

VI . CONCLUSION :

Managing encrypted data with deduplication is most significant in practice for running a cloud storage service which is secure and dependable, especially for data processes. Future work includes efficient data ownership verification, scheme optimization with hardware acceleration at IoT devices for practical deployment, and development of a flexible solution to support deduplication and data access controlled by either the data owner or its representative agent.

REFERENCES

[1] Deduplication on Encrypted Big Data in Cloud Zheng Yan, Senior Member, IEEE, Wenxiu Ding, Xixun Yu, Haiqi Zhu, and Robert H. Deng, Fellow, IEEE Deduplication on Encrypted Big Data in Cloud IEEE TRANSACTIONS ON BIG DATA, VOL. 2, NO. 2, APRIL-JUNE 2016.

[2] J. Li, Y. K. Li, X. F. Chen, P. P. C. Lee, and W. J. Lou, “A hybrid cloud approach for secure authorized deduplication,” IEEE Trans. Parallel Distrib. Syst., vol. 26, no. 5, pp. 1206–1216, May 2015.

[3] Mozy, Mozy: A File-storage and Sharing Service. (2016).

[4] G. Wallace, et al., “Characteristics of backup workloads in production systems,” in Proc. USENIX Conf. File Storage Technol., 2012, pp. 1–16.

[5] Z. O. Wilcox, “Convergent encryption reconsidered,” 2011.

[6] G. Ateniese, K. Fu, M. Green, and S. Hohenberger, “Improved proxy re-encryption schemes with applications to secure distributed storage,” ACM Trans. Inform. Syst. Secur., vol. 9, no. 1, pp. 1– 30, 2006, doi:10.1145/1127345.1127346.

[7] M. Lillibridge, K. Eshghi, and D. Bhagwat, “Improving restore speed for backup systems that use inline chunk-based deduplication,” in Proc. USENIX Conf. File Storage Technol., 2013.

[8] L. J. Gao, “Game theoretic analysis on acceptance of a cloud data access control scheme based on reputation,” M.S. thesis, Xidian University, State Key Lab of ISN, School of Telecommunications Engineering, Xi’an, China, 2015.