REMOVING DUPLICATION DATA IN CLOUD WITH SECURE ACCESSIBILITY

B.Shireesha
B.Tech IV CSE Student, VITS (N6), Karimnagar, JNTUH, Hyderabad, TS, INDIA
bajjurishirisha@gmail.com

G.Akhila
B.Tech IV CSE Student, VITS (N6), Karimnagar, JNTUH, Hyderabad, TS, INDIA
akhilagandra5@gmail.com

K .Shravya
B.Tech IV CSE Student, VITS (N6), Karimnagar, JNTUH, Hyderabad, TS, INDIA
shravyakonduri6@gmail.com

P.Pradeep Kumar
HOD-CSE, Department of CSE, VITS, Karimnagar, JNTUH, Hyderabad, TS, INDIA
pkpuram@yahoo.com

ABSTRACT

In recent days, cloud computing offering various resources over Internet. Among those resources, one of the important services is data storage. In order to preserve the privacy of users, the data is stored in cloud in an encrypted form. Deduplication becomes crucial and a challenging task when the data is stored in encrypted form, which also leads to complexity in storing large data and processing in cloud. A traditional deduplication method does not work on encrypted data. Existing solutions available for deduplicating encrypted data has various security issues. Cloud does not provide access control and revocation in terms of storage. Hence, the deduplication schemes are not mostly deployed in practice. In this paper, we propose a technique to deduplicate encrypted data stored in cloud based on access control, thereby avoiding redundant storage. It integrates cloud data deduplications with access control. The result of our scheme shows superior efficiency and has potential for practical deployment in the case of huge data storage**.** We will retrieve files based on fast keyword searching and at the time of downloading will provide security for the file with keys.

Keywords: *Deduplication, encrypted data, secured access control, cloud computing.*

I.INTRODUCTION

With the potentially infinite storage space offered by cloud providers, users tend to use as much space as they can and vendors constantly look for techniques aimed to minimize redundant data and maximize space savings. A technique which has been widely adopted is cross-user deduplication. The simple idea behind deduplication is to store duplicate data (either files or blocks) only once. Therefore, if a user wants to upload a file (block) which is already stored, the cloud provider will add the user to the owner list of that file (block). Deduplication has proved to achieve high space and cost savings and many cloud storage providers are currently adopting it. Along with low ownership costs and flexibility, users require the protection of their data and confidentiality guarantees through encryption. Unfortunately, deduplication and encryption are two conflicting technologies. Cloud computing provides various services by rearranging the resources over the Internet. In order to preserve the security of these data, they are often stored in an encrypted form. Encrypted data create new challenges for cloud deduplication which becomes crucial for big data storage and processing in cloud. A traditional deduplication scheme does not work on encrypted data. Therefore in this project we introduce a scheme to deduplicate encrypted data in could based on ownership to deduplicate multiple copies of same data. We aim to solve the issues in deduplication that are being faced by data holders by providing privacy for accessing the file. The results show superior efficiency and effectiveness of the scheme for practical deployment in cloud. The contributions of this paper can be summarized as follows.

We propose methods to save cloud storage without revealing the privacy of data holders by providing a scheme to deduplicate and manage encrypted data. The scheme manages data deduplication with data sharing even in the absence of the data holder while preserving their privacy. We combine cloud data deduplication with data access control in a simple way.

II.RELATED WORK

1.Role of Convergent Encryption in Elimination of Redundant Data

Now a days, most of the clients would like to store their data to public cloud servers (PCS) along with the rapid development of cloud computing. New security problems have to be solved in order to help more clients to process their data in public cloud. So to provide security symmetric key algorithm is one of the technique. Symmetric-key algorithms are those which use the same cryptographic keys for both encryption of plaintext and decryption of cipher text. If the same keys are used for encryption using the same algorithm, it will produce a same cipher value. Otherwise the cipher values will be different. The cipher values should be same on encrypting the file for proper deduplication of files.

Another way is Asymmetric cryptography, which is also known as public key cryptography, uses public keys to encrypt and private keys to decrypt the files. The keys used in asymmetric encryption are usually large values that have been paired together but are not identical (asymmetric). The key that can be shared with everyone is called public key and the other key in the pair is kept secret called the private key. Either of the keys can be used to encrypt a message another key from the one used to encrypt the message is used for decryption. Most commonly public keys are used for encryption and private keys are used when authentication is required.

Convergent encryption, also known as content hash keying. It is a method that produces identical cipher text from identical plaintext files based on the symmetric encryption but the content hashes are used as a key. Symmetric encryption is a part of convergence encryption since it uses the same key to encrypt and decrypt. Whereas, in terms of asymmetric encryption it is difficult to generate key pairs from hash value. Convergence encryption is useful in cloud computing for removing duplicate files from storage.

2.Encrypted Data Deduplication

Data deduplication which is a specialized technique for removing duplicate copies of redundant data. It is used in cloud storage to improve the storage utilization in terms of storing huge amount of data. This method also reduces the amount of data to be transferred over a network as the number of bytes are reduced by avoiding duplication.

The deduplication process mainly consists of identifying the binary value of each data and stored during the process of analysis. In the next stage of analysis, the stored value is compared with the upcoming data to be stored and if a match occurs, the redundant chunk is replaced by providing a small reference to the already stored chunk. Given that the same set of data may occur in huge copies, the amount of data to be stored or transferred can be highly reduced.

Encryption is the method used to encode a message or information in such a way that only authorized persons will get access to it. Adding deduplication over encrypted data is a challenging task as the data will be stored in a binary format after encryption.

C. Data Ownership Verification and Others

Data ownership is the process of getting legal rights and complete control for a single piece or set of data elements. It provides the information about the rightful owner of data sets, how to use and distribute the particular data as implemented by its owner. The data ownership is a data governing process which describes an organization's legal ownership of enterprise wide data. The data owner is a specific organization that has the ability to create, edit, modify, share and access the data. It also defines the data owner's ability to share, assign or surrender all of the privileges to a third party. This concept of data ownership is implemented in medium to large enterprises with huge repositories of centralized or distributed data elements. The data owner has copyrights to ensure their control over the data and has the ability to take legal action if is illegally breached by any entity.

III.ACCOMPLISHMENTS

EXISTING SYSTEM

The existing solutions for de-duplication that are available cannot flexibly provide access control and revocation at the same time. Most of the solutions cannot ensure reliability, security and privacy for the data. In real time, it is hard to allow a data holder to manage de-duplication due to a number of reasons. At first, the data holders may not be always available for managing the data. In case if they perform de-duplication manually, it could cause delay in storing the data. Second, de-duplication becomes complicated in terms of computations involved in the process. Third, it may violate the privacy of data in the process of analyzing the de-duplicated data. Forth, a shared key is provided to the users to access the same data. Whenever a user removes the data from his account, the keys must be reassigned to the remaining users who have access to the data. This reassignment which is a complex task, if not carried out properly, the removed user may have illegal access to that data.

Disadvantages:

- Increase Network round trips.
- Lose of the network bandwidth.
- Increase the cost due to increase the storage space.

PROPOSED SYSTEM

We propose methods to save cloud storage without revealing the privacy of data holders by providing a scheme to de-duplicate and manage encrypted data. The scheme manages data de-duplication with data sharing even in the absence of the data holder while preserving their privacy. We combine cloud data de-duplication with data access control in a simple way.

Advantages:

- Eliminating duplicate copies of identical data.
- Save the storage space and network bandwidth.
- Reduce the network round trips.
- We will retrieve files based on fast keyword search.
- At time of downloading we will provide security for the file with keys.

IV.SYSTEMARCHITECTURE

1.Deduplication while uploading the file

Whenever the user uploads the file F into the cloud, the hash value is generated for that file HF = H (F). The hash value is used as a key to encrypt the file. The hash value will be unique for every file (a small change in one bit of the file will result in multiple bit changes in its hash value generated. This is known as avalanche effect). A random key will be generated which is used to encrypt the hash value of the file. Encrypted hash will be stored in the database and generated key will be given to the user. Original file name will be stored in the database and a hash will be generated again for the previously generated hash X= H (HF). If a file named X already exists in the data storage, the file will not be stored. The user will be provided access to the file from above steps and upload count will be incremented by one. Otherwise, file will be renamed as the X value that is generated and stored in the cloud with a new upload count as one.

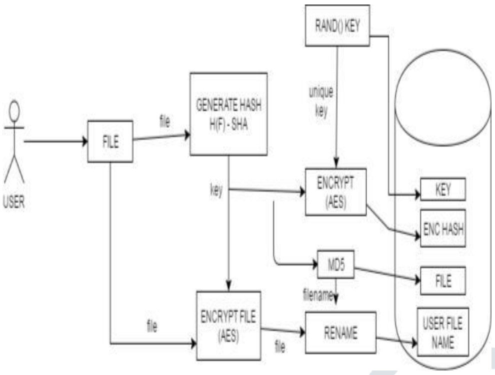


Fig: Architecture for Uploading the File

2. Retrieval of file

The key provided to the user during encryption is used to decrypt the encrypted hash that is stored in the database. The decrypted hash will be hashed again and it is used as the file name to search for the particular file in the data storage. Then the file will be renamed to its original name saved in the database.

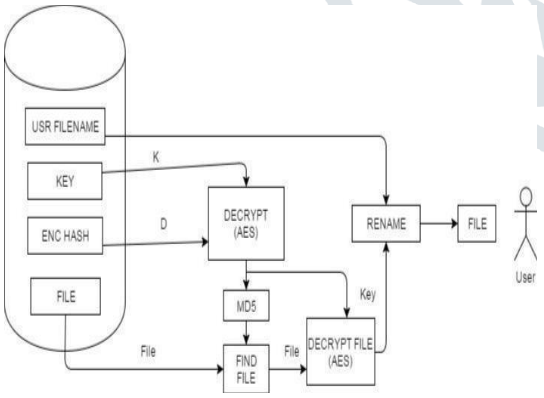


Fig: Architecture for Retrieving the File

3. Removal of a file and access revocation

If a data holder deletes the file from data storage, the upload count is decremented by one and the encrypted hash value provided to the user will be removed for that file. The file will be available as long as at least one user is available to access the file.

4. Database Schema

There are three tables are required for deduplicationg the data in a secure manner, those are shown below. (user information table, file information table, user file mapping table). Eventhough there is a table, file cannot be identified or decrypted since altogether stored in an encrypted form. So that the security and privacy of the users is improved.

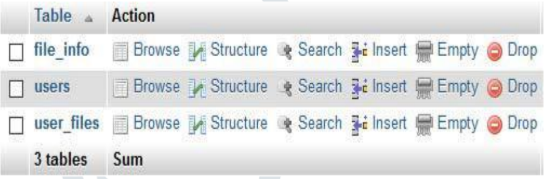
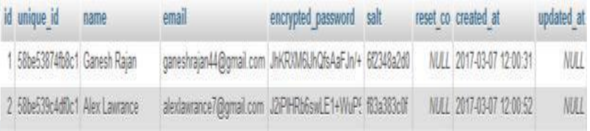


	Table ▲	Action
☐	file_info	Browse Structure Search Insert Empty Drop
☐	users	Browse Structure Search Insert Empty Drop
☐	user_files	Browse Structure Search Insert Empty Drop
	3 tables	Sum

Snapshot for Database Schema

A.User Information Schema

In user information schema users authentication information like username, password and other information about the users are stored Snapshot for User Information Schema

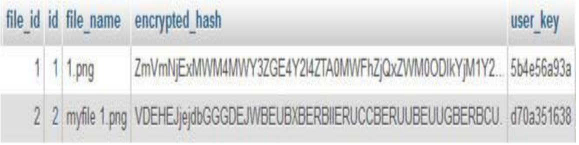


id	unique_id	name	email	encrypted_password	salt	reset_co	created_at	updated_at
1	58be53874fb8c1	Ganesh Rajan	ganeshrajan44@gmail.com	JhKRXM6JhQfsAaFJn/+	6f2348a2d0	NULL	2017-03-07 12:00:31	NULL
2	58be539c4df0c1	Alex Lawrance	alexlawrance7@gmail.com	J2lPlHRb6swLE1+WuP!	f83a383c0f	NULL	2017-03-07 12:00:52	NULL

Snapshot for User Information Schema

B. User File Mapping Schema

It Maps the user with their encrypted hash value and a original file name which is uploaded by the user.

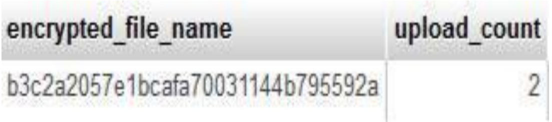


file_id	id	file_name	encrypted_hash	user_key
1	1	1.png	ZmVmNjExMWM4MWY3ZGE4Y2I4ZTA0MWFhZjQxZWM0ODIkYjM1Y2...	5b4e56a93a
2	2	myfile 1.png	VDEHEJjejdbGGGDEJWBEUBXBERBIIERUCCBERUUBEUUGBERBCU...	d70a351638

Snapshot for use File Mapping Schema

C. File Information Schema

This Schema is used to determine the number of users sharing the same file and it is also used for proper removal of deduplicated data from the storage when all the users are revoked access to the file or when they removed from their allocated storage.



encrypted_file_name	upload_count
b3c2a2057e1bcafa70031144b795592a	2

Snapshot for File Information Schema

5. File Stored in Data Storage

Even though the same file is uploaded by the multiple users, only a single copy of a file is maintained in the cloud data storage in encrypted form

V.CONCLUSION

Managed encrypted data with deduplication is important and significant in practice for achieving a successful cloud storage service, especially for big data storage. In this paper,

we proposed a scheme to manage the encrypted files in a cloud with deduplication based on ownership. Our scheme can flexibly support data update and sharing with deduplication. Encrypted data can be securely accessed only by authorized data holders can obtain the symmetric keys used for data decryption.

REFERENCES

1. A secure data deduplication framework for cloud environments, authors: Fatema Rashid, Ali Miri, Isaac Woungang

2. Attribute-Based Storage Supporting Secure Deduplication of Encrypted Data in Cloud, authors: Hui Cui, Robert H. Deng, Yingjiu Li

3. Achieving lightweight, time-specific and secure access control in cloud storage, authors: Yanchao Wang,Fenghua Li, Ben Niu

4. Design and implementation of various file deduplication schemes on storage devices, authors: Yong-Ting Wu , Min-Chieh Yu , Jenq-Shiou Leu , Eau-Chung Lee,TianSong

5. T. T. Wu, W. C. Dou, C. H. Hu, and J. J. Chen, “Service mining for trusted service composition in cross-cloud environment,” IEEE Systems Syst. J., vol. PP, no. 99, pp. 1–12, 2014, doi:10.1109/ JSYST.2014.2361841.

6. Liu, C. Yang, X. Y. Zhang, and J. J. Chen, “External integrity verification for outsourced big data in cloud and iot: A big picture,” Future Generation Comput. Syst., vol. 49, pp. 58–67, 2015

7. W. Tsai, C. F. Lai, H. C. Chao, and A. V. Vasilakos, “Big data analytics: A survey,” J. Big Data, vol. 2, no. 1, pp. 1–32, 2015, doi:10.1186/s40537-015-0030-3.

8. L. F. Wei, et al., “Security and privacy for storage and computation in cloud computing,” Inf. Sci., vol. 258, pp. 371–386, 2014, doi:10.1016/j.ins.2013.04.028.

9. “Deduplication on Encrypted Big Data in Cloud” by Zheng Yan, Senior Member, IEEE, Wenxiu Ding, Xixun Yu, Haiqi Zhu, and Robert H. Deng, Fellow, IEEE

10. M. Ali, S. U. Khan, and A. V. Vasilakos, “Security in cloud computing: Opportunities and challenges,” Inf. Sci., vol. 305, pp. 357– 383, 2015, doi:10.1016/j.ins.2015.01.025.

11. M. Ali, et al., “SeDaSC: Secure data in clouds,” IEEE Syst. J., vol. PP, no. 99, pp. 1–10, 2015, doi: 10.1109/JSYST.2014.2379646. R. D. Pietro and A. Sorniotti, “Boosting efficiency and security in proof of

ownership for deduplication," in Proc. 7th ACM Symp. Inf. Comput. Commun. Secur., 2012, pp. 81–82, doi:10.1145/ 2414456.2414504.

12. W. K. Ng, Y. Wen, and H. Zhu, "Private data deduplication protocols in cloud storage," in Proc 27th Annu. ACM Symp. Appl. Comput., 2012, pp. 441–446.

13 .C. W. Tsai, C. F. Lai, H. C. Chao, and A. V. Vasilakos, "Big