Improving security in File Level Deduplication over Cloud

Miss M.K.Takarkhede
SGBAU, Amravati
Maharashtra, India.
mktakarkhede@gmail.com

Dr. V.M.Thakare
SGBAU, Amravati
Maharashtra, India
vilthakare@yahoo.co.in

ABSTRACT-

Deduplication is excellent method to decrease the used size of cloud server and with the help of 'deduplication of data 'approach which removes repeated data from cloud storage. This paper presents a new file level deduplication scheme "AES with secure deduplication system" which generates hash value by using MD5 algorithm and also apply Advanced Encryption Standard algorithm (AES). This scheme improves the data confidentiality, storage utilization, reliability and removes redundancy of data in file level deduplication. In the proposed security model, Security analysis demonstrates that these deduplication systems are secure.

***Index Terms* — Data privacy, Reliability, Data Redundancy, Storage utilization.**

I) INTRODUCTION

The cloud facility providers offer low cost for both huge space of storage and computing resources. Every day enormous number of data is put up on the cloud and which is mutually shared by number of users having specified rights. The managing of the ever-increasing size is serious problems with cloud. Deduplication is one best solution that makes storage management scalable. Its idea is to remove the storage of repeated messages that have identical same content, by keeping only one message copy and referring other repeated messages to copy through small-size pointers. Deduplication is shown to effectively reduce disk storage space for some workloads, such as backup data, disk space.In deduplication approach which are going to develop file-level deduplication systems which are secure and reducing data redundancy. In file-level deduplication as long as any part of a file is modified, the entire file is considered as a new file and therefore it will be stored [1]. Every incoming file is either divided into fixed or variable sized chunks that are Static Chunking (SC) of some fixed chunk size. File level de-duplication, as the name suggests, is always performed over a single file. Identification of same hash value of two or more files determines that the files are similar [2, 3]. Data deduplication relate to techniques that saves only a one copy of redundant data, and provide links to that copy instead of saving other original copies of this data. By saving and transmitting only a one copy of duplicate data, deduplication saves both network bandwidth and storage space [4, 5]. This approach eliminates data storage costs and realizes storage savings of 50-90%.

This paper, discusses five different file level deduplication schemes such as authorized duplicate check scheme, attribute-based storage system, Server-aided encryption schemes, Encrypted Data scheme, Secure system architecture.These file level deduplication schemes have some limitation such as high storage cost, computation time, increase cost. So to overcome such problems improved version of file level deduplication scheme is proposed here i.e. **"AES with secure deduplication system".** The idea behind this approach is reducing relevant data from server and improves the data confidentiality and storage utilization.

II) BACKGROUND

Many techniques and technologies which support the file level deduplication schemes have been done to reduced redundant data in cloud server in recent few decades. Such schemes are as follows: In Hybrid cloud approach author proposed authorized duplicate check scheme which provides stronger security by encrypting the file with differential privilege keys. In such manner, the users without corresponding privileges cannot execute the duplicate check. So that, such unauthorized users cannot decrypt the cipher text [1]. In attribute-based storage system author used concept of attribute-based encryption (ABE) with hybrid cloud where a private cloud is dependable for duplicate detection and a public cloud handles the storage. This scheme can be used to distribute data confidentially with users by specifying access policies rather than sharing decryption key [2]. In decentralized server aided encryption scheme author proposed inter-KS deduplication algorithm, in which a cloud storage service provider can execute deduplication over cipher texts from different KSs within a tenant. In such way, this scheme synchronously offers flexibility of KS management and cross-tenant deduplication over encrypted data [3]. In encrypted big data author used scheme to deduplicated encrypted data stored in cloud based on ownership challenge and proxy re-encryption. It integrates cloud data deduplication with access control. This approach verifies data ownership and check duplicate storage with secure challenge and big data support [4]. In secure system architecture which enables the cloud with the critical deduplication functionality to completely reduce the additional storage space and bandwidth cost, which would have been incurred by hosting encrypted videos from different entities. It supports media applications that perform media files with scalable structures [5].

This paper introduces five different file level deduplication schemes such as duplicate check scheme, attribute-based storage system, Server-aided encryption schemes, Encrypted Data scheme, secure system architecture which are organizes as follows. **Section I** Introduction. **Section II** discusses Background. **Section III** discusses previous work. **Section IV** discusses existing scheme. **Section V** analysis and discusses scheme results. **Section VI** proposed method. **Section VII** includes outcome result possible. **Section VIII** Conclude this review paper. **Section IX** discusses Future Scope.

III) PREVIOUS WORK DONE

Jin Li et. al. (2015) [1] have proposed duplicate check scheme which support authorized duplicate check in hybrid cloud architecture, in this scheme the duplicate tokens of files are generated by the private cloud server with private keys. This scheme protects data security.

Hui Cui et. al. (2016) [2] have proposed Attribute-based encryption (ABE) has been greatly used in cloud computing in which data providers outsource their encrypted data to the cloud and can distribute the data with users controlling specified credentials. This storage system is built under a hybrid cloud architecture, in which a private cloud manipulates the computation and a public cloud handles the storage.

Youngjoo Shin et. al. (2017) [3] have proposed decentralized server aided encryption scheme which supports encrypted deduplication for cloud services. By implementing the inter-KS deduplication algorithm, the proposed scheme provide flexibility in managing a KS to tenants, while at the same time allowing a CSP to perform cross-tenant deduplication over encrypted data. This scheme offers high deduplication scalability and efficiency for storage services in the cloud environment, while still guaranteeing the strongest data confidentiality.

Zheng Yan et. al. (2016) [4] has a practical scheme to handle the encrypted big data in cloud storage with deduplication based on proxy re- encryption and ownership challenge. This scheme can flexibly support data update and distributing with deduplication even when the data holders are offline. This scheme can efficiently perform big data deduplication. It saves the storage space and flexibly supports access control on encrypted data with deduplication.

Yifeng Zheng et. al. (2017) [5] have proposed secure system framework enabling secure deduplication which extremely protecting the video confidentiality. It is resistant to the adversaries in the bounded leakage setting, and the adversaries launching brute-force attacks over predicted videos, respectively.

IV) EXISTING METHODOLOGIES

A.AUTHORIZED DUPLICATE CHECK SCHEME:

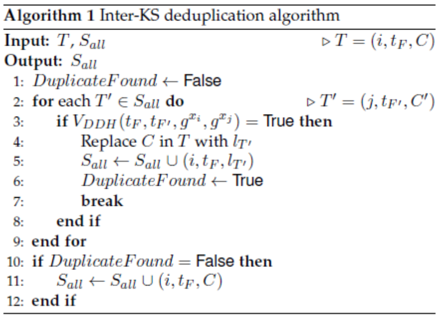
In hybrid cloud approach implements authorized duplicate check scheme which contains three entities such as private cloud, users and server cloud service provider in public cloud. In this system tokens of files are generated by the private cloud with private keys. This system security analysis shows that these schemes are secure in terms of outsider and insider attacks specified in this proposed security model. A user calculates and sends duplicate check tokens to the public cloud server for authorized duplicate check. This approach uses SHA-1 hash function [1].

B. ATTRIBUTE-BASED STORAGE SYSTEM:

The attribute-based storage system which provides secure deduplication in a hybrid cloud, where a private cloud is reliable for duplicate detection and a public cloud handles the storage. In attribute-based encryption (ABE) which include a user's private key is collaborated with an attribute set, a message is encrypted under an access policy over a set of attributes. As well as a user can decrypt a cipher text with his private key if his set of attributes satisfies the access policy collaborated with this cipher text. The private cloud is provided with a trapdoor key collaborated with the respective cipher text [2].

C. SERVER-AIDED ENCRYPTION SCHEMES:

In decentralized server-aided encryption proposed inter-KS algorithm in which the cloud service provider can validate whether two cipher texts encrypted under the different secret keys of different KSs have the same plaintext or not.



Algorithm 1 Inter-KS deduplication algorithm

```
Input: $T$, $S_{all}$                                   ▷ $T = (i, t_F, C)$
Output: $S_{all}$
$DuplicateFound \leftarrow$ False
for each $T' \in S_{all}$ do                            ▷ $T' = (j, t_{F'}, C')$
    if $V_{DDH}(t_F, t_{F'}, g^{x_i}, g^{x_j}) =$ True then
        Replace $C$ in $T$ with $l_{T'}$
        $S_{all} \leftarrow S_{all} \cup (i, t_F, l_{T'})$
        $DuplicateFound \leftarrow$ True
        break
    end if
end for
if $DuplicateFound =$ False then
    $S_{all} \leftarrow S_{all} \cup (i, t_F, C)$
end if
```

By utilizing a blind signature scheme, the inter-KS deduplication algorithm recognized cross-tenant data deduplication without exposing any sensitive data except the cipher text itself [3].

D. ENCRYPTED DATA SCHEME:

In encrypted data deduplication proposed a practical approach to manage the encrypted big data in cloud server with deduplication based on proxy re-encryption and ownership challenge. This scheme can flexibly support data update and distributing with deduplication even when the data holders are offline. Encrypted data can be securely accessed because only authorized data holders can obtain the symmetric keys used for data decryption. These proposed schemes can efficiently implementing on big data deduplication. This scheme save storage space which reduced cost [4].

E. SECURE SYSTEM ARCHITECTURE:

In Secure system architecture an encrypted cloud video hosting service including three different entities which are the user, the agency server and cloud media center. Cloud serves as a video hosting platform saving encrypted videos outsourced by users. It enforces deduplication to remove the storage and bandwidth redundancy, and is required to adaptively deliver the encrypted videos to heterogeneous devices and networks.

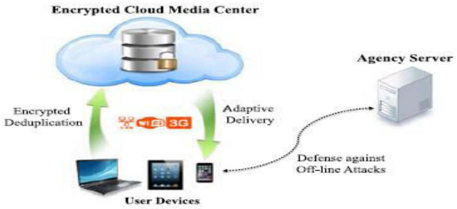


Fig. 1: Secure system architecture

After outsourcing the encrypted videos, the user may delete them at local, and later access own videos at cloud. The agency server, hosted by a third party, facilitates system to defend against offline brute-force attacks [5].

V) ANALYSIS AND DISCUSSION

In hybrid cloud approach proposed authorized duplicate check scheme which includes the duplicate-check tokens of files are generated by the private cloud server with private keys. Security analysis shows that these schemes are secure in terms of outsider and insider attack specified in this security model. This approach is used to improve storage utilization and eliminates redundant data. [1].

The attribute-based storage system is built under a hybrid cloud architecture, where a private cloud manipulates the computation and a public cloud handles the storage space. The private cloud is provided with a trapdoor key associated with the corresponding cipher text, with which it can transfer the cipher text over one access policy into cipher texts of the same plaintext under any other access policies without being aware of the underlying plaintext [2].

In Server-aided encryption proposed inter-KS deduplication algorithm scheme which utilize the parallelism supported by modern CPU architecture. With aid of a KS, the proposed scheme offers the strongest data confidentiality in cloud storage against any users who do not have valid ownership of the data, as well as an honest-but-curious CSP and KS [3].

In encrypted data deduplication showed that these scheme is secure and efficient security model and very suitable for big data deduplication. This scheme can flexibly support data update and sharing with deduplication even when the data holders are offline. Encrypted data can be securely accessed because only authorized data holders can obtain the symmetric keys used for data decryption [4].

A secure system framework enables secure deduplication while strongly protecting the video confidentiality. This system design achieves strong protection of the video confidentiality as well as greatly improves the storage efficiency and dissemination scalability. But due to the considerable amount of storage and bandwidth overhead, increasing the capital cost of using cloud services [5].

File level deduplication scheme	**Advantages**	**Disadvantages**
Authorized Duplicate Check Scheme	This approach is used to improve storage utilization and eliminates redundant data.	This system does not support differential authorized deduplicated check.
Attribute-Based Storage Scheme	It can be used to confidentially share data with other users by specifying an access policy rather than sharing the decryption key. .	The standard ABE systems do not support secure deduplication, which makes them costly to be applied in some commercial storage services.
Server-aided Encryption Scheme	The proposed scheme provides security in terms of data confidentiality, data integrity and collusion resistance.	This scheme argues the feasibility of the proposed scheme by undergoing rigorous security analysis and performance evaluation.

Encrypted Data Scheme	This scheme can flexibly support data update and sharing with deduplication even when the data holders are offline.	The only drawback is that in practically, it is hard to allow data holders to manage deduplication.
Secure System Framework	This system design achieves strong protection of the video confidentiality as well as greatly improves the storage efficiency and dissemination scalability.	Due to the considerable amount of storage and bandwidth overhead, increasing the capital cost of using cloud services.

TABLE 1: Comparisons between different File level deduplication scheme.

VI) PROPOSED METHODOLOGY

AES with secure deduplication system

The proposed methodology aims at providing a new deduplication system with higher reliability by splitting the file and performing both file-level duplication checking. The cloud computing provides an individual user abundant storage space, availability of data and accessibility anytime at anywhere. Due to this it increases data redundancy and computation time on cloud server which reduced space by integrating data deduplication into cloud storage. So that this paper are implementing "**AES with secure deduplication**" on file level deduplication. This scheme ensures the security by means of ownership verification. In this scheme when user upload file ownership verification is done by the Cloud server to identify the valid user when uploading or downloading the file from preventing against attackers. Then users select the file which is to be uploaded and stored by using the application. This approach uses the MD5 hash function to calculate the file's hash value. As well as also maintain all file hash values as index table. Then perform clustering for partitioning cloud data into set of meaningful sub classes. Later compare uploaded file's hash value with the existing hash value in these sub classes. Due to clustering it increase computational speed for matching pair of data. If uploaded file does not present, a new hash value will be saved in the index table, and then user will upload the file into the cloud by applying Advanced Encryption Standard (AES) algorithm which improves high security and consistency of data stored. If uploaded file does exist, then will get the message that 'File already exist' which means deduplication found.

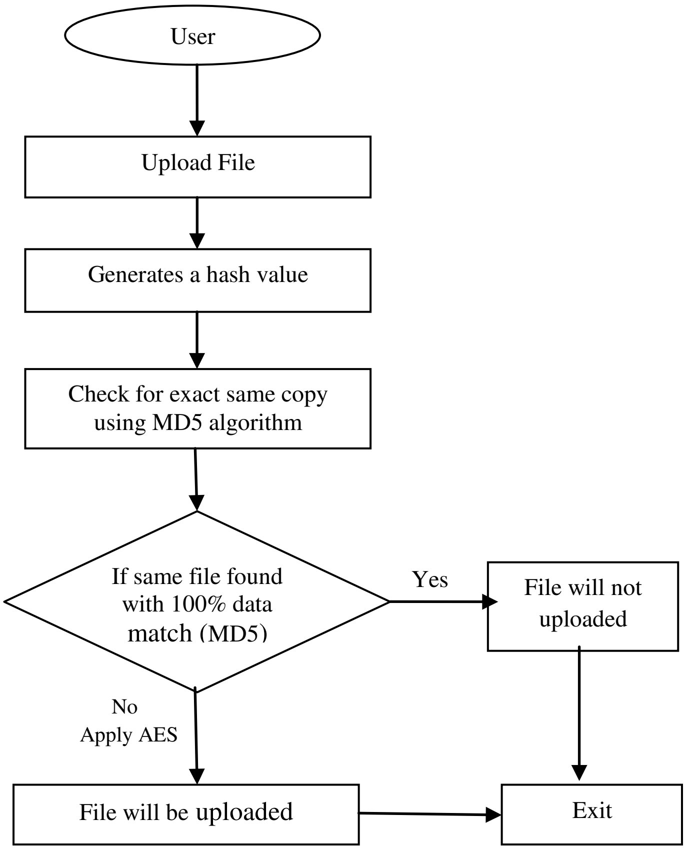


Fig: Working process of File level deduplication.

ALGORITHM:

Step1. The input file uploaded is processed by the server.

Step2. Generate a hash value for the file using MD5.

Step3. Check for exact same copy using MD5 algorithm.

If same file found with 100% data match

a. File will not upload.

Else

a. Perform AES algorithm.

b. File will be uploaded.

Step4. End.

In this way "**AES with secure deduplication system**" improves the data confidentiality, storage utilization, reliability and removes redundancy of data in file level deduplication in cloud storage.

VII) OUTCOME AND POSSIBLE RESULT

This paper performs file-level deduplication and use MD5 algorithm to generate hash value to perform the ownership verification and matching the content in the file. Due to MD5 algorithm it improves speed of file matching time. Later this scheme uses clustering for partitioning cloud data into group of meaningful sub classes. As well as also apply Advanced Encryption Standard algorithm (AES) using more secure and reliable against hackers. This scheme eliminates the redundant data and save storage space.

VIII) CONCLUSION

This paper proposed a simple approach for the file level deduplication for eliminating redundant data in the file. So that this paper proposed "AES with secure deduplication system" which uses MD5 algorithm to generate hash value for matching data in file and also apply Advanced Encryption Standard algorithm (AES) when file is uploaded on the server. The result demonstrates that the duplicated data space can be saved and the upload its performance is not affected by the integrated schemes significantly. That is, this proposed approach actually reduces the storage space consumption by removing redundant files. Thus a new deduplication system is achieved with higher reliability, data confidentiality, storage utilization and security of data in cloud storage.

IX) FUTURE SCOPE:

Future work may include research on how to save storage space for index and time for creation of block for deduplication. Also more improve data confidentiality.

REFERENCES

[1] Jin Li, Yan Kit Li, Xiaofeng Chen, Patrick P.C. Lee, and Wenjing Lou," A Hybrid Cloud Approach for Secure Authorized Deduplication", IEEE TRANSACTIONS ON PARALLEL AND DISTRIBUTED SYSTEMS, vol. 26, no. 5, pp. 1206-1216, MAY 2015.

[2] Hui Cui, Robert H. Deng, Yingjiu Li, and Guowei Wu," Attribute-Based Storage Supporting Secure Deduplication of Encrypted Data in Cloud", JOURNAL OF LATEX CLASS FILES , pp. 1-13, 2016.

[3] Youngjoo Shin, Dongyoung Koo, Joobeom Yun and Junbeom Hur," Decentralized Server-aided Encryption for Secure Deduplication in Cloud Storage", IEEE Transactions on Services Computing, pp. 1-14, 2017.

[4] Zheng Yan, Wenxiu Ding, Xixun Yu, Haiqi Zhu, and Robert H. Deng, Fellow," Deduplication on Encrypted Big Data in Cloud ", IEEE TRANSACTIONS ON BIG DATA, vol. 2, no. 2, pp. 138-150, APRIL-JUNE 2016.

[5] Yifeng Zheng, Xingliang Yuan, Xinyu Wang, Jinghua Jiang, Cong Wang, Xiaolin Gui," Toward Encrypted Cloud Media Center With Secure Deduplication", IEEE TRANSACTIONS ON MULTIMEDIA, vol. 19, no. 2, pp. 251-265.,FEBRUARY 2017.