Deduplication of Encrypted Data in Cloud

Suhasini V. Padwekar
Computer Science and Engineering
Government College Of Engineering Amravati
Amravati, India

Prof. Milind B. Waghmare
Computer Science and Engineering
Government College Of Engineering Amravati
Amravati, India

Abstract—**Cloud computing technology is gaining popularity now days. Cloud stores quite a large amount of data. Cloud services are widely used providing users a way to store process and manage data. The data uploaded on the cloud is increasing each day. The cost required to maintain, process and store is quiet high. The amount of space required to store such huge data is also very large. It is been observed that much of data stored on the cloud is duplicate data. Handling of duplicate data requires more cost that handling original data. So, Deduplication of data is a way to reduce cost and increase efficiency of the system.**

Deduplication is method to eliminate duplicate data on the cloud and maintain unique data. Authorized deduplication of data provides secure and efficient process to allow only the authorized user to access data. The system is based on convergent algorithm and Merkle hash tree. Convergent algorithm is encryption algorithm to maintain privacy of data and achieve authorized deduplication. Merkle hash tree is intoduced .We conduct the security analysis and the simulation experiment to demonstrate the security and efficiency of our proposed system.

Keywords-Data Deduplication; Data privacy; Data confidentiality;

I. INTRODUCTION

Cloud Computing is storing, sharing and accessing resources with availability of internet connection. Cloud computing provide various services to the user. The user can upload their huge data on the cloud. Hardware as well as software requirement is also fulfilled by cloud. Enterprises and personals are shifting their storage requirement to cloud from client. The data stored to cloud increasing day by day. As data on cloud increases, processing power requirement also increases. With the increase in power requirement, cost also increases. Efficiency of the system should be maintained by decreasing power requirement and cost.

As data stored on the cloud is increasing daily. It is been observed that half of the data stored on the cloud is duplicate data. This duplicate data increases processing power and cost. So, there is need to handle duplicate data. Duplicate data can beprocessed by using data deduplication technique.

Data Deduplication is a technique to remove redundant data and maintain unique data. data Deduplication technique is widely used by Cloud Service Provider(CSP). Data Deduplication removes all duplicate data on the cloud. This increases efficiency of the system. Confidentiality of the user is also maintained while performing data deduplication.

II. DATA DEDUPLICATION CLASSIFICATION

A. *Based on data implimentation*

Data deduplication based on data implimentation can be classified as server-side deduplication, client-side deduplication and cross-user deduplication. The cross-user deduplication saves more storage space and is widely used. The rate at which deduplication ia achieved is up to 90%-95%. Client-side deduplication is also known as source-based deduplication. In source-based deduplication duplication copies of file are removed before sending it to the target machine. It sends only single copy of file resulting in reduction of bandwidth consumption. Server-side deduplication is known as target-based deduplication. All data is send from client machine to server machine and process of data deduplication is carried at target machine. Sending of all data increases bandwidth comsumption and cost but performance is much better as compared to client-side deduplication.

B. *Based on processing Unit*

Data Deduplication can be classified into file-level data deduplication and block-level data deduplication. In file-level data deduplication whole file is send form server to target rather than dividing it into multiple blocks. Complete file encryption is done by generating one key of authorized access of file. In block-level data deduplication file is divided into multiple chunks. Each block is the encrypted to obtain keys. all keys are shared with user to allow authorized user to access file. Blocks can be of fixed or variable size.

III. RELATED WORK

In [1] the authors have introduced SRRS system. Secure role re-encryption algorithm comprises of convergent algorithm and role re-encryption algorithm. Convergent algorithm performs encryption and decryption of file to maintain confidentiality of the data. Role re-encryption algorithm performs authorized deduplication. As cloud service provider is assumed to be curious management center

is introduced in the system. Management center manage keys, user's roles by reducing load over the CSP. The system reduces storage requirement and cost.

In [2] authors have proposed novel Attribute-Based Storage system which supports secure and efficient deduplication. Drawback of standard Attribute-based encryption technique is explained in the paper. Attribute-based encryption technique does not support secure deduplication. The system prposed here is based on hybrid cloud environment where private cloud is in charge of identical copies detection and public cloud opts for managing storage.

The system has two major advantages

1) Data Confidentiality is maintained while sharing data by specifying access policy.

2) Data security is maintained at higher level.

In [3] author explained ABE (Attribute Based Encryption) technique used to eliminate duplicate data and reduce storage space. This makes sharing of data efficient. Since the system is based on attributes so if attributes of user matches only then the user is allowed to access data.

In [4] authors have introduced convergent encryption technique to secure data in process of deduplication of data. The data outsourced is converted to cipher text before performing deduplication. The authors have also introduced different privileges to the users.

In [5], authors have introduced (MLE) which provide secure deduplication. This scheme is best for large files as this needs schema perpetuation at servers. As large files needs better maintenance scheme suits it. This scheme supports both file-level and block-level deduplication.

In [6] authors have introduced updatable block-level deduplication which provides deduplication on encrypted data and easy updation of data. The issue in file level deduplication of effective updating of data is overcome here. Some challenges are overcome by MLE and others are effectively dealt by $UBLD_e$ protocol. Dynamic Ownership management challenge is fulfilled here.

In [7] authors initiate idea to reduce the cost of updation of data. The existing MLE solution does not provide effective and secure updation of encrypted data to the user. The cost of updating single bit of data is quite high. So, the authors have introduced Updatable block-level message locked encryption technique which aims to reduce computation cost logarithm to file size. It has also introduced proof-of-ownership to users for access of files.

In [8], the author has introduced scheme which uses Symmetric Encryption algorithm, Hashing technique, Convergent encryption algorithm and token generation scheme to provide authorized duplication of data. Here the user data confidentiality and security is maintained. The data is protected both form passive and active attacks.

In [9] authors have introduced PoW (Proof-of-ownership) with data deduplication to support dynamic ownership management. This system support file-level, cross-user and block-level data deduplication. This scheme effectively carries out secure deduplication and maintains data confidentiality, consistency. It also reduces load of key management and storage space.

In [10] author has surveyed various methodologies and technologies for implementing data deduplication. They have also shown comparison of various technologies. The data confidentiality is compromised at different extent while performing data deduplication is depicted in the paper.

In [11] authors have introduced PoW (Proof-of-ownership) with data deduplication to support dynamic ownership management. This system support file-level, cross-user and block-level data deduplication. This scheme effectively carries out secure deduplication and maintains data confidentiality, consistency. It also reduces load of key management and storage space.

IV. PROPOSED SYSTEM

The proposed system has 3 units: A Cloud Service Provider (CSP), Users (U) and a Management Center (MC), as shown in Figure 1. The user makes request to MC for encrypting file uploaded by user for maintaining data confidentiality. Further the enciphered file is forwarded to the CSP [1].

A. Entities of System Model

1) User

User belongs to different role groups having different role keys. Depending on control policies and role keys users can download and upload files from CSP. The creator of file is is special user and also unique.

2) CSP

Cloud service provider is responsible for data management, storage and verification. The file uploaded by the user is stored and managed by the CSP. CSP verify the user's identity and prevent unauthorized access.

3) MC

Management Center is trusted party and is responsible for role key management and user authorization.

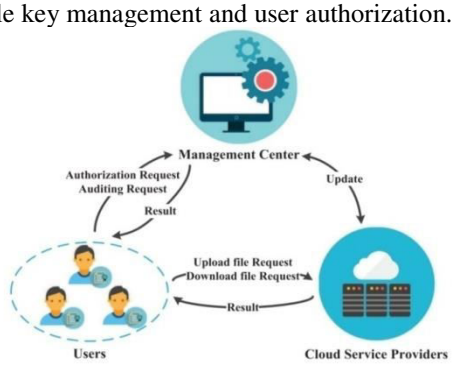

Fig. 1 Posposed System

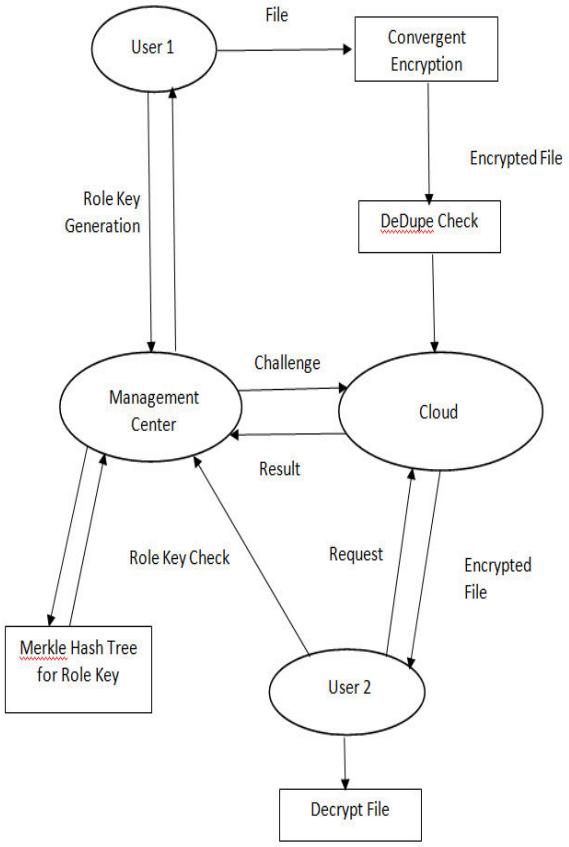

Fig. 2 Flow of the system

B. CONVERGENT ENCRYPTION

Convergent encryption algorithm is a symmetrical encryption algorithm. This algorithm encrypts the user file at the management center to maintain confidentiality of data while performing deduplication process [4], [17]. In this strong hash value is used to obtain convergent key from the original file. Hash value is is applied on the original file to obtain key. This key along with encryption algorithm is applied on file to get cipher text. Identical users who have identical files get identical hash values and identical keys to obtain identical cipher text.

Given original text file f, Encryption file Enc, ciphertext C, hash function h and convergent key k, we obtain

$$k = h(f)$$

$$C = Enck(f)$$

Convergent algorithm is used in SRRS to maintain user privacy and achieve secure deduplication of data in the cloud.

C. Merkle Tree

A hash tree or merkle tree is a tree structure in which each leaf node is a hash of a block of data and each non-leaf node is a hash of its children. This result in a single hash called the Merkle root. If every node has two children, the tree is called a binary hash tree. This merkle hash tree allows secure and efficient stoarge and handling of data. Since each node is generated from leaf node data cannot be altered or damaged thus maintaining confidentiality of data.

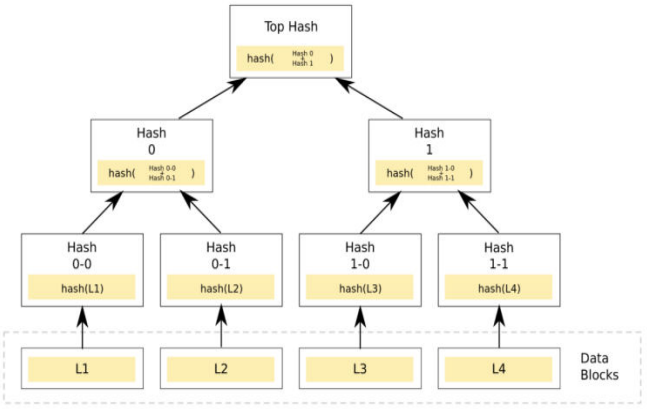

Fig. 3 Merkle Tree

D. Role Authorized Tree

In order to maintain role keys, role authorized tree is defined based on merkle hash tree, as shown in Figure 2. Management Center maintains this role authorized tree in order to manage authorized requests and confidentiality of data in the cloud. This tree has 2 nodes: leaf nodes and internal nodes. A leaf node contains data information and non-leaf or internal node does not have this information. Root node is considered under internal node.

E. Proof of ownership

In order to mainatin privacy of sensitive data proof of ownership is genearted. In case if any user supplies hash value to access the file PoW is implimented by CSP.

V. RESULT

In this section, we give the overall performance of the system to establish secure deduplication of encrypted data on the cloud. The graph below shows performances.

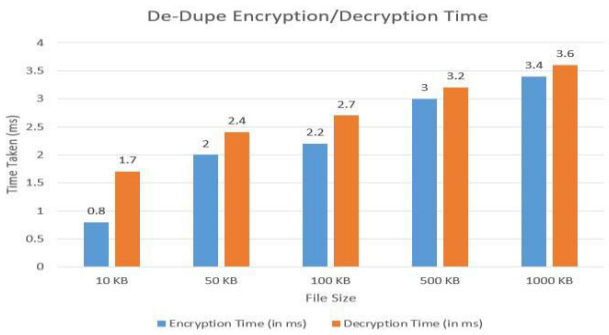

Fig. 4 Encryption/Decryption Time

The Figure 3 shows time required by the system to encrypt and decrypt data. When the user uploads data management center firstly encrypts data and the sends it to the CSP. When user wants to access data it if decrypted first then downloaded by the user. Figure 5 shows hash generation/integrity check time. Figure 6 shows MHT generation/check time.

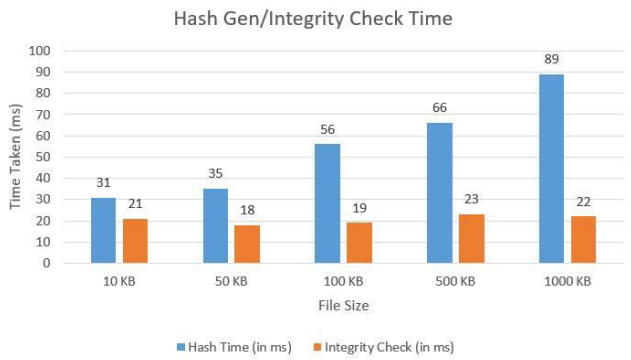

Fig. 5 Hash Generation/Integrity check time

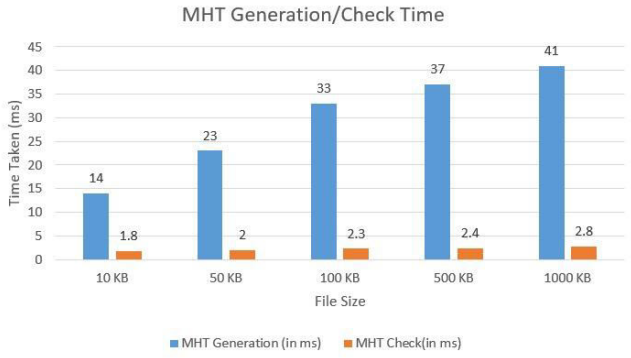

Fig. 6 MHT Generation/Check Time

VI. CONCLUSION

In this paper data deduplication is employed to maintain efficient working of cloud by reducing duplicate data. This also preserves data privacy and allow only authorized user to access data. Storage and bandwidth requirement in cloud is reduced. This system helps to maintain confidentiality of data. In the proposed system convergent algorithm performs encryption and decryption of file to accomplish authorized deduplication. Merkle hash tree secure user data from unauthorized access. The security and performance analysis shows that system is effective.

VII. REFERNCES

[1] Jinbo xiong, Yuanyuan zhang, Shaohua tang, Ximengl liu and Zhiqiang Yao, "Secure encrypted data with authorized deduplication in cloud" IEEE Access, vol. 7, pp. 75090–75104, Jun.2019.

[2] Hui cui , Robert H. deng, Yingjiu Li , Member and Guowei Wu, ‘‘Attribute-based storage supporting secure deduplication of encrypted data in cloud,’’ IEEE transactions on big data, vol. 5, no. 3, July-September 2019.

[3] Hua Ma1 , Ying Xie 1 , Jianfeng Wang2 , Guohua Tian1 , And Zhenhua Liu1, ‘‘Revocable attribute-based encryption scheme with efficient deduplication for e-health systems,’’ Volume 7, 2019.

[4] 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,. 2015.

[5] Chen, R., Mu, Y., Yang, G., & Guo, F., ” BL-MLE: Block-level message-locked encryption for secure large file deduplication”, IEEE Transactions on Security, 2015.

[6] Yongjun Zhao and Sherman S. M. Chow," Updatable block-level Message-locked encryption" *Proc.* IEEE Transaction on Dependable and secure computing, vol. xx, no. y, MAY 2019.

[7] Maozhen Liu, Chao Yang, Qi Jiang, Xiaofeng Chen, Jianfeng Ma, Jian Ren, School of Cyber Engineering, Xidian University, Xi’an, Shaanxi, " Updatable block-level deduplication with dynamic ownership management on encrypted data".

[8] Waghmare, V., & Kapse, S., ”Authorized deduplication: An approach for secure cloud environment, 2016.

[9] Hyungjune shin, Dongyoung koo, Youngjoo shin, and Junbeom hur," Privacy-preserving and updatable block-level data deduplication in cloud storage services" *Proc.* 2018 IEEE 11th International Conference on Cloud Computing.

[10] Nipun Chhabra and Manju Bala,"A Comparative study of data deduplication strategies," in *Proc.* 2018 First International Conference on Secure Cyber Computing and Communication (ICSCCC).

[11] Shunrong Jiang , Tao Jiang and Liangmin Wang," Secure and Efficient cloud data deduplication with ownership management" *Proc.* IEEE Transaction, 2017.

[12] Dapeng Wu, Hang Shi, Honggang Wang, Ruyan Wang, Hua Fang, ‘‘A feature-based learning system for Internet of Things applications,’’ IEEE Internet Things J., vol. 6, no. 2, pp. 1928–1937, Apr. 2019.

[13] J. Xiong, Y. Zhang, X. Li, M. Lin, Z. Yao, and G. Liu, ‘‘RSE-PoW: A role symmetric encryption pow scheme with authorized deduplication for multimedia data,’’ Mobile Netw. Appl., vol. 23, no. 3, pp. 650–663, 2018.

[14] W. Xia, H. Jiang, D. Feng, F. Douglis, P. Shilane, Y. Hua, M. Fu, Y. Zhang, and Y. Zhou, ‘‘A comprehensive study of the past, present, and future of data deduplication,’’ Proc. IEEE, vol. 104, no. 9, pp. 1681–1710, Sep. 2016.

[15] J. Li, C. Qin, P. P. C. Lee, and X. Zhang, ‘‘Information leakage in encrypted deduplication via frequency analysis,’’ in Proc. 47th Annu. IEEE/IFIP Int. Conf. Dependable Syst. Netw., Jun. 2017, pp. 1–12.