ISSN: 0975-8585

Research Journal of Pharmaceutical, Biological and Chemical Sciences

Encryption of Big Data in Cloud using Deduplication Technique.

G Usha Devi*, and Supriya G.

School of Information Technology, Vellore Institute of Technology, Vellore-632014, Tamil Nadu, India

ABSTRACT

Data Deduplication plays one of the major roles in reducing the data storage cost required in the cloud. Data Deduplication is the process of compressing the data that are repeatedly stored; it helps in maintaining the unique data by eliminating its duplicate, thus reducing the storage space. It reduces the storage space by checking the data size through its binary value of the data in the database and it checks with a database for similar data chunks and finally, it either deduplicates the data or rehydrates the data using the pointer by maintaining the index for the same. This paper presents a survey of different techniques that are used to deduplicate the data by compressing the storage space required for storing similar data in a cloud environment.

Keywords: Data Deduplication, Index, Rehydrate, Cloud Computing.

Corresponding author

ISSN: 0975-8585

INTRODUCTION

Cloud computing has become a promising platform by providing various technology services, in that technology services one of the main and popular services is storage services in the cloud.This helps in storage and is responsible for easy accessing and available of data. The client availing those storage services either buy it or lease the storage services from the service provider based on their need.

Cloud Service Provider (CSP) helps in maintains the data that has been uploaded by clients. The CSP maintains the uploaded data through data centers. Since the data that is uploaded tocloud is not confidential and trustworthy because the intruders may easily find the ways to manage the data that is stored. This may lead to the data security risk and also the data privacy of the clients to avoid this trust break with cloud services it is better to upload the encrypted data in the cloud. Encryptions of data help in maintain the privacy and security of the data as well. Even though the cloud has enormous space in cloud storage, the process of storing the similar data file in cloud leads in wasting the cloud space. So to avoid this process of data deduplication is introduced. Data deduplication helps in storing the same data file with different names and hence by storing the space and money for storage services. But the deduplication technique does not support when it comes to encrypted data because traditional encryption techniques require different cipher key values hence it is difficult to deduplicate the data files.

Data Deduplication Technique

Data deduplication technique helps in reducing the storage disk tapes it also helps in lowering the bandwidth of the network [7]. The main aim of data deduplication is to maintain only one unique file with unique data content if more than one is found only the instance of the pointer is replaced for reference purpose.

File-level data deduplication

File-level data deduplication is also called as single instance storage. If the client or user uploads any file it first checks for repetition of the same file is present or not if data content doesn't match it will store the file and as well as it will update the index. In case if any repetition of the same content file is found it maintains a "stub" and copies the alternative where the stub actually points to the original file where it actually resides.

Block-level data deduplication

Block-level data deduplication [1] [4] here it splits the data file into n number of data blocks and the uniqueness of data content is found in each and every block using the unique signature at the time of implementing hash algorithm all the data blocks are maintained using the hash table so that the pointer value of hash table can reduce the time required for data block deduplication.

Byte-level data deduplication

In byte-level data deduplication, the uniqueness of data is checked in byte –by byte level. In this technique, the accuracy rate is considered to be higher [9]. The data check for deduplication with the older backup files whether it matches with similar data files that have been received in the past.

Zheng Yan et al. [1] proposed a scheme to deduplicate the encrypted data that is stored in the cloud using proxy re-encryption and ownership challenge with the access control technique to deduplicate the data file in cloud for minimising the storage space that is required in maintaining the duplicate files, hence here the concept of unique data file is entertained. The scheme undergoes the following process:Encrypted data upload, Data deduplication, Data deletion, Data owner management, Encrypted data update. By undergoing all these processes this paper ensures that the data are deduplicated and it also supports the deduplication process even when the user is offline.

Tiawei Yuan et al. [2] proposed the concepts of auditing with deduplication techniques for data integrity auditing it uses the Proof Of Retrievability (POR) and Proof Of Data Possession (PODP) scheme and for the storage purpose it uses the proof of ownership challenge .it also ensures that both the integrity and

storage are achieved simultaneously. The security of proposed work is startedusing:Diffie-Hellman,Static Diffie-Hellman, T-Strong Diffie-Hellman. Polynomial and homomorphic linear authentication tags are used for authentication of integrity auditing and storage. In this polynomial authentication is independent when compared to all other related applications.

Chao Yang et al. [3] proposed to minimize the bandwidth and storage space required for deduplication by using the client-side deduplication technique. It uses provable ownership of the file in client level; it also minimizes the vulnerable attacks by the intruder by getting full access to the data by having the partial information about the file like the hash value of file etc. So to avoid those attacks even if the user wants to access the partial data they must undergo the provable ownership of the file for whole data file preserve the security of data file by spot checking for verification purpose. This paper proposes efficient and cryptographically secure scheme for the client to prove the server for the possession of original data.

Pasquale Puzio et al. [4] proposed block-level data deduplication and also the confidentiality of data [5]. It first uses the convergent encryption technique and it also uses a special and additional encryption layer technique for access control mechanism and key management for each and every block Metadata is used for key management purpose in Metadata manager it containsFile table, Pointer table, Signature table, Linked list. This mechanism is considered as the additional mechanism for encryption, it uses proof of Retrievability and integrity checking for storage and secure services using ClouDedup.

Youngjoo Shin et al. [5] proposed a novel based scheme called SEED (Serverless and Efficient Encrypted Deduplication) with that it provides a high level of confidentiality to the outsourced file. It uses pairing based encryption in which the user generates its own public or private keys hence they share their keys directly, this kind of sharing the key directly helps in maintaining the interactive encryption process and the SEED scheme actually supports lazy encryption technique and thus helps in reducing deduplication from the client side. This method also shows that it is very much efficient when compared to PAKE (Password Authenticated Key Exchange) scheme [19]. For file encryption, it uses symmetric encryption algorithm and the SEED scheme has the capability to withstand brute force attack.

Hui Cui et al. [6] proposed an Attribute-Based Encryption (ABE) in which it uses the hybrid cloud for security and storage it uses both public and private cloud. It uses the public cloud for storage and private cloud is responsible for duplicate detection through this it provides higher data confidentiality. The main advantage of this paper is that it uses a Ciphertext-PolicyAttribute-Based Encryption (CP-ABE) for secure encryption while the standard attribute based encryption scheme does not support secure encryption. In private cloud, it uses trapdoor key management by which the duplicates can be found easily.

Chun- I –Fan et al. [7] propose a hybrid deduplication technique in a cloud environment. In the hybrid mechanism, it uses three blocks namely cipher blocks, check blocks and enabling blocks. In enabling block AES key is used to encrypt the data and the same data is stored in enabling block. Check block is used to find whether the encrypted file has repetition or not. And finally,the session key is used for encryption. This scheme uses SHA-2 for hash function and AESalgorithm for a public key generation. This hybrid scheme mainly helps in eliminating the duplicate files mainly in multimedia files.

Junbeom Hur et al. [8] proposed server side scheme for deduplication it ensures the data integrity and access control of the data. It uses convergent encryption method and proof of ownership challenge (POW), server-side deduplication is efficient in identifying the duplicates. Even if the ownership of the file keeps changing dynamically. It uses group key management in every ownership group. The group key is distributed in the form of a binary tree to the valid user only it ensures the confidentiality to the owner data without leaking any data with the help of group key management and POW challenge. The data are not even leaked within the same user community. The server maintains the ownership list in order to manage the dynamic ownership change. This scheme uses convergent encryption technique and KES algorithm for secure deduplication.

Tao Jiang et al. [9] proposed a new scheme by introducing a new primitive µR-MLE2 to reduce the overhead that is caused in R-MLE2. This µR-MLE2 especially works for the very large amount of data. It undergoes equality testing using the decision tree. This scheme is interactive between the client-server.In this paper instead of linear pair comparison, it adopts equality testing for verifying client and server side it uses

two algorithms static deduplication decision tree and dynamic deduplication decision tree. It uses SHA-1 for the hash function. Data are efficiently deduplicated using decision tree balanced based on the equality sharing.

Kirubakaran et al. [10] propose that the deduplication for the cloud is weighed by calculating the deduplication rate using the simple formula i.e. the data size before deduplication is divided by the data size after deduplication and the final answer is taken in percentage to identify the exact rate of duplicated files. This scheme uses an SHA-224 algorithm for hash generation.

Siewei Luo et al. [11] propose a scheme for deduplication using a coalescing algorithm for chunks. In general, the sub-chunks are merged into super chunks with no particular specifications but in this paper, the sub-chunks are merged into super chunks with their number of chunks with minimum and maximum on the basis of coalescing algorithm. This process is done during deduplication that is after the encryption. It uses content-defined chunking algorithm for data chunks which uses Rabin's fingerprint it compares its evaluation technique with finger diff algorithm and also sparse indexing for backup with disk-to-disk. They actually name their algorithm as min-max coalescing algorithm particularly to minimize the amount of cost required for chunking. With this, the input and output performance of the chunks have been evaluated in this paper in order to deduplicate in a secured and efficient manner for the data.

Shengmei Luo et al. [12] proposed a new scheme called Boafft which is used for storage in the cloud. In this data's are stored in the cloud by using distributed manner. In this method, it maintains many servers and here the servers are not only responsible for storage but also it do the data deduplication process in parallel.

This scheme uses data routing algorithm to efficiently identify the location of data. Boafft helps in organizing the local deduplication it also uses fingerprint cache to verify the data access rights on each server. Boafft is generally called or maintained as cluster based deduplication. It uses min-hash algorithm the main aim of Boafft is to find the similarities in the local servers and do check parallel for deduplication process.

G Madhubala et al. [13] proposed deduplication technique with the help of nature inspired thing. In this paper, it verifies the data for text matching using the algorithm and the genetic programming model to identify the ideology. It uses sequence matching algorithm and Levenshtein algorithm for text matching and for deduplication process it uses genetic programming algorithm with this genetic programming approach the cloud users will be benefited more.

Sejun Song et al. [14] proposed SEACOD i.e. Selective Encryption and Component-Oriented Deduplication. It collaboratively uses three cloud computing, fog computing, mobile edge computing and cloud computing. The seacod technique helps in identifying the repetition of data and make the storage space very efficiently it implements block level deduplication. SEACOD maintains android smartphones in the collaborative cloud environment and also the fog computing technology has been used in collaborative cloud computing. Performs block level deduplication and performs an evaluation through 3DES and blowfish algorithm for secure deduplication.

Jan Stanek et al. [15] proposed a multi-layered crypto-system. The cryptosystemis used with threshold concepts it is providing a high level of security using Diffie-Hellman Symmetric External algorithm. This enhanced security is maintained in oracle model by maintaining two entities namely index repository and identity provider. Threshold convergent cryptosystem algorithm is used to encrypt with pairing Diffie-Hellman Symmetric External algorithm. It is effective for secure storage than identifying the duplicates and the process of deduplication.

Table 1 shows different methods and features used in different papers for deduplication technique.

CONCLUSION

The concept of deduplication in cloud paves the way for minimizing the storage space in the cloud and the encryption technique for those data helps in maintain the privacy of the data. Although encryption does not support deduplication technique since it generates same cipher text.This survey describes the various

methodologies that are proposed in the different paper for encryption of data in the cloud using deduplication technique and this opens a wide area of research.

Table 1: Methods and features

s.no	Method	Features
1.	Proxy re-encryption and ownership challenge	Proxy re-encryption helps in maintaining security and using ciphertext, it allowsdecrypting. It helps in revealing either of the keys. It uses two functions delegation and transitivity
2	Deduplication with auditing,POR,PODP,AES algorithm	To minimize both storage capacity and network bandwidth. AES has 128bits as fixed block size.
3	Provable ownership of the file by spot checking	Detects the client misbehavior, efficient in identifying the ownership.
4	Block-level deduplication, convergent encryption, metadata manager(KMIP)	It helps in protecting,organizing,backup and store encryption keys.
5	SEED(Serverless and efficient encrypted deduplication)	SEED is efficient when compared to PAKE(password Authentication key Exchange) scheme. It withstands brute force attack.
6	CP-ABE	Helps to overcome the overhead in standard attribute based encryption. It provides secure encryption using trapdoor key management to find duplicates.
7	SHA-2,AES,hybrid deduplication	SHA-2 is implemented using SSL protocol. Secure password hashing.
8	Server side deduplication,POW,KES algorithm	Helps in identifying the change in ownership. even does not leakwithin the group community.
9	µR-MLE2,SHA-1	Overcome the overhead in R-MLE2,helps in equality testing,also supportsthevery large amount of data.
10	SHA-224 algorithm	Overcome computational complexity, helps in handling a large number of hashes.
11	Coalescing algorithm, content-defined chunking algorithm	Increase the execution speed comparatively. Studies on multiple chunking.
12	Boafft, data routing algorithm	Supportscluster-based deduplication to find similarities in local servers.
13	Genetic programming algorithm, Levenshtein algorithm	Nature inspired technique verifies text matching to find ideology.
14	SEACOD	Assures security in collaborative computing.
15	Symmetric Diffie Hellman algorithm	Effective storage in a secure manner

REFERENCES

[1] Yan, Zheng, et al. "Deduplication on encrypted big data in cloud." *IEEE Transactions on Big Data* 2.2 (2016): 138-150.

[2] Yuan, Jiawei, and Shucheng Yu. "Secure and constant cost public cloud storage auditing with deduplication." *Communications and Network Security, 2013 IEEE Conference on*. IEEE, 2013.

[3] Yang, Chao, Jian Ren, and Jianfeng Ma. "Provable ownership of files in deduplication cloud storage." *Security and Communication Networks* 8.14 (2015): 2457-2468.

[4] Puzio, Pasquale, et al. "ClouDedup: secure deduplication with encrypted data for cloud storage." *Cloud Computing Technology and Science, 2013 IEEE 5th International Conference on*. Vol. 1. IEEE, 2013.

[5] Shin, Youngjoo, et al. "SEED: Enabling Serverless and Efficient Encrypted Deduplication for Cloud Storage."

[6] Cui, Hui, et al. "Attribute-Based Storage Supporting Secure Deduplication of Encrypted Data in Cloud." *IEEE Transactions on Big Data* (2017).

[7] Fan, Chun-I., Shi-Yuan Huang, and Wen-Che Hsu. "Hybrid data deduplication in cloud environment." *Information Security and Intelligence Control (ISIC), 2012 International Conference on*. IEEE, 2012.

[8] Hur, Junbeom, et al. "Secure data deduplication with dynamic ownership management in cloud storage." *IEEE Transactions on Knowledge and Data Engineering* 28.11 (2016): 3113-3125.

[9] Jiang, Tao, et al. "Secure and Efficient Cloud Data Deduplication With Randomized Tag." *IEEE Transactions on Information Forensics and Security* 12.3 (2017): 532-543.

[10] Kirubakaran, R., C. Mano Prathibhan, and C. Karthika. "A cloud based model for deduplication of large data." *Engineering and Technology (ICETECH), 2015 IEEE International Conference on*.IEEE, 2015.

[11] Luo, Siwei, and MengshuHou. "A novel chunk coalescing algorithm for data deduplication in cloud storage." *Applied Electrical Engineering and Computing Technologies (AEECT), 2013 IEEE Jordan Conference on*. IEEE, 2013.

[12] Luo, Shengmei, et al. "Boafft: Distributed deduplication for big data storage in the cloud." *IEEE Transactions on Cloud Computing* (2015).

[13] Madhubala, G., et al. "Nature-Inspired enhanced data deduplication for efficient cloud storage." *Recent Trends in Information Technology (ICRTIT), 2014 International Conference on*.IEEE, 2014.

[14] Song, Sejun, Baek-Young Choi, and Daehee Kim. "Selective encryption and component-oriented deduplication for mobile cloud data computing." *Computing, Networking and Communications (ICNC), 2016 International Conference on*.IEEE, 2016.

[15] Stanek, Jan, and Lukas Kencl. "Enhanced Secure Threshold Data Deduplication Scheme for Cloud Storage." *IEEE Transactions on Dependable and Secure Computing* (2016).