IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727,
PP 26-29
www.iosrjournals.org

Encrypted Clou Dedup: Data Management

Mayuri U. Bedis[1], Prashant P. Bhavsar [2], Rahul H. Bilade[3],
Smital V. Vikhankar[4]

[1,]B.E Student, Department of Computer Engineering
LokneteGopinathjiMunde Institute of Engineering Education & Research (LoGMIEER) Nashik-02, India
[2]B.E Student, Department of Computer Engineering
LokneteGopinathjiMunde Institute of Engineering Education & Research (LoGMIEER) Nashik-02, India
[3] B.E Student, Department of Computer Engineering
LokneteGopinathjiMunde Institute of Engineering Education & Research (LoGMIEER) Nashik-02, India
[4]B.E Student, Department of Computer Engineering
LokneteGopinathjiMunde Institute of Engineering Education & Research (LoGMIEER) Nashik-02, India

Abstract: *Cloud storage is a storage service provide to user, where they can transfer their data anytime and any- where. With the great development of cloud computing in modern era, the massively increasing number of data, the chunk of information storage and the application requirement for high availability of data, network backup is facing an extraordinary challenge. On the other hand, a cloud server normally performs a data compression technique (data deduplication) to eliminate duplicate data because the storage is not infinite. Data deduplication, which makes possible for data holder to share a copy of the duplicate data, can be performed to reduce the consuming of storage space. Due to the issues, there is a some research on encrypted data deduplication. Data deduplication is one of fundamental data compression techniques for eliminating multiple copies of repeating data, and has been widely used in cloud storage to miniaturize the volume of storage space and bandwidth. Deduplication is storage optimization technique that avoids keeping duplicate copies of data. Presently, to provide security, data stored in cloud also other large storage areas are in an encrypted format and one issue with that is, we cannot apply deduplication over such an encrypted data. Thus, performing deduplication on the encrypted data in cloud surface to be a challenging task.*

Keywords -*Deduplication, Cloud storage, Avoid Duplication in Cloud System*

I. Introduction

The In modern days, cloud computing has become an important topic and gives many advantages through various services. The many database, device, hardware, and operating system can be handled by a cloud server. Users only need some devices, which can connect to the cloud. But, in the environment, the cloud server can gather and control all the transmitted data because all the data are stored the cloud. The security and privacy are major cloud computing. Cloud computing gives a new way to transport services by reorganize resources over the Internet and providing them to users on requirement. It plays a vital role in supporting data storage, processing, and managing in the Internet of Things (IoT). Numerous cloud service providers (CSPs) offer huge size of storage to maintain and manage IoT data, which can combine videos, photos, and personal health records.

We mainly focus on the cloud storage. Users can save their data (videos, photos, and personal health records) in the cloud storage and download the stored data anyplace. Even if users impoverish their own storage spaces, the cloud server can increase their storage spaces without spoiling the stored data. However, the fast improvement of storage requirements responsibility the cloud storage, which is not infinite. The cloud storage server normally applies the data deduplication to reduce the utilization of storage space.

Cloud service providers such as Drop box, one drive and Google drive massively rely on data deduplication to save storage by only saving one copy of each that uploaded file. While recent studies report that whole file deduplication can achieve up to 50% storage space, users do not directly profit from these savings as there is no clear relation between the prices offered to the users and effective storage costs.

Also, the use of social media sites such as Facebook, Instagram, YouTube to name a few and digital cameras, have given to a fast growth of data which known as Big Data problem. In, it has been said that the global data supply hiked 3.7 trillion GB in 2013 - but just 0.5% of it was used for analysis. In the same having, the volumes of data are projected to reach 50 ZB per individual by 2020. A study has also tell that only 25% of data is exclusive in data warehouses; and only 35 GB of the entire data for each individual user are exclusive and the other ones are similar data that are shared with various cloud users.

At the equivalent time, the goal of encryptions is to save information secret and make it hard to analyse the encrypted data (i.e., ciphertexts) from random values. If an encryption is protected, it would be difficult to obtain information from ciphertexts. Hence, encrypted data deduplication is converted to a challenge because the beginning ofdata deduplication is to search for same data. If the cloud server can separate out whether the contents of two ciphertexts are the same, it means that the ciphertexts lose some information about the plaintexts.

Data deduplication is a data compression technique which makes all the data owners, who upload the same data, shared a single copy of duplicate data and removes the duplicate copies in the storage. When users transmit their data, the cloud storage server wills analysis whether the transmitted data have been saved or not. If the data have not been saved, it will be actually written in the storage; any other way, the cloud storage server only stores a pointer, which points to the first saved copy, rather than saving the whole data. Hence, it can avoid the same data being saved frequently. Normally, data deduplication can be divided into two basic approaches: the target-based data deduplication and the source-based data deduplication.

The two approaches of data deduplication are defined as follows:

Target-Based Data Deduplication: In this approach, users simply transfer their data, and the steps of data deduplication are hold by the cloud storage server. The target-based approach can upgrade the storage utilization, and users do not have to shift their habits of using cloud storage services. However, the target based approach mainly focuses on avoiding saving duplicate data. Those duplicate data are still transferred repeatedly. Therefore, it cannot improve the size of transmissions.

Source-Based Data Deduplication: Source-Based Data Deduplication: In this approach, users have to transmit the identification of their own data and query the cloud storage server whether the data are stored in the cloud storage before really transmitting them. If the data have not been saved, users need to transmit the whole data, and the cloud storage server completely saves them. Otherwise, users need to transmit only the metadata, and the cloud storage server commonly creates a pointer, which points to the first saved copy. Therefore, the source-based approach can improve both the utilization of the storage and the bandwidth. Nevertheless, it changes the familiar process of cloud storage services. When users want to transmit their data, they must query the cloud storage server for the existence of the data first. Since all the data are transmitted to the cloud storage, users attend to the privacy problem in cloud storage service.

We consider two types of attackers in data deduplication as follows:

Malicious User: It is a special attacker in source-based data deduplication. In a source-based data deduplication scheme, each user queries the cloud storage server whether the data have been uploaded by another. The cloud storage server responds ”Yes “or ”No” honestly to avoid the duplicate data being uploaded repeatedly. Therefore, a malicious user can use the response of the cloud storage server to obtain private information about the existence of data.

Cloud Storage Server: In cloud storage service,

The cloud storage server can obtain and control all the uploaded data. In addition, if encrypted data deduplication can be performed, even though all the data are encrypted, the cloud storage server can still analyse the encrypted data by using the encrypted data deduplication and try to obtain information about the plaintexts.

II. Related Works

There have been several of deduplication technologies proposed in recent times. Most researchers observant on text deduplication like. They suggest a scheme to address the key management in deduplication system. Technique based on the file-level deduplication is to remove the same file to reduce the data storage capacity, save storage space capacity. It uses a hash function for each file which in server to compute a hash value. Any two files with the same hash value are considered to be the duplicate file. For example, FarSite, SIS, EMC Centre systems use this way. Technique based on the block-level deduplication is to remove the same data block to reduce storage space. This method is to split a file into some data blocks or chunk , and uses hash functions compute the hash value, which be named as block or chunk fingerprint. Any two data block or chunk with the same block fingerprint are defined duplicate data block or chunk. Based on the Deduplication remove time, Deduplication technology could separate to on-line deduplication and post-processing deduplication.

On-line deduplication is to remove the duplicate data before storing; the storage service always stores a unique data copy. Post6-processing deduplication required additional storage buffer to realize remove repeated data. Based on the deduplication remove place, it can be divided to client deduplication and service deduplication. Client deduplication is before sending the data copy to cloud server, user check and remove duplicate data. Service deduplication is executing duplicate data check and removes with services resource in cloud server. However, multi-media data like videos, images are larger than text. So deduplication is becoming more significant. Researchers have given attention to this field like Compression technique save the space of

cloud storage in some way, but deduplication will give attention this problem from the root. Many data deduplication mechanisms have been proposed recently. In this section, we research into the encrypted data deduplication and discuss some security issues.

A. Data Deduplication on Ciphertext

The first step of data deduplication is to search duplicate data in cloud storage, and the cloud storage server only saves one copy of the duplicate data, which can be shared by the data owners. This process is easy when all the stored data are unencrypted. However, when the data owners encrypt their data using their secret keys for privacy consideration, the above mentioned data deduplication becomes infeasible. There are two ways that can launch data deduplication on encrypted data:

- Using data deduplication on ciphertexts is infeasible because ciphertexts are distinct even though the corresponding plaintexts are the same. Therefore, the first way is to add a header. The header can let the cloud storage server be able to identify whether two different ciphertexts correspond to the same plaintext. However, the duplicate ciphertexts, which correspond to the same plaintext, were generated by different data owners using different secret keys. There is another problem: how can the cloud storage server save a shared ciphertext such that all the data owners can decrypt it, in which the cloud storage server cannot know the corresponding plaintext?
- The second way is to let the same plaintexts correspond to the same ciphertexts. In order to avoid the cloud storage server obtaining the plaintexts, the secret keys must be generated from the plaintexts, and thus the input of the encryption algorithm has no random factor. It means that the same plaintexts generate the same secret keys, and using the same secret keys to

III. Architecture

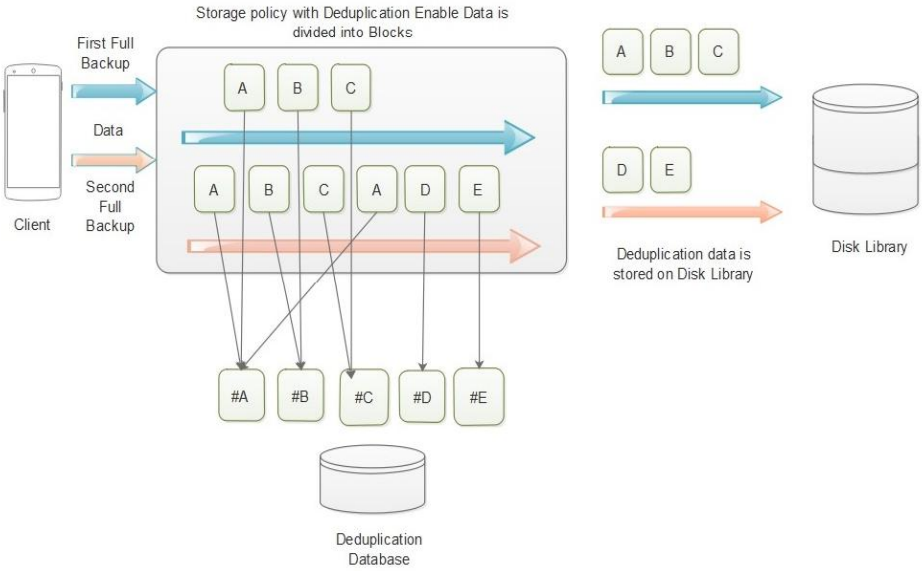

Fig. Architecture

A block of facts is examine from the supply that is uploaded by way of consumer.

- Then this uploaded fact is separated out the usage of tag which is very beneficial for diagnosed for data kind.
- A signature for the block of statistics is generated using hash set of rules. Signatures are unique for special data blocks.
- The signature is as compared towards a database of current signatures for records blocks already in destination storage. The database has containing the signatures is known as the Deduplication Database (DDB).
- If the signature already exists, the DDB up to date to reflect another use of a present information has block on destination storage capacity. The assigned media agent writes the index records and the duplicate statistics block is discarded.
- If the signature does now not exist, the DDB is up to date with the new signature.

Metadata storage space is estimated by taking into account four main data structures:

- **File table:** The file table stores one record for each file and contains the file id (256 bits), file name (256 bits), user id (32 bits) and the id of the first data block (256 bits).
- **Pointer table:** The pointer table stores one record for each block and contains the block id (256 bits) and the id of the actual block stored at the cloud storage provider (64 bits).
- **Signature table:** The signature table stores one record for each block (non-deduplicated) and contains the block id (256 bits), the file id (256 bits) and the signature (2048 bits for the first block, 128 bits for the remaining blocks).
- **Linked list:** The linked list contains one node (256 bits) and zero or more links for each block. A link contains a pointer (64 bits) to a successor block for a given file and stores additional information such as encrypted block keys (256 bits) and file id (256 its).
- **Proof Of Ownership :**Deduplication works by means of computing cryptographic hash characteristic on to facts and the use of this hash price to decide similar statistics. once a replica copy is found then new records is not uploaded but pointer to report possession is updated for that reason saving storage and bandwidth. in relation to purchaser aspect deduplication, hash values of records are computed at patron and send for reproduction check. An attacker, who profits get entry to to hash price of a facts which now not authorized to him/her, may claim deduplication of file and thereby gaining access to the record. To protect such an attack, a proof Of ownership (PoW) has been proposed in, and various works like, and so on adapted this method. PoW works as an interactive algorithm between two parties - a prover and verifier to prove the ownership of the file. Verifier computes a quick cost of statistics M whereas, a prover want to compute short fee of M and send it to verifier for claiming ownership of M.

IV. Conclusion

Deduplication is storage optimization technique that avoids keeping duplicate copies of data. But, deduplication is less beneficial with encrypted data since, different key encryption convert same data into various formats. In this paper different methods are examined where deduplication methods are implementing on encrypted data in a large storage area. Most of the technique studied here work on the basis of concurrent encryption, which is a simple way that makes deduplication adaptable with encrypted data. In this information intense world, we cannot compromise on both security and duplication of data across storage areas. A procedure needs to be formulated which will increase storage optimization without negotiating on encryption method; by giving deduplication technique in data storage servers where the available data is encrypted.

Acknowledgements

I would like to my special thanks of honour to my teacher and Head of Department Prof.K.V.Ugale as well as our principal Dr.A.K.Dwivedi who gave me the golden opportunity to do this wonderful paper on the topic 'Encrypted ClouDedup: Data Management with Deduplication in Cloud database server', which also helped me in doing a lot of Research and I came to know about so many such new things I am really thankful to them. Secondly I would also like to thank my parents and partner who had been helping me a lot in finalizing this paper within the limited time frame.

References

[1] Amazon EC2. http://aws.amazon.com/ec2/.
[2] Amazon Glacier. http://aws.amazon.com/glacier/.
[3] AWS Cloud HSM. http://aws.amazon.com/cloudhsm/.
[4] Dropbox. http://www.dropbox.com.
[5] Google Drive. http://drive.google.com/
[6] Opendedup. http://opendedup.org/.
[7] Zheng Yan, Mingjun Wang, and Yuxiang Li "Encrypted Data Management with Deduplication in Cloud Computing" Xidian University, China, IEEE 2016
[8] .D.T. Meyer and W.J. Bolosky, "A Study of Practical Deduplication," ACM Trans. Storage, vol. 7, no. 4, 2012, pp. 1–20
[9] M. Bellare, S. Keelveedhi, and T. Ristenpart, "Message-Locked Encryption and Secure Deduplication," Advances in Cryptology (EUROCRYPT 13), LNCS 7881, 2013, pp. 296–312
[10] J. Li et al., "A Hybrid Cloud Approach for Secure Authorized Deduplication," IEEE Trans. Parallel Distributed Systems, vol. 26, no. 5, 2015, pp. 1206–1216
[11] J. Harauz, L. M. Kaufman, and B. Potter, "Data security in the world of cloud computing," IEEE Security and Privacy, vol. 7, no. 4, pp. 61–64, 2009.
[12] D. Zissis and D. Lekkas, "Addressing cloud computing security issues," Future Generation Computer Systems, vol.28, no.3, pp.583–592, 2012.