IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

A Survey on Encrypted Data Deduplication Techniques in Cloud Computing

[1]Gayathri, [2]Ashrittha, [3]Soni, [4]Dr.S.Srinivasan
[1]student, [2]student, [3]student, [4]associate professor
[1]RMD ENGINEERING COLLEGE,
[2]RMD ENGINEERING COLLEGE,
[3]RMD ENGINEERING COLLEGE,
[4]RMD ENGINEERING COLLEGE

Abstract

In cloud storage services, deduplication technology is essential to reduce the space and bandwidth requirements of services . It ensures that there is no redundant data and stored only one copy of data. Deduplication is useful when multiple users upload the same data to the cloud storage. At the same time, it leads to several issues relating to security and ownership. Commonly, Proof of Ownership schemes are used to prove the ownership of files. This paper proposes several deduplication schemes to eliminate the duplicate copies one of the challenges in deduplication is that it should consider the dynamic changes in the ownership of the data that occur frequently.

Introduction

Cloud storage as one of the most important services of cloud computing. It helps the users to store any amount of data irrespective of storage availability. Typically cloud data are always encrypted before uploaded due to security concerns. However, encrypted data leads to cloud storage wastage and complicate data sharing. It complicates both data storage and management with duplicate data. Traditionally, deduplication is completely controlled by either data owners or cloud servers. Perhaps, some of them are well suited to the cloud environment.

Data De-duplication methods are classified into two broad categories: File –Level Deduplication and Block level deduplication. In the File-level De-duplication, one copy of a file is kept in the cloud storage in all circumstances. The duplication of the file is detected by checking whether the file is identical or not. Such files are eliminated to optimize cloud storage. In the second type Block-level Deduplication, that segments of each file is stored and managed separately. Hence, it stores solely one copy of every data block.

Many deduplication methods have been proposed to control redundancy in cloud storage server. This paper presents the outline of key deduplication methods and their merits and demerits. It is organized as follows. Section 2 describes the deduplication techniques and Section 3 presents the challenges in deduplication. Finally, Section 4 concludes the paper.

2. DEDUPLICATION METHODS

Jiawei Yuan and Shucheng Yu have applied Proof of Ownership (POW) to improve cloud storage efficiency. According to this method, the deduplication is detected using data chunks(blocks). Whenever the new file is uploaded to the cloud, the cloud server divides the file into many data chunks and stored them. When the user tries to upload a duplicate file, the server first checks whether the user owns the file or not. To find out the ownership, it randomly chooses‘d’ number of data chunks to the user. Upon receiving the data chunks, the user side application compares the data chunks. If the d data chunks are matched, it trusts the user allows him to upload the file. It finally allows updating of data chunks that are not exists in the cloud server. They have proved that their technique performs better than Proof of Retrievability (POR) and Proof of Data Possession (PDP) techniques in terms of data integrity for cloud storage services of Amazon AWS.

Bellare et al. proposed a method named DupLESS to avoid the brute-force attacks on Convergent Encryption based earlier methods [16]. In this method, cloud users encrypt the data using the keys which are generated based on the data using the Pseudo Random Function (PRF) protocol by the Key Server (KS). Since KS and Storage Service (SS) are separated, data owners first authenticate themselves to the KS without sharing the data. This method is more secure as KS is not accessible to attackers. Even though both KS and SS are attacked, DupLESS can still protect the stored data using Message Locked Encryption (MLE). They key limitation of this method is that it requires the data owners to authorize a third party to control their data.

Wu et al have proposed a deduplication scheme using the index name servers (INS)[11]. According to this technique, the file is divided into many smaller chunks. It then finds a unique 128-bit hash code of each chunk using MD5 algorithm. The key is called as signature/fingerprint of the chunk. The signature is an unique identity of a data chunk. The signatures of all data chunks in the cloud storage is maintained in the INS. Whenever a file is uploaded, the file is divided into data chunks and signature for each chunk is computed. The INSs will checks whether the data chunk of the same fingerprint presents in the storage. If the one is found, INS labels the data chunk as duplicate and does not store physically. However, it is not found, the INS system allows the uploading procedure and stores the data chunk. In addition to deduplication, it helps to reduce file storage, server load balancing, file compression and real-time feedback control. The key advantages of this approach, deduplication is achieved in micro level. However, this process is computationally very expensive

Kavade and Lomte proposed a new technique , called Dupkey , to find out duplication in both file and data chunk level[12]. This technique uses a concept called Convergent Encryption for both encryption and decryption. Whenever a user uploads a data, convergent key is computed using the cryptological hash price of the content of the information. The key is stored across multiple servers using Load Equalization Algorithm. So when copy of the same data is uploaded, the same convergent key will be generated by the system. The key matching helps to identify the duplicate files which are subsequently eliminated by the system. Such a system is very useful in online cloud storage service providing applications such as Dropbox. The main advantage of this method is that it assures secure deduplication. Moreover, it maintains and manages keys on the servers. It can be implemented at block level to improve security.

Zheng Yan et al proposed a method by which multiple clouds are checked to detect the duplications[18]. The method has used Authorized Party to generate a key for the encrypted data which is subsequently used to search multiple clouds for duplications. If the duplicate is found, it will not allow users to upload the file. It generates the key by considering the attributes of the users unlike the earlier methods. It significantly improves both the security and privacy of cloud data.

Guo and Jiang have proposed a technique in which Cloud Service Provider(CSP) detects the duplication. It generates the tag for a file using Elliptic Curve Cryptography (ECC). When the user uploads the data, it generates the tag and checks whether it has the same key of the encrypted file. If it does not exists, it allows the users to upload the new file. If such a file exists, it directs the user requests to the Authorized Party (AP) which is then tests the ownership of the file. If the request comes from the same user, it asks the user to upload the revised file. Otherwise; it cancels the storage the existing file of the user.

Conclusion

Data deduplication is essential in cloud computing especially for managing and controlling storage. It is important to save the cloud storage space and simplify the access. This paper presents the different techniques to detect duplicate files in cloud. Moreover, the deduplication techniques are coupled with data secutiry and used many encryption algorithms for key generations. It is found that Proof of Ownership is now predominantly used to ensure data integrity and deduplication methods. However, the high computational and communicational cost of the methods remains the big challenge in the field of deduplications.

References:

[1] R. Chow, et al., “Controlling data in the cloud: outsourcing computation without outsourcing control,” in Proc. ACM Workshop Cloud Computing. Security., 2009, pp. 85–90.

[2] S. Kamara, and K. Lauter, “Cryptographic cloud storage,” Finance. Crypto. Data Security., 2010, pp. 136–149.

[3] Q. Liu, C. C. Tan, J. Wu, and G. Wang, “Efficient information retrieval for ranked queries in cost-effective cloud environments,” in Proc. IEEE INFOCOM, 2012, pp. 2581–2585.

[4] M. Kallahalla, E. Riedel, R. Swaminathan, Q. Wang, and K. Fu, “Plutus: scalable secure file sharing on untrusted storage,” in Proc. USENIX Conf. File Storage Technol., 2003, pp. 29–42.

[5] E.-J. Goh, H. Shacham, N. Modadugu, and D. Boneh, “SiRiUS: securing remote untrusted storage,” in Proc. Netw. Distrib. Syst. Security. Symp., 2003, pp. 131–145.

[6] J. Bethencourt, A. Sahai, and B. Waters, “Ciphertext-policy attribute- based encryption,” in Proc. IEEE Symp. Secuity. Privacy, 2007, pp. 321–334.

[7] V. Goyal, O. Pandey, A. Sahai, and B. Waters, “Attribute-based encryption for fine-grained access control of encrypted data,” in Proc. 13th ACM Comput. Communication. Security., 2006, pp. 89–98.

[8] S. Muller, S. Katzenbeisser, and C. Eckert, “Distributed attributebased encryption,” in Proc. 11th Annu. Int. Conf. Inf. Security. Crypto., 2008, pp. 20–36.

[9] A. Sahai, and B.Waters, “Fuzzy identity-based encryption,” in Proc. 24th Int. Conf. Theory App. Cryptographic Tech., 2005, pp. 457–473.

[10] S. C. Yu, C. Wang, K. Ren, and W. J. Lou, “Achieving secure, scalable, and fine-grained data access control in cloud computing,” in Proc. IEEE INFOCOM, 2010, pp. 534–542.

[11] T. Wu, J. Pan and C. Lin, "Improving Accessing Efficiency of Cloud Storage Using De-Duplication and Feedback Schemes," in *IEEE Systems Journal*, vol. 8, no. 1, pp. 208-218, March 2014, doi: 10.1109/JSYST.2013.2256715.

[12] Madhuri Kavade, A.C. Lomte,” Secure De-Duplication using Convergent Keys (Convergent Cryptography) for Cloud Storage”, International Journal of Computer Applications (0975 – 8887) , pp.5-9, Volume 126 – No.10, September 2015

[13] J. Yuan and S. Yu, "Secure and constant cost public cloud storage auditing with deduplication," *2013 IEEE Conference on Communications and Network Security (CNS)*, National Harbor, MD, 2013, pp. 145-153, doi: 10.1109/CNS.2013.6682702

[14] Yongan Guo, Chunlei Jiang, "High Efficient Secure Data Deduplication Method for Cloud Computing," Journal of Internet Technology, vol. 21, no. 2 , pp. 557-564, Mar. 2020.

[15] Meixia Miao, Jianfeng Wang, Hui Li, Xiaofeng Chen,”Secure multi-server-aided data deduplication in cloud computing,”, Pervasive and Mobile Computing,Vol. 24, Pages 129-137,2015

[16] M. Bellare, S. Keelveedhi, and T. Ristenpart, “DupLESS: server aided encryption for deduplicated storage,” in Proc. 22nd USENIX Conf. Secur., 2013, pp. 179–194

[17] J. Liu, N. Asokan, and B. Pinkas. “Secure deduplication of encrypted data without additional independent servers,” in Proc. 22nd ACM SIGSAC Conf. Comput. Commun. Secur., 2015, pp. 874– 885.

[18] Z. Yan, L. Zhang, W. DING and Q. Zheng, "Heterogeneous Data Storage Management with Deduplication in Cloud Computing," in *IEEE Transactions on Big Data*, vol. 5, no. 3, pp. 393-407, 1 Sept. 2019, doi: 10.1109/TBDATA.2017.2701352.