Review:Analysis of Various Data De-Duplication Algorithms in Cloud Storage Optimization

Madhu Bala[1], Mr. Navtej Singh Ghumman[2]
M.Tech (Scholar), Assistant Professor
Department of Computer Science Engineering

Abstract **-** Cloud computing offers a novel way of service establishment by re-arranging various possessions over the Internet. The most important and popular cloud service is data storage. In order to preserve the privacy of data holders, data are often stored in cloud in an encrypted form. However, encrypted data present new challenges for cloud data de-duplication, which develops crucial for big data storage and processing in cloud. Traditional de-duplication schemes cannot work on encrypted data. Existing solutions of encrypted data de-duplication suffer from security weakness. They cannot flexibly support data access control and revocation. Deduplication is a one such storing optimization technique that avoids storage duplicate copies of data. Currently, to ensure security, data stored in cloud as well as other large storage areas are in an encrypted format and one problem with that is, we cannot apply deduplication technique over such an encrypted data. De-duplication steadily over the encoded data in cloud appears to be a challenging task. Various methods that address this challenge are studied in this review paper.

Keywords – Cloud computing, cloud service, encryption techniques and de-duplication.

I. INTRODUCTION

Cloud computing is model of the distribution of the information services in which the resources are the retrieved from the web through some of the interfaces and applications, instead forming direct connections to the server. The fast expansion in information sources has mandatory for the users to make use of some of the storage systems for storing their secret data. Cloud storage systems provide the management of the ever increasing quantity of data by keeping in mind factors like reduce occupation storage space and the network bandwidth [1]. To make the scalable and consistent management of the data in the cloud computing, deduplication technique plays an important role. Data deduplication also helps to improve the results in efficiency term and searches are quicker. Data deduplication may happen as file level deduplication or as block level data deduplication. Instead of maintaining numerous duplicate copies of file or the data with alike content, deduplication senses and remove the redundant data by keeping original physical copy [2]. The data generation rates are increasing, it is a tedious task for cloud storage providers to provide efficient storage. Cloud storage providers uses different techniques to improve storage efficiency and one of leading technique employed by them is deduplication, which claims to be saving 90 to 95% of storage. Data Deduplication technique evolved as an simple storage optimization technique in secondary then widely adapted in primary storage as well as larger storage areas like cloud storage area. Now, data deduplication is widely used by various cloud storage providers like Drop box, Amazon S3 [3] , Google Drive ,etc. Data once deployed to cloud servers, its beyond the security premises of the data owner, thus most of them prefer to outsource their in an encrypted format. Data encryption by data owners eliminates cloud service providers chance of deduplication it since encryption and deduplication techniques have conflicting strategies, i.e., data encryption with a key converts data into an unidentifiable format called cipher text thus encrypting, even the same data, with different keys may result in different cipher texts , making deduplication less feasible .However, performing encryption is essential to make data secure, at the same time, performing deduplication is essential for achieving optimized storage. Therefore, deduplication and encryption need to work in hand to hand to ensure secure and optimized storage [4].

II. DE-DUPLICATION OVER ENCRYPTED DATA

In order to preserve data privacy against inside cloud server as well as outside adversaries, users may want their data encrypted. However, conventional encryption under different users' keys makes cross user deduplication impossible, since the cloud server would always see different cipher texts, even if the data are the same, regardless of whether the encryption algorithm is deterministic. Deduplication is basically a compression technique for removing redundant data. Figure no. 1 explains the deduplication process before storing data onto memory. Deduplication can be categorized as file level deduplication and block level deduplication based on granularity. File level deduplication takes into account the entire file, thus even small update or append makes the file different from previous version of it and thereby reducing deduplication ratio.Where as in case of block level deduplication data blocks are considered for deduplication. Deduplication can further categorized based on location of deduplication i.e., as client side deduplication and as source side deduplication. Performing deduplication at client side is ensures bandwidth saving since only hash value of file is sent to server, if duplicate is existing[5] .Deduplication is widely is used various applications like backup, metadata management, primary storage, etc. for storage optimization. De-duplication presents two mechanisms:

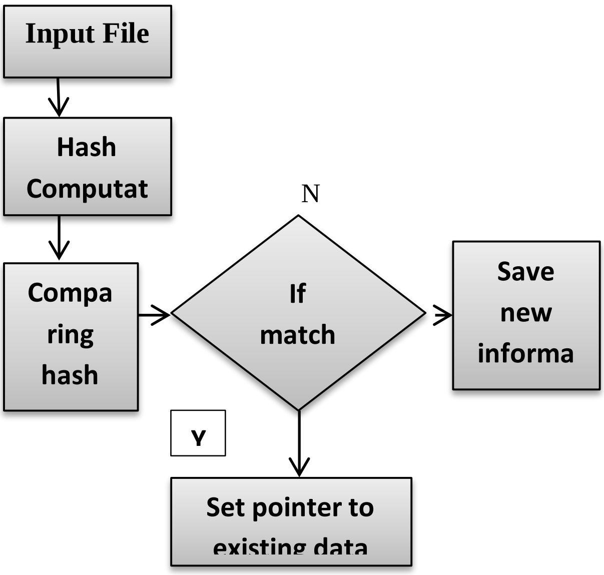

Fig.1: De-duplication Flow chart

A. ***Convergent Encryption:*** It enables duplicate files to coalesce into space of single files, even if the files are encrypted with different user's keys. It produces identical cipher text files from identical plaintext files irrespective of encryption keys. Convergent encryption enables identical encrypted files to be recognized as identical but there remains the problem of performing this identification across large no of machines in decentralized manner. This problem solved by storing location of file & content information in distributed data structure it is nothing but SALAD.

B. ***SALAD (Self Arranging Lossy Associative Database):*** It aggregate file content and location information in decentralized, scalable, fault tolerant manner. Collectively these components called as DFC (Duplicate File Coalescing) sub system of Farasite [6].

III. LITERATURE REVIEW

Youngjoo Shin et al., 2017 [7] described that the state-of-the-art secure deduplication techniques for each approach that deal with different security issues under specific or combined threat models, which include both

cryptographic and protocol solutions. It discusses and compares each scheme in terms of security and efficiency specific to different security goals. Finally, it identifies and discusses unresolved issues and further research challenges for secure deduplication in cloud storage**. R. Shobana et al., 2016 [8]** described that cloud architecture the intrusion detection and prevention was performed automatically by defining rules for the major attacks and alert the system automatically. The major attacks/events includes vulnerabilities, cross site scripting (XSS), SQL injection, cookie poisoning, wrapping. Data deduplication technique allows the cloud users to manage their cloud storage space effectively by avoiding storage of repeated data's and save bandwidth. The data were finally stored in cloud server namely Cloud Me. To ensure data confidentiality the data are stored in an encrypted type using Advanced Encryption Standard (AES) algorithm. **[9] Wen Xia et al., 2016** review the background and key features of data deduplication, then summarize and classify the state-of-the-art research in data deduplication according to the key workflow of the data deduplication process. The summary and taxonomy of the state of the art on deduplication help identify and understand the most important design considerations for data deduplication systems. In addition, it discusses the main applications and industry trend of data deduplication, and provides a list of the publicly available sources for deduplication research and studies.[10] **Junbeom Hur et al., 2016** discussed novel server-side deduplication scheme for encrypted data. It allows the cloud server to control access to outsourced data even when the ownership changes dynamically by exploiting randomized convergent encryption and secure ownership group key distribution. This prevents data leakage not only to revoked users even though they previously owned that data, but also to an honest-but-curious cloud storage server. In addition, the proposed scheme guarantees data integrity against any tag inconsistency attack.

IV. SCOPE OF DE-DUPLICATION

This survey presents an extensive overview of secure deduplication systems and explains their design decisions and main challenges. This article is focused on secure deduplication in cloud storage that is, we do not address general deduplication techniques without any security design or network deduplication, although some secure deduplication systems to which we refer do offer storage deduplication for plain data or network deduplication as a basic technique [11].

V. MESSAGE DEPENDENT ENCRYPTION

In cloud storage systems, data owners lose control of their data after outsourcing it to cloud storage. In practical scenarios, however, the CSP cannot be fully trusted (e.g., it can leak the clients' private data for monetary purposes or manipulate it to ease storage costs). Therefore, data owners may encrypt their data before outsourcing for the sake of data confidentiality. Unfortunately, conventional data encryption makes deduplication very difficult for the following two reasons:

(1) The CSP finds it difficult to identify whether the underlying plain data is the same or not, as data encryption under different keys selected by different users generates different cipher texts, and

(2) Even though the CSP can determine the identity of the underlying data, it is difficult to store two different cipher texts in a storage-efficient manner so that any data owner can retrieve the plain data after outsourcing it [12].

Message-dependent encryption is considered a promising cryptographic primitive for ensuring data privacy during deduplication. In message dependent encryption, a data owner computes an encryption key from the data about to be encrypted and encrypts the data under the key. Formally, it can be defined as the following four deterministic algorithms (key generation, encryption, decryption, and tag generation):

Key Generation Function : The key generation algorithm takes a data content F as input and outputs the symmetric key ckF of F. Basically, the key ckF is generated by rendering a cryptographic hash function h,

Ckf = h(F)

Encrypt : The encryption algorithm takes the key ckF and the data content F as input and outputs a ciphertext ctF encrypted under ckF.

Decrypt : The decryption algorithm takes the key ckF and the encrypted content ctF and outputs the restored plain data F.

TagGen(F): The tag generation algorithm takes the data content F as input and outputs a tag tagF of F [13].

All of the generated values from the preceding four algorithms are the same for the same F in generic message–dependent encryption [15]. Therefore, multiple data owners of the same data content are allowed to generate the same keys (and eventually the same cipher text) without additional communication for key agreement before outsourcing [14].

VI. CONCLUSION

Using Secure distributed deduplication technique for the IT Industries provides lot of benefits with the use of both public and private clouds and also provide storage benefits at lower costs. Deduplication is a method available in cloud storage for saving bandwidth and storage capacity. But, deduplication is less feasible with encrypted data since, different key encryptions convert same data into different formats. In this paper various methods are discussed where deduplication methods are carried out on encrypted data in a large storage area. Most of the methods studied here work on the basis of convergent encryption, which is a simple approach that makes deduplication compatible with encrypted data. In this information dense world, we cannot compromise on both security and duplication of data across storage areas. A strategy needs to be formulated which will enhance storage optimization without negotiating on encryption method; by providing deduplication technique in data storage servers where the available data is encrypted.

VII. REFERENCES

[1]. M. W. Storer, K. Greenan, D. D. E. Long, and E. L. Miller, “Secure data deduplication,” in Proc. of StorageSS, 2008.

[2]. P. Anderson and L. Zhang, “Fast and secure laptop backups with encrypted de- duplication,” in Proc. of USENIX LISA, 2010.

[3]. M. Bellare, S. Keelveedhi, and T. Ristenpart, “Dupless: Server aided encryption for deduplicated storage,” in USENIX Security Symposium, 2013.

[4]. Science, IEEE, 1997.Jin Li, Yan Kit Li, Xiaofeng Chen, Patrick P. C. Lee, Wenjing Lou, “A Hybrid Cloud Approach for Secure Authorized Deduplication”, IEEE Transactions on Parallel and Distributed Systems, Volume: PP, Issue:99, Date of Publication :18.April.2014

[5]. C. Yang, J. Ren, and J. F. Ma, “Provable ownership of file in deduplication cloud storage,” in Proc. IEEE Global Commun. Conf., 2013, pp. 695–700, doi:10.1109/GLOCOM.2013.6831153.

[6]. T. Y. Wu, J. S. Pan, and C. F. Lin, “Improving accessing efficiency of cloud storage using de-duplication and feedback schemes,” IEEE Syst. J., vol. 8, no. 1, pp. 208–218, Mar. 2014, doi:10.1109/ JSYST.2013.2256715.

[7]. Shin, Youngjoo, Dongyoung Koo, and Junbeom Hur. "A Survey of Secure Data Deduplication Schemes for Cloud Storage Systems." ACM Computing Surveys (CSUR) 49, no. 4 (2017): 74.

[8]. Shobana, R., K. Shantha Shalini, S. Leelavathy, and V. Sridevi. "De-Duplication of Data in Cloud." International Journal of Chemical Sciences 14, no. 4 (2016).

[9]. Xia, Wen, Hong Jiang, Dan Feng, Fred Douglis, Philip Shilane, Yu Hua, Min Fu, Yucheng Zhang, and Yukun Zhou. "A comprehensive study of the past, present, and future of data deduplication." Proceedings of the IEEE 104, no. 9 (2016): 1681-1710.

[10]. Hur, Junbeom, Dongyoung Koo, Youngjoo Shin, and Kyungtae Kang. "Secure data deduplication with dynamic ownership management in cloud storage." IEEE Transactions on Knowledge and Data Engineering 28, no. 11 (2016): 3113-3125.

[11]. Kalyani, Zodge, and Amruta Amune. "REVIEW ON SECURE DISTRIBUTED DEDUPLICATION SYSTEMS WITH IMPROVE RELIABILITY." Global Journal of Advanced Engineering Technologies Volume 5, Issue 1- 2016 ISSN (Online): 2277-6370 & ISSN (Print):2394-0921.

[12]. Akhila, K., Amal Ganesh, and C. Sunitha. "A Study on Deduplication Techniques over Encrypted Data." Procedia Computer Science 87 (2016): 38-43.

[13]. Yan, Zheng, Wenxiu Ding, Xixun Yu, Haiqi Zhu, and Robert H. Deng. "Deduplication on encrypted big data in cloud." IEEE transactions on big data 2, no. 2 (2016): 138-150.

[14]. Rashid, Fatema, Ali Miri, and Isaac Woungang. "A Secure Video Deduplication Scheme in Cloud Storage Environments Using H. 264 Compression." In Big Data Computing Service and Applications (BigDataService), 2015 IEEE First International Conference on, pp. 138-146. IEEE, 2015.

[15]. Kaaniche, Nesrine, and Maryline Laurent. "A secure client side deduplication scheme in cloud storage environments." In New Technologies, Mobility and Security (NTMS), 2014 6th International Conference on, pp. 1-7. IEEE, 2014.