Efficient Client-Side Deduplication of Encrypted Data With Public Auditing In Cloud Storage

Priyanka Sable, Nilima Sharmale, Rucha Kokate, Prof. S.N.Lohar
Sinhgad Institute of Technology Lonavala, Pune

Abstract: Cloud computing offers a brand new approach of service provision by re-arranging numerous resources over the web. The most important and widespread cloud service is information storage. So as to preserve the privacy of information holders, information square measure usually hold on in cloud in associate degree encrypted kind. However, encrypted information introduces new challenges for cloud information deduplication that becomes crucial for large information storage and process in cloud. Ancient deduplication schemes cannot work on encrypted information. Existing solutions of encrypted data deduplication suffer from security weakness. They can't flexibly support information access management and revocation. Therefore, few of them may be without delay deployed in follow. During this paper, we tend to propose a theme to deduplicate encrypted information hold on in cloud supported ownership challenge and proxy re-encryption. It integrates cloud information deduplication with access management. We tend to valuate its performance based on in depth analysis and pc simulations. The results show the superior potency and effectiveness of the theme for potential sensible readying, particularly for large information deduplication in cloud storage.

Keywords: **Access control, big data, cloud computing, data deduplication, proxy re-encryption**

I. INTRODUCTION

Cloud computing offers a new way of Information Technology services by rearranging various resources (e.g., storage, computing) and providing them to users based on their demands. The most important and popular cloud service is data. Storage service. Cloud users upload personal or confidential data to the data centre of a Cloud Service Provider (CSP) and allow it to maintain these data. Since intrusions and attacks towards sensitive data at CSP are not avoidable. It is prudent to assume that CSP cannot be fully trusted by cloud users. Due to the rapid development of data mining and other analysis technologies, the privacy issue becomes serious. Hence, a good practice is to only outsource encrypted data to the cloud in order to ensure data security and user privacy. But the same or different users may upload duplicated data in encrypted form to CSP, especially for scenarios where data are shared among many users. Although cloud storage space is huge, data duplication greatly wastes network resources, consumes a lot of energy, and complicates data management. The development of numerous services further makes it urgent to deploy efficient resource management mechanisms. Consequently, deduplication becomes critical for big data storage and processing in the cloud. Deduplication has proved to achieve high cost savings, e.g., reducing up to 90-95 percent storage needs for backup applications and up to 68 percent in standard file systems. Obviously, the savings, which can be passed back directly or indirectly to cloud users, are significant to the economics of cloud business. How to manage encrypted data storage with deduplication in an efficient way is a practical issue. However, current industrial deduplication solutions cannot handle encrypted data. Existing solutions for deduplication suffer from brute-force attacks. Deduplication has proved to achieve high cost savings, e.g., reducing up to 90-95 percent storage needs for backup applications and up to 68 percent in standard file systems. Obviously, the savings, which can be passed back directly or indirectly to cloud users, are significant to the economics of cloud business. How to manage encrypted data storage with deduplication in an efficient way is a practical issue.

However, current industrial deduplication solutions cannot handle encrypted data. Existing solutions for deduplication suffer from brute-force attacks .They cannot flexibly support data access control and revocation at the same time. Most existing solutions cannot ensure reliability, security and privacy with sound performance. In this paper, we propose a scheme based on data ownership challenge and Proxy Re-Encryption (PRE) to manage encrypted data storage with deduplication. We aim to solve the issue of deduplication in the situation where the data holder is not available or difficult to get involved. Meanwhile, the performance of data deduplication in our scheme. Is not influenced by the size of data, thus applicable for big data.

II. History and Background

Encrypted Data Deduplication Cloud storage service providers such as Dropbox, Google Drive, Mozy, and others perform deduplication to save space by only storing one copy of each file uploaded. However, if clients conventionally encrypt their data, storage savings by deduplication are totally lost. This is because the encrypted data are saved as different contents by applying different encryption keys. Existing industrial solutions fail in encrypted data deduplication. For example, DeDu is an efficient deduplication system, but it cannot handle encrypted data. Reconciling deduplication and client-side encryption is an active research topic. Message-Locked Encryption (MLE) intends to solve this problem. The most prominent manifestation of MLE is Convergent Encryption (CE), introduced by Douceur and others. CE was used within a wide variety of commercial and research storage service systems. Letting M be a file's data, a client first computes a key K HðMÞ by applying a cryptographic hash function H to M, and then computes ciphertext C EðK;MÞ via a deterministic symmetric encryption scheme. A second client B encrypting the same file M will produce the same C, enabling deduplication. However, CE is subject to an inherent security limitation, namely, susceptibility to offline brute-force dictionary attacks. Knowing that the target data M underlying the target ciphertext C is drawn from a dictionary S ¼ fM1; . . .;Mng of size n, an attacker can recover M in the time for n ¼ jSj off-line encryptions: for each i ¼ 1; . . . ; n, it simply CEencrypts Mi to get a ciphertext denoted as Ci and returns Mi such that C ¼ Ci. This works because CE is deterministic and keyless. The security of CE is only possible when the target data is drawn from a space too large to exhaust.

Another problem of CE is that it is not flexible to support data access control by data holders, especially for data revocation process, since it is impossible for data holders to generate the same new key for data re-encryption. An image deduplication scheme adopts two servers to achieve verifiability of deduplication. The CE-based scheme described in combines file content and user privilege to obtain a file token with token unforgeability. However, both schemes directly encrypt data with a CE key, thus suffer from the problem as described above. To resist the attack of manipulation of data identifier, Meye et al. proposed to adopt two servers for intra-user deduplication and inter deduplication. The ciphertext C of CE is further encrypted with a user key and transferred to the servers. However, it does not deal with data sharing after deduplication among different users. Cloud Dedup also aims to cope with the inherent security exposures of CE, but it cannot solve the issue caused by data deletion. A data holder that removes the data from the cloud can still access the same data since it still knows the data encryption key if the data is not completely removed from the cloud.

III. Design Issues

Math or Equation

Mathematical Model:

S={ I,O,P,F,s,Ic)

Identify set of input as I

Let I ={Set of outsourced data sets by corresponding data user}

3. Identify set of output as O

Let O={store unique file on cloud server .}

4) Identify the set of processes as P

PRE= proxy re-encryptionv.

TPA=Third Party Auditor.

Uo=set of owners.

SE=Symmetric Encryption

CSP=Cloud Service Provider

Sk=Symmetric Key

Op= Output of System

5. Identify failure cases as F

F=store duplicate file on cloud server and unable to find file ownership.}

6. Identify success as s.

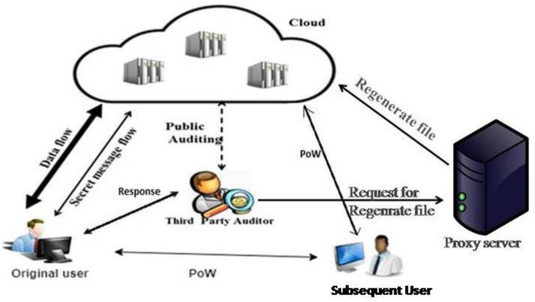
s={check duplicate file that is already store on cloud server If file already exist then duplicate file is not stored on cloud only give reference to new file.}

7. Identify the initial condition as Ic

Ic={ Outsourced data with its privacy privileges to be maintain)

Figure and Table

Architecture Diagram:



IV. Literature Survey

Paper 1. A Verifiable Data Deduplication Scheme in Cloud Computing

Author Name: Z. C. Wen, J. M. Luo, H. J. Chen, J. X. Meng, X. Li, and J. Li

Description: Deduplication is an important technique to save the storage cost at the cloud storage server. Image is an important data type stored in cloud, but rarely discussed in previous work on deduplication. This paper studies the problem of validating the deduplication of image storage in cloud. In particular, we consider the task of allowing a cloud server to verify the correctness of deduplication. Our scheme consists of several advantages over the previous work, whose framework can be described through the following algorithms. Firstly, before each user uploads an encrypted image, he calculates its hash value as the fingerprint. Secondly, the fingerprint is sent to both cloud servers for checking duplicates. If the storage and verification servers both reply to the user with 'no deduplication', the user transfers his data to the servers. Otherwise, once the fingerprint is consistently found, the user gives up uploading data for deduplication. Specially, when the fingerprint is only found in one server, it implies that the results are inconsistent and at least one of servers is invalid. The security and efficiency analysis is also presented in this paper.

Paper 2. A hybrid cloud approach for secure authorized deduplication

Author Name: J. Li, Y. K. Li, X. F. Chen, P. P. C. Lee, and W. J. Lou

Description: Data deduplication is one of important data compression techniques for eliminating duplicate copies of repeating data, and has been widely used in cloud storage to reduce the amount of storage space and save bandwidth. To protect the confidentiality of sensitive data while supporting deduplication, the convergent encryption technique has been proposed to encrypt the data before outsourcing. To better protect data security, this paper makes the first attempt to formally address the problem of authorized data deduplication. Different from traditional deduplication systems, the differential privileges of users are further considered in duplicate check besides the data itself. We also present several new deduplication constructions supporting authorized duplicate check in a hybrid cloud architecture. Security analysis demonstrates that our scheme is secure in terms of the definitions specified in the proposed security model. As a proof of concept, we implement a prototype of our proposed authorized duplicate check scheme and conduct test bed experiments using our prototype. We show that our proposed authorized duplicate check scheme incurs minimal overhead compared to normal operations.

Paper 3. Reducing impact of data fragmentation caused by in-line deduplication

Author Name: M. Kaczmarczyk, M. Barczynski, W. Kilian, and C. Dubnicki

Description: Deduplication results inevitably in data fragmentation, because logically continuous data is scattered across many disk locations. In this work we focus on fragmentation caused by duplicates from previous backups of the same backup set, since such duplicates are very common due to repeated full backups containing a lot of unchanged data. For systems with in-line dedup which detects duplicates during writing and avoids storing them, such fragmentation causes data from the latest backup being scattered across older backups. As a result, the time of restore from the latest backup can be significantly increased, sometimes more than doubled.

We propose an algorithm called context-based rewriting (CBR in short) minimizing this drop in restore performance for latest backups by shifting fragmentation to older backups, which are rarely used for restore. By selectively rewriting a small percentage of duplicates during backup, we can reduce the drop in restore bandwidth from 12-55% to only 4-7%, as shown by experiments driven by a set of backup traces. All of this is achieved with only small increase in writing time, between 1% and 5%. Since we rewrite only few duplicates and old copies of rewritten data are removed in the background, the whole process introduces small and temporary space overhead.

Paper 4. DeyPoS: Deduplicatable Dynamic Proof of Storage for Multi-User Environments

Author Name: Kun He, Jing Chen, Ruiying Du, Qianhong Wu, Guoliang Xue, and Xiang Zhang

Description: Dynamic Proof of Storage (PoS) is a useful cryptographic primitive that enables a user to check the integrity of outsourced files and to efficiently update the files in a cloud server. Although researchers have proposed many dynamic PoS schemes in single user environments, the problem in multi-user environments has not been investigated sufficiently. A practical multi-user cloud storage system needs the secure client-side cross-user deduplication technique, which allows a user to skip the uploading process and obtain the ownership of the files immediately, when other owners of the same files have uploaded them to the cloud server. To the best of our knowledge, none of the existing dynamic PoSs can support this technique. In this paper, we introduce the concept of deduplicatable dynamic proof of storage and propose an efficient construction called DeyPoS, to achieve dynamic PoS and secure cross-user deduplication, simultaneously. Considering the challenges of structure diversity and private tag generation, we exploit a novel tool called Homomorphic Authenticated Tree (HAT). We prove the security of our construction, and the theoretical analysis and experimental results show that our construction is efficient in practice.

Paper 5. Provable ownership of files in deduplication cloud storage
Author Name: Chao Yang1,2, Jian Ren2* and Jianfeng Ma1
Description: With the rapid adoption of cloud storage services, a great deal of data is being stored at remote servers, so a new technology, client-side deduplication, which stores only a single copy of repeating data, is proposed to identify the client's deduplication and save the bandwidth of uploading copies of existing files to the server. It was recently found, however, that this promising technology is vulnerable to a new kind of attack in which by learning just a small piece of information about the file, namely its hash value, an attacker is able to obtain the entire file from the server. In this paper, to solve this problem, we propose a cryptographically secure and efficient scheme for a client to prove to the server his ownership on the basis of actual possession of the entire original file instead of only partial information about it. Our scheme utilizes the technique of spot checking in which the client only needs to access small portions of the original file, dynamic coefficients and randomly chosen indices of the original files. Our extensive security analysis shows that the proposed scheme can generate provable ownership of the file and maintain high detection probability of client misbehavior. Both performance analysis and simulation results demonstrate that our proposed scheme is much more efficient than the existing schemes, especially in reducing the burden of the client.

V. Conclusion

Interoperability between hospitals not only help improve patient safety and quality of care but also reduce time and resources spend on data format conversion. Interoperability is treated more important as the number of hospitals participating in HIE increases .if one hospital does not support interoperability, the other hospitals are required to convert data format of their clinical information to exchange data for HIE. When the number of hospitals that do not support interoperability, complexity for HIE inevitably increase in proportion. The advantage of API service as ours are at the amount of resources that hospitals need to allocate for interoperability is only minimal. Therefore, offering system that supports interoperability by relying on a cloud computing platform may be good and we provide the QR code security for patient's data that stored on cloud.

VI. References

[1] Z. C. Wen, J. M. Luo, H. J. Chen, J. X. Meng, X. Li, and J. Li, "A verifiable data deduplication scheme in cloud computing," in Proc. Int. Conf. Intell. Netw. Collaborative Syst., 2014, pp. 85–90, doi:10.1109/INCoS.2014.111.

[2] J. Li, Y. K. Li, X. F. Chen, P. P. C. Lee, and W. J. Lou, "A hybrid cloud approach for secure authorized deduplication," IEEE Trans. Parallel Distrib. Syst., vol. 26, no. 5, pp. 1206–1216, May 2015, doi:10.1109/TPDS.2014.2318320.

[3] P. Meye, P. Raipin, F. Tronel, and E. Anceaume, "A secure twophase data deduplication scheme," in Proc. HPCC/CSS/ICESS, 2014, pp. 802–809, doi:10.1109/HPCC.2014.134.

[4] J. Paulo and J. Pereira, "A survey and classification of storage deduplication systems," ACM Comput. Surveys, vol. 47, no. 1, pp. 1–30, 2014, doi:10.1109/HPCC.2014.134.

[5] Y.-K. Li, M. Xu, C.-H. Ng, and P. P. C. Lee, "Efficient hybrid inline and out-of-line deduplication for backup storage," ACM Trans. Storage, vol. 11, no. 1, pp. 2:1-2:21, 2014, doi:10.1145/2641572.

[6] M. Fu, et al., "Accelerating restore and garbage collection in deduplication-based backup systems via exploiting historical information," in Proc. USENIX Annu. Tech. Conf., 2014, pp. 181–192.

[7] M. Kaczmarczyk, M. Barczynski, W. Kilian, and C. Dubnicki, "Reducing impact of data fragmentation caused by in-line deduplication," in Proc. 5th Annu. Int. Syst. Storage Conf., 2012, pp. 15:1–15:12, doi:10.1145/2367589.2367600.

[8] M. Lillibridge, K. Eshghi, and D. Bhagwat, "Improving restore speed for backup systems that use inline chunk-based deduplication," in Proc.USENIX Conf. File Storage Technol., 2013, pp. 183–198.

[9] L. J. Gao, "Game theoretic analysis on acceptance of a cloud data access control scheme based on reputation," M.S. thesis, Xidian University, State Key Lab of ISN, School of Telecommunications Engineering, Xi'an, China, 2015.

[10] Z. Yan, X. Y. Li, M. J. Wang, and A. V. Vasilakos, "Flexible data access control based on trust and reputation in cloud computing," IEEE Trans. Cloud Comput., vol. PP, no. 99, Aug. 2015, doi:10.1109/ TCC.2015.2469662, Art. no. 1..