Deduplication on Encrypted Big Data in Cloud

Nehal Pandey[1] Nishant Jain[2] Nikshay Jain[3] Ishita Mattoo[4] Prof. Neha Hajar[5]
[1,2,3,4,5]School of Computer Engineering MIT Academy of Engineering Alandi, Pune-412105, India

Abstract— Cloud computing offers a brand new approach of service provision by re-arranging numerous resources over the web. The most major and widespread cloud service is information storage. So as to maintain the privacy of information holders, information square measure usually hold on cloud in associated degree encrypted kind. However, encrypted information introduces provocation for cloud information duplication that becomes crucial for large information storage and process in cloud. Ancient replicating schemes cannot work on encrypted information[1].They cannot exile support in- formation access management and abrogation. That is why, not of them may be without delay deployed in follow. During this paper, we gravitate to put forward a theme to deduplicate encrypted information hold on in cloud supported freehold challenge and proxy re-encryption[6]. 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[2].

Keywords: Access Control, Big Data, Cloud Computing, Data Deduplication, Proxy Re-Encryption

I. INTRODUCTION

Cloud computing provide a new method of Information Technology service by rearranging various resources (e.g. storage, computing) providing them to users based on their demands. The critical and desired cloud service is data. Storage service[3]. Cloud users upload their personal or confidential data to the data center of a Cloud Service Provider (CSP) and allows it to maintain these type of data. Since intrusions and attacks to-wards sensitive data at CSP are not avoidable. It is sagacious to take for granted that CSP cannot be trustworthy by cloud users[4]. Due to the fast growth of data processing and other inspection technologies, the privacy matter becomes serious. Hence, the good practice is only to outsource the encrypted data to the cloud in order to protect data security and user privacy. But similar or dissimilar users may upload duplicated data in the encrypted form to CSP, mostly in scenarios where data are shared among many users[5]. Although storage of cloud space is vast, data duplication considerably misspend network resources, absorb a large amount of energy, and complexes data management. The development of numerous services further makes it extreme to deploy systematic resource management mechanisms. Respectively, duplication becomes censorious for big data storage and processing in the cloud[6]. Reduplication has proved to attain high cost savings, e.g. reducing upto 90-95 percent of the storage needed for backup applications and up to 68 percent in standard in systems. Obviously, the savings, which can be passed back straightly or incidentally to cloud users, are significant to the economics of cloud business[3]. How to manage the encrypted data storage with the duplication in an efficient way is a workable issue. However, belonging to the present time industrial duplication solution cant handle the encrypted data. Present solution for duplication endure from brute-force attacks. Reduplication has proved to attain high cost savings, e.g. reducing up to 90-95 percent storage needed for backup applications and up to 68 percent in the standard systems[7]. Evidently, the savings, which can be returned back straightly or incidentally to the cloud users, are remarkable to the economics of cloud business. How to manage the encrypted data storage with deduplication in a well-ordered way is an actual issue. However, present industrial duplication solutions can't handle the encrypted data. Present solutions for duplication endure from brute-force attacks[8].They cannot adjust data access control and take back at the same time. Present solutions cannot reliability, security and privacy with sound performance. In this paper, we had presented a scheme build on data freehold, stand against and Proxy Re-Encryption (PRE) to control encrypted data storage with duplication. We have intention to resolve the issue of duplication in the situation where the data holder is unavailable or difficult to get involved. Meanwhile, the performance of data duplication in our scheme. Is not developed by the size of data, thus applicable for big data[9].

II. HISTORY AND BACKGROUND

Encrypted Data Reduplication Cloud storage service provider such as Mozy, Dropbox, Google Drive and others perform duplication to consume less space by keeping a single copy of each uploaded. However, if the clients regularly encrypt their data, storage savings by duplication are completely lost[2]. It is because the encrypted data are stored as different constituents by applying various encryption keys. Existing industrial solutions fail in encrypted data duplication. For example, DeDu is an efficient duplication system, but it cannot manage encrypted data. Restore duplication and client-side encryption is a bustling research topic. Message-Locked Encryption (MLE) intends to solve this problem. The most important of showing MLE is Convergent Encryption (CE), found by Douceur and other[4]s. CE was used within a wide range of commercial and research storage service systems. Let M be a less data, a client first computes a key K HM by using a cryptographic hash function H to M, and then computes cipher text C EK;M via an oppressive symmetric encryption schemes. A second client B will encrypt the same M and it will produce the same C, which enables duplication. However, CE is subject to a fundamental security drawback, namely, susceptibility to down brute-force dictionary attacks. Knowing that the target data M underlying the target cipher text C is drawn from a dictionary S fM1; . . .; Mng of size n, an attacker can recapture M in the time for n jSj off-line encryptions: for each i 1; . . . ; n, it simply Encrypts Mi to receive a cipher text indicated as Ci and returns Mi such that C Ci. This works because CE is oppresive and keyless[5]. The security of CE can only be done when the target data is drawn from a huge space to exhaust. Other problem of CE is that it is not flexible to bear data access control by data holders, mainly for data revocation process, since it is impractical for

All rights reserved by *www.ijsrd.com*

data holders to generate the same new key for data re-encryption. An image duplication scheme adopts two servers to achieve variability of duplication[6]. The cipher text C of CE is encrypted with the help of a user key and then send to the servers. It does not deal with the data sharing after duplication is done among different users. Cloud duplication also focuses to cope up with the internal security exposures of CE, but it cannot solve the problem caused due to data deletion. A data holder who takes the data from the cloud can still access the same data because the data holder knows the data encryption key if the data is not deleted from the cloud completely.

III. MATHEMATICAL MODEL

A. Mathematical Model:

1) S= I,O,P,F,s,Ic)
2) 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-encryption v.
AP=Authorized Party. 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.

Identify the initial condition as Ic Ic= Outsourced data with its privacy privileges to be maintain)

IV. LITERATURE SURVEY

A Veritable Data Reduplication Scheme in Cloud Computing Author Name: Z. C. Wen, J. M. Luo, H. J. Chen, J. X. Meng, X. Li, and J. Li

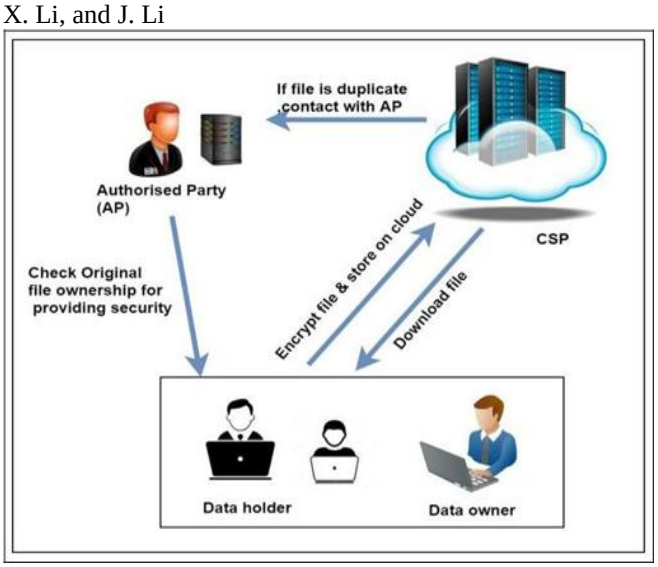

Fig. 1: Architecture Diagram:

Description: A key technique which is used to conserve the storage cost at the cloud storage server is called reduplication. Image is one of the crucial data type which is stored in cloud, but is not frequently discussed in previous works done on reduplication. This paper research on the issue and also recognizes the reduplication of image storage which is in the cloud. In this we consider a task in which we allow a cloud server to check the accuracy of reduplication. Our scheme consists of various advantages over the previous work done, whose framework can be described with the help of following algorithms. At First, before each of the user uploads an encrypted image, he will calculate its hash value which act as a fingerprint. Then, the fingerprint will be sent to both the cloud servers for scrutinizing the duplicates. If the Storage and verification servers both will respond to the user that there is no duplication, then the user can easily transfer his data to the servers. Else, if the fingerprint is consistently found, then the user will give up uploading the data for duplication. Especially, when the fingerprint is only matched with one server, it indicates that the results are conflicting and at least one of server is not valid. The analysis of security and efficiency is also presented in this paper.

A hybrid cloud undertakes to secure the authorized duplication Author Name: J. Li, Y. K. Li, X. F. Chen, P. P. C. Lee, and W. J. Lou

Description: Data duplication is one of the main data compression techniques which is used for removing duplicate copies of same data, and it is mostly used in cloud storage in order to reduce the storage space in cloud and also save bandwidth. In order to protect the confidentiality of susceptive data while using duplication, an encryption technique is put forward to encrypt the data before outsourcing it. So to protect the data security, this paper makes the first attempt to solve the problem of authorized data duplication. As it is different from traditional duplication systems, the differential benefits of users are also considered in duplicate check apart from the data itself. We also present various new duplication constructions which support validated duplicate check in a hybrid cloud architecture. Security analysis is used for demonstrating that our scheme is protected in terms of the definitions which are specified in the suggested security model.

As a proof of our concept, we implemented a prototype of our suggested and authorized duplicate check and conducted a test of experiments using our prototype. We showed that our suggested authorized duplicate check scheme incurs minimum overhead as compared to normal operations.

Diminishing impact of data fragmentation produced by in-line duplication Author Name: C. Dubnicki, W. Kilian, M. Barczynski, and M. Kaczmarczyk

Description: Deduplication results in data fragmentation, because logically there is a continuous data which is spread over many disk locations. This work is focused on fragmentation which is created by duplicates from previous backups of the same backup set, also these duplicates are very regular due to frequent full backups which contains a lot of unchanged data. For systems which have in-line dedup, they detect duplicates during writing and also avoids saving them, such fragmentation causes data from the latest backup being spread across the older back- ups. As a result of this, the duration of restore from the new backup can

All rights reserved by *www.ijsrd.com*

be signicantly increased, and sometimes can be higher than doubled. We suggest an algorithm which is known as context-based rewriting (CBR) to minimize this drop in restore performance for new backups by shifting these fragments to earlier backups, which are barely used for restoring. By selecting and rewriting a few per-centage of duplicates during backup, we can decrease the drop in restore bandwidth from 12-55 percent to only 4-7 percent, as shown by inspection driven by a set of backup traces. All this achieved only with small increase in writing time, between 1 and 5 percent. Since we rewrite only rear duplicates and previous copies of rewritten data are eliminated from the background, the whole process introduces little and temporary space overhead.

DeyPoS: Deduplicatable Dynamic Proof of Storage for Multi-User Environments Author Name: Xiang Zhang, Ruiying Du, Jing Chen, Guoliang Xue, Qianhong Wu, and Kun He Description: Dynamic Proof of Storage (PoS) is a useful cryptographic primitive that give permission to user to check the integrity of outsourced document and also to update the data efficiently into a cloud server. Although there are many researchers who have suggested many dynamic PoS schemes in a single user environments, the problem in a multi-user environments is not been examined properly. A practical multi-user cloud storage system is the one which needs a secure client-side cross-user deduplication technique, which give license to a user to skip the uploading process and gain the ownership, when the other owners have uploaded the same data into the cloud server. As per foremost of our knowledge, none of the existing dynamic PoSs can bear this technique. In this paper, idea of deduplicatable dynamic proof of storage is proposed and also proposed an efficient construction called as DeyPoS, to achieve the dynamic PoS and also secure cross-user deduplication, concurrently. By taking the challenges of structure diversity and private tag generation, we bulid a novel tool called Homomorphic Authenticated Tree (HAT). It helped us to verify the security of our creation, and also the theoretical analysis and experimental result showed that our creation is efficient.

Provable ownership of files in deduplication cloud storage Author Name: Chao Yang1,2, Jian Ren2* and Jianfeng Ma1 Description: With the rapid usage of cloud storage services, a large amount of data is being saved at remote servers, so a latest technology, client-side deduplication, which is used for storing only one copy of repeating data, is preferred which is used to identify the clients deduplication and also helps to store the bandwidth of uploading copies of already current files to the server. It was recently found, that this promising technology is vulnerable to some new kind of attack in which by learning just a small piece of information in the file that is its hash value, an attacker is able to acquire the entire file from the server. In this paper, in order to resolve this problem, we prefer a cryptographically secure and efficient scheme for a client to validate to the server his ownership on the basis of actual possession of the entire original file instead of only partial information about it. The scheme that we are using 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. This huge security analysis shows that the suggested scheme can help to produce provable ownership of the file and it also maintains big detection chances of client misbehavior. Both performance analysis and simulation result shows that our suggested scheme is much more efficient than the existing schemes, especially it helps to reduce the burden of the client.

V. Conclusion

Interoperability between hospitals does not only help in improving patient security and quality of care but it also reduces time and resources that are spend on data format conversion. Interoperability is most important that the number of hospitals which are taking part in HIE increases if single hospital does not help interoperability, the remaining hospitals are needed to convert data format of their clinical knowledge to exchange data for HIE. When the number of hospitals that does not help interoperability, complication for HIE is increased in proportion. The advantage of API service as our at the amount of resources that hospitals require to allocate for interoperability is only minimum. Therefore, offering system that helps interoperability by relying on a cloud computing platform may be good and we provide the QR code security for patients data that stored on cloud.

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. 8590, 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. 12061216, 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. 802809, 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. 130, 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 i deduplication-based backup systems via exploiting historical information, in Proc. USENIX Annu. Tech. Conf., 2014, pp. 181192.

[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. 15:115:12, 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. 183198.

[9] L. J. Gao, Game theoretic analysis on acceptance of a cloud data access control scheme based on reputation, M.S. thesis, Xidian Univer-sity, State Key Lab of ISN,

All rights reserved by *www.ijsrd.com*

School of Telecommunications Engineering, Xian, 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..
[11] P. Meye, P. Raipin, F. Tronel, and E. Anceaume, A secure twophase data deduplication scheme, in Proc. HPCC/CSS/ICESS, 2014, pp. 802809, doi:10.1109/HPCC.2014.134.
[12] J. Paulo and J. Pereira, A survey and classification of storage deduplication systems, ACM Comput. Surveys, vol. 47, no. 1, pp. 130, 2014, doi:10.1109/HPCC.2014.134.
[13] 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.
[14] 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. 181192.
[15] 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. 15:115:12, doi:10.1145/2367589.2367600.
[16] 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. 183198.
[17] L. J. Gao, Game theoretic analysis on acceptance of a cloud data access control scheme based on reputation, M.S. thesis, Xidian Univer-sity, State Key Lab of ISN, School of Telecommunications Engineering, Xian, China, 2015.
[18] 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.

All rights reserved by *www.ijsrd.com*