Secure Distributed De-duplication Systems with Block-Level Message-Locked Encryption

Priyanka Misal[1], Bhagyashree Kshirsagar[2], Monica Tiple[3], Jayashree Bhosale[4]
Department of Computer Engineering
[1, 2, 3, 4]Anantrao Pawar College Of Engineering and Research,Parvati Pune ,savitribai Phule Pune University

Abstract*-Data de-duplication is a technique for eliminating duplicate copies of data, and has been widely used in cloud storage to reduce storage space and upload bandwidth. However, there is only one copy for each file stored in cloud even if such a file is owned by a huge number of users. As a result, de-duplication system improves storage utilization while reducing reliability Furthermore , the challenge of privacy for sensitive data also arises when they are outsourced by users to cloud. Aiming to address the above security challenges, this paper makes the first attempt to formalize the notion of distributed reliable de-duplication system. We propose new distributed de-duplication systems with higher reliability in which the data chunks are distributed across multiple cloud servers. Here we also use TPA i.e. Third Party Auditor for checking user files are present on server or not. The security requirements of data confidentiality and tag consistency are also achieved by introducing a deterministic proof of ownership in distributed storage systems, using convergent encryption as in previous de-duplication systems. Security analysis demonstrates that our de-duplication systems are secure in terms of the definitions specified in the proposed security model. As a proof of concept, we implement the proposed systems and demonstrate that the incurred overhead is very limited in realistic environments.*

Keywords-cloud security, data de-duplication, proof of ownership, distributed system, file level, block level

I. INTRODUCTION

The purpose of the design is to secure de-duplication systems with higher reliability in cloud computing We introduce the distributed cloud storage servers into de-duplication systems to provide better fault tolerance. To further protect data confidentiality, the secret sharing technique is utilized, which is also compatible with the distributed storage systems. In more details, a file is first split and encoded into fragments by using the technique of secret sharing, instead of encryption mechanisms.

These shares will be distributed across multiple independent storage servers. Furthermore, to support de-duplication, a short cryptographic hash value of the-content will also be computed and sent to each storage server as the fingerprint of the fragment stored at each server. Only the data owner who first uploads the details required to compute and distribute such secret shares while all following users who own the same data copy do not need to compute and store these shares any more To recover data copies, users must access a minimum number of storage servers through authentication and obtain the secret shares to reconstruct the data. In other words, the secret shares of data will only be accessible by the authorized users who own the corresponding data copy.

II. PROPOSED SYSTEM

Secure de-duplication systems are proposed to provide efficient de-duplication with high reliability for file-level and block-level de-duplication, respectively. The secret splitting technique, instead of traditional encryption methods, is utilized to protect data confidentiality. Data are split into fragments by using secure secret sharing schemes and stored at different servers.

III. ADVANTAGES AND PROPOSED SYSTEM

1. High Reliability where data chunks are distributed across multiple servers
2. Security for the Data

IV. LITERATURE SURVEY

1. Secure Distributed De-duplication system with Improved reliability.

Author: Jin Li, Xiaofeng Chen, Xinyi Huang,

With the explosive growth of digital data, de-duplication techniques are widely employed to backup data and minimize network and storage overhead by detecting and eliminating redundancy among data. Instead of keeping multiple data copies with the same content, de-duplication eliminates redundant data by keeping only one physical copy and referring other redundant data to that copy. De-duplication has received much attention from both academia and industry because it can greatly improves storage utilization and save storage space, especially for the applications with high de-duplication ratio such as archival storage systems.

A number of de-duplication systems have been proposed based on various de-duplication such as client side or server-side de-duplications, file-level or block-level de-duplications. A brief review is given in Section 6. Especially, with the event of cloud storage, data de-duplication techniques become more attractive and critical for the management of ever-increasing volumes of data in cloud storage services which motivates enterprises and organizations to outsource data storage to third-party cloud providers, as evidenced by many real-life case studies [1]. According to the analysis report of IDC, the volume of data in the world is expected to reach 40 trillion gigabytes in 2020. Today's commercial cloud storage services, such as Drop box, Google Drive and Mozy, have been applying de-duplication to save the network bandwidth and the storage cost with client-side de-duplication.

There are two types of de-duplication in terms of the size:

(i) file-level de-duplication, which discovers redundancies between different files and removes these redundancies to reduce capacity demands, and
(ii) block-level de-duplication, which discovers and removes redundancies between data blocks. The file can be divided into smaller fixed-size or variable- size blocks. Using fixed-size blocks simplifies the computations of block boundaries, while using variable-size blocks provides better de-duplication efficiency.

2. BL-MLE: Block-Level Message-Locked Encryption for Secure Large File Deduplication

AUTHORS: Rongmao Chen, Yi Mu, Guomin Yang

According to the analysis of the International Data Corporation (IDC), the volume of data in the world will reach 40 trillion gigabytes in 2020 [1]. In order to reduce the burden of maintaining big data, more and more enterprises and organizations have chosen to outsource data storage to cloud storage providers. This makes data management a critical challenge for the cloud storage providers. To achieve optimal usage of storage resources, many cloud storage providers perform de-duplication, which exploits data redundancy and avoids storing duplicated data from multiple users. De-duplication Strategy.According to the architecture and the granularity of data processing, de-duplication strategies can be mainly classified into the following types.

In terms of de-duplication granularity, there are two main de-duplication strategies.

(1) File-level de-duplication: the data redundancy is exploited on the file level and thus only a single copy of each file is stored on the server.
(2) Block-level de-duplication each file is divided into blocks, and the severe xploits data redundancy at the block level and hence performs a more fine-grained de-duplication. It is worth noting that for block-level de-duplication, the block size can be either fixed or variable in practice, and each method has its advantages and disadvantages . In this work, we focus on the block-level de-duplication with fixed block size. From the perspective of de-duplication architecture, there are also two strategies.

(1) Target-based de-duplication: users are unaware of any de-duplication that might occur to their out sourced files. They just upload the files to the data storage server which then performs de-duplication upon receiving the data.
(2) Source-based de-duplication: unlike target-based de-duplication, before uploading the data, the user first send san identifier/tag of the data (e.g., a hash value of the data and thus much shorter) to the server for redundancy checking and thus duplicated data would not be sent over the network.

3. De-duplication and Encryption in Cloud Storage

AUTHORS: Joshi Vinay Kumar1, V Ravi Shankar

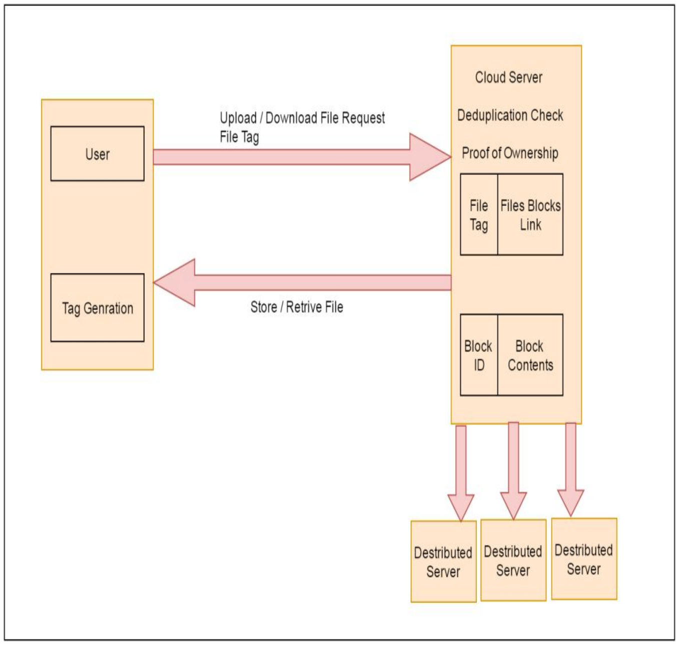
data set multiple times for recovery purposes over 30- to 90-day periods. As a result, enterprises of all sizes rely on backup and recovery with de-duplication for fast, reliable, and cost-effective backup and recovery. De-duplication segments an incoming data stream, uniquely identifies data segments, and then compares the segments to previously stored data. If the segment is unique, it's stored on disk. However, if an incoming data segment is a duplicate of what has already been stored, a reference is created to it and the segment isn't stored again. For example, a file or volume that's backed up every week creates a significant amount of duplicate data. De-duplication algorithms analyze the data and store only the compressed, unique segments of a file. This process can provide an average of 10 to 30 times reduction in storage capacity requirements, with average backup retention policies on normal enterprise data. This means that companies can store 10 TB to 30 TB of backup data on 1 TB of physical disk capacity, which has huge economic benefits as follows. Eliminating redundant data can significantly shrink storage requirements and improve bandwidth efficiency. Because primary storage has gotten cheaper over time, enterprises typically store many versions of the same information so that new workers can reuse previously done work. Some operations like backup store extremely redundant information. De-duplication lowers storage costs as fewer disks are needed. It also improves disaster recovery since there's far less data to transfer. Backup and archive data usually includes a lot of duplicate data. The same data is stored over and over again, consuming unnecessary storage space on disk or tape, electricity to power and cool the disk or tape drives, and bandwidth for replication. This creates a chain of cost and resource inefficiencies within the organization.

4. A Novel Hybrid Cloud Methodology to De-duplicate Authorization Protected Redundancy

Author: Priyanka Bhopale, R B Singh

As cloud computing becomes current, associate increasing quantity of knowledge is being stored within the cloud and shared by users with mere privileges, that outline the access rights of the hold on data. One essential challenge of cloud storage services is the management of the ever-increasing volume of knowledge. To make information management climbable in cloud compute. Cloud computing is associate raising service model that has computation and storage resources on the web. One engaging practicality that cloud computing can give is cloud storage. People and enterprises square measure typically needed to remotely archive their information to avoid any data loss just in case there square measure any hardware or software failures or unforeseen disasters. Rather than buying the required storage media to stay information backups, people and enterprises will merely source their information backup services to the cloud service suppliers, which give the mandatory storage resources to host the info backups. Whereas cloud storage is engaging, a way to give security guarantees for outsourced information becomes a rising concern. One major security challenge is to produce the property of assured deletion, i.e., information files square measure for good inaccessible upon requests of deletion. Keeping information backups for good is undesirable, as sensitive data could also be exposed within the future owing to information breach or incorrect management of cloud operators. Thus, to avoid liabilities, enterprises and government agencies sometimes keep their backups for a finite variety of years and request to delete (or destroy) the backups subsequently. Though information de-duplication brings lots of advantages, security and privacy considerations arise as users' sensitive information square measure liable to each corporate executive and outsider attacks. Ancient coding, whereas providing information confidentiality, is incompatible with information de-duplication. Specifically, ancient coding needs totally different users to write their information with their own keys. Thus, identical information copies {of totally different of various} users can cause different cipher texts, creating de-duplication not possible. Merging coding [8] has been projected to enforce information confidentiality whereas creating de-duplication possible. It encrypts/ decrypts an information copy with a merging key that is obtained by computing the scientific discipline hash price of the content of the info copy. When key generation and encryption, users retain the keys and send the cipher text to the cloud. Since the coding operation is settled and comes from the info content, identical information copies can generate a similar merging key and thence a similar cipher text. To stop unauthorized access, a secure proof of possession protocol [11] is additionally required to produce the proof that the user so owns a similar file once a reproduction is found. When the proof, ulterior users with a similar file are provided a pointer from the server while not having to transfer a similar file.

V. ARCHITECTURE OF PR0POSED SYSTEM



VI. CONCLUSION

Examinations demonstrate that our plans make it down to earth to check possession of vast data sets. Past plans that don't permit testing are not commonsense when PDP is utilized to demonstrate possession of a lot of data, as they force a significant I/O and computational weigh on the server. We designed a system which achieves confidentiality and enables block-level de-duplication at the same time. Our system is built on top of convergent encryption. We showed that it is worth performing block-level de-duplication instead of file level de-duplication since the gains in terms of storage space are not affected by the overhead of metadata management, which is minimal. Additional layers of encryption are added by the server. Thanks to the features of these components. As the additional encryption is symmetric, the impact on performance is negligible. We also showed that our design, in which no component is completely trusted, prevents any single component from compromising the security of the whole system. Our solution also prevents curious cloud storage providers from inferring the original content of stored data by observing access patterns or accessing metadata. Furthermore, we showed that our solution can be easily implemented with existing and widespread technologies. Finally, our solution is fully compatible with standard storage APIs and transparent for the cloud storage provider, which does not have to be aware of the running de-duplication system. Therefore, any potentially non trusted cloud storage provider such as Amazon, Dropbox and Google Drive, can play the role of storage provider. As part of future work, Cloud De-dup may be extended with more security features such as proofs of retrievability [7], data integrity checking [9] and search over encrypted data.

In this paper we mainly focused on the definition of the two most important operations in cloud storage are storage and retrieval. We plan to define other typical operations such as edit and delete. After implementing a prototype of the system, we aim to provide a full performance analysis. Furthermore, we will work on finding possible optimizations in terms of bandwidth, storage space and computation

FUTURE SCOPE

The distributed de-duplication system is to improve the reliability of data. Four constructions were proposes to support file-level,block-level data de-duplicaion.

REFERENCES

[1] J. Li, X. Chen, M. Li, J. Li, P. Lee, and W. Lou, “Secure deduplication with efficient and reliable convergent key management,”IEEE Trans. Parallel Distrib. Syst., vol. 25, no. 6, pp. 1615–1625, Jun.2014.

[2] S. Halevi, D. Harnik, B. Pinkas, and A. Shulman-Peleg, “Proofs of ownership in remote storage systems,” in Proc. ACM Conf. Comput. Commun. Secur., 2011, pp. 491–500.

[3] J.R. Douceur, A. Adya, W.J. Bolosky, D. Simon, and M. Theimer, ‘‘Reclaiming Space from Duplicate Files in a Serverless Distributed File System,’’ in Proc. ICDCS, 2002, pp. 617-624.

[4] M. Abadi, D. Boneh, I. Mironov, A. Raghunathan, and G. Segev. Message-locked encryption for lock-dependent messages. In R. Canetti and J. A. Garay, editors, CRYPTO 2013, Part I, volume 8042 ofLNCS, pages 374{391. Springer, Aug. 2013. (Cited on page 3, 4, 5, 6, 9, 13.)

[5] Landon P Cox, Christopher D Murray, and Brian D Noble. Pastiche: Making backup cheap and easy. ACM SIGOPS Operating Systems Review, 36(SI):285–298, 2002.

[6] Is Convergent Encryption really secure? http://bit.ly/Uf63yH.Mihir Bellare, Sriram Keelveedhi, and Thomas Ristenpart. Message locked encryption and secure de-duplication. In Advances in Cryptology–EUROCRYPT 2013, pages 296–312. Springer, 2013.

[7] Point cheval, D., Johansson, T. (eds.): Advances in Cryptology | EUROCRYPT 2012, 31st Annual International Conference on the Theory and Applications of Cryptographic Techniques, Cambridge, UK, April 15-19, 2012. Proceedings, Lecture Notes in Computer Science, vol. 7237. Springer (2012)

[8] Bellare, M.Fischlin, M., O'Neill, A., Ristenpart, T.: Deterministic encryption: De_nitional equivalences and constructions without random oracles. In: Wagner [31], pp. 360{378}

[9] Bellare, M., Keelveedhi, S., Ristenpart, T.: Message-locked encryption and secure de-duplication. In: Johansson and Nguyen [21], pp. 296{312}.