International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2013): 4.438

A Secure Approach for Deduplication using Hybrid Cloud

Yusuf Aliyu Adamu

Faculty of Sciences and Humanities, Department of Information Technology, SRM University
Kattankulathur, 603203, Chennai, India

Abstract: ***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 study 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, also several new deduplication constructions supporting authorized duplicate check in hybrid cloud architecture. Security analysis demonstrates that our scheme is secure in terms of the definition specified in the proposed security model and shows that the proposed authorized duplicate check scheme incurs minimal overhead compared to normal operations.***

Keywords: convergent encryption, deduplication, key generation, storage service provider, proof of ownership.

1. Introduction

Cloud computing provides seemingly unlimited "virtualized" resources to users as services across the whole Internet, while hiding rostrum and implementation details. Today's cloud service providers offer available storage and massively parallel computing re-sources at low costs. As cloud computing becomes widespread, an increasing amount of large volume of data is being stored in the cloud and shared by users with specified honour, which define the access rights of the stored data. One challenge of cloud storage services is the management of the large amount of data. To make data management scalable in cloud computing, deduplication has been a known technique and has attracted more attention recently.

Data deduplication is a specialized data compression technique for removing duplicate copies of repeating data in cloud storage. The technique is used to improve storage utilization and can also be applied to network data transfers to reduce the number of bytes that must be sent. Instead of storing multiple data copies with the same content, deduplication removed the redundant data and keep only one physical copy and referring other redundant data to that copy. Deduplication can either be at the block level or the file level. For file-level deduplication, it eliminates duplicate copies of the same file. It also takes place at the block level, which removes duplicate blocks of data that occur in non-identical files.

To avoid this duplication of data and to maintain the confidentiality in the cloud the concept of Hybrid cloud is used which is a combination of public and private cloud. Hybrid cloud storage combines both the advantages of public and private by providing reliability, scalability and rapid deployment and potential cost savings of public cloud storage with the security and full control of private cloud storage.

2. Existing System

Data deduplication is one of the specialized data compression techniques for eliminating duplicate copies of same data in storage. This technique is used to improve storage utilization and also be applied to network data transfers to reduce the amount of bytes that must be sent. Instead of storing multiple data copies with the same content, deduplication removes redundant data by keeping only one physical copy and referring other redundant data to that copy.

One critical challenge of cloud storage services is the management of the large amount of data. To make data management scalable in cloud computing, deduplication has been a well adopted technique. Although data deduplication brings a lot of advantage, security and privacy issues arise as users' sensitive data are susceptible to both inside and outside attacks. Identical data copies of different users will lead to different cipher texts, making deduplication difficult and impossible

3. Proposed System

A system is implemented that includes the public cloud and the private cloud and also the hybrid cloud which is a combination of both public and private cloud. In general public cloud can't provide security to private data and hence private data will be loss so that we have to provide the security to our data for that we make a use of private cloud. The private cloud provides greater security. In this system we also provide the data deduplication technique which is used to eliminate duplicate copies of data. Users can be able to upload and download the files from public cloud but private cloud provides the security for that data in which only the authorized users can upload and download the files from the public cloud at the same time files can be shared to different users.

In this study, aiming at efficiently solving the problem of deduplication with differential privileges in cloud computing, hybrid cloud architecture is used consisting of a public and a private cloud. Different from existing data deduplication systems, the private cloud is involved to allow data owner/users to securely perform duplicate check with differential privileges.

3.1 Post-Process Deduplication

With post-process deduplication, data is first stored on the storage device and then process at a later time and analyzed the data looking for duplication. The advantage is that there is no need to wait for the hash calculations and lookup to be completed before storing the data thereby ensuring that store performance is not reduced. The Implementations policy based operation can give users the ability to defer optimization on active files, or to process files based on their type and location. One potential deficiency is that you may unnecessarily store duplicate data for a short time which affect the storage capacity.

3.2 In-line Deduplication

This process is where the deduplication hash calculations are created on the target device as the data enters the device in real time. If the device spots a block that already stored on the system it does not store the new block but refers to the existing block. The advantage of in-line deduplication over post-process deduplication is that it requires less storage as data is not duplicated. On the other side, it is always argued that because hash calculations and lookups takes so long, it can mean that the data takes can be slower thereby reducing the backup throughput of the device. However, certain companies with in-line deduplication have demonstrated equipment with similar performance to their post-process deduplication counterparts. Post-process and in-line deduplication techniques are often heavily debated.

3.3 Source versus Target Deduplication

Another way to think about data deduplication is by where it occurs. When the deduplication occurs close to where data is created, it is often referred to as "source deduplication." When it occurs near where the data is stored, it is commonly called "target deduplication." Source deduplication ensures that data on the data source is deduplicated. This takes place directly within a file system. The file system will periodically scan new files creating hashes and compare them to hashes of existing files.

When files with same hashes are found then the file copy is removed and the new file points to the old file. Unlike hard links however, duplicated files are considered to be separate entities and if one of the duplicated files is later modified, then using a system called Copy on-write a copy of that file or changed block is created. The deduplication process is transparent to the users and backup applications. Backing up a deduplicated file system will often cause duplication to occur resulting in the backups being bigger than the source data. Target deduplication is the process of removing duplicates of data in the secondary store. Generally this will be a backup store such as a data repository or a virtual tape library.

One of the most common forms of data deduplication implementations works by comparing chunks of data to detect duplicates. For that to happen, each chunk of data is assigned identification, calculated by the software, typically using cryptographic hash functions. In many implementations, the assumption is made that if the identification is identical, the data is identical, even though this cannot be true in all cases due to the pigeonhole principle; other implementations do not assume that two blocks of data with the same identifier are identical, but actually verify that data with the same identification is identical. If the software either assumes that a given identification already exists in the deduplication namespace or actually verifies the identity of the two blocks of data, depending on the implementation, then it will replace that duplicate chunk with a link. Once the data has been deduplicated, upon read back of the file, wherever a link is found, the system simply replaces that link with the referenced data chunk. The deduplication process is intended to be transparent to end users and applications.

Secure Deduplication with the advent of cloud computing, secure data deduplication has attracted much attention recently from research community. Yuan et al. [1] proposed a deduplication system in the cloud storage to reduce the storage size of the tags for integrity check. To enhance the security of deduplication and protect the data confidentiality, Bellare et al. [2] showed how to protect the data confidentiality by transforming the predictable message into unpredictable message. In their system, another third party called key server is introduced to generate the file tag for duplicate check. Stanek et al. [3] presented a novel encryption scheme that provides differential security for popular data and unpopular data. For popular data that are not particularly sensitive, the traditional conventional encryption is performed.

Another two-layered encryption scheme with strongest security while supporting deduplication is proposed for unpopular data. In this way, they achieved better Tradeoff between the efficiency and security of the outsourced data Li et al. [4] addressed the key management issue in block-level deduplication by distributing these keys across multiple servers after encrypting the files.

3.4 Convergent Encryption

Convergent encryption [5] ensures data privacy in deduplication. Bellare et al. [6] expressed this primitive as message-locked encryption (MLE), and explored its application in space efficient secure outsourced storage. Xu et al. [7] and also addressed the problem and showed a secure convergent encryption for efficient encryption, without taking account in to issues of the key-management and block-level deduplication. There are also many implementations of convergent implementations of different convergent encryption variants for secure deduplication (e.g., [8], [9]). It is known that some commercial cloud storage providers, such as Bitcasa, also deploy convergent encryption.

Convergent encryption provides data confidentiality in de-duplication. A user derives a convergent key from each original data copy and encrypts the data copy with the convergent key.

The basic idea of convergent encryption (CE) is to derive the encryption key from the hash of the plain text. The simplest implementation of convergent encryption can be defined as follows:

User derives the encryption key from the message M such that K = H (M), where H is a cryptographic hash function; User can encrypt the message with the key,
C = E (K; M) = E (H (M); M)
Where E stand for block cipher.

By using this technique, two users with two identical plaintexts will obtain two identical cipher texts since the encryption key is the same; therefore the cloud storage provider will be able to perform deduplication on such cipher texts. The encryption keys are generated, retained and protected by users. As the encryption key is generated from the plaintext, users do not have to relate with each other for establishing an agreement on the key to encrypt a given plaintext. So, convergent encryption is a good scheme for the encryption and deduplication in the cloud storage field. Furthermore, the user derives a tag for the copy data in which the tag will be used to detect duplicates.

A convergent encryption technique can be defined with four primitive functions:

KeyGen (M) → K implies key generation algorithm that maps a data copy M to a convergent key K.

Encrypt(K,M) → C is the symmetric encryption algorithm that takes both the convergent key K and the data copy M as inputs and then outputs a cipher text C

Decrypt (K, C) → M is the decryption algorithm that takes both the cipher text C and the convergent key K as inputs and then outputs the original data copy M

TagGen (M) →T (M) is the tag generation algorithm that maps the original data copy M and outputs a tag T(M). Also allow TagGen to generate a tag from the corresponding cipher text by using T (M) =TagGen(C), where C=Encrypt (K, M).

3.5 Proof of Ownership

Halevi et al. [10] proposed the notion of "proofs of ownership" (POW) for deduplication systems, such that a user can simply prove to the cloud storage server that he/she owns the data without uploading it itself. Many POW constructions based on the Merkle-Hash Tree are proposed to enable client-side deduplication, which carry the bounded leakage setting. Pietro and Sorniotti [11] proposed another efficient POW technique by selecting the projection of a file onto some randomly selected bit-positions as the file proof. Note that all the above system do not consider data privacy. Recently, Ng et al. [12] prolonged POW for encrypted file but they do not address how to minimize the key management hanging.

3.6 Twin Cloud Architecture

Recently, Bugiel et al. [13] provided an architecture consisting of twin clouds for secure outsourcing of data and arbitrary computations to an untrusted commodity cloud. Zhang et al. [14] also conferred the hybrid cloud techniques to support privacy aware data intensive computing. Authorized deduplication problem over data in public cloud is considered. The security mode is similar to those related work, where the private cloud is expect to be honest but curious.

Pinkas, and A. Shulman-Peleg[15] – To prevent illegal access, a secure proof of ownership protocol is also needed to provide the proof that the user really owns the same file when a duplicate is found. After the proof, consecutive users with the same file will be provided a pointer from the server without need to upload the same file. A user can download the encrypted file with the pointer from the server, which can only be decrypted by the matching data owners with their convergent keys.

S. Halevi, D. Harnik, B. Pinkas, and A. Shulman-[10] Proposes a proof of ownership (POW) technique which user can freely prove to the cloud storage server that he/she owns a file without uploading it and also proposes the Merkle-Hash Tree to enable client-side deduplication, which involved the bounded leakage setting. The proposed scheme is focusing only on the data ownership and not on the data privacy.

4. Implementation

Three entities are used which are the user, public and private cloud. First if the users want to upload the files on the public cloud then user first encrypt that file with the convergent key and then send it to the public cloud at the same time user also generates the key for that file and sends that key to the private cloud for the purpose of security. In the public cloud one algorithm for deduplication is used to avoid the duplicate copies of files which are entered in the public cloud and it also minimizes the bandwidth. That means we require the less storage space for storing the files on the public cloud. In the public cloud unauthorized user can also access or store the data so we can conclude that in the public cloud the security is not provided. In general for providing more security private cloud is used instead of using the public cloud. Users generate the key at the time of uploading file and store it to the private cloud When user wants to downloads the file that he/she upload a request is sent to the public cloud. Public cloud provides the list of files that are uploaded and the user can download the file successfully.

4.1 Roles of Entities

4.1.1S-CPS

The S-CSP provides the data outsourcing service and stores data on behalf of the users. The S-CSP eliminates the

redundant data via deduplication and keeps only unique data to reduce the storage cost. The S-CSP entity is used to reduce the storage cost and has abundant storage capacity and computational power. When user send respective token for accessing his file from public cloud S-CSP matches this token with internally stored data if it matched otherwise an abort signal is sent to user. After receiving the file, user use convergent key K_F to decrypt the file.

4.1.2 Data User

A user is an individual that wants to outsource data storage to the S-CSP and access the data later. The user is allowed to upload unique data. And he/she don't have any right to upload any duplicate data that may be owned by the same or different users.

Each file is protected by convergent encryption key and can access by only authorized user. In our system user must need to register in private cloud for storing token with respective file which are store on public cloud. When user want to access that file he/she access respective token from private cloud and then access his files from public cloud token consist of file content F and convergent key K_F.

4.1.3Private Cloud

In general for providing more security user can use the private cloud instead of public cloud. Users stored the generated key in private cloud. During downloading a system ask the key to download the file. User can't store the secrete key internally. For providing proper protection to key a private cloud is use and store only the convergent key with respective file. When user want to access the key it first check the authentication of the user and then provide the key.

4.1.4Public Cloud

Public cloud is used for the storage purpose. User upload the files in public cloud same with S-CSP ,When the user want to download the files from public cloud it will be ask the key which is generated or stored in private cloud. When the user's key matched with files key at that time user can download the file but without key user can not access the file. Only legal user can access the file, in public Cloud all data's are stored in encrypted format to avoid illegal access by an unauthorized users, in which it cannot be hack without secrete or convergent key the original file can't be access.

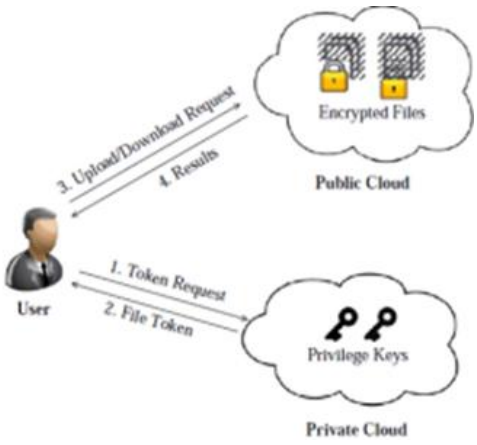

Figure 1

5. Result

In this study, aiming at efficiently solving the problem of deduplication with differential privileges in cloud computing, a secure hybrid cloud architecture consisting of a public and a private cloud was developed. Different from existing data deduplication systems, the private cloud is involved to allow data owner/users to securely perform duplicate check with differential privileges

6. Limitations and Future Works

Though the above solution supports the differential privilege duplicate, it is subject to brute force attacks launched by the public server, which can recover files falling into a known set. More specifically, knowing that the target file space underlying a given cipher text C is drawn from a message space S = {F1, · · · ,Fn} of size n, the public cloud server can recover F after at most n off-line encryptions. That is, for each i = 1, · · · ,n, it simply encrypts Fi to get a cipher text denoted by Ci If C = Ci , it means that the underlying file is Fi, Security is only possible when such a message is unpredictable. This type of traditional convergent encryption will be insecure for predictable file.

7. Conclusion

In this study, the notion of authorized data de-duplication was proposed to protect the data security by including differential privileges of users in the duplicate check also new de-duplication constructions supporting authorized duplicate check in hybrid cloud architecture in which the duplicate check the tokens of files that are generated by the private cloud server with private keys. Security analysis demonstrates that this scheme is secure in terms of insider and outsider attacks.

Cloud computing has reached a maturity that leads it into a productive stage. This means that most of the main issues with cloud computing have been addressed to a degree that clouds have become interesting for full commercial sector. However, does not mean that all the problems listed above have actually been solved. Cloud computing is therefore still as much a research topic, as it is a market offering. For better confidentiality and security in cloud computing we have proposed new de-duplication constructions supporting authorized duplicate check in hybrid cloud architecture.

References

[1] J.Yuan and S. Yu. Secure and constant cost public cloud storage auditing with deduplication. IACR Cryptology ePrint Archive 2013:149, 2013.
[2] M.Bellare, S. Keelveedhi, and T. Ristenpart. Dupless: Serveraided encryption for deduplicated storage. In USENIX Security Symposium, 2013.
[3] J. Stanek, A. Sorniotti, E. Androulaki, and L. Kencl. A secure data deduplication scheme for cloud storage. In Technical Report, 2013.
[4] J. Li, X. Chen, M. Li, J. Li, P. Lee, and W. Lou. Secure deduplication with efficient and reliable convergent key

management. In IEEE Transactions on Parallel and Distributed Systems, 2013.
[5] 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 ICDCS, pages 617–624, 2002.
[6] M.Bellare, S. Keelveedhi, and T. Ristenpart. Message-locked encryption and secure deduplication. In EUROCRYPT, pages 296–312, 2013
[7] J. Xu, E.-C.Chang, and J. Zhou. Weak leakage-resilient client-side deduplication of encrypted data in cloud storage. In ASIACCS, pages 195–206, 2013.
[8] P. Anderson and L. Zhang. Fast and secure laptop backups with encrypted de duplication. In Proc. of USENIX LISA, 2010.
[9] A. Rahumed, H. C. H. Chen, Y. Tang, P. P. C. Lee, and J. C. S.Lui. A secure cloud backup system with assured deletion and version control.In 3rd International Workshop on Security in Cloud Computing, 2011.
[10] R. D. Pietro and A. Sorniotti. Boosting efficiency and security in proof of ownership for deduplication. In H. Y. Youm and Y. Won, editors, ACM Symposium on Information, Computer and Communications Security, pages 81–82. ACM, 2012.
[11] S. Halevi, D. Harnik, B. Pinkas, and A. Shulman-Peleg.Proofs of ownership in remote storage systems. In Y. Chen, G. Danezis, and V. Shmatikov, editors, ACM Conference on Computer and Communications Security, pages 491–500. ACM, 2011.
[12] W. K. Ng, Y. Wen, and H. Zhu. Private data deduplication protocols in cloud storage. In S. Ossowski and P. Lecca, editors, Proceedings of the 27th Annual ACM Symposium on Applied Computing, pages 441–446. ACM, 2012.
[13] S. Bugiel, S. Nurnberger, A. Sadeghi, and T. Schneider.Twin clouds: An architecture for secure cloud computing.In Workshop on Cryptography and Security in Clouds (WCSC 2011), 2011.
[14] K. Zhang, X. Zhou, Y. Chen, X. Wang, and Y. Ruan.Sedic: privacyaware data intensive computing on hybrid clouds. In Proceedings of the 18th ACM conference on Computer and communications security, CCS'11, pages 515–526, New York, NY, USA, 2011. ACM.
[15] Pinkas, and A. Shulman-Peleg."Proofs of ownership in remote storage systems." In Y. Chen, G. Danezis, and V.Shmatikov, editors, ACM Conference on Computer and Communications Security, pages 491–500. ACM, 2011

Author Profile

Yusuf Aliyu Adamu B.sc Computer science at Kano state university of science and technology wudil Nigeria and M.sc in Information Technology from SRM University Tamil Nadu India in 2010 and 2015, respectively. Also work with North West University Kano, Nigeria.