An Efficient Data Deduplication Methodology in a Hybrid Cloud

Naveentha[#1], Sangeetha priyalakshmi [#2]
Final year, Computer Science Department,
Ultra College of engineering and Technology for women,
Madurai, India.

Abstract

Data Deduplication is a specialized data compression technique for eliminating duplicate copies of repeating data. By this technique the Storage Utilization in Cloud can be improved and saves bandwidth. To make the cloud storage more Secured during the process of deduplication, Convergent Encryption technique has been used to encrypt the data before sending to the cloud. Different from Conventional systems, the Advanced Deduplication System support authorized duplication check based on differential privileges of user besides data itself. Hybrid cloud architecture has been used where new deduplication constructions supporting authorized duplication check is performed. Deduplication is done both at file level and block level. This paper addresses the problem of achieving reliable and efficient key management in secured deduplication.

Index Terms -*Hybrid cloud, Deduplication, Convergent encryption, key management*

I. INTRODUCTION

Unlimited "Virtualized" resources are being provided by Cloud Computing across the whole world as a service, where platform and implementation details are hided.At low cost, cloud service provider provide parallel computing resources and storage. Nowadays cloud computing has become a vital role, so large amount of data are stored in cloud and shared by users with certain privileges which describes the access of the stored data. One of the most critical challenges in cloud is about managing the large volume of data.

Deduplication is a special technique used to manage the data in the cloud. Data Deduplication is a technique where the data is being compressed to avoid duplicate copies of redundant data. It is used to improve the storage and can be applied in network data transfer. Deduplication eliminates redundant data by keeping only one physical copy and referring other Redundant data to that copy instead of using multiple copies of data with same content. Data deduplication takes place at two levels. File level and block level. In file level, duplicate copies of same file are being eliminated. And at block level duplicate blocks of data that occur in non-identical files are eliminated.

Even though deduplication has lot of advantages, users sensitive data are susceptible to both insider and outsider attacks. Conventional encryption has been used so far where data confidentiality is maintained but incompatible with deduplication. The traditional encryption needs different users to encrypt their data with their own key. So different users with identical data copies will lead to different cipertext.This makes data deduplication impossible. Convergent Encryption has been used where both data confidentiality and data deduplication is being maintained. This encryption encrypts the data copy with convergent key that is obtained by computing the cryptographic hash value of the content of the data copy. After key generation and data encryption, users retain the keys and send the cipher text to the cloud. The encryption operation is deterministic so it is obtained from data content, identical data copies will produce same convergent key and hence the same ciphertext.A secure proof of ownership protocol is used to provide the proof that the user owns the same file when a duplicate is found to avoid unauthorized access. Subsequent users with the same file will be provided a pointer from the server without needing to upload the same file after the proof .Now the specified user can download the encrypted file with pointer from the server, which can only be decrypted by the corresponding data owners with their own convergent keys. The deduplication system which was used previously did not support differential authorization duplicate check, which is important in many applications. In authorized deduplication system, each user is specified with a set of privileges during initialization. The uploaded file in the cloud is also specified with a set of privileges to check which kind of user is allowed to perform duplicate check. The user has to take his files and his privileges as inputs before submitting his duplicate check request for some file. The user is able to find a duplicate for this file if and only if there is a copy of this file and a matched privilege stored in cloud.

A. Contributions

Aiming at efficiently solving the problem of deduplication with differential privileges in cloud computing, we consider a hybrid cloud architecture consisting of a public cloud and a private cloud. The private cloud is involved as a proxy to allow data owner/users to securely perform duplicate check with

differential privileges. The data owners only outsource their data storage by utilizing public cloud while the data operation is managed in private cloud. The user is only allowed to perform the duplicate check for files marked with the corresponding privileges. We present an advanced scheme to support stronger security by encrypting the file with differential privilege keys. In this way, the users without corresponding privileges cannot perform the duplicate check.

B. Preliminaries

The main encryption used in this paper is Symmetric encryption and convergent encryption.

1) Symmetric Encryption:

A secret key k is used to encrypt/decrypt the data in symmetric encryption. It consists of three main functions.

KeyGenSE (1): κ is the key generation algorithm that generates κ using security parameter 1

EncES (κ, M): C is the symmetric encryption algorithm that takes the secret κ and message M and then outputs the cipher text C

DecSE (κ, C): M is the symmetric decryption algorithm that takes the secret κ and cipher text C and then outputs the original message M.

2) Convergent Encryption:

This encryption provides data confidentiality in data deduplication. The user encrypts the data with the convergent key derived from the original data. The user also produces an *tag* which is used to check the duplicates. To check duplicates, the user initially sends the tag to the server side to check if the identical copy has been already saved. Note that both the convergent key and the tag are independently derived and the tag cannot be used to deduce the convergent key and compromise data confidentiality. Both the encrypted data copy and its corresponding tag will be stored on the server side. The convergent encryption has four main functions:

KeyGenCE (M): K is the key generation algorithm that maps a data copy M to a convergent key K;

EncE (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

DecCE (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).

C. Proof of Ownership

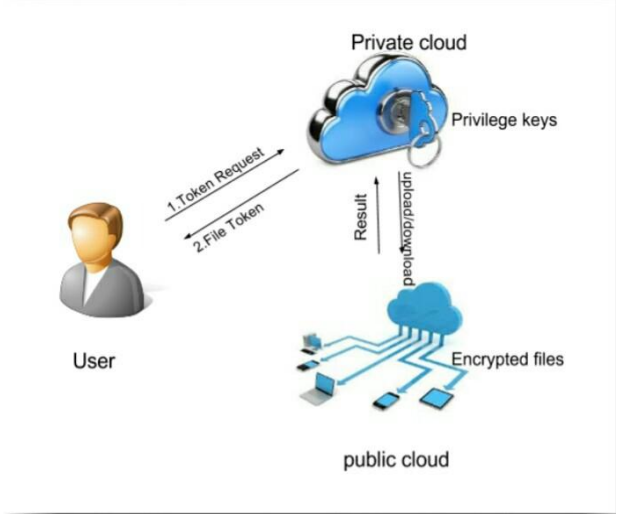
The proof of ownership is denoted by PoW that enables prove their ownership of data copies to the cloud. PoW is used as an interactive algorithm which is run by a prover and verifier. The prover is known to be the user and verifier is the cloud server. The verifier derives a short value S(M) from a data copy M. To prove the ownership of the data copy M, the prover needs to send S′ to the verifier such that S′ = S (M). The formal security definition for PoW roughly follows the threat model in a content distribution network, where an attacker does not know the entire file, but has accomplices who have the file.

D. Identification Protocol

The identification protocol consists of two phases. They are Proof and verify. In Proof, a prover who is the user U can demonstrate his identity to a verifier by performing some identification proof related to his identity. The input of the user is his private key skU that is sensitive information such as private key of a public key in his certificate or debit card number etc. that he would not like to share with the other users. The verifier performs the verification with input of public information pkU related to skU. At the conclusion of the protocol, the verifier outputs either accept or reject to denote whether the proof is passed or not.

II. SYSTEM MODEL

This consists of number of data providers or user who stores their data public cloud server through the private cloud. For example the employees working in a company are data providers or user and the private cloud is the owner of the company and the public cloud includes the public cloud servers. In this setting, deduplication can be frequently used in these settings for data backup and disaster recovery applications while greatly reducing storage space. Such systems are widespread and are often more suitable to user file backup and synchronization applications. There are three main components defined in our system, that is, users, private cloud and S-CSP in public cloud as shown in Fig.S-CSP is defined as Storage Cloud Service Provider. The access right to a file is defined based on a set of privileges. The exact definition of a privilege varies across applications. For example, we may define a role based privilege, according to job positions or we may define a time-based privilege that specifies a valid time period. *Token* is formed which is the short message of each privilege. Each file is associated with some file tokens, which denote the tag with specified privileges. A user computes and sends duplicate-check tokens to the public cloud for authorized duplicate check.



Hybrid Architecture for Secure Deduplication

A. *S-CSP*

It provides the data storage service in the public cloud. The S-CSP provides the data outsourcing service and stores data on behalf of the users. The S-CSP eliminates the storage of redundant data via deduplication and keeps only unique data to reduce the storage cost.

B. *Data Users*

A user is one of the components that want to outsource data storage to the S-CSP and access the data. Each user is issued a set of privileges in the setup of the system in the authorized deduplication system. The file of the respective user has been protected with the convergent encryption key and privilege keys to realize the authorized deduplication.

C. *Private Cloud*

The private cloud is involved as a proxy to allow data owner/users to securely perform duplicate check with differential privileges. Private cloud is able to provide data user/owner with an execution environment and infrastructure working as an interface between user and the public cloud. The private cloud manages the private key for the privileges who answers the file token requests from the users.

III. EARLIEST MODEL

In the existing systems, the deduplication operation took place but the level of data confidentiality was less. In some of the deduplication systems the data are stored in multiple cloud servers but it consumes more time to retrieve it. In some systems the methodology used increases the storage in cloud by applying deduplication on convergent keys. In other conventional system normal deduplication takes place but there will be no privilege key.

IV. OUR PROPOSED MODEL

In our proposed system, we introduce a hybrid cloud architecture in which private cloud servers are used to manage the private key. Here the private keys based on the privileges, are not directly issued to the user. So, the user is cannot share these private keys of privileges and this prevents the sharing of privilege keys among the users. User sends a request to the private cloud servers to get token for a file. The process can be described as follows. Firstly, the user gets the file token from the private cloud servers to check duplication for some files. The private cloud server authenticates the user by checking the user's identity and then issues the corresponding file token to the user. The user performs the authorized duplication check for this file before uploading them, with the public cloud server(S-CSP). It either uploads this file or runs proof of ownership (PoW) based on the duplication check.

A. *System Setup*

A symmetric key for each privilege in the set of privileges are selected and these set of keys will be sent to the private cloud. Each user is assumed to have a secrete key to identify themselves with the servers. The PoW protocol for the file ownership proof has been initiated by the user having the privilege set. A table is maintained by the private cloud server that stores each user's public information and their corresponding privilege set

B. *File Uploading*

If a data owner wants to share or upload a file, before performing the duplication check with the public cloud server, they interact with the private cloud. The data owners prove their identity with private key by performing identification. The private cloud server finds the corresponding privileges of the user from its already stored table list, if user passes the identification. The user computes the file tag and sends it to the private cloud server, which will send the file tag along with the private key, to the public cloud server(S-CSP) to perform duplication check.

1) If duplicate file is found, the user proves the file ownership by running PoW protocol with the S-CSP. The user will be given a pointer for the file, if the proof passes. The user sends the proof along with the privilege sets for a file to the private cloud server. After receiving the request, the private cloud server checks the proof with S-CSP. If it passes, the private cloud server computes the file tag and the private key and sends them to the S-CSP with the signature.

2) If no duplicate file is found, S-CSP will return a proof. The user sends the proof as well as the privilege sets for a file to the private cloud server. After receiving the request, the private cloud server checks the proof with S-CSP. If it passes, the private cloud server computes the file tag and the private key

and sends them to the S-CSP with the signature. Finally, the user encrypts the file with the convergent key and uploads the encrypted file to the S-CSP.

C. File Retrieving

If a user wants to download a file, it sends a request and the file name to the S-CSP. After receiving the request and the file name, S-CSP checks whether the user is eligible to download the requested file. If the user is eligible, S-CSP sends the corresponding cipher texts. After receiving the encrypted form from the S-CSP, the user uses locally stored key to decrypt or to recover the original file.

V. RELATED WORKS

In the traditional storage stack comprising applications, file systems and storage hardware, each of the layers contains different kinds of information about the data they manage and such information in one layer is typically not available to any other layers. Code design for storage and application is possible to optimize deduplication based storage system when the lower-level storage layer has extensive knowledge about the data structures and their access characteristics in the higher-level application layer.

A. Convergent Encryption

Convergent encryption data privacy in deduplication. Bellare et al. [4] formalize this primitive as a message lock encryption, and analyze its application in space-efficient secure outsourced storage. There are also several implementations of different convergent encryption variants for secure deduplication. It is known that some commercial cloud storage providers, such as Bitcasa, also deploy convergent encryption.

VI. CONCLUSION

New deduplication constructions supporting authorized duplicate check in hybrid cloud architecture, in which the duplicate-check tokens of files are generated by the private cloud server with private keys. Security analysis demonstrates that our schemes are secure in terms of insider and outsider attacks specified in the proposed security model. We showed that our authorized duplicate check scheme incurs minimal overhead compared to convergent encryption and network transfer.

REFERENCES

[1] "Secure De-duplication with Efficient and Reliable Convergent Key Management" Jin Li, Xiaofeng Chen, Mingqiang Li, Jingwei Li, Patrick P.C. Lee, and Wenjing Lou in IEEE Transactions On Parallel And Distributed Systems.

[2] Li, J., Chen, X., Huang, X., Tang, S., Xiang, Y., Hassan, M., Alelaiwi A, " SecureDistributed Deduplication Systems with Improved Reliability, "IEEE Transactions on Computers, Volume PP, Issue No. 99, Pages 1, February 2015, DOI - 10.1109/TC.2015.2401017.

[3] P. Anderson and L. Zhang, ''Fast and Secure Laptop Backups with Encrypted De-Duplication,'' in Proc. USENIX LISA, 2010, pp. 1-8.

[4] M. Bellare, S. Keelveedhi, and T. Ristenpart, ''Message-Locked Encryption and Secure Deduplication,'' in Proc. IACR Cryptology ePrint Archive, 2012, pp. 296-3122012:631.

[5] NesrineKaaniche, Maryline Laurent " Client Side Deduplication Scheme Cloud Storage Environments" 6TH International Conference On New Technologies, And Security Year 2014

[6] S. Halevi, D. Harnik, B. Pinkas, and A. Shulman-Peleg, ''Proofs of Ownership in Remote Storage Systems,'' in Proc. ACM Commun. Security, Y. Chen, G. Danezis, and V. Shmatikov, Eds., 2011, pp. 491

[7] R.D. Pietro and A. Sorniotti, ''Boosting Efficiency and Security in Proof of Ownership for Deduplication,'' in Proc. ACM Symp. Inf., Comput. Commun. Security, H.Y. Youm and Y. Won, Eds., 2012, pg.81-82.

[8] "Data Deduplication in Cloud Explained" http://www.computerworld.com/article/2474479/da ta-center/data-deduplication