International Journal of
Emerging Research in Management &Technology
ISSN: 2278-9359 (Volume-5, Issue-5)

Research Article | **May 2016**

Data Deduplicating and Auditing in Cloud

Sahana Kumari B
Student, SCEM Adyar,
Mangalore, Karnataka, India

Dr. Rajni
Associate Professor, SCEM Adyar,
Mangalore, Karnataka, India

Abstract—

Nowadays, cloud computing provides high amount of storage space and massive parallel computing at effective cost. As cloud computing becomes prevalent, excessive amount of data being stored in the cloud. However, exponential growth of ever increasing volume of data has raised many new challenges. De-duplication technique is specialized data compression technique which eliminates redundant data as well as improves storage and bandwidth utilization. Convergent encryption technique is proposed to enforce confidentiality during deduplication, which encrypt data before outsourcing. To better protect data security, we present different privileges of user to address problem of authorized data de-duplication. We also present several new de-duplication constructions supporting authorized duplicate check in hybrid cloud architecture, which incurs minimal overhead compared to normal operation.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 deduplication system. In this paper new distributed deduplication systems with higher reliability in which the data chunks are distributed across multiple cloud servers is being proposed.

Keywords— Cloud Storage, Data de-duplicating, Secure auditing, Data Integrity, Cost effective.

I. INTRODUCTION

Cloud computing is computing in which large groups of remote servers are networked to allow centralized data storage and online access to computer services or resources [9]. With cloud computing, large pools of resources can be connected through private or public network. In public cloud, services (i.e. applications and storage) are available for general use over the internet. A private cloud is a virtualized data centre that operates within a firewall. In this research introduce mix of public and private cloud, hybrid cloud.

Cloud computing provides computation and storage resources on the Internet. Increasing amount of data is being stored in the cloud and it is shared by users with specified privileges, which defines special rights to access stored data. Managing the exponential growth of ever-increasing volume of data has become a critical challenge. According to IDC cloud report 2014, companies in India are making a gradual move from on premise legacy to different forms of cloud. While the process is gradual, it has started by migrating certain application workloads to cloud [1]. To make scalable management of stored data in cloud computing, de-duplication [2] has been well known technique which becomes more popular recently. De-duplication is a specialized data compression technique, which reduce storage space and upload bandwidth in cloud storage. In de-duplication, only one unique instance of the data is actually on the server and redundant data is replaced with a pointer to the unique data copy. deduplication can take place either at file level or block level.

From the user perspective, security and privacy concerns are arise as data are susceptible to both insider and outsider attack. We must properly enforce confidentiality, integrity checking, and access control mechanisms both attacks. De-duplication does not work with traditional encryption. User encrypts their files with their individual encryption key, different cipher text would emerge even for identical files. Thus, traditional encryption is incompatible with data de duplication.

Convergent encryption [3] is a widely used technique to combine the storage saving of de-duplication to enforce confidentiality. In convergent encryption, the data copy is encrypted under a key derived by hashing the data itself. This convergent key is used for encrypt and decrypt a data copy. After key generation and data encryption, users retain the keys and send the cipher text to the cloud. Since encryption is deterministic, identical data copies will generate the same convergent key and the same cipher text. This allows the cloud to perform de duplication on the cipher texts. The cipher texts can only be decrypted by the corresponding data owners with their convergent keys.

Differential authorization duplicate check is an authorized de-duplication technique where each user is issued a set of privileges during system initialization. This set of privileges specifies that which kind of users is allowed to perform duplicate check and access the files.

Hiding platform and implementation details unlimited virtualized resources provided to the users as a service is a cloud computing. Presently cloud service provided to the users offered high available storage and massively parallel

computing of resources at relatively low costs. But the question is about the cloud users with different privileges store data on cloud is a most challenge issue in managing cloud data storage system.

II. EXISTING SYSTEM

A number of deduplication systems have been proposed based on various deduplication strategies. such as client side or server-side deduplications , file-level or block-level deduplication. Bellare et al formalized this primitive as message-locked encryption, and explored its application in space efficient secure outsourced storage. Li addressed the key-management issue in block-level deduplication by distributing these keys across multiple servers after encrypting the files Bellare et al showed how to protect data confidentiality by transforming the predictable message into unpredictable message. The first problem is integrity auditing. The cloud server is able to relieve clients from the heavy burden of storage management and maintenance. The most difference of cloud storage from traditional in-house storage is that the data is transferred via Internet and stored in an uncertain domain ,not under control of the clients at all, which inevitably raises clients great concerns on the integrity of their data. The second problem is secure deduplication. The rapid adoption of cloud services is accompanied by increasing volumes of data stored at remote cloud servers. Among these remote stored files, most of them are duplicated: according to a recent survey by EMC, 75% of recent digital data is duplicated copies. Unfortunately, this action of deduplication would lead to a number of threats potentially affecting the storage system, for example, a server telling a client that it (i.e., the client) does not need to send the file reveals that some other client has the exact same file, which could be sensitive sometimes. These attacks originate from the reason that the proof that the client owns a given file (or block of data) is solely based on static, short value (in most cases the hash of the file).

Disadvantage

- Data reliability is actually a very critical issue in a deduplication storage system because there is only one copy for each file stored in the server shared by all the owners.
- Most of the previous deduplication systems have only been considered in a single-server setting.
- The traditional deduplication methods cannot be directly extended and applied in distributed and multi-server systems.

III. PROBLEM STATEMENT

The problem is to determine how to design secure deduplication systems with higher reliability in cloud computing. Hence it is been proposed in the distributed cloud storage servers into deduplication systems to provide better fault tolerance. To protect data confidentiality, the secret sharing technique is utilized, which is also compatible with the distributed storage systems. To support deduplication, 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.

IV. PROPOSED SYSTEM

In this paper, It has shown how to design secure deduplication systems with higher reliability in cloud computing. By introducing the distributed cloud storage servers into deduplication 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 deduplication, 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 data is 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. Four new secure deduplication systems are proposed to provide efficient deduplication with high reliability for file level and block-level deduplication, respectively. The secret splitting technique, instead of traditional encryption methods, is utilized to protect data confidentiality. Specifically, data are split into fragments by using secure secret sharing schemes and stored at different servers.

Advantage

- Distinguishing feature of our proposal is that data integrity, including tag consistency, can be achieved.
- No existing work on secure deduplication can properly address the reliability and tag consistency problem in distributed system.
- Security analysis demonstrates that the proposed deduplication systems are secure in terms of the definitions pecified in the proposed security model. In more details, confidentiality, reliability and integrity can be achieved in proposed system. Two kinds of collusion attacks are considered in our solutions. These are the collusion attack on the data and the collusion attack against servers. In particular, the data remains secure even if the adversary controls a limited number of storage servers.

- This deduplication systems has been implemented using the Ramp secret sharing scheme that enables high reliability and confidentiality.

V. SYSTEM MODEL

The architecture involves three main entities cloud clients, cloud server and auditor.

Fig 1 .System architecture

Cloud clients have large data files to be stored and rely on the cloud for data maintenance and computation. They can be either individual consumers or commercial organizations Cloud servers virtualize the resources according to the requirements of clients and expose them as storage pools. Typically, the cloud clients may buy or lease storage capacity from cloud servers, Auditor which helps clients upload and audit their outsourced data maintains a MapReduce cloud and acts like a certificate authority. This assumption presumes that the auditor is associated with a pair of public and private keys. Its public key is made available to the other entities in the system.

The SecCloud system supporting file-level deduplication includes the following three protocols respectively highlighted by red, blue and green in Fig.1

File Uploading Protocol: This protocol aims at allowing clients to upload files via the auditor. Specifically, the file uploading protocol includes three phases:

I) Phase 1 (cloud client → cloud server): Client takes the duplicate check with the cloud server to confirm if such a file is stored in cloud storage or not before uploading a file. If there is a duplicate, another protocol called Proof of Ownership will be run between the client and the cloud storage server. Otherwise, the following protocols (including phase 2 and phase 3) are run between these two entities.

II) Phase 2 (cloud client → auditor): Client uploads files to the auditor, and receives a receipt from auditor.

III)Phase 3 (auditor → cloud server): Auditor helps generate a set of tags for the uploading file, and send them along with this file to cloud server.

Integrity Auditing Protocol: It is an interactive protocol for integrity verification and allowed to be initialized by any entity except the cloud server. In this protocol, the cloud server. plays the role of prover, while the auditor or client works as the verifier. This protocol includes two phases:

I) Phase 1 (cloud client/auditor → cloud server): Verifier (i.e., client or auditor) generates a set of challenges and sends them to the prover (i.e., cloud server).

II) Phase 2 (cloud server → cloud client/auditor):

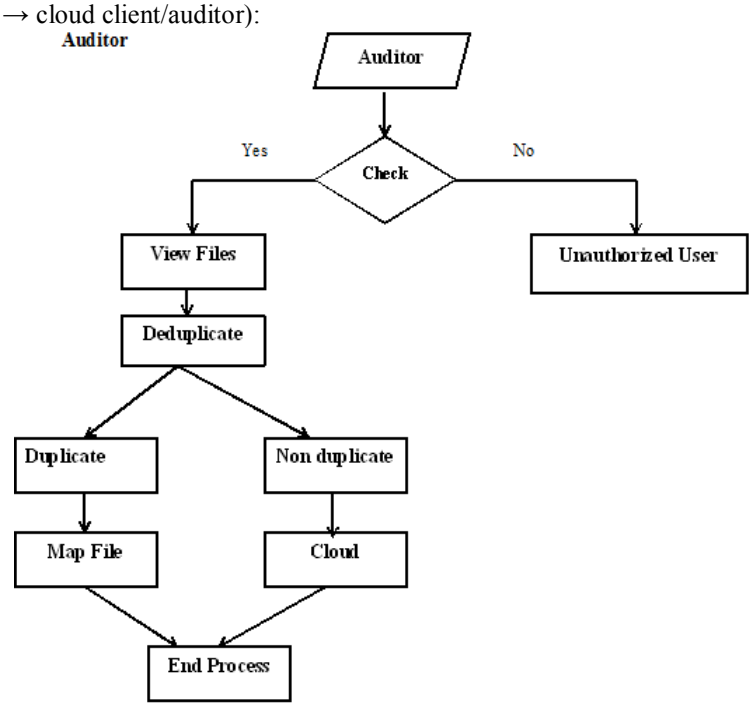

Fig 2: Auditor functionalities

Based on the stored files and file tags, prover (i.e., cloud server) tries to prove that it exactly owns the target file by sending the proof back to verifier (i.e., cloud client or auditor).

Proof of Ownership Protocol: It is an interactive protocol initialized at the cloud server for verifying that the client exactly owns a claimed file. This protocol is typically triggered along with file uploading protocol to prevent the leakage of side channel information. On the contrast to integrity auditing protocol, in PoW the cloud server works as verifier, while the client plays the role of prover.

I) Phase 1 (cloud server → client): Cloud server generates a set of challenges and sends them to the client.
II) Phase 2 (client → cloud server): The client responds with the proof for file ownership, and cloud server finally verifies the validity of proof.

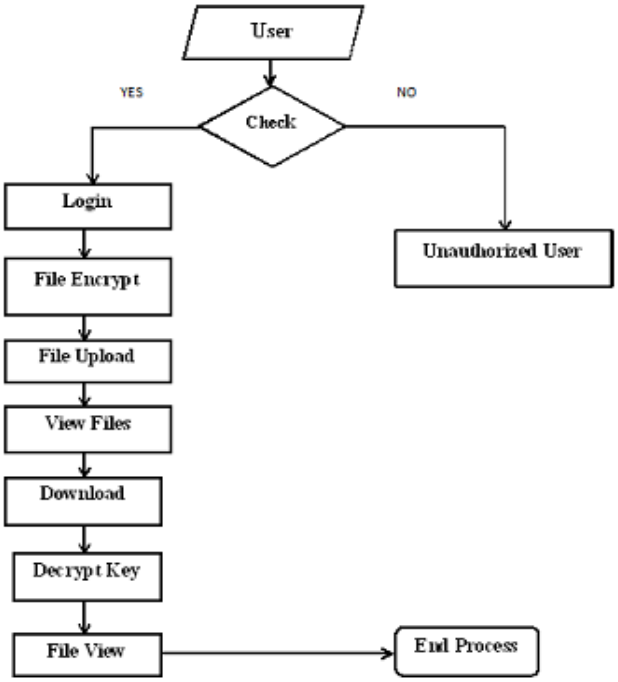

Fig 3: User functionalities

VI. CONCLUSION

Aiming at achieving both data integrity and deduplication in cloud, we propose SecCloud and SecCloud+. SecCloud introduces an auditing entity with maintenance of a MapReduce cloud, which helps clients generate data tags before uploading as well as audit the integrity of data having been stored in cloud. In addition, SecCoud enables secure deduplication through introducing a Proof of Ownership protocol and preventing the leakage of side channel information in data deduplication. Compared with previous work, the computation by user in SecCloud is greatly reduced during the file uploading and auditing phases. SecCloud+ is an advanced construction motivated by the fact that customers always want to encrypt their data before uploading, and allows for integrity auditing and secure deduplication directly on encrypted data.

ACKNOWLEDGEMENT

I dedicate all my paper work to my esteemed guide, Dr. Rajni , whose interest and guidance helped me to complete the work successfully. This experience will always steer me to do my work perfectly and professionally. I also extend my gratitude to Mr. Sudheer Shetty (H.O.D. Computer Engineering Department) Prof B.S.Umashankar (P.G Coordinator) who has provided facilities to explore the subject with more enthusiasm.

REFERENCES

[1] Komal Puri, "IDC Cloud Report 2014: Cross-Vertical Demand Side Perspective", Nov 19, 2014, [Online] Available

[2] S. Quinlan and S. Dorward. "Venti: a new approach to archival storage". In Proc. USENIX FAST, Jan 2002.

[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] S. Halevi, D. Harnik, B. Pinkas, and A. Shulman-Peleg, ,,,,Proofs of Ownership in Remote Storage Systems,"" in Proc. ACM Conf. Comput. Commun. Security, Y. Chen, G. Danezis, and V. Shmatikov, Eds., 2011, pp. 491-500..

[5] M. Bellare, S. Keelveedhi, and T. Ristenpart. Dupless: Serveraided encryption for deduplicated storage. In USENIX Security Symposium, 2013

[6] M. Bellare, S. Keelveedhi, and T. Ristenpart. Message-locked encryption and secure deduplication. In EUROCRYPT, pages 296–312,2013.

[7] M. Bellare, C. Namprempre, and G. Neven. Security proofs for identity based identification and signature schemes. J. Cryptology, 22(1):1–61, 2009.

[8] M. Bellare and A. Palacio. Gq and schnorr identification schemes: Proofs of security against impersonation under active and concurrent attacks. In CRYPTO, pages 162–177, 2002.

[9] 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.

[10] 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

[11] J. Li, X. Chen, M. Li, J. Li, P. Lee, andW. Lou. Secure deduplication with efficient and reliable convergent key management. In IEEE Transactions on Parallel and Distributed Systems, 2013.

[12] libcurl. http://curl.haxx.se/libcurl/.

