International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 02 Issue: 05 | Aug-2015 www.irjet.net p-ISSN: 2395-0072

A Survey on Secure Authorized Deduplication Systems

Mr. Dama Tirumala Babu[1], Prof.Yaddala Srinivasulu[2]

[1]M.Tech student, Department of Computer Science, Audisankara College of Engineering (Autonomous), Andhra Pradesh, India.
[2]Assistant Professor, Department of Computer Science, Audisankara College of Engineering (Autonomous), Andhra Pradesh, India

---

Abstract- ***Cloud Storage Systems are becoming increasingly popular with the continuous and exponential increase of the number of users and the size of data. Data deduplication becomes more and more a necessity for cloud storage providers. Data deduplication is one of the important data compression technique for eliminating duplicate copies of repeating data. It has been widely used in the cloud storage to reduce the amount of storage space and save bandwidth. The advantage of deduplication unfortunately come with high cost in terms of new security and privacy challenges .The proposed scheme in this paper not only the reduces the cloud storage capacity but also improves the speed of data deduplication. To protect confidentiality of sensitive data while supporting deduplication the convergent encryption technique has been proposed to encrypt the data before outsourcing. This paper makes the first attempt to address the problem of authorized data deduplication. Deduplication system is different from the traditional system, because the differential privileges of users are further considered in duplicate check besides the data itself. Security analysis demonstrate that our scheme is secure in terms of the definitions specified in the proposed security model. We show that our proposed system is authorized duplicate check scheme incurs minimal overhead ,compared to normal operations and also show that encryption for deduplicatedstorage can achieve performance and space saving close to that using the storage service.***

Keywords: *Deduplication, authorized duplicate check, confidentiality, hybrid cloud.*

1. INTRODUCTION

Cloud computing provides a low-cost, scalable, location-independent infrastructure for data management and storage. Owing to the population of cloud service and the increasing of data volume, more and more people pay attention to economize the capacity of cloud storage than before .Therefore how to utilize the cloud storage capacity well becomes an important issue nowadays.

Data deduplication [10] is a specialized data compression technique for eliminating duplicate copies of repeating data.A Hybrid Cloud is a combined form of private clouds and public cloudsin which some critical data resides in the enterprise's private cloud while other data is stored in and accessible from a public cloud. Public cloud or eternal cloud describes cloud computing in the traditional main stream sense ,whereby resources are dynamically provisioned on a fine-grained ,self –services basics over the internet via web application /web services from an out-site third party provider who shares resources and bill on a fine grained utility computing basis. Private cloud and internal cloud are neologisms that some vendors have recently used to describe offerings the emulate cloud computing on private network. Hybrid clouds seek to deliver the advantages of scalability, reliability, rapid deployment and potential cost savings of public clouds with the security and increased control and management of private clouds.

Deduplication strategy can be categorized into two main strategies as follow, differentiated by the type of basic data units.

1. File-level deduplication: A file is a data unit when examining the data of duplication, and it typically uses the hash value of the file as its identifier. If two or more files have the same hash value, they are assumed to have the same contents and only one of these files will be stored.
2. Block-level deduplication: This strategy segments a file into several fixed-sized blocks or variable-sized blocks, and computes hash value for each block for examining the duplication blocks.

Data deduplication has certain benefits: Eliminating redundant data can extensively shrink storage requirements and improve bandwidth efficiency. Since primary storage has gotten cheaper over time, typically store many versions of the same information so that new workers can reuse previously done work. Some of the

operations like backup store extremely redundant information.

Deduplication lowers storage costs as fewer disks are needed. It improves disaster recovery since there's far less data to transfer. Backup/archive data usually includes a lot of duplicate data. The similar data is stored over and over again, consuming unwanted storage space on disk or tape, electricity to power and cool the disk/tape drives and bandwidth for replication. This will create a chain of cost and resource inefficiencies within the organization.While providing data confidentiality, traditional encryption is incompatible with data deduplication. Specifically, it requires different users to encrypt their data with their own keys. Thus, indistinguishable data copies of different users will lead to different cipher texts, making deduplication unfeasible.

Convergent encryption has been proposed to impose data confidentiality while making deduplication feasible. It encrypts and decrypts a data copy with a convergent key, which is obtain by computing the cryptographic hash value of the content of the data copy. After key generation and data encryption, users preserve the keys and send the cipher text to the cloud. Because the encryption operation is deterministic and is derived from the data content, indistinguishable data copies will generate the same convergent key and hence the same cipher text.

To avoid unauthorized access, a secure PoW (proof of ownership protocol) is also needed to provide the confirmation that the user indeed owns the same file when a duplicate is found. After the confirmation, consequent users with the same file will be provided a pointer from the server without needing to upload the similar file.

A user can download the encrypted file with the pointer from the server, which can only be decrypted by the equivalent data owners with their convergent keys. Thus, convergent encryption will allow the cloud to do deduplication on the cipher texts and the proof of ownership (PoW) prevents the unauthorized user to access the file.

2.LITERATURE SURVEY

2.1. Fast and secure laptop backups with encrypted de- duplication

Back up algorithm (2010).
The data which is common between users to increase the speed of backup and reduce the storage requirement [2]. Supports client-end per user encryption is necessary for confidential personal data.
Disadvantages:
1. Network bandwidth can be a bottle-neck.
2. Backing up directly to a cloud can be very costly.
Conclusion: This provides the potential to significantly decrease backup times and storage requirement.

2.2. SecureDedup:Server Deduplication with Encrypted Data for cloud storage.

Deduplication unfortunately come with a high cost in terms of new security and privacy challenges [3].
Disadvantages: .Does not impact the overall storage and computational cost.
Conclusion: A system achieves confidentiality and enables block level deduplication at the same time.

2.3. Proofs Of Ownership in Remote storage System

Present solution based on Merkle Trees and Specific encoding we identify attacks that exploit client side deduplication attempts to identify reduplication [11].
Disadvantages: It is impossible to verify experimentally the assumption about the input distribution.
Conclusion: implemented a prototype of the new protocol and ran it to evaluate performance and asses the PoW scheme benefits.

2.4.Weak leakage –resilient client side Deduplication of encrypted data in cloud storage.

Propose a secure client –side deduplication scheme [8]. Disadvantages: Convergent encryption and custom encryption methods are not semantically secure.
Conclusion: Addressed an important security concern in cross-user client –side deduplication.

3. METHODS USED IN SECURE DEDUPLICATION

3.1. Symmetric Encryption

Symmetric encryption uses a common secret key k to encrypt and decrypt information. A symmetric encryption scheme made up of three primary functions.

- KeyGen SE (1λ) → k is the key generation algorithm that generates k using security parameter 1λ;
- Enc SE (k, M)→ C is the symmetric encryption algorithm that takes the secret k, and message M and then outputs the cipher text C, and
- Dec SE (k, C) → M is the symmetric decryption algorithm that takes the secret k and cipher text C and then outputs the original message M.

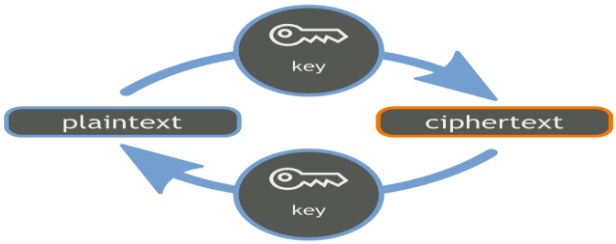

Fig -1. Symmetric key Encryption

3.2.Convergent encryption

Convergent encryption [4], provides data confidentiality in deduplication. A user (or data owner) derives a convergent key from each original data copy and encrypts the data copy with the convergent key. In addition, the user also derives a tag for the data copy, such that the tag will be used to detect duplicates. Here, we assume that the tag correctness property [4] holds, i.e., if two data copies are the same, then their tags are the same. To detect duplicates, the user first sends the tag to the server side to check if the identical copy has been already stored.

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. Formally, a convergent encryption scheme can be defined with four primitive functions:

- KeyGenCE(M) → K is the key generation algorithm that maps a data copy M to a convergent key K;
- EncCE(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 ciphertext C;
- DecCE(K,C) → M is the decryption algorithm that takes both the ciphertext C and the convergent key
  K as inputs and then outputs the original data copy M; and
- TagGen (M) → T (M) is the tag generation algorithm that maps the original data copy M and outputs a tag T (M).

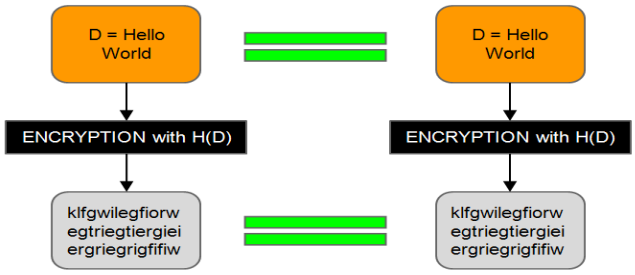

Fig -2. Convergent encryption

3.3. Proof of ownership

The notion of proof of ownership (PoW) [9] enables users to prove their ownership of data copies to the storage server. Specifically, PoW is implemented as an interactive algorithm (denoted by PoW) run by a user and a storage server. The verifier derives a short value ϕ (M) from a data copy M. To prove the ownership of the data copy M, the user needs to send ϕ′ to the server such that ϕ′ = ϕ (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 partners who have the file. The partners follow the "bounded retrieval model", such that they can help the attacker obtain the file, subject to the constraint that they must send fewer bits than the initial min-entropy of the file to the attacker [9].

3.4. Identification Protocol.

An identification protocol II can be described with two phases: Proof and Verify. In the stage of Proof, a prover/user U can demonstrate his identity to a verifier by performing some identification proof related to his identity. The input of the prover/user is his private key skU that is sensitive information such as private key of a public key in his certificate or credit 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. There are many efficient identification protocols in literature, including certificate-based, identity-based identification etc. [5], [6].

Acronym	Description
S-CSP	Storage-cloud service provider
PoW	Proof of Ownership
(pk_U, sk_U)	User's public and secret key pair
k_F	Convergent encryption key for file F
P_U	Privilege set of a user U
P_F	Specified privilege set of a file F
$\phi'_{F,p}$	Token of file F with privilege p

Table-1. Notations used in this paper.

3.4. Twin Clouds Architecture

Bugiel et al. [7] provided an architecture consisting of twin clouds for secure outsourcing of data and arbitrary computations to an untrusted commodity cloud. Zhang et al. [12] also presented the hybrid cloud techniques to support privacy-aware data-intensive computing. In our work, we consider to address the authorized deduplication problem over data in public cloud. The security model of our systems is similar to those related work, where the private cloud is assume to be honest but curious.

4. PROPOSED SYSTEM

4.1. Hybrid cloud architecture

Architecture for data deduplication in cloud computing, which consists of a twin clouds(i.e., the public cloud and the private cloud). Actually, this hybrid cloud setting has attracted more and more attention recently. For example, an enterprise might use a public cloud service, such as Amazon S3, for archived data, but continue to maintain in-house storage for operational customer data. Alternatively, the trustedprivate cloud could be a cluster of virtualized crypto- graphic co-processors, which are offered as a service by a third party and provide the necessary hardware- based security features to implement a remote execution environment trusted by the users.

There are three entities defined in our system, they are users, private cloud and S-CSP in public cloud.

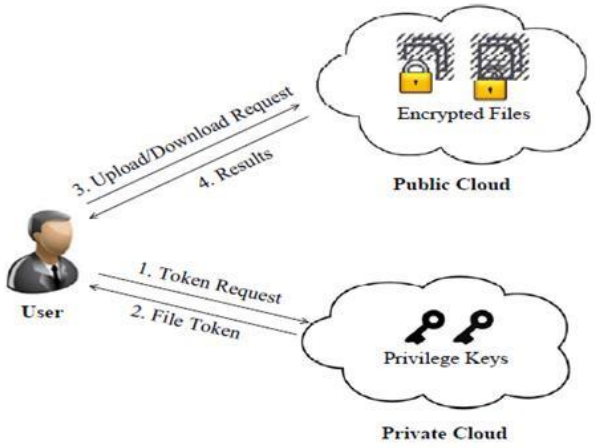

Fig-3. Architecture for Authorized Deduplication

• **S-CSP.** This is an entity that provides a data storage service in public cloud. The S-CSP provides the data outsourcing service and stores data on behalf of the users. To reduce the storage cost, the S-CSP eliminates the storage of redundant data via deduplication and keeps only unique data. In this paper, we assume that S-CSP is always online and has abundant storage capacity and computation power. • **Data Users.** A user is an entity that wants to out- source data storage to the S-CSP and access the data later. In a storage system supporting deduplication, the user only uploads unique data but does not upload any duplicate data to save the upload bandwidth, which may be owned by the same user or different users. In the authorized deduplication system, each user is issued a set of privileges in the setup of the system. Each file is protected with the convergent encryption key and privilege keys to realize the authorized deduplication with differential privileges.

• **Private Cloud.** Compared with the traditional deduplication architecture in cloud computing, this is a new entity introduced for facilitating user's se- cure usage of cloud service. Specifically, since the computing resources at data user/owner side are restricted and the public cloud is not fully trusted in practice, 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 keys for the privileges are managed by the private cloud, who answers the file token requests from the users. The interface offered by the private cloud allows user to submit

files and queries to be securely stored and computed respectively.

4.2. Proposed System Framework Communication

In the proposed system we are achieving the data deduplication by providing the proof of data by the data owner. This proof is used at the time of uploading of the file. Each file uploaded to the cloud is also bounded by a set of privileges to specify which kind of users is allowed to perform the duplicate check and access the files. Before submitting his duplicate check request for some file, the user needs to take this file and his own privileges as inputs. 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.

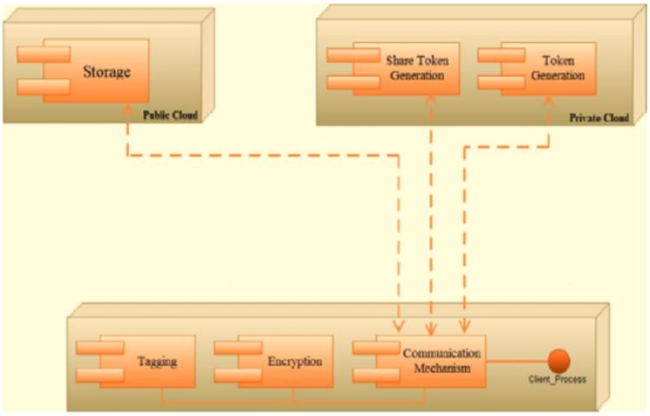

Fig-4. Proposed System Framework Communication

In this work Public Cloud is used for storage data ,Private Cloud is used for performance the operations like share token generation ,token generation, also Client is performed operations like tagging of file ,encryption of file, communication between private cloud and public cloud.

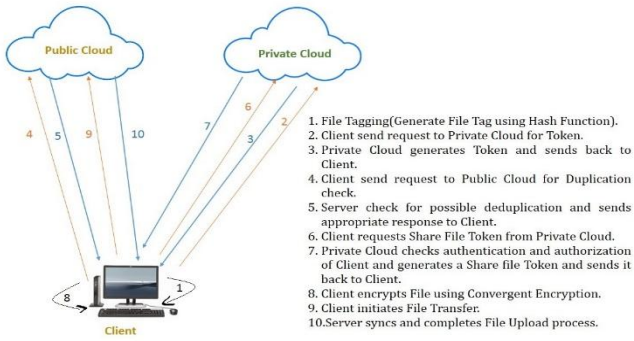

Fig-5. Proposed System Operations.

4.3. Security Analysis

For solving the problem of privacypreserving deduplication in cloud computing and propose a new deduplication system supporting for

• **Differential Authorization.** Each authorized user is able to get his/her individual token of his file to perform duplicate check based on his privileges. Under this assumption, any user cannot generate a token for duplicate check out of his privileges or without the aid from the private cloud server.

• **Authorized Duplicate Check.** Authorized user is able to use his/her individual private keys to generate query for certain file and the privileges he/she owned with the help of private cloud, while the public cloud performs duplicate check directly and tells the user if there is any duplicate.

The security requirements considered in this paper lie in two folds, including the security of file token and security of data files. For the security of file token, two aspects are defined as unforgeability and indistinguishability of file token. The details are given below.

• **Unforgeability of file token/duplicate-check token.** Unauthorized users without appropriate privileges or file should be prevented from getting or generating the file tokens for duplicate check of any file stored at the S-CSP. The users are not allowed to collude with the public cloud server to break the unforgeabilityof file tokens. In our system, the S-CSP is honest but curious and will honestly perform the duplicate check upon receiving the duplicate request from users. The duplicate check token of users should be issued from the private cloud server in our scheme.

• **Indistinguishability of file token/duplicate-check token.** It requires that any user without querying the private cloud server for some file token, he cannot get any useful information from the token, which includes the file information or the privilege information.

• **Data Confidentiality.** Unauthorized users without appropriate privileges or files, including the S-CSP and the private cloud server, should be prevented from access to the underlying plaintext stored at S-CSP. In another word, the goal of the adversary is to retrieve and recover the files that do not belong to them. In our system, compared to the previous definition of data confidentiality based on convergent encryption, a higher level confidentiality is defined and achieved.

5. CONCLUSION

In this paper, we have surveyed and presented the methods used in authorized data deduplication system

provides the secure data deduplication on cloud storage. The hybrid cloud approach is presented for the security purpose which has system with a differential privileges to different users accordingly. Data Deduplication eradicates the redundant data by storing only the single copies of data. It uses the convergent encryption technique to encrypt the data with the convergent key. It also provides Differential Authorized duplicate check, so that only authorized user with specified privileges can perform the duplicate check. The concept deduplication save the bandwidth and reduce the storage space.

REFERENCES

[1] OpenSSL Project. http://www.openssl.org/.
[2] P. Anderson and L. Zhang. Fast and secure laptop Backups with encrypted de-duplication. In Proc. of USENIX LISA, 2010.
[3] PasqualoPuzio, RefikMolva, MelekOnen,"CloudDedup: SecureDeduplication with Encrypted Data for CloudStorage", SecludIT and EURECOM, France.
[4] M. Bellare, S. Keelveedhi, and T. Ristenpart. Message-locked encryption and secure deduplication. In EUROCRYPT, pages 296– 312, 2013.
[5] M. Bellare, C. Namprempre, and G. Neven. Securityproofs for identity-based identification and signature schemes.J. Cryptology, 22(1):1–61, 2009.
[6] M. Bellare and A. Palacio. Gq and schnorr identification schemes: Proofs of security against impersonation under active and concur- rent attacks. In CRYPTO, pages 162–177, 2002.
[7] S. Bugiel, S. Nurnberger, A. Sadeghi, and T. Schneider. Twin clouds: An architecture for secure cloudcomputing. In Workshop on Cryptography and Security in Clouds (WCSC 2011), 2011.
[8] Weak Leakage-Resilient Client –Side deduplication ofEncrypted Data in Cloud Storage" Institute for Info Comm Research, Singapore, 2013
[9] 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, ACMConference on Computer and CommunicationsSecurity, pages 491–500. ACM, 2011.
[10] S. Quinlan and S. Dorward. Venti: a newapproach toArchival storage. In Proc. USENIX FAST, Jan, 2002.
[11] K. Zhang, X. Zhou, Y. Chen, X. Wang, and Y. Ruan. Sedic: privacy- aware data intensive computing on hybrid clouds. In Proceedings of the 18th ACMconference on Computer and communications security, CCS'11, pages 515–526, New York, NY, USA, 2011. ACM.

BIOGRAPHIES

Mr.Dama Tirumala Babu, M.Tech Student, Department of CSE, Audisankara College of Engineering, Gudur. Received B.Tech in Computer Science and Engineering from NBKR Institute of Science and Technology,Vakadu affiliated to SVU. Interesting Areas are Cloud Computing and Core Java.

Mr. Yaddala Srinivasulu, M.Tech,MISTE. Assistant Professor, Department of CSE, Audisankara College of Engineering, Gudur. Received B.Techin Computer Science and Engineering fromV.R.SIDDARTHA College of Engineering affiliated to Acharya Nagarjuna University. Received M.Tech in Computer science and Engineering from NIST, Vijayawadaaffiliated to JNTU Kakinada. Having 09 years ofteaching experience. Interesting Areasarecloud computing and Formal Methods.