Secure and Open Cloud for Deduplication using HMAC Algorithm

Aparna[1] Mohammed Shakura Banu[2] Manickavasagan[3]
[3]Assistant Professor
[1,2,3]Department of Computer Science and Engineering
[1,2,3]Velammal Institute of Technology, Tamilnadu, India

Abstract— The elimination of duplicate data particularly in computer related data is called as data deduplication. According to deduplication, we introduce a method that eliminate redundant encrypted data owned by different users. This mechanism uses the concept of secure cloud where user is able to generate data tags before storing data on cloud which helps during performing audit to check integrity of data.

Key words: Deduplication, Secured Data Deduplication, Convergent Encryption, Hybrid Cloud, HMAC

I. INTRODUCTION

Data deduplication is a process or a technique of eliminating redundant copies of data, and has been widely used in cloud storage to reduce storage space and for the easy access of the data. With the rapid growing of cloud storage services, such as cloud storage encryption becomes an important technique for protecting the confidentiality of data. Although data encryption provides an important guarantee for the security and privacy of clients' data, it limits the manners of the accessibility and availability of the encrypted data. The limitation of schemes with encrypted data is that, when some special processing applications over the data are needed, such as cross-client data deduplication query sorting over encrypted data the schemes usually becomes inefficient due to the frequent date encryption and decryption operations. Thus, it is important to design efficient schemes to support secure and efficient computation outsourcing and storage outsourcing. This has applications in cloud computing to remove duplicate files from storage.

II. OBJECTIVE

To enhance the security of deduplication and protect the data confidentiality showed how to protect the data confidentiality by transforming the predictable message into an unpredictable message. In their system, a third party called key server is introduced to generate the file tag for duplication check. Addressed the key management issue in block-level deduplication by distributing these keys across multiple servers after encrypting the files.

III. RELATED WORK

Secure deduplication is interesting for both industrial and research communities; therefore, several secure deduplication schemes have been proposed. To support data integrity, two concepts, PDP (provable data possession) and POR (proof of retrievability), have been introduced. PDP for ensuring that the cloud storage providers actually possess the files without retrieving or downloading the entire data. It is basically a challenge-response protocol between the verifier (a client or TPA) and the prover (a cloud). Compared to PDP, POR not only ensures that the cloud servers possess the target files, but also guarantees their full recovery.

IV. PROPOSED SYSTEM

However, concerning the challenging problem described, the scheme is not efficient in the deduplication process because of the comparison of the randomized tag introduced in their paper. It is important to maintain tags for sub-linear deduplication time, since for large data sets linear scans are prohibitive, particularly if they involve a linear number of cryptographic operations. In this paper, we ask whether the scheme can be much more efficient in data deduplication for large database while also keep the security properties of the deduplication scheme We adopt client-server interaction based on randomly balanced tree, mutable tree and self-generation tree to improve the efficiency of our schemes, and design two efficient schemes. Both of the designed schemes support efficient data equality test while keeping the security of clients' data by allowing a small number of interaction.

After key generation and data encryption, users retain the keys and send the cipher text to the cloud. Since the encryption operation is deterministic and is derived from the data content, identical information copies will generate the same convergent key and hence the same ciphertext.

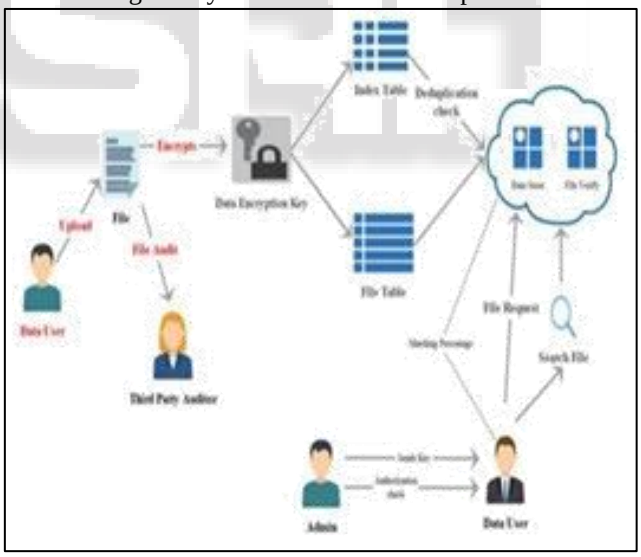

Fig. 1: System Architecture

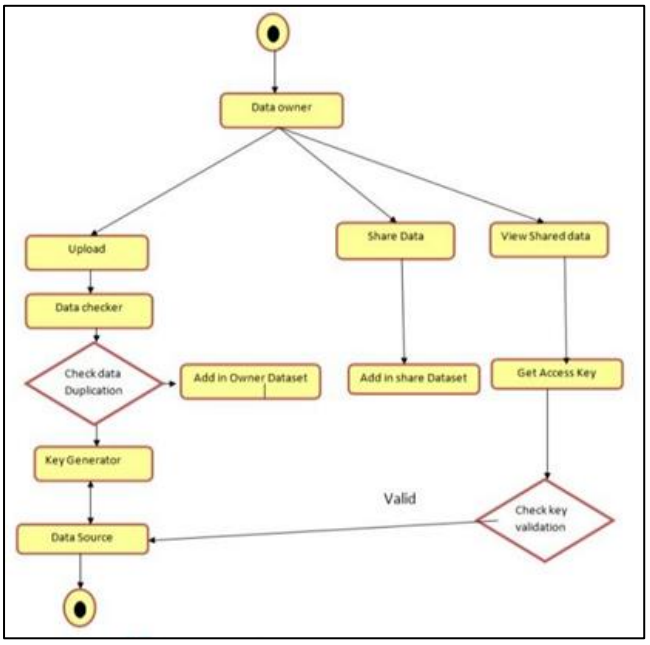

Fig. 2: Use Case Diagram

V. SECURE DATA DEDUPLICATION

Mostly Small Scale to Medium Scale companies outsource their data to S-CSP. Because Small Scale to Medium Scale companies cannot offer the storage cost. So service providers provides storage space to these companies on rent basis. S-CSP provides authority on the basis of their privileges. There are three entities defined in our system, that is, users, private cloud and S-CSP in public cloud. The S-CSP performs deduplication by checking if the contents of two files are the same and stores only one of them. 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 designation (e.g., Senior Engineer, Team Lead and Manager), or we may define a time based privilege that specifies a valid time period within which a file can be accessed.", so that she can access any file whose access role is "Manager" and accessible time period covers . Each privilege is represented in the form of a short message called token. Each file is associated with some file tokens. A user computes and sends duplicate-check tokens to the public cloud for authorized duplicate check. Users have access to the private cloud server, a semi trusted third party which will aid in performing duplicable encryption by generating file tokens for the requesting users. Users are also provisioned with per-user encryption keys and credentials. To prevent unauthorized access, a secure proof of ownership (POW) protocol is also needed to provide the proof that the user indeed owns the same file when a duplicate is found.

VI. MODULE DESCRIPTION

A. Data Owners

Data owner can upload data's, that data are split into part data then send to trusted data checker, job of the data checker is to generate signature key from MD5 and compare with previous keys, if mismatch then that data send to Key generator Server, Job of the key generator are generate encryption key as user specified algorithm, finally encrypt then store in Database.

B. Owner Dataset

In this Module We create data owner dataset, this dataset only map owner with our upload data's, we maintain common database for effectively find duplications. The files will be uploading only once. If another data owner going to upload the same file in database means they will get the notification (the data is already uploaded in database).So data owner can save cost and time.

C. Third Party Verifier

In these modules, the third party auditor checks for the file integrity. If the file contains the same word as was in the file previously saved in the cloud then file will not store instead it shows error. The TPA will filter the file. If the file has some updating with uniqueness then TPD will accept the file and encrypt the file and stored to the cloud.

D. Shared Dataset

Share Dataset is a lightweight dataset that only contain mapping file metadata information, in our project we maintain one common big data database instead of unique because efficiently find duplication and memory management, if data owner share our data to client that data not replicate instead map client name. Data deduplication enables data storage systems to find and remove duplication within data without compromising its availability.

E. Security

We are implementing "Dynamic Encryption key Generation". It means all shared data only view with data owner permission, so we can avoid from unknown access.

Social users are group members they can only view and share the data. If want show the data mean they need to get permission to data owner then data owner will send Encryption key after they can view the data. If data owner does not provide the KEY mean user cannot view the file. Data encryption provides an important guarantee for the security and privacy of clients' data, it limits the manners of the accessibility and availability of the encrypted data.

VII. CONCLUSION

Cloud computing is viewed as the next generation architecture of IT companies. As promising as it is, cloud computing also brings forth many new security issues when users outsource sensitive data to cloud servers. To keep sensitive users' data confidential against untrusted servers, existing solutions usually apply cryptographic methods. With data encryption, the same file will become different from each other, thus deduplication which is widely adopted by cloud storage service providers meets some challenges. Current method to solve the problem is to make use of some information computed from the shared file to achieve deduplication of encrypted data, say convergent encryption. But this piece of information which is computable from the file via a deterministic public algorithm is not really meant to be secret. To this end, we propose a scheme to address the deduplication of encrypted data efficiently and securely with the help of ensuring the ownership of the shared file, encrypting data using keys at user's will and realizing the anonymous store through the digital credential.

REFERENCES

[1] Ms. Deepali C. Ghosalkar "Implementation Idea for Secure Data Deduplication Using Hybrid Cloud Approach"- International Journal of Computer Science Trends and Technology (IJCST) — Volume 4 Issue 1, Jan – Feb 2016
[2] S. Arun Kumar, G.Sandeep, N.Deepath Prof Mr. B. Hariharan "A Secured Authorized Data DE Duplication in Cloud using Hash based Message Authentication CodeMs"-International Journal of Science Technology & Engineering Volume 2 | Issue 09 | March 2016
[3] P Jagadeesh, K Narayana, P Chandra Prakas "Secure Data Deduplication for Cloud Server using HMAC Algorithm-International Research Journal of Engineering and Technology Volume: 05 Issue: 11 | Nov 2018

