International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 4 Issue 05, May-2015

Client Side Deduplication Scheme for Secured Data Storage in Cloud Environments

Naveen A N
M.Tech student, Dept. of CSE,
Siddaganga Institute of Technology,
Tumkur, Karnataka, India

V Ravi
Assistant Professor, Dept. of CSE,
Siddaganga Institute of Technology,
Tumkur, Karnataka, India

***Abstract*—Cloud storage services commonly use duplication, which is useful in storing only a unique file or block by eliminating duplicate copies of data. Deduplication is helpful in saving network bandwidth and storage space which is an advantage to the users or clients in cloud. A new client side deduplication technique is used for data storage in a secured way and sharing data with other users in a public cloud is to be implemented in order to solve issues in security and privacy.**

keywords: Deduplication, Merkle hash tree, convergent encryption, Proof of ownership

I. INTRODUCTION

In Farsite distributed file system propose of recovering storing space is to identify duplicate data files and eliminate them. It is a server less file system which functions as a central file server and this system will be present in a physically distributed fashion in a network which is a collection of workstations.

In the real world whether its enterprise world, business world or home the biggest problem is able to store the data that is currently being created. In recent days, the speeding growth related to digital contents is gearing up to raise requirement of storage space with an efficient utilization of this space and bandwidth to transfer the data. The clients are moving their certain positions of their environment to the cloud as they are looking for cost efficiency and cloud storage provides cost efficient architectures.

Deduplication is most efficient technique, a process of identifying and eliminating redundant data. In the client side deduplication data is duplicated at the client side where client sends only new, unique data across the network which results in reduced storage capacity and network bandwidth savings. The benefits of deduplication include reduced infrastructure costs, reduced management costs, many cloud storage providers such as Dropbox, Memopal and Mozy use client side deduplication in order to save resources which results in avoiding storage of redundant data in cloud storage servers and network bandwidth savings by eliminating transmission of same contents several times.

Even though they are advantages in client side deduplication it has issues related to security, for example attackers will mainly target the bandwidth and confidentiality which is related to privacy of legitimate cloud users.

In order to solve these concerns, Proof of Ownership (PoW) schemes are introduced where they allow the storage server check a client data ownership, based on a hash value. Even though existing schemes deals with different properties of security but there is a still need of careful consideration of potential attacks which includes data Leakage and poison attacks, which mainly target on privacy preservation and data confidentiality. In the baseline approach which is a Proof of Ownership (PoW) scheme a new cryptographic method which uses the Merkle-based Tree and convergent encryption, which results in efficient data deduplication while providing data security in cloud storage systems and providing dynamic sharing between users.

In the Ramp secret sharing scheme it stores only one key resulting in saving resources.

II. REALTED WORK

Douceur et al. describes the problem deduplication problem present in multi-tenant environment[2]. The authors describe about the convergent encryption where keys are derived from the hash of data and use of convergent encryption. Then, in 2008 storer et al. proposed two approaches for secured data deduplication. Where it has disadvantages like in security and in deduplication open areas for exploration exists, multiple levels of permissions can be utilized by future designs.

Proof of Ownership (PoW) is proposed by Halevi to avoid private data leakage, three different schemes were introduced based on performances and security. In these the client has to prove to server that he has accurate sibling paths which are derived from the leaves of the Merkle based tree. Erasure coding is applied by the first scheme on source file contents then the Merkle tree is constructed by this encoded version as a input to it. The second scheme is a substitute of erasure coding, where data file is prepossess using an universal hash function. The third scheme dealt with assumptions on security which lead to designing of efficient families of hash. Unfortunately the assumption of the proof was a appropriate distribution samples the data file.

The Proof of Ownership (PoW) is proposed by Halevi. It is a challenge-response protocol that enables a storage server to verify whether a request is from the data owner, which is based on a hash value. Whenever client or user upload a data file to the cloud server , he has to compute a hash value and sends this value to the cloud storage. The cloud storage server has a database of short values of all stored files, and checks up the short value(hash value). If the hash value is present in cloud server then the file is already

outsourced and it inform the data owner that uploading of file is not required.

A. Security Analysis

Despite having the significant resource saving advantages, PoW schemes comes along with number of security challenges that may create an dangerous environment for sensitive data.

• Data confidentiality disclosure – If attacker knows the hash value of the data file which is present in cloud storage then he can get access to the data file easily by submitting hash value to the cloud storage server, Data confidentiality is an important concern.

• Privacy violation – Sensitive data leakage is a major critical challenge that was not addressed by Halevi et al. The cloud storage should not build user profiles and access the data stored by the user in cloud.

• Poison attack – The data file is encrypted by using some random encrypted key. Now the cloud server cannot verify the uploaded file and the hash value present in its database as values are different and attacker can easily replace enciphered original file with a malicious file.

III. LITERATURE SURVEY

The architecture of cloud data storage consists:

• Cloud Service Provider (CSP): A CSP has sufficient resources to manage its database servers and to govern distributed cloud storage servers. Virtual infrastructure is provided by CSP to host application services , client can use these services in order to manage the data stored in the cloud storage servers.

• Client: A client may refer to an individual or an enterprise client. CSP resources are utilized by client to manage and sharing data with a group of users.

• Users: Based on the permissions granted by the client the user can access the data stored in the cloud storage servers.

Baseline approach

The client deduplication scheme for secured data uses convergent encryption. In cloud storage servers the data owner stores enciphered file by generating the enciphering key . Data encrypting key can be derived by applying SHA-256 on data content. The data owner after encrypting the data file and before uploading the data file to cloud, he has to generate a identifier of the data, so that the identifier is unique which will be compared with the identifiers present in cloud database servers. By applying Merkle tree over the encrypted data file a unique data identifier can be generated. Data owner cannot upload same file to the cloud again and he has to prove his ownership by providing the root value and a sibling value of Merkle tree along with his private key whenever he wants to access the data file that has been already outsourced into the cloud.

A. Methodology

convergent encryption: Produces identical cipher text from identical plain files, convergent encryption is used to remove duplication files from the cloud storage while the cloud service provider has no access to encryption keys. Convergent encryption is deriving keys from the hash of plain text and provides a security model for secure data deduplication

Merkle hash tree : The data file is broken down into leaves of the tree which are grouped together, these are hashed till the root value of the Merkle tree is generated.

Interactive Proof System: It is an interactive interaction between client and cloud where the data owner has to prove his identity

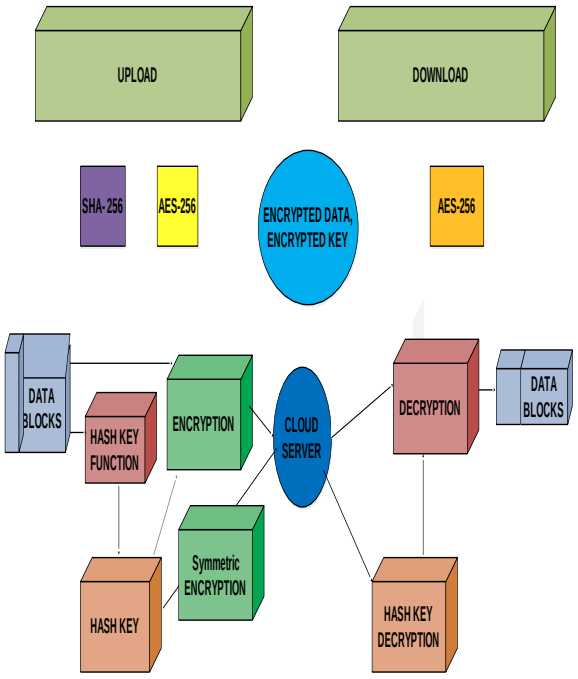
Cloud data storage: The data owner starts storage procedure by sending a client request verification message in order to check the uniqueness of the file by comparing it with data identifier in cloud database.

Advantages

- The data file is enciphered using symmetric encryption and use of asymmetric encryption for meta data files
- The Merkle tree properties is to support data deduplication as it employs pre-verification of data existence
- Unauthorized users cannot access data as one has to prove his ownership to the cloud.

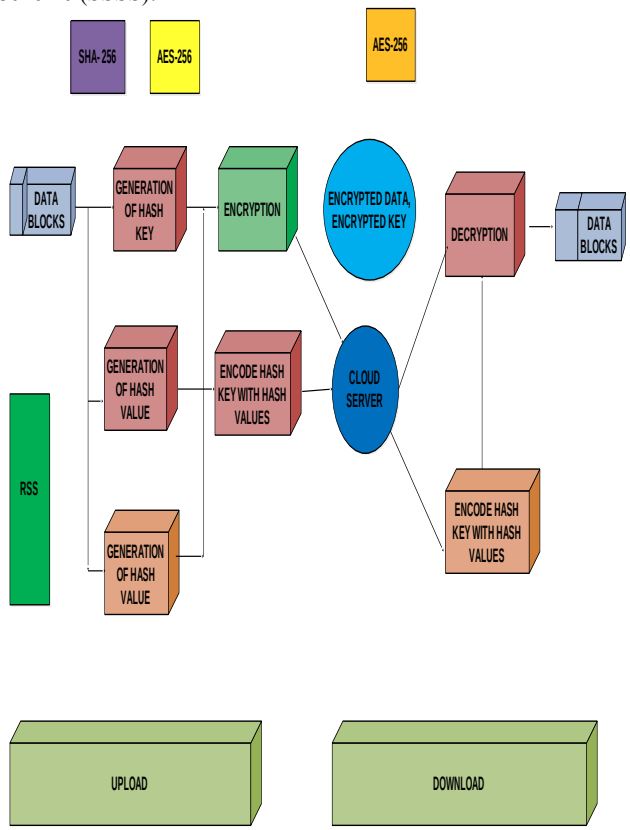
Disadvantages

- Each user has to store his private key in the cloud, if there are more number of users then it requires some amount of storage.



RSSS approach

The Ramp secret sharing scheme (RSSS) is used to store the convergent keys. Especially, the (n,k,r)-RSSS (where n>k>r≥0) generates n shares from a secret sharing where the secret can be obtained by using any k shares 2) The secret information cannot be considered from any of the r shares. when r=0, the (n,k,0)-RSSS becomes the (n,k,) Rabin's Information Dispersal Algorithm (IDA); when r=k-1, the (n,k,k-1)-RSSS becomes the (n,k,) Shamir's Secret Sharing Scheme (SSSS).



The (n,k,r)-RSSS builds on two primary functions:

- A secret is divided by the share into (k,r) pieces of equal size which produces random pieces r, and non systematic k of n erasure coding is used to encode k pieces into same size of shares n.
- The genuine secret can be obtained by recovering any k out of n shares.

Generated shares are made appropriate for deduplication by replacing random pieces with pseudorandom pieces in the implementation of this approach.

III. CONCLUSION

Client Side Deduplication eliminate duplicate which results in effective utilization of the resources such as storage space and bandwidth consumption instead of transmitting same data repeatedly. Deduplication has benefits like reduced infrastructure costs, management costs and reduced downtime. By using convergent encryption and merkle based deduplication can be done in a secured and efficient way.

REFERENCES

(1) S. Halevi, D. Harnik, B. Pinkas, and A. Shulman-Peleg. Proofs of ownership in remote storage systems. New York, NY, USA, 2011.
(2) M. W. Storer, K. Greenan, D. D. Long, and E. L. Miller "Secure data deduplication" In Proceedings of the 4th ACM International Workshop on Storage Security and Survivability, New York, NY, USA, 2008.
(3) C. Wang, Z. guang Qin, J. Peng, and J. Wang. A novel encryption scheme for data deduplication system. In Communications, Circuits and Systems (ICCCAS), 2010 International Conference on, pages 265–269, 2010.
(4) M. Dutch. Understanding data deduplication ratios. SNIA White Paper, June 2008.
(5) D. Harnik, B. Pinkas, and A. Shulman-Peleg "Side channels in cloud services: Deduplication in cloud storage". IEEE Security And Privacy, 2010.
(6) Nesrine Kaaniche, Maryline Laurent "A secure Client Side eduplication Scheme Cloud Storage Environments" 6TH International Conference On New Technologies, And Security Year 2014
(7) Jin Li, Xiaofeng Chen, Mingqiang Li, Jingwei Li, Patrick P.C. Lee, and Wenjing Lou "Secure Deduplication with Efficient and Reliable Convergent Key Management" IEEE Transactions On Parallel And Distributed Systems, JUNE 2014.