Client Side Deduplication Scheme on Encrypted Data in Cloud Environments

RJ. Poovaraghan[1], Bh. Krishna Abhishek[2], V. Srikar Varma[3], Y. Sai Kiran[4], B. Nikhileshwar Reddy[5]
[2,3,4,5]B.Tech (IV) Year Student , Department of Computer Science and Engineering
SRM Institute of Science and Technology, Chennai, India
[1,]Assistant Professor, Department of Computer Science and Technology
SRM Institute of Science and Technology, Chennai, Tmail Nadu, India

Abstract:- Cloud Storage services commonly utilize duplication, that will be beneficial in storing a block or file from reducing duplicate copies of data. De-duplication is beneficial in saving network bandwidth and storage space that's an edge to those clients or users from cloud. A new client-side de-duplication technique is useful for storage at a secure way and sharing data with different users at 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 indicate of regaining saving space is always to spot duplicate data files and expel them. It's a host not as file system that operates like a central fileserver and also this system will probably undoubtedly be there at a physically dispersed mode at a system that is a more set of workstations. From the true World if its venture planet, firm community or home exactly the largest problem can put away the info that's presently being generated. Lately, the speeding growth associated with digital contents is now gearing around raise demand of space for storage using an efficient use with this bandwidth and space to transport the data. The customers are moving their certain sections of these environment on the Cloud as they're seeking cost efficacy and cloud storage offers inexpensive architectures.

De-duplication Is most effective procedure, an activity of pinpointing and removing redundant info. From the customer side de-duplication data is replicated in the client side by which client sends simply fresh, specific data throughout the system that contributes to reduced memory capacity and network bandwidth savings. The advantages of de-duplication include paid down infrastructure expenses, reduced management fees, lots of cloud storage providers like drop-box, Memopal and Mozy utilize client-side de-duplication as a way to save funds that contributes to avoiding storage of data that is redundant in cloud-storage servers and network bandwidth savings from eliminating transmission of the same contents a few times. Although they're advantages in client-side de-duplication its problems associated with security, such as Fans will mostly aim the confidentiality and bandwidth that's linked to solitude of valid users. So as to address those questions, proof Settlement (PoW) approaches are introduced at which they allow the Storage server assess a customer data possession, primarily 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. RELATED 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 suggested by Halevi in order to steer clear of private data footprint, three distinct approaches were introduced primarily based on security and performances. In such customer has to establish to host he has true sibling avenues that are based on the leaves of this Merkle tree. Erasure programming is put by the very first strategy on origin file contents afterward your Merkle tree is assembled with this encoded form for a input. The next plot is an alternative for erasure programming, in which data document is prepossess working with an international hash functionality. The 3rd plot handled assumptions on security that result in designing of categories. Alas, the premise of this proof was a proper supply samples that the information document.

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 obtaining the substantial resource saving edges, PoW plans comes alongside numerous security challenges which can make an dangerous atmosphere 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 could be derived by Employing SHA 256 on 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

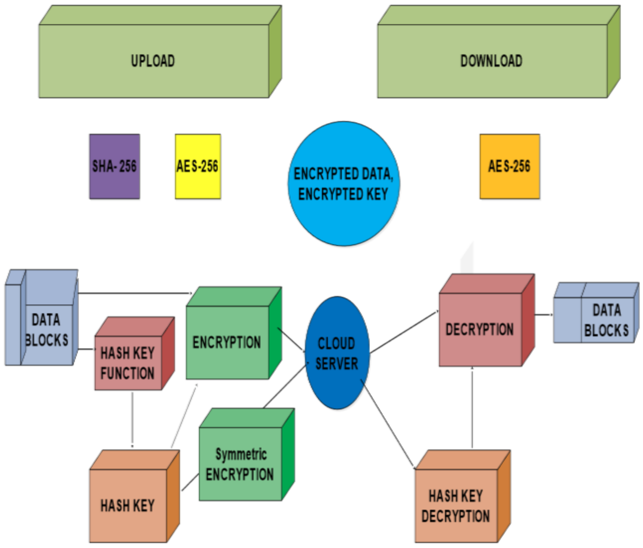
- *Convergent Encryption:* Convergent Encryption produces identical cipher text from identical plain files, convergent encryption is used to remove duplication files from the cloud storage while the cloud service provider has no access to encryption keys. Convergent encryption Has Been deriving keys by the Hash of text and offers a security tool for protected information de-duplication.
- *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.

B. Advantages

- The Info Document is enciphered using symmetric encryption and use of asymmetric encryption to get Meta Data files.
- The Merkle tree attributes would be to encourage information De-duplication since it uses Pre-verification of information presence.
- Unauthorized users cannot access data as one has to prove his ownership to the cloud.

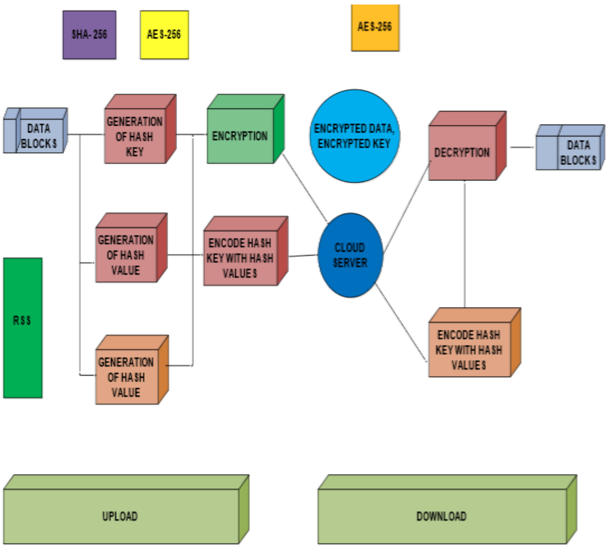
C. Disadvantages

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.



IV. 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.

V. 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. De-duplication has benefits like lower infrastructure costs, control expenses and reduced downtime. By employing convergent encryption and Merkle Based Tree established de-duplication can be carried out at a bonded and successful way.

REFERENCES

[1]. S. Halevi, D. Harnik, B. Pinkas, and A. Shulman-Peleg. Proofs of ownership in remote storage systems. 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 deduplication 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.