A Secure and Efficacious Data De-duplication System at Client Side in Cloud Computing

Payal A. Mulik*

Student, Department of Computer Science and Engineering, Center for PG Studies, Visvesvaraya Technological University, Belagavi, Karnataka, India

Abstract – Abstract – The world today is moving to digitization, it leads to exponential increase and growth of data. Cloud based services use de-duplication technique to remove duplicate copies of data. Deduplication technique is useful for clients of public cloud to decrease storage space and network bandwidth consumption. A new client side de-duplication system is proposed to securely store data in public cloud and only unique copies of contents are stored in a cloud database. To protect the privacy of contents while providing support to de-duplication, a convergent encryption and merkle-based tree concepts are used. The authorized users can download the file from cloud storage and secure search on encrypted data is provided.

Keywords- Cloud Storage, Deduplication, Security, Privacy, Proof of Ownership

---------------------------♦-----------------------------

I. INTRODUCTION

Recent years, the tremendous growth of digital data continues which leads to rise in new storage space and network bandwidth to transfer the data. Now-a-days, a bulky amount of data need to backup from mobile devices. To prevent from natural disaster backing up of data is necessary. As the technology is improving the data generated in enterprise area, business world or home is in huge amount and the major problem is to store the data. There is a demand for storage which is not more costly and has less bandwidth consumption.

The clients are moving to use secluded storage systems such as cloud storage as it provides cost reduce architectures. A technology called cloud computing is used to store numerous resources over the internet. The user does not worry about how to maintain and manage all the resources. Cloud computing allows access to large amount of content, without tackle of keeping huge storage and computing devices. It makes it possible to control, store and share huge amount of digital content over the internet. It provides highly manageable and scalable virtual servers, virtual networks, computing power and network bandwidth according to user need.

To save resources utilization in network bandwidth and storage capacities, many cloud service providers such as Wuala and Memopal provide the data de-duplication technique to the client side.

Deduplication is a technique which is most generally used for removing replicate copies of data in the cloud data storage to minimize the storage space and the network bandwidth. It is a process to store distinctive copy of replica contents. The benefits of this process are to reduce infrastructure costs, reduce network bandwidth consumption.

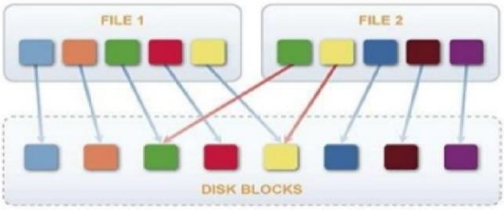

Fig 1: Deduplication

As we can see in Fig 1, recurring copies of data block is not store. It store distinct copy of spare contents. The link is provided to duplicate block. Client side Deduplication helps to eliminate recurring copies of data in cloud storage and reduces less transfer of data over network bandwidth.

Apart from these significant advantages client side De_duplication has issues related to security. For Example- Attackers may attack either bandwidth utilization or privacy and confidentiality of authorized cloud users.

To overcome these issues, Proof of Ownership (PoW) schemes are introduce in which base on the hash value, cloud server can test for the owner contents. The existing PoW scheme address some security problems but still we need to consider possible attacks such as the poison-attack and the Data outflow.

In secluded storage server, an owner stores his contents on the cloud database, retrieve the contents whenever he needs. Since the public cloud is not secure, the data is encrypted to prevent privacy of the user. But to access encrypted contents became complex. To deal with such situation searchable-encryption can be used.

Encryption makes it difficult to attackers to get the contents although it removes an entire search facility from the user. One method to get over it is, allow the cloud server to decrypt the contents. Server executes the query and outputs are sent to the user. But through this method server can get the plaintext contents and then encryption will not be helpful. Searchable encryption helps to have all different search options on the cloud server without decrypting the plaintext data.

Deduplication scheme use convergent encryption. It produces the same cipher data from the same plain data. In this, derivation of encryption key is done from hash of plain data.

In this project, a secure Client Side De_duplication technique is applied to eliminate recurring copies of data. A secure Proof-of-Ownership (PoW) model is developed by the means of Merkle-Based Tree and convergent encryption for providing protection in cloud storage systems. Efficient sharing is provided between users and only authoritative users can get access to the contents. To search over encrypted data on cloud system searchable encryption technique is used in this project.

The idea of project is to use Merkle-Based Tree over encrypted data which generate unique identifier of information and this identifier helps to check whether same data is present in cloud server and apply searchable encryption to securely search over encrypted data in cloud server.

II. LITERATURE SURVEY

2.1 Related Work on Deduplication

J.R. Doucer et al [3] in 2002 proposed a far-site model which helps to get security and the reliability by storing an encrypted duplicate of every file on the various stand-alone systems. The authors proposed the convergent encryption model. In far-site system if a use of conventional cryptosystem was made to encrypt the files, then the machine cannot identify whether the files are unique or duplicate because similar encrypted files will have diverse cipher-text. They introduce a cryptosystem named as convergent encryption. It generates same cipher-text from same plaintext files. The key is derived from hash of plaintext.

M.W. Storer el al [6] presented an efficient model for secure Deduplication of data. They provide a solution that provides space effectiveness and protection of data in single server and scattered storage systems. But there are some open areas in which security of data is needed to be considered. Their model does not provide any level of access such as user can only access the file if he has key with him. The model does not consist of multiple levels of permissions for users. It does not consider of leakage of data against malevolent users.

S. Halevi et al [4] introduce the idea of Proof-of-Ownership (PoW) in this client provides the proof that he is an owner of the requested file. It is a scheme in which server identifies whether the request is from data owner based on hash-value. If a user needs to upload to the cloud some information such as data file (D), he computes hash value and forward it to cloud storage. A server maintains hash-value of files received in its database. If any new file needs to be uploaded, cloud server verifies that the hash-value is present in cloud database. If not present in that case it allows the owner to upload the file. In database the hash value is already presented and user needs to upload file then cloud server tags that user as an owner of the file and there is no requirement to send that file.

R. Di. Pietro et al [2] address the problem of an adversary in which if partial part of original file is with user and claims to hold such a file. They implement an efficient proof of ownership technique. But, the disadvantage is privacy violation of snooping storage servers.

2.2 Related Work on Searchable Encryption

Song et al [7] started the research on searchable encryption and demonstrate the first cryptographic scheme to seek on encrypted text. They provide a secure encryption in which un-trusted server cannot know about plaintext when cipher-text is provided and un-trusted server can only know only search result not the entire plaintext. They provided solution using symmetric key technique.

Boneh et al [8], they first address to search the keywords using public key encryption technique. They introduce a scheme known as keyword search using public key encryption in which public key can be used by anyone but only authorized users he having secret key can search the data. But, the scheme search consists of only single keyword explore.

Golle et al [9] presents a gateway to learn each unique keyword based on result of conjunctive query. They demonstrate the first solution for keyword explores using symmetric conjunctive technique. In this, encrypted keywords are associated with encrypted file. A user can search several keywords using a trapdoor.

Lai et al [10] based on attribute – based encryption scheme presented a new searchable public key encryption scheme. But, in this scheme the trapdoor generated reveals the search keywords which allow the server to learn the keywords which are presented in encrypted data and not in the trapdoor.

2.3 Security Analysis

Different type of users may have identical important files even they do not know each other. For Example, they might have copyright file from same source. The some amount of information may be leaked by various channels. Apart from saving resource, Proof of Ownership protocol has several security issues which lead to leakage of sensitive data of users.

1) Data privacy leak –

An attack on hash-as-a-proof reveals important information. If the attacker comes to know by some means the short hash value of files which are present in the cloud storage then attacker can easily fool the cloud server that he is a possessor of the file by providing the hash value. The assailant can easily get access to important data.

The file when uploaded to server hash value of it is calculated and store in database for Deduplication check if attacker comes to know hash value of uploaded file then he can easily get the sensitive information of user.

2) Data privacy violation by cloud storage servers-

Confidentiality of user data is important and it must be secluded against cloud storage servers itself. Users consider that access of their data is unable by cloud storage servers. But, we to make sure that storage server should not access the data.

3) Collision Attack-

Suppose an encrypted File F, encryption key is chosen arbitrarily on client side, then the server could not validate among uploaded file and proof hash value. For example given a pair $(H_F\ ,C_F)$ where a server, is not capable to recognize whether an encryption key K and file F is present such that hash value H_F is from original file F. An attacker may replace an original cipher-text C_F. Consider a user Alice wants to send the similar file F to cloud storage. The cloud will display message to Alice that File F is already in cloud and uploading of file will be prevented. The user Alice may in future to save her local storage will delete her local file F and will download it whenever she required from cloud storage. But, she may download from cloud, malicious file and her original file is lost. In this attack, hash function collision is not needed and may take place if Deduplication is not implemented correctly on encrypted data.

III. ARCHITECTURE DIAGRAM

1) Cloud Service Provider (CSP)-

Cloud service provider has various facilities to manage database servers and to maintain cloud servers storage. A virtual environment is provided by CSP to host services. The services provided by CSP can be used to control contents stored by client in storage cloud servers.

2) Client-

It makes proper use of whatever resources are provided by cloud service provider. It uses resources to download, upload and share the content with several users.

3) Users-

They can access, data stored in cloud based on their authorization rights. They can write, read and store modified content in cloud.

Client Service Provider (CSP) gives a network interface for client to store its contents in cloud servers and that interface is used by users to download the content provided on cloud.

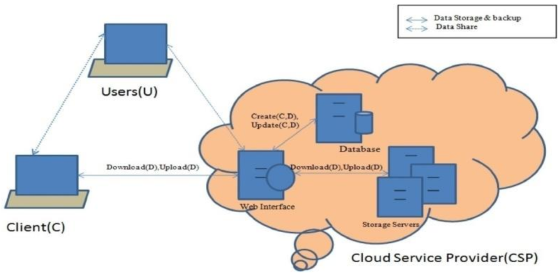

Fig 2: Cloud Storage Architecture

IV. PROPOSED SYSTEM

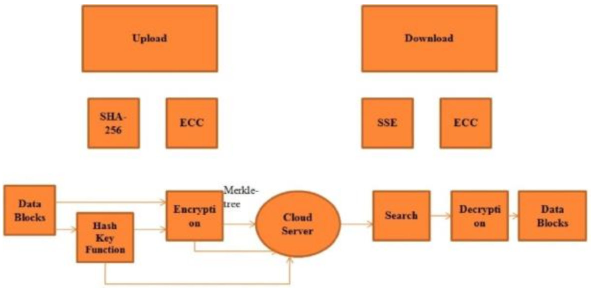

Fig 3: Block level diagram of proposed system

The process of proposed model is shown in above Fig 3. At the client side/data owner, the file (F) is selected by the owner which he needs to upload to cloud server. The file is separated into number of the blocks and conflict-resistant hash function is applied on it. The encryption key is derived after applying hash function. The key is encrypted with public key of cloud user. The hash function used is SHA-256. Then, on original content (F) asymmetric encryption is applied. The Elliptic Curve Cryptography (ECC) is used for encrypting of data. Then, on encrypted content merkle-tree is run which generate unique identifier which helps to check the uniqueness of file in cloud database.

Cloud server checks the identifier is already present in cloud database or not. If the identifier is already there then there is no requirement to upload the file by client. The cloud server displays message file already present or found. The cloud server link the another owner of that file. If the identifier is not present then cloud server allows the owner to upload the file.

It helps to save the storage space in cloud database and less consumption of network bandwidth. At the user side, he can download the file only if he is authorized user. Secure searchable encryption (SSE) is run to search the required file so that server cannot decrypt the file. After searching the required file, the decryption of content is done by applying Elliptic Curve Cryptography decryption technique and required blocks of information is obtained.

V. CONCLUSION

The requirement for secured storage space services in the cloud, amazing use of the merkle tree and the convergent encryption leads to data outsourcing in a secure and an effective manner.

The result consists of cryptographic use of asymmetric encryption which is make to encrypt the file and provide security against several intrusions and most important work is use of Merkle tree properties which helps for data Deduplication. It helps to eliminate replica copies of data and result in resource utilization in an effective manner such as bandwidth consumption and storage space. De-duplication consists of benefits such as reduce management costs, reduce infrastructure costs, and reduce storage space and low bandwidth consumption of network. A apply of secure searchable encryption helps to search over encrypted content and download the content in a secure way.

The secure uploading of file and removing replica copies of data and only authoritative users can download the file and search over encrypted contents is successfully implemented.

REFERENCES

https://github.com/openstack/swift.

R. Di Pietro and A. Sorniotti. Boosting efficiency and security in proof of ownership for Deduplication. In Proceedings of the 7th ACM Symposium on Information, Computer and Communications Security, ASIACCS'12, pages 81-82, New York, NY, USA, 2012. ACM.

J.R. Douceur, A. Adya, W.J. Bolosky, D. Simo, and M. Theimer. Reclaiming space from duplicate files in a serverless distributed file system. In Proceedings of 22nd International Conference on Distributed Computing Systems (ICDCS), 2002.

S. Halevi, D. Harnik, B.Pinkas and A. Shulman-Peleg. Proofs of ownership in remote storage systems. In Proceedings of the 18th ACM conference on Computer and communications security, CCS'11, pages 491-500, New York, NY, USA, 2011. ACM.

D. Hankerson, A.J. Menezes, and S.Vanstone. Guide to elliptic Curve Cryptography. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2003.

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, StorageSS '08, pages 1-10, New York, NY, USA, 2008. ACM.

Song, D.X., Wagner, D., Perrig, A.: Practical techniques for searcheson encrypted data. In: IEEE Symposium on Security and Privacy, S & P 2000, PP. 44-55(2000).

Boneh, D., Di Crescenzo, G.,Ostrovsk, R., Persiano, G.: Public Key encryption with keyword search. In: Cachin, C., Camenisch, J. (eds.) EUROCRYPT 2004. LNCS, vol.3027, pp. 506-522. Springer, Heidelberg (2004).

Golle, P., Staddon, J., Waters, B.: Secure conjunctive keyword Search over encrypted data. In: jakobsson, M., Yung, M. (eds.) ACNS 2004. LNCS, Vol. 3089, pp.31-45. Springer, Heidelberg (2004).

Lai, J., Zhou, X., Deng, R.H., Li, Y., Chan, K.: Expressive search on encrypted data. In: ACM ASIACCS 2013, pp. 243-252(2013).

Corresponding Author

Payal A. Mulik*

Student, Department of Computer Science and Engineering, Center for PG Studies, Visvesvaraya Technological University, Belagavi, Karnataka, India

E-Mail – payalmulik777@gmail.com