Confidential Deduplication with Effective and Trustworthy Convergent Key Management

Mr. Sudhir D. Chaskar[1]
[1] Department Of Computer Engineering
[1] P.K. Technical Campus, Chakan (Pune), India

Abstract- *Data outsourcing raises security and privacy issue of outsourced data. We must trust third-party (cloud providers) to properly enforce data confidentiality, integrity checking, and access control mechanisms against any theft attacks. On the other hand, by using deduplication improvement in storage and bandwidth efficiency is in compatible with traditional encryption. Now days, encryption requires different users to encrypt their data with their own keys. That's why identical data of different users will generate different cipher texts, which make deduplication impossible. While Convergent encryption provides a viable option to enforce data confidentiality while realizing deduplication. It encrypts/decrypts a data copy with a convergent key, which is generated by the cryptographic hash value of the blocks of the plain text itself. The cipher text is send to the cloud after key generation and data encryption & users retain the keys. Since in this encryption method identical data generate identical convergent keys & cipher text.*

Keywords- Deduplication, Convergent Encryption/Decryption Key, Proof of ownership and Storage Cloud Service Provider (S-CSP).

I. INTRODUCTION

Data Deduplication-

Data deduplication is enables companies to save a lot of money on storage costs to store the data and on the bandwidth costs to move the data when replicating it offsite for data retrieval. This is great news for cloud providers, because if you store less, you need less hardware utilization. In deduplication better utilization of storage space is achieved. If you store less, you also backup less, this again means less hardware and backup Media. The deduplication process removes blocks that are not unique

Deduplication simply put the process consists of four steps:

1. Divide the input data into blocks.
2. Calculate the Hash value for each block of data.
3. Use these Hash values to determine the duplication.
4. Replace the duplicate data with a reference to the object already in the database.

Convergent Encryption-

Convergent encryption is also known as content hash keying as the key generated from plain text by using hash key generation. In this cryptosystem identical cipher text produced from identical plaintext files. In cloud computing service provider can remove duplicate files without accessing encryption keys this is done using convergent encryption. In convergent encryption file attack is possible in which attacker can encrypt unencrypted file or plain text and then compare that encrypted file with files stored on cloud. This attack is possible while multiple users storing duplicate data on cloud, and that data would be available publically or already present on cloud. We can avoid file attack by simply using padding mechanism. In padding mechanism some unique piece of characters are added at the start and at the end of data string. If we are storing the data by creating blocks then we can apply padding mechanism to each block this will give more security to our data and file attack can be avoided.

II. RELATED WORK

In M. Bellare's Message Locked Encryption (MLE) and Secure Deduplication [1] Encryption & Decryption are done by using the key generated from the message itself. In this paper secure deduplication is achieved on the outsourced data. This also provides definition for privacy and integrity to a Message. As here encryption done by using conventional method so we can provide security to stored data but deduplication is impossible.

In A. T. Clements Decentralized Deduplication(DEDE) in SAN Cluster File System [2] solution provided against the problem of wasted storage space and increased storage array cache footprint due to the duplicate blocks present in virtual machines hosted by file system. Dedpulication is well implemented in centralized data cluster while Clements provide deduplicetion in decentralized data cluster. In this we don't require any central coordination among the decentralized data clusters. Host keeps summary of their data write to cluster file system, merge that summaries with other host. In DEDE deduplication done on metadata using general file system without understanding whole file so

recovery of original data is not achieved that's the big disadvantage of DEDE method.

In Cloud storage deduplication is used to reduce space and bandwidth requirements. Side channels in cloud services [3] make this phenomenon more effective by applying it across multiple users. In this paper Harnik propose the mechanism to reduce the risk of data leakage from cloud data stored by multiple users.

M. Li's Information Dispersal Algorithms [4] have been widely applied to reliable and secure storage and transmission of data files in distributed systems. An IDA is a method that encodes a file F of size L = |F | into n unrecognisable pieces F1, F2…. Fn, each of size L/m (m < n), so that the original file F can be reconstructed from any m pieces.

This paper makes a systematic study on the confidentiality of an IDA and its connection with the adopted erasure code. Two levels of confidentiality Weak confidentiality and strong confidentiality of data. This paper constructs an IDA with strong confidentiality from a Reed-Solomon code.

In this paper of W. K. Ng, a new notion which we call private data deduplication protocol [5] for private data storage is introduced. A private data deduplication protocol allows a client or user who holds a private data, proves to a server who holds a summary string of the data that he/she is the owner of that data. As this protocol gives security to data storage but recovery of data is critical task.

III. MATHEMATICAL MODEL

System S= {U, F, H, SE, CE, K, TG, PoW, SS}
Input:
U= Set of users.
F= Input File= FBig = FB1, B2. . .BNg
H=Hash key generation service.
H (F) =K
Where, F= Input file.
H (F) =Hash value of file F.
K= Hash Key.
SE= Symmetric Encryption/Decryption service.
CE=Convergent Encryption/Decryption service.
E (K, F) =C. . . . . (Encryption)
D (K, C) =F. . . . . .. (Decryption)
Where
F- File,
K-Hash Key,
E- Encryption
C-Cipher Text
TG=Tag generation service.
TG (F) =T. . . . . .
(TagGeneration)
Where
F- File
T-Tag of File F
PoW= Proof of ownership service.
SS=Storage Service (upload, download)

IV. EXISTING SYSTEM

Convergent encryption provides a viable option to enforce data confidentiality while realizing deduplication. In convergent encryption we are using convergent key for encryption/decryption which is cryptographic hash value of content of data copy itself. Once the encryption done user sends cipher text to cloud and takes the key. As here convergent key is generated from original data the convergent key & cipher text is identical for identical data copies. That's why we can apply deduplication on cipher text easily. The decryption can only done by the corresponding data owners with their convergent keys. That is, the plain text is first encrypted with a convergent key derived by the original text itself, and the convergent key is then encrypted by a master key. Master will be kept at local server securely by each user. The encrypted convergent keys are then stored, with the corresponding cipher text, in cloud storage. By using master key we can decrypt convergent key and then cipher test. Simply, each user only needs to keep the master key and the metadata about the outsourced data.

V. PROPOSED SYSTEM

As in our system Encryption/Decryption is done by convergent key and that was generated using plain text itself by using cryptographic hash function. After key generation and data encryption, users retain the keys and send the cipher text to the cloud. Encryption process generates identical encrypted text/cipher text and convergent key also. This allows the cloud to perform deduplication on the cipher texts. The Encrypted texts can only be decrypted and again plain text will be generated by the corresponding data owners with their convergent keys.

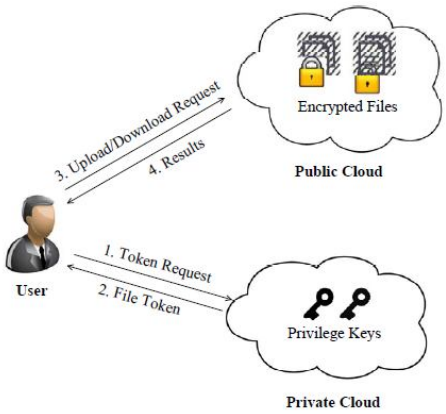

Figure 1. PROPOSED SYSTEM

As shown in above figure 1 our system has three entities that are user, Private cloud & S-CSP. Duplicate files or data is removed that is deduplication is done by S-CSP. The access to file is given by Token which is in the form of short Message. That token also works as Tag with specified Privileges. User can access private cloud server.

S-CSP

The data storage service in public cloud is provided by S-CSP on behalf of the User. Actually deduplication is done by S-CSP by removing redundant data which reduce the cost of storage. In this paper we assume s-CSP is online through out with very huge storage capacity and computation power too.

Data User-

A User is an entity which stores data to Cloud and access saved data for later use. In normal storage system deduplication is achieved by uploading only unique data which may be owned by one or multiple users. Uploading unique data saves upload bandwidth.

Private Cloud-

Private cloud is a new mechanism came into existence for cloud users. User requires cloud services when resources available at data user's or owner's end are restricted or limited and when public cloud is not trusted in practice then private cloud is best option to go. Private cloud also works as interface between User & Public cloud. Private cloud manages Private keys & allows users to submit files and queries.

Deployment models

a. Private cloud

Private cloud is cloud used in single organisation, which is managed and hosted internally or by a third party.

b. Public cloud

When cloud is open for public is called Public cloud. Public cloud provides services in zero cost or user has to pay as per usage. Public and private clouds are not much different. Only security considerations varies for different service like application, storage and other resources that are made available by service provider for public users & communication is effected over non-trusted network.

c. Hybrid cloud [Private + Public] Architecture

In cloud architecture multiple cloud components communicates with each other over messaging queue mechanism which is loose coupling mechanism.

VI. CONCLUSION

The proposed system is used for deduplication purpose with efficient and reliable key management. Dekey method is used to achieve deduplication among convergent keys. Security of convergent keys and confidentiality of outsourced data is preserved by servers where convergent keys are distributed.

REFERENCES

[1] L. Zhang and P. Anderson, ''Fast and Secure Laptop Backups with Encrypted Deduplication,'' inProc. USENIX LISA, 2010, pp. 1-8.

[2] A. Shulman-Peleg, S. Halevi, D. Harnik and B. Pinkas, ''Proofs of Ownership in Remote Storage Systems,'' in Proc.ACM Conf. Comput.Commun. Security, Y. Chen, G. Danezis, and V. Shmatikov, Eds., 2011, pp. 491-500.

[3] T. Ristenpart and M. Bellare, S. Keelveedhi, ''Message-Locked Encryption and Secure Deduplication,'' inProc. IACR CryptologyePrint Archive, 2012, pp. 296-3122012:631.

[4] A. Shulman-Peleg, D. Harnik and B. Pinkas, ''Side Channels in Cloud Services: Deduplication in Cloud Storage,'' IEEESecurityPrivacy, vol. 8, no. 6, pp. 40-47, Nov. /Dec. 2010.

[5] K. Lauterand S. Kamara, ''Cryptographic Cloud Storage,'' inProc. Financial Cryptography: Workshop Real-Life Cryptograph. Protocols Standardization, 2010,

pp. 136-149.

[6] D. Reinsel and J. Gantz, The Digital Universe in 2020: Big Data, Bigger Digital Shadows, Biggest Growth in the Far East, Dec. 2012. [Online]. Available: http://www.emc.com/collateral/analystreports/idc-the-digital-universe-in-2020.pdf.

[7] M. Vilayannur, J. Li, I. Ahmad and A.T. Clements, ‘‘Decentralized Deduplication in San Cluster File Systems,’’ in Proc.USENIX ATC, 2009.

[8] Amazon Case Studies. [Online]. Available: https://aws.amazon.com/solutions/case-studies/#backup.

[9] M. Li, ‘‘On the Confidentiality of Information Dispersal Algorithms and their Erasure Codes,’’ inProc. CoRR, 2012, pp. 1-4abs/1206.4123.