Securing Data Deduplication Via Hybrid Cloud

Anjali U

***Abstract*—Deduplication is the method of eliminating replicas of data. This technique is widely used in cloud computing to reduce storage space and bandwidth. Even though data deduplication brings a lot of benefits, it does not provide any security measures for users sensitive data. To protect these sensitive data from insiders and outsiders attacks, a hybrid cloud approach is used. Convergent encryption technique has been proposed to protect the confidentiality of sensitive data while supporting deduplication, which encrypt the data before outsourcing. The unauthorized access will be prevented by proof of ownership protocol.**

***Index Terms* —Convergent Encryption, Deduplication, Hybrid cloud, Proof of Ownership.**

I. INTRODUCTION

Cloud computing provides unlimited resources to users as a services across the Internet by hiding their platform and implementation details. The cloud computing contains large amount of data's and shared by users with different privileges. One important challenge of cloud storage is the management of large volume of data. Deduplication is considered as a well identified technique for the scalable management of data in cloud computing. Data deduplication is a data compression technique for eliminating duplicate copies of repeating data in cloud storage. The technique is used to improve storage utilization and to reduce bandwidth. By keeping a single physical copy and referring other data related to that copy deduplication removes the redundant data.

In deduplication, multiple data copies with the same content are not saved. Either file level or block level can take place in deduplication. For file level depuplication duplicate copies of same file will be removed. In block level deduplication, duplicate blocks of data which occur in non identical files will be eliminated. Here we can use the concept of hybrid cloud, which will remove data duplication and maintains confidentiality in a better way. Hybrid cloud, a combination of public and private Cloud combines the advantages of scalability and reliability. It also supports potential cost savings of public cloud storage with the security and full control of private cloud storage.

II. PRELIMINARIES

A. Symmetric Encryption

Symmetric encryption is also referred to as conventional encryption or single-key encryption in which both encryption and decryption can be performed using a same key. Symmetric encryption scheme can be described by three primitive functions:

Manuscript received March 02, 2015.

Anjali U, Department Of Computer Science And Engineering, M . G University, Mount Zion College of Engineering, Pathanamthitta, India, 9544581881.

• KeyGenSE(1^{λ}) -> κ is the key generation algorithm that generates secret key κ using security parameter 1^{λ}

• EncSE(κ,*M*) -> *C* is the symmetric encryption algorithm that transforms original message M into ciphertext *C* using a secret key κ

• DecSE(C,κ) -> M is the symmetric decryption algorithm that transforms ciphertext C into the original message M by using key κ.

B. Convergent Encryption

Convergent encryption is also referred to as content hash keying, in which identical ciphertext are produced from identical plaintext files. This convergent encryption scheme is used in cloud computing to remove duplicate copies of repeating data's in cloud storage and provides data confidentiality in deduplication. A user (or data owner) be capable of deriving a convergent key from each original copy of data and encrypts the data copy with the convergent key. Also the user can derive a *t*ag for the data copy, to detect duplicates. If two data copies are the same, then their tags are also same. Both convergent key and the tag are independently derived.

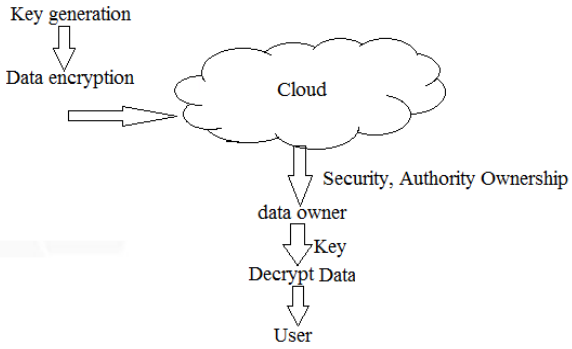

Fig: Confidential Data Encryption

A convergent encryption scheme can be defined by four primitive functions:

• KeyGenCE(*M*)->*K* is the key generation algorithm that maps a data copy *M* to a convergent key *K*;

• EncCE(*K*,*M*) -> *C* is the symmetric encryption algorithm that generate cipher text *C as output* by encrypting both the convergent key *K* and the data copy *M* as inputs

• DecCE(*C*,*K*) -> *M* is the decryption algorithm that generate the original data copy *M* by decrypting the convergent key *K* and the cipher text *C* .

• TagGen(*M*) -> *T*(*M*) is the tag creation algorithm that maps the original data copy *M* and outputs a tag *T*(*M*).

C. Proof Of Ownership

To prevent unauthorized access, the proof of ownership (PoW) protocol enables users to prove their ownership of data copies to the storage server. The PoW is implemented as an interactive algorithm, run by a user (prover) and a storage server (verifier). The verifier derives a value $\phi(M)$ from the

data copy *M*. To prove the ownership of the data copy *M* the Prover needs to send ϕ' to the verifier, where $\phi' = \phi(M)$.

D. Identification Protocol

An identification protocol can be described in two stages: Proof and Verify. In the stage of Proof, prover/user U needs to demonstrate his identity by performing some identification proof related to his identity to a verifier. The input provided by the prover/user is his private key sk_U that he would not like to share with the other users, such as private key of a public key in his certificate or credit card number etc. The verifier performs the verification with input of public information pk_U related to private key sk_U. Finally the verifier outputs either accept or reject to denote whether the proof is passed or not.

III. SYSTEM MODEL

A. Hybrid Architecture For Secure Deduplication

Hybrid cloud approach is introduced to solve the problems of deduplication along with differential privilages in cloud computing environment. This hybrid cloud contains of both private and public cloud. Private cloud acts as a proxy cloud to allow data owners and users to securely perform duplicate using differential privileges. User will store their data on public cloud while the data operation will be controlled by private cloud. Only the users with corresponding privilages can perform duplicate check. Users have access to private cloud, which perform duplicable encryption by generating file tokens for requesting users. User can upload and download the files from public cloud but private cloud provides the security for that data, i.e., only the authorized person can upload and download the files from the public cloud.

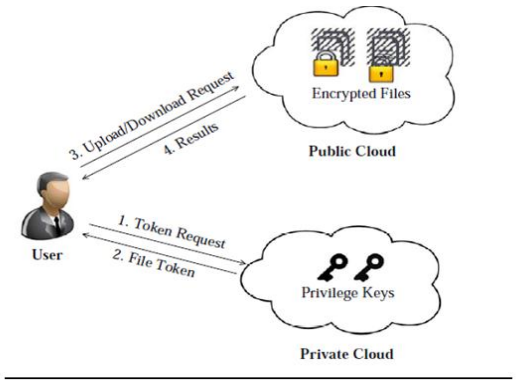

Fig. Architecture for Authorized Deduplication

There are three entities defined in our hybrid cloud architecture, they are Storage-Cloud Service Provider(S-CSP) in Public Cloud , Data users and Private Cloud.

S-CSP: This entity provides the data storage service in public cloud. On behalf of the users the S-CSP provides the data outsourcing service and stores data. The S-CSP eliminates the storage of redundant data via deduplication and keeps only unique data to reduce the storage cost. S-CSP is always considered as online and has abundant storage capacity and computation power.

Data Users: In this system user is an entity who wants to outsource data into S-SCP and to access the data or files from S-SCP. In a storage system supporting deduplication, the user only uploads unique data to save the upload bandwidth. Each file is protected by the convergent encryption key and privilege keys to recognize the authorized deduplication. Each user is assigned with a set of privileges for example, if we define a role based privilege according to job positions (eg. Director, technical lead and engineer), or we define a time-based privileges that indicate validity for accessing a file. A user, say Bob, may be assigned two privileges “Engineer” and “access right valid till "2016-12-10”, so that Bob can access any file with access role “Engineer” and accessible till "2016-12-10".

Private Cloud: To facilitate user’s secure use of cloud services this new entity is introduced. Private cloud manages the private keys for privilege, which supply the file token to users. Private cloud is able to provide data user/owner with an execution environment and infrastructure working as an interface between user and the public cloud as the computing resources at data user/owner side are restricted and the public cloud is not fully trusted. This interface allows user to submit files and queries that are securely stored and computed respectively.

Public Cloud: This entity provides the data storage service. It reduces storage cost by using data deduplication technique.

B. Working

File Uploading: For uploading a file to the public cloud, the user first needs to encrypt his file or data copy with a convergent key, by computing the cryptographic hash value of the data copy. Users must retain the keys after key generation and data encryption, and then send the generated ciphertext to the public cloud. Since the encryption operation is deterministic, identical data copies generate the same convergent key and hence the same ciphertext is obtained.In such an authorized deduplication system, each file uploaded to the cloud is based on a set of privileges to specify which users are able to perform the duplicate check and access the files.

The user needs to take this file and his own privileges as inputs, before submitting his duplicate check request for some file. The user is capable to find a duplicate for this file if and only if there exists a same copy of this file and a corresponding privilege stored in cloud. Before performing duplicate check with the S-CSP, the data owner needs to interact with the private cloud. In particular, the data owner performs an identification to prove its identity with private key. If the data owner’s identification is valid then the proof is accepted and the private cloud server will discover the corresponding privileges *PU* of the user from its stored table list. The user calculates and sends the file tag ϕ to the private cloud server. To prove the ownership of the data copy, who will return back token ϕ' to the user. To interact with public cloud the user send the file token ϕ' to S-CSP.

The user needs to run the PoW protocol with the S-CSP to prove their ownership of file, If a file duplicate is found. If the proof is accepted a pointer for the file will be provided to the user along with a signature and a time stamp as a proof from

the S-CSP. In addition to this proof the user sends a privilege set for the file to the private cloud server. After receiving the request from user, the private cloud server first verifies the proof from the S-CSP. If the proof is passed, the private cloud server computes token and sends to user. The user will also uploads these tokens of the file to the private cloud server. Otherwise, a proof (signature and a time stamp) from the S-CSP will be returned, if no duplicate is found. In addition to the proof the user also sends the privilege set to the private cloud server. First the private cloud server verifies the proof from the S-CSP, after receiving the request. If it is approved, the private cloud server generates token. At last, by using the convergent key user encrypt the file and uploads with privilege.

File Downloading: User first need to sends a request and the file name to the S-CSP To download a file. After receiving the request and file name from the user, the S-CSP checks whether the user is qualified to download the file. If the user is not eligible to download the file, the S-CSP sends an abort message to the user to indicate the download failure. Otherwise, the S-CSP sends back the corresponding cipher text of the file .Upon receiving the cipher text from the S-CSP, the user decrypt the original data copy or file with the help of same convergent key used for encryption.

IV. CONCLUSION

The main idea of our system is to achieve data compression by eliminating repeating data's along with better confidentiality and security in cloud computing by using the concept of hybrid cloud architecture. The convergent encryption technique has been proposed to protect the confidentiality of sensitive data by encrypting the user's data before outsourcing. This paper makes the first attempt to formally address the problem of authorized data, to enhance data security. Security analysis proved that our system is secure in terms of both insiders and outsiders attacks .It also provide a secure proof of ownership protocol to prevent illegal access.

ACKNOWLEDGEMENT

With immense pleasure, I am publishing this paper as a part of the curriculum of M.Tech. Computer Science And Engineering. It gives me proud privilege to complete this paper work under the valuable guidance of Principal for providing all facilities and help for smooth progress of paper work. I would also like to thank all the Staff Members of Computer Science Department, Management, friends and family members, for providing the necessary guidance and serious advice for the preparation of this paper.

REFERENCES

[1] OpenSSL Project. http://www.openssl.org/.
[2] P. Anderson and L. Zhang. Fast and secure laptop backups with encrypted de-duplication. In Proc. of USENIX LISA, 2010.
[3] M. Bellare, S. Keelveedhi, and T. Ristenpart. Dupless: Serveraided encryption for deduplicated storage. In USENIX Security Symposium, 2013.
[4] GNU Libmicrohttpd. http://www.gnu.org/software/libmicrohttpd/.
[5] libcurl. http://curl.haxx.se/libcurl/.