A NOVEL MECHANISM TO CHECK DATA DEDUPLICATION IN A SECURED CLOUD

Shivganga S Mujgond Student
Information Technology
Sinhgad College of Engineering
Pune, India
smujgond@gmail.com

Dr. Nilesh J Uke Professor & Head
Information Technology
Sinhgad College of Engineering
Pune, India
Nilesh.uke@gmail.com

***Abstract*-- With the enormous creation of data in the day to day life, storing it costs a lot of space, be it on a personal computer, a private cloud, a public cloud or any reusable media. The storage and transfer cost of data can be reduced by storing a unique copy of duplicate data. This gives birth to data deduplication, is one of the important data compression techniques and has been widely used in cloud storage to reduce the amount of storage space and save bandwidth. However it leads to high cost in terms of new security and privacy challenges while protecting sensitive data. The proposed system uses convergent encryption technique which assures block level deduplication and data confidentiality at a same time. As a requirement for data duplication at block level raises an issue with block creation, block comparison and key management system suggests including new components in order to implement these issues for each block together with the actual deduplication operation. In addition to this system uses another component which takes care of authenticity of users and data confidentiality. The proposed system shows that overhead introduced by these components is minimal and does not impact the overall storage and computational costs. The results were compared against file level data deduplication and encryption system which refers to data copy as a whole a file and it eliminates the storage of any redundant files. However this system cannot identify two or more files with slightly modified data thus making redundant copies of the identical data which can be overcome by using block level data deduplication where user performs block level duplicate check and identify unique blocks to be uploaded by encrypting them.**

Keywords- cloud; deduplication; convergent encryption

I. INTRODUCTION

With the potentially infinite storage space offered by cloud, users tends to use as much space as they can and vendors constantly look for techniques aimed to minimize redundant data and maximize space savings. A well know technique which has been widely adopted is data deduplication, is a data compression technique for eliminating duplicate copies of repeating data. The technique is used to improve storage utilization and can be also applied to network data transfers to reduce the no of bytes that must be sent. Simple idea behind deduplication is to store duplicate data only once. Deduplication can take place at either file level or block level. For a file level deduplication it eliminates the duplicate copies of the same file. For block level deduplication it eliminates duplicate block of data that occur in non identical files.

Along with low ownership cost and flexibility, users require the protection of their data and confidentiality guarantees through encryption. Unfortunately deduplication and encryption are two conflicting technologies. While the aim of deduplication is to detect identical data segments and store them only one, the result of encryption is to make to identical data segments indistinguishable after being encrypted. This means that if data are encrypted by users in a standard way the cloud storage provider cannot apply deduplication since two identical data segments will be different after encryption. On the other hand, if data are not encrypted by users, confidentiality cannot be guaranteed and data are not protected against curious cloud storage providers.

A technique which has been proposed to meet these two conflicting requirements is convergent encryption (CE). It encrypts or decrypts a data copy with a convergent key, which is obtained by computing the cryptographic hash value of the content of the data copy. After key generation and data encryption, users retain the keys and send the cipher text to the cloud. In this scheme, identical data copies will generate the same convergent key and hence the same cipher text. To prevent unauthorized access a secure proof of ownership protocol is also needed. To provide the proof that the user indeed owns the same file when duplicate is found. After the proof, subsequent users with the same file will be provided a pointer from the server without needing to upload the same file. A user can download the encrypted file with the pointer from the server, which can only be decrypted by the corresponding data owners with their convergent keys. Thus, convergent encryption allows the cloud to perform deduplication on the cipher texts and the proof of ownership prevents the unauthorized user to access the file.

II. LITERATURE REVIEW

A. *Cloud Computing*

Cloud computing is the delivery of computing as a service rather than a products, whereby shared resources, software and information are provided to computers and other devices as a utility over a network. Clouds can be classified as public, private or hybrid. Cloud offers features such as on demand capabilities, resource pooling, broad network access, rapid elasticity, measured service. Cloud offers following service models:

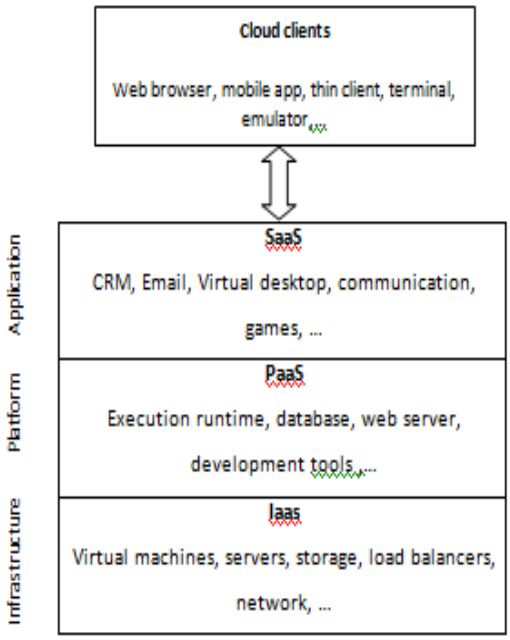

Figure.1 Cloud Computing offers three different service models

B. *Features of Cloud Computing*

- On-demand capabilities: A business will secure cloud-hosting services through a cloud host provider which could be your usual software vendor. You have access to your services and you have the power to change cloud services through an online control panel or directly with the provider. You can add or delete users and change storage networks and software as needed. Typically, you are billed with a monthly subscription or a pay-for-what-you-use scenario. Terms of subscriptions and payments will vary with each software provider.
- Broad network access: Your team can access business management solutions using their smart phones, tablets, laptops, and office computers. They can use these devices wherever they are located with a simple online access point. This mobility is particularly attractive for businesses so that during business hours or on off-times, employees can stay on top of projects, contracts, and customers whether they are on the road or in the office. Broad network access includes private clouds that operate within a company's firewall, public clouds, or a hybrid deployment.
- Resource pooling: The cloud enables your employees to enter and use data within the business management software hosted in the cloud at the same time, from any location, and at any time. This is an attractive feature for multiple business offices and field service or sales teams that are usually outside the office.
- Rapid elasticity: If anything, the cloud is flexible and scalable to suit your immediate business needs. You can quickly and easily add or remove users, software features, and other resources.
- Measured service: Going back to the affordable nature of the cloud, you only pay for what you use. You and your cloud provider can measure storage levels, processing, bandwidth, and the number of user accounts and you are billed appropriately. The amount of resources that you may use can be monitored and controlled from both your side and your cloud provider's side which provides transparency.

C. *Deduplication*

According to the data granularity, deduplication strategies can be categorized into two main categories: file level deduplication and block level deduplication, which is now days, a most common strategy. In block based deduplication, the block size can either fixed or variable. Another categorization criterion is the location at which deduplication is performed: if data are deduplicated at the client then it is called source based deduplication, otherwise target based. In source based deduplication the client first hashes each data segment he wishes to upload and sends those results to the storage provider to check whether such data are already stored: those only "un deduplicated" data segments will be actually uploaded by the user.

D. *Convergent Encryption*

- Advanced Encryption Standard: Symmetric encryption uses a common secret key κ to encrypt and decrypt information. AES is a symmetric block cipher that can encrypt data blocks of 128 bits using symmetric keys 128,192, or 256. AES encrypt the data blocks of 128 bits in 10, 12, and 14 rounds depending on the key size. Brute force attack is the only effective attack known against this algorithm. AES encryption is fast and reliable.
- Convergent Encryption: Convergent Encryption provides data confidentiality in deduplication. A user derives a convergent key from each original data copy and encrypts the data copy with the convergent key. In addition, the user also derives a tag for the data copy, such that the tag will be used to detect duplicates. Here, we assume that the tag correctness property holds, i.e., if two data copies are the same, then their tags are the same. To detect duplicates, the

user first sends the tag to the server side to check if the identical copy has been already stored. Note that both the convergent key and the tag are independently derived and the tag cannot be used to deduce the convergent key and compromise data confidentiality. Both the encrypted data copy and its corresponding tag will be stored on the server side. Formally, a convergent encryption scheme can be defined with four primitive functions:

- ✓ $keyGen_{CE}(M) \rightarrow K$ is the key generation algorithm that maps a data copy M to a convergent key K;
- ✓ $Enc_{CE}(K,M) \rightarrow C$ is the symmetric encryption algorithm that takes both the convergent key K and the data copy M as inputs and then outputs a cipher text C;
- ✓ $Dec_{CE}(K,C) \rightarrow M$ is the decryption algorithm that takes both the cipher text C and the convergent key K as inputs and then outputs the original data copy M; and
- ✓ $TagGen(M) \rightarrow T(M)$ is the tag generation algorithm that maps the original data copy M and outputs a tag T(M).

E. Identification Protocol:

An identification protocol II can be described with two phases: Proof and Verify. In the stage of Proof, a prover/user U can demonstrate his identity to a verifier by performing some identification proof related to his identity. The input of the prover/user is his private key skU that is sensitive information such as private key of a public key in his certificate or credit card number etc. that he would not like to share with the other users. The verifier performs the verification with input of public information pkU related to skU. At the conclusion of the protocol, the verifier outputs either accept or reject to denote whether the proof is passed or not. There are many efficient identification protocols in literature, including certificate-based, identity-based identification etc.

III. PROBLEM DEFINITION

Aiming at efficiently solving the problem of deduplication with differential privileges in cloud computing, we consider a hybrid cloud architecture consisting of a public cloud and a private cloud. Previous deduplication systems cannot support differential authorization duplicate check, which in important in many applications. In such an authorized deduplication system, each user is issued a set of privileges during system initialization, each file uploaded to the cloud is also bounded by a set of privileges to specify which kind of users is allowed to perform the duplicate check request for some file, the user needs to take this file and his own privileges as inputs. The user is able to find a duplicate for this file if and only if there is a copy of this file and a matched privilege stored in cloud. For example, in a company, many different privileges will be assigned to employees. In order to save cost and efficiently management, the data will be moved to the storage server provided (S-CSP) in the public cloud with specified privileges and the deduplication technique will be applied to store only on copy of the same file. Because of privacy consideration, some files will be encrypted and allowed the duplicate check by employees with specified privileges to realize the access control. Traditional deduplication systems based on convergent encryption, although providing confidentiality to some extent; do not support the duplicate check with differential privileges. In other words, no differential privileges have been considered in the deduplication based on convergent encryption technique. It seems to be contradicted if we want to realize both deduplication and differential authorization duplicate check at the same time.

IV. COMPONENTS OF PROPOSED SYSTEM

Unlike existing data deduplication systems, the private cloud is involved as a proxy to allow data owner/users to securely perform duplicate check with differential privileges; such architecture is practical and has attracted much attention from researchers. The data owners only outsource their data storage by utilizing public cloud while the data operation is managed in private cloud. A new deduplication system supporting differential duplicate check is proposed under this hybrid cloud architecture where the S-SCP resides in the public cloud. The use in only allowed to perform the duplicate check for files marked with the corresponding privileges.

Furthermore, system will be enhanced in security specifically using an advanced scheme to support stronger security by encrypting there file with differential privilege keys. In this way, the users without corresponding privileges cannot perform the duplicate check. Furthermore, such unauthorized users cannot decrypt the cipher text even collude with the S-CSP. Security analysis demonstrates that our system is secure in terms of the definitions specified in the proposed security model.

Finally, a prototype of the proposed authorized duplicate check will be implemented and testbed experiments will be conducted to evaluate the overhead of the prototype. Proposed system shows that the overhead is minimal compared to the normal convergent encryption and file upload operations.

A. Hybrid System Architecture for Secure Deduplication

At a high level, our setting of interest is an enterprise network, consisting of a group of affiliated clients who will use the S-CSP and store data with deduplication technique. In this setting, deduplication can be frequently used in these settings for data backup and disaster recovery applications while greatly reducing storage space. Such systems are

widespread and are often more suitable to user file backup and synchronization applications than richer storage abstraction. There are three entities defined in our system, that is, users, private cloud and S-CSP in public cloud as shown in Fig. 2 The S-CSP performs deduplication by checking if the contents of two files are the same and stores only one of them.

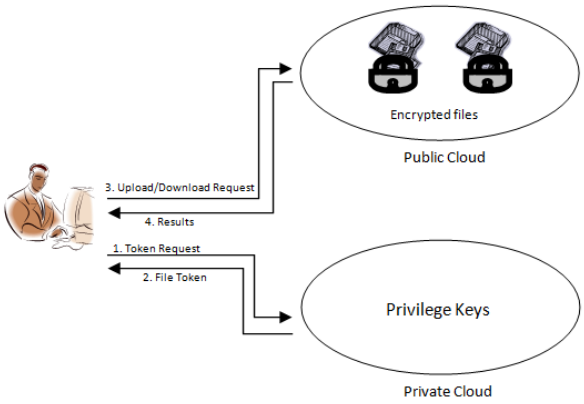

Figure.2 Architecture for Authorized Deduplication

The access right to a file is defined based on a set of privileges. The exact definition of a privilege varies across applications. For example, we may define a role-based privilege according to job positions (e.g. Director, Project Lead and Engineer), or we may define a time-based privilege that specifies a valid time period within which a file can be accessed. A user, say Alice, may be assigned two privileges "Director" and "access right valid on 2014-01-01", so that she can access any file whose access role is "Director" and accessible time period covers 2014-01-01. Each privilege is represented in the form of a short message called token. Each file is associated with some file tokens, which denote the tag with specified privileges. A user computes and sends duplicate-check tokens to the public cloud for authorized duplicate check.

Users have access to the private cloud server, which will aid in performing deduplication encryption by generating file tokens for the requesting users. We will explain further the role of the private cloud server below. Users are also provisioned with per-user encryption key and credentials. (e.g. user certificates). In this system, block level deduplication is considered which eliminates the storage of any redundant files. Each data copy is associated with a token for the duplicate check.

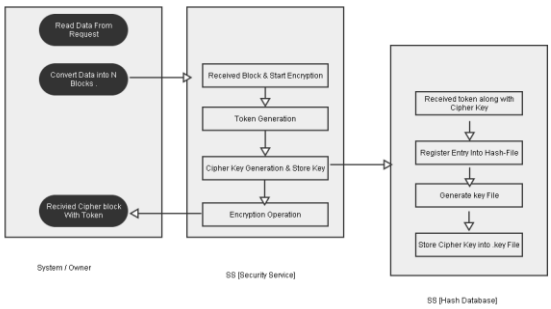

Figure.3 System Workflow

- S-CSP (Storage Management): This is an entity that provides a data storage provider in public cloud. The S-CSP provides the data outsourcing service and stores data on behalf of the users. To reduce the storage cost, the S-CSP eliminates the storage of redundant data via deduplication and keeps only unique data. In this system it is assumed that S-CSP always be online and has abundant storage capacity and computation power.
- Data Users: A user is an entity that wants to outsource data storage to the S-CSP and access the data later. In a storage system supporting deduplication, the user only uploads unique data but does not upload any duplicate data to save the upload bandwidth, which may be owned by the same user or different users. In the authorized deduplication system, each user is issued a set of privileges in the setup of the system. Each file is protected with the convergent encryption key and privilege keys to realize the authorized deduplication with differential privileges.
- Security Service (Private Cloud): Compared with the traditional deduplication architecture in cloud computing, this is a new entity introduced for facilitating users secure usage of cloud service. Specifically, since the computing resources at data user/owner side are restricted and the pubic cloud is not fully trusted in practice, 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. The private keys for the privileges are managed by the private cloud, who answers the file token requests from the users. The interface offered by the private cloud allows user to submit files and queries to be securely stored and computed respectively.
- Hashmap: HashMap works on the principal of hashing. It stores values in the form of key,value pair and to access a value you need to provide the key. For efficient use of HashMap the 'key' element should implement equals() and hashcode() method. equals() method define that two objects are meaningfully equal. hashcode() helps HashMap to arrange elements separately in a bucket. So elements with same hascode are kept in the same bucket together. So when

we want to fetch a element using get(K key), HashMap first identifies the bucket in which all elements of the same hascode as the hashcode of the 'key' passed are present. Than it uses the equals() method to identify the actual object present in the bucket.

B. *Dynamic Operations*

Below are the dynamic operations present in the proposed system

- Insertion

We have received the request from client to insert file (F). data owner wants to divide the multiple blocks ΣB_{ij} as system as decided that single block (B_i) is 4KB i.e. total no of blocks per file (F) = size (F) / 4096

Once the request has been received from the client file (F) is divided onto the F = {b_1, b_2, b_3, b_4,.....b_n}

For each block (B_i) we perform encryption operation and generate below response.

- ✓ Cipher Text (B_i)
- ✓ Token (Ti Bi) [16-bit, unique token for Block]
- ✓ Private Key (PKi) [Key is used for encryption and decryption mechanism]

After all the information has been generated PK_i is stored into internal database of SS (security service)

The main idea behind to hide PK_i is to provide security to Cipher Text (B_i), so no one else can use the key and try to decrypt the block. System needs to store Cipher Text (B_i) to the CSP along with Token (T_i B_i). System uses the referral data integrity algorithm to associate the B_i -> T_i B_i.

Along with the Cipher Text (B_i) and Token (T_i B_i) system starts the generation of the metadata. Metadata contains following fields.

- ✓ Logged user info (U-info)
- ✓ File Name (Fname)
- ✓ Token Collection (TC = { T1 B1, T2 B2, T3 B3, …… Tn Bn})

Once all the block (ΣB_{ij}) has been processed successfully metadata stored into CSP and TTP database respectively.

- Data Retrieval

User has received the request for retrieval of a File (F) from CSP database. System sends the request to CSP (CSP-Metadata) to validate the request. i.e. F is present onto CSP or not.

If F is available, system will retrieve the F-Metadata, verify TC != 'TEMPER', which gives assurance that data is not corrupted on CSP.

So the TC contains TC = { T_1 B_1, T_2 B_2, T_3 B_3, …… T_n B_n} as TC is stored into sequential format, system will be able find to find block sequence. System will not start the decryption process for blocks. System will also initialize the buffer to hold the Plaintext (B). i.e. F = Σ Plaintext (B_{ij})

Plaintext (B_{ij}) = SS (Cipher Text(B_i),T_iB_i)

- Access Control Provider

To implement role based system the system uses the access control list for each file being uploaded to the cloud storage. The access control list includes the parameters such as access type, filename, token, userlist. Suppose the system has user u1, u2, u3 of which user u3 is the admin of cloud storage having all the access to the cloud storage. User u1 has file sample.txt to be uploaded to the cloud storage. While uploading file the access control list for file sample.txt is also uploaded. The access control list will look like as below

Access type: Grant or Revoke

Filename: sample.txt

Token: file identifier

Userlist: u2, with which the file sample.txt is been shared

V. FLOW OF IMPLEMENTATION

As shown in figure 4 whenever user tries to upload a file to cloud storage the file is divided into blocks of 4 kb size and

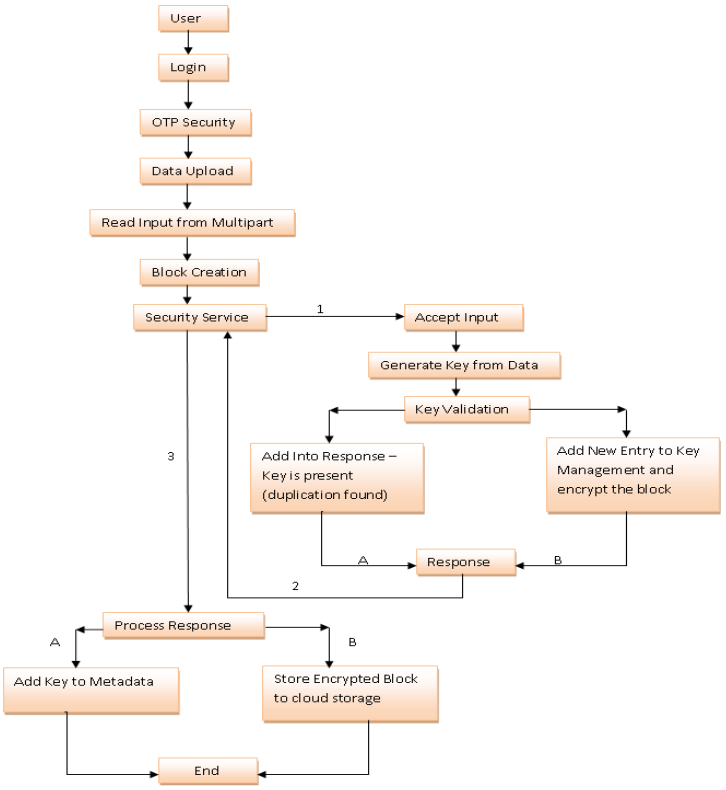

Figure 4: Detailed workflow of the proposed system

sent to security service for creation of public key and data duplication detection. The security service is responsible

for generating keys for each block and validates it to see if it exists in the database. If the key exists in the database it just adds that key to the metadata of a file being uploaded and does not store the block to the cloud storage. If a key is not present then security service encrypts the block with the generated key and stores the encrypted block to the cloud storage and adds the key to the metadata of a file. In this way security service is responsible for identifying duplicate data and avoids it storing to cloud storage.

As discussed in chapter IV the use of access control list is to implement a role based system. The security service is responsible for checking access control list for each file and detecting data duplication in a group of users who has permissions to access the file stored on cloud storage

Results:

The implemented system uses block level encryption and data deduplication for effective space utilization. To show that this system really helps in reducing storage space cost the results are compared against the system which uses file level encryption and data deduplication mechanism. This comparison is explained using following example. Let's say user user1 has uploaded a file f1 containing data "This day is Sunday" using file level encryption mechanism. However another user user2 tries to upload the file f2 containing data "This day is Sunday and it's nice to be here". In this scenario both f1 and f2 are considered to be different files even though the contents are similar up to some extent and two different files will be stored onto cloud storage.

However, using block level encryption and data duplication, the file is divided into blocks such that block b1 contains text "this day" block b2 contains text "is Sunday" and so on. In this case when user1 and user2 tries to upload the file f1 and f2 only one copy of blocks b1, b2 will be saved as both are identical and remaining blocks for text "and it's nice to be here" will be stored onto cloud. So the data deduplication is achieved by storing only one unique copy of redundant data and by maintaining the metadata for each file containing references of its blocks plus the blocks which are stored uniquely onto cloud.

VI. CONCLUSION

In the proposed system, the notion of authorized data deduplication was proposed to protect the data security by including differential privileges of users in the duplicate check. The implementation of this approach shows that better deduplication can be achieved by using block level deduplication mechanism as compared to file level deduplication. This system also achieves confidentiality using convergent encryption. Furthermore, this system shows that the solution proposed here can be easily implemented with existing and widespread technologies. Finally the solution is fully compatible with standard storage APIs and transparent for the cloud storage provider, which does not have to be aware of the running deduplication system.

Currently the system is implemented to support text based block level encryption and deduplication this can be extended to all format like multimedia files, image files etc.

VII. REFERENCES

[1]M. Bellare, S. Keelveedhi, and T. Ristenpart. Message-locked encryption and secure deduplication. In EURPCRYPT, pages 296-312, 2013

[2]Pasquale Puzio, et al "ClouDedup: Secure Deduplication with Encrypted Data for cloud Storage", Cloud Computing Technology and Science (CloudCom), 2013 IEEE 5th International Conference DOI 10.1109/CloudCom.2013.54

[3]Dutch T Meyer, et al, "A study of practical deduplication". ACM Transactions on Storage (TOS), 7(4):14, 2012.

[4]P. Anderson, et al, "Fast and secure laptop backups with encrypted de-duplication". In Proc. of USENIX LISA, 2010.

[5]M. Bellare, et al, "Dupless: Server- aided encryption for deduplicated storage". In USENIX Security Symposium, 2013.

[6]D. Ferraiolo, et al. "Roe Based access controls". In 15th NIST-NCSC National Computer Security Conf., 1992.

[7]J. Li, et al. "Secure deduplication with efficient and reliable convergent key management". In IEEE Transactions on Parallel and Distributed Systems, 2013.

[8]W. K. Ng, et al. "Private data deduplication protocols in cloud storage". In S. Ossowski and P. Lecca, editors, Proceedings of the 27th Annual ACM Symposium on Applied Computing, pages 441-446, ACM, 2012.

[9]R. S. Sandhu, et al. "Role-based access control models". IEEE Computer, 29:38-47, Feb 1996.

[10]J. Stanek, et al. "A secure data deduplication scheme for cloud storage". In Technical Report, 2013.

[11]Is Convergent Encryption really secure?

http://bit.ly/Uf63yH.

[12] AMAZON. Amazon Elastic Compute Cloud (EC2).

http://aws.amazon.com/ec2.

[13]Amazon S3. http://aws.amazon.com/S3

[14]Amazon Web Services. http://aws.amazon.com/.