Data Confidentiality Enabled Deduplication Scheme Using Convergent Keys

Deduplication Scheme in Cloud Storage

[1]Deepa.D [2]Revathi.M
[1]Student, [2]Assistant Professor,
Department of Computer Science and Engineering,
Kingston Engineering College, Vellore

***Abstract* - Cloud computing is the emerging technology in the computer science. Cloud provides different service to us such as Infrastructure as a Service, Platform as a Service, Storage as a Service. Storage as a service is most widely used service for all common users, small, mid, large size organization. But the main issue is privacy. There is no security for data in the cloud server. Deduplication is the technique to eliminate the same data stored in the server. The same data copies are identified by comparing the data content of the users. In this the cloud server may leak the information or even to be hacked. For this issue the data is encrypted before upload to the server. If different user encrypt the data with their own encryption algorithm means the same data copy will generate the different cipher text so the deduplication is impossible. Convergent encryption is used to encrypt the data and it will produce the same cipher text for same content. The Ramp Secret Sharing Scheme is used to share the convergent keys between the different users. The tag value is independently derived from the data content this tag value is used to identify whether the data is already stored or not.**

Index terms – Convergent, block, tag, duplication, and encryption

I. INTRODUCTION

Nowadays most of the organization and the enterprise upload the data in the third party server due to less maintenance. According to the analysis of the IDE the number of data stored in the cloud will reach 200 trillion gigabytes. If data is rapidly increased in the cloud server the maintenance and the accessing data is more difficult. So this deduplication technique is used to eliminate the same data copies stored in the server. This technique reduces the 80% space by eliminating the redundant data stored in the cloud. The main goal of this deduplication reduces the storage space and the uploading bandwidth. Convergent encryption is used to encrypt the data and it will produce the same ciphertexts for the same data. The data confidentiality is also maintained no information will be hack by the attackers. Ramp Secret Sharing Scheme is used to produce the convergent keys and those keys is shared to the different users who are all having the same data copies. Advance Encryption standard is used to encrypt the data. This AES is use 16bit key to encrypt the data. AES is more secure compared to other encryption algorithm.

Deduplication that can be done in two ways file level and the block level. In the file level the whole data file encrypted once and give one ciphertexts. The block level encryption is dividing the data block into small chucks of the same size. Block level encryption is more efficient than the file level encryption. Suppose if the different store the same data with some modification means that block only stored and remaining block is appended from the existing file. Storage space and the uploading bandwidth reduced in this deduplication technique. The data maintenance is easily handled.

II. PRELIMINARIES

In this section we define the cryptographic primitives used in the deduplication scheme.

2.1 Symmetric Encryption

Symmetric encryption is uses a common secret key between different users who are all having the same data copies. Same key is used to encrypt and decrypt the data and stored in the server. Advance Encryption Standard is used to encrypt the data to generate the ciphertexts. AES is more secure compared to other encryption algorithm. AES is meant to protect from disclosure to third parties. It used to take three important function to encrypt the data is (i) What you are encrypting (ii) Key (iii) Initial vector. Block level is used in this deduplication process so AES is best for the block cipher. Advance Encryption is mainly used for the security problem. More difficult to decrypt the original plain text in this symmetric advanced encryption standard

2.2 Convergent Encryption

Convergent encryption provides data confidentiality in deduplication. First user derives the convergent key from their original data copy. That convergent key is used to encrypt and decrypt the data. The same convergent key is shared between the different users who are all having the same data. The convergent encryption that consist of four functions

(i) Key Generation: Convergent key is derived from the data copy of the users original data.
(ii) Encryption: Convergent key is used to encrypt the data it will produce the same cipher text for the same data.

(iii) Decryption: The cipher text is decrypt from the original plain text from the server using the convergent keys.

(iv) Tag Generation: The tag value is independently derived from the content and this tag value is used to find the duplicate copies.

III. CONSTRUCTION

This section describes the basic approach that realizes the convergent encryption in deduplication and explains the user activities performed normally without any internal activities.

3.1 File Upload

The user file uploading is done based on the following steps

Step1: User uploads the file
Step 2: The user share the data with the RSSS and generate the key for the data. This key is called convergent key.
Step 3: The user’s data file is divided into number of blocks.
Step 4: Each block is encrypted with the convergent key with the use of Advanced Encryption Standard.
Step 5: Tag value is derived from the original data file and append the value to the file.
Step 6: The encrypted data and the tag value is uploaded to the server.
Step 7: The server check the tag value whether the particular file is already stored or not.
Step 8: If the file is not stored the server save the file
Step 9: Else if the data is already present then the file is not stored and the server give the link to the particular file.

3.2 File Download

To download the file from the server the activities done in the following ways

Step 1: User send the request to the user by specifying the particular file name.
Step 2: The server verifies the particular user is authorized or not
Step 3: If the user is not authorized the server abort the signal
Step 4: If the user is authorized then the server send the requested file to the user.

IV. IMPLEMENTATION DETAILS

The implementation of the deduplication scheme is done in the following process

4.1 Block Creation

Firs the user file is divided into number of blocks. The block size is constant for the entire file. We can choose the block size based on our wish. Suppose the file size is 100kb means 10blocks is created with 10kb each.

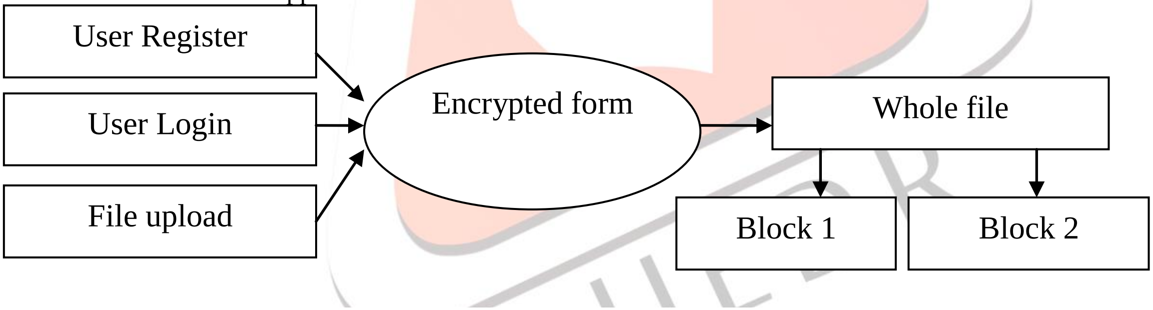

Figure 1 Block Creation

Example for Block Creation: The original message size is 800bytes; the file is divided into 8blocks with each 100bytes.

Data deduplication is a technique for eliminating duplicate copies of data, and has been widely used in cloud storage to reduce storage space and upload bandwidth. Instead of keeping multiple data copies with the same content, deduplication eliminates redundant data by keeping only one physical copy and referring other redundant data to that copy. However, deduplication, while improving storage and bandwidth efficiency, is incompatible with traditional encryption. Specifically, traditional encryption requires different users to encrypt their data with their own keys. Thus, identical data copies of different users will lead to different cipher texts, making deduplication impossible. Convergent encryption provides a same ciphertexts for the same data. It encrypts/decrypts a data copy with a convergent key.

Block 1: Data deduplication is a technique for eliminating duplicate copies of data, and has been widely us
Block 2: in cloud storage to reduce storage space and upload bandwidth. Instead of keeping multiple data c
Block 3: ies with the same content, deduplication eliminates redundant data by keeping only one physical c
Block 4: and referring other redundant data to that copy. However, deduplication, while improving storage
Block 5: d bandwidth efficiency is incompatible with traditional encryption. Specifically, traditional encry
Block 6: ption requires different users to encrypt their data with their own keys. Thus, identical data copie
Block 7: s of different users will lead to different cipher texts, making deduplication impossible. Convergen
Block 8: t encryption provides a same ciphertexts for the same data. It encrypts/decrypts a data copy with a convergent key

4.2 Key Generation

The key generation is done using the Ramp Secret Sharing Scheme. If two or more users upload the same data, that data is shared with the RSSS. The RSSS generate the convergent key from the data copy and send to the user. That key is called the convergent key. The convergent key is used to produce the same cipher text for the same data. The convergent key is used to encrypt and decrypt the data. The secret key is shared with the different users. Suppose if one user lost the key means that key is retrieved from the other shares. The privacy is maintained in this secret sharing scheme.

4.3 Convergent Encryption

The convergent key is used to encrypt the data. Advanced Encryption Standard is used to encrypt the data. AES is used for encryption because that is suitable for the block cipher and provide more secure compared to other encryption algorithm. AES encryption is taking the basic functions as (i) what you are encrypting (ii) Key (iii) Initial vector. So this encryption is best for providing from hacking. Normally for convergent encryption hashing algorithm is used this does not provide any security. The hash function is not to encrypt the data. The AES use the 16bit key to encrypt the data. The user retains the key and uploads the data copy only to the server. So the data confidentiality is maintained in this encryption process. Unless knowing the key value no one can decrypt the data from the server.

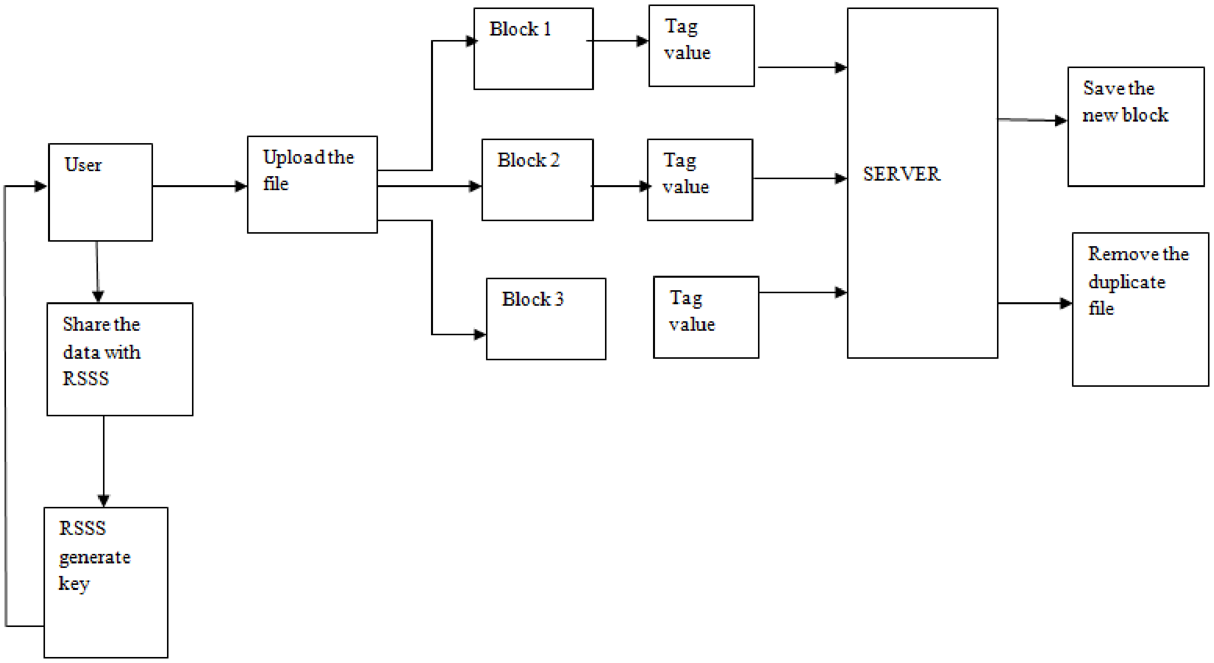

Figure 2 Deduplication process

4.4 Tag Generation

Tag value is used to identify the duplicate data in the storage. The identical data copies will generate the same tag value, so the duplicate file easily identified with the tag value. The tag value is identically derived from the data copy. The tag value is generated by the following functions

1. Block Size: first the block size is calculated, so if different user mentions the different block size means this will easily identify the same block.
2. Number of Words in the block: the total number of words in the block is calculated.
3. Word size: each word size in the block is calculated and mention in the tag code
4. Word pattern: the word pattern is defining that taking the first letter of each word in the block as a sequence.

This combination of that value is considered as a tag value. That consists of User ID, Block Number, and the TAG value is specified. The server identifies the duplicate copies by checking their tag vale and eliminates the duplicate copies.

The tag code generation is done based on this process. Suppose the word pattern only is taking for the tag value means many block will generate the same tag code for different data block. For those values the duplicate file will not be saved in the server. So the combination of all functions such as block size, word size, number of words in the block and word pattern. This type of code generation will produce the same tag for the same data block. The difference between the word pattern tag code generation and the combination of all functions are given in the table.

Table 4.1 Comparison of tag generation

Word pattern	**Combination of all functions**

This tag generation is done by using the word pattern (pattern refers the first letter of each word in the document). Example for word pattern is given below.

Data deduplication is the process of eliminating duplicate copies

Tag code for this sentence is: **dditpoedp**

In this function more duplication may occur due to similarities of the combination of the words.

EXAMPLE

this is sample block for the tag generation
BLOCK 1: Hello World
BLOCK 2: Hai Word

Tag code : HW

The two blocks are having the different words but it produce the same tag value. So this method is reduce the efficiency and reduce the duplication process.

This combination of all words is producing the efficient tag code for same data. Those functions are

1. Block Size
2. Number of words in the block
3. Word size
4. Word pattern

So these combinations produce the efficient code for only the same data blocks uploaded by the different customers.

The word pattern only did not produce the efficient code so these combinations are used for the tag generation. The tag value is shuffled and produces the combination of all functions.

EXAMPLE

BLOCK 1: Data deduplication is the process of eliminating the duplicate copies of same data.

Tag code: 27548kweskuorhcevcr71738

This code is the combination of word size, block size, number of words in the block, word pattern.

4.5 Deduplication Process

First the user uploads the file then the server checks the tag value, if the tag value is already stored then the server not store the data copy else if the tag value is not present in the server then the data is stored in the storage. This deduplication scheme that reduce the storage space and the uploading bandwidth

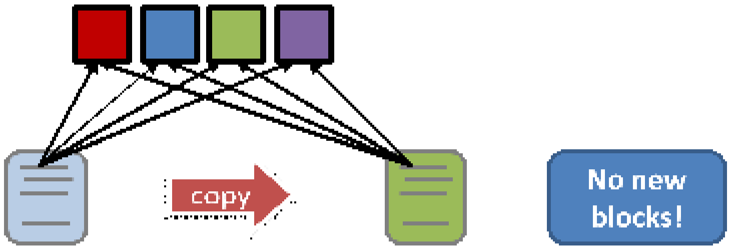

Figure 3 Deduplication Link

V. PERFORMANCE EVALUATION

The performance evaluation is done by comparing the storage space required for the traditional storage and the deduplication storage. In deduplication more storage space is reduced by eliminating the duplicate copies of the same data uploaded by the different users.

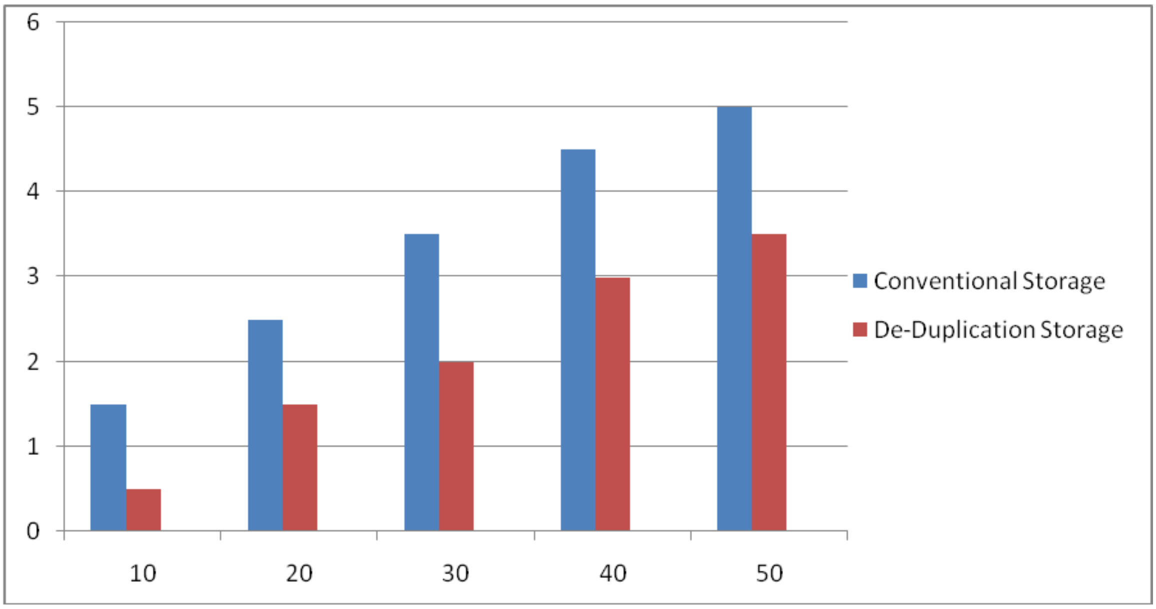

Figure 4 Comparison of storage space

In this graph X-axis represents number of files stored in the server. Y-axis represents storage space required to store the file in the server.

In conventional storage space required more to store the data in the server. Because the duplicate file also need separate space to store the data. So storage space is increased. If number of the duplicate copy is increased the storage space is also rapidly increased.

Deduplication scheme that reduce the storage space with eliminating the duplicate copy of the same data and stored once. If number of files increase the duplicate file is also increased in conventional encryption more storage is required to save the same data file again and again. Deduplication process that reduce the storage space and the uploading bandwidth. For any organization and the enterprise this technique is efficient to utilize storage space efficiently.

VI. CONCLUSON

Deduplication reduces the storage space and increases the efficiencies of data management and the easy access. If same data is again and again stored in the server means rapidly increase the storage space and also increase the uploading and downloading bandwidth. By eliminating the same data stored in the server more storage space is reduced and maintains the data confidentiality of the data. There is no chance to hack the original data from the cloud server because they are in encrypted format. The Ramp Secret sharing Scheme is generated and maintains the secret key shared by the different users. The convergent encryption that efficiently generate the same ciphertexts for the same data uploaded by the different users

REFERNCES

[1] Jin Li, Xiaofeng Chen, Mingqiang Li, Jingwei Li, Patrick P.C. Lee, and Wenjing Lou (2014) Secure Deduplication with Efficient and Reliable Convergent Key Management IEEE transactions on parallel and distributed systems, Volume 25, Issue 6, pp. 666-678.
[2] J.R. Douceur, A. Adya, W.J. Bolosky, D. Simon, and M. Theimer, ‘‘Reclaiming Space from Duplicate Files in a Serverless DistributedFile System,’’ in Proc. ICDCS, 2002, pp. 617-624.
[3] D.T. Meyer and W.J. Bolosky, ‘‘A Study of Practical Deduplication,’’ in Proc. 9th USENIX Conf. FAST, 2011, pp. 1-13.
[4] A. Rahumed, H.C.H. Chen, Y. Tang, P.P.C. Lee, and J.C.S. Lui, ‘‘A secure Cloud Backup System with Assured Deletion and Version Control,’’ in Proc. 3rd Int’l Workshop Security Cloud Comput., 2011, pp. 160-167.
[5] P. Anderson and L. Zhang, ‘‘Fast and Secure Laptop Backups with Encrypted De-Duplication,’’ in Proc. USENIX LISA, 2010, pp. 1-8.
[6]A.D. Santis and B. Masucci, ‘‘Multiple Ramp Schemes,’’ IEEE Trans. Inf. Theory, vol. 45, no. 5, pp. 1720-1728, July 1999.
[7] A. Shamir, ‘‘How to Share a Secret,’’ Commun. ACM, vol. 22, no. 11, pp. 612-613, 1979.
[8] Y. Tang, P.P. Lee, J.C. Lui, and R. Perlman, ‘‘Secure Overlay Storage with Access Control and Assured Deletion,’’ IEEE Trans. Dependable Secure Comput., vol. 9, no. 6, pp. 903-916, Nov./Dec. 2012.
[9] W. Wang, Z. Li, R. Owens, and B. Bhargava, ‘‘Secure and Efficient Access to Outsourced Data,’’ in Proc. ACM CCSW, Nov. 2009, pp. 55-66
[10] R. Geambasu, T. Kohno, A. Levy, and H.M. Levy, ‘‘Vanish: Increasing Data Privacy with Self-Destructing Data,’’ in Proc. USENIX Security Symp., Aug. 2009, pp. 316-299.