International Journal of Scientific Research in Computer Science, Engineering and Information Technology
 | Volume 2 | Issue 5 | ISSN : 2456-3307

Enhancing Cloud Data Storage Performance by Secure Data Deduplication

Monika Sharma*, Neeraj Shrivastava
Department of Computer Science & Engineering, IES IPS Academy, Indore, Madhya Pradesh, India

ABSTRACT

In Recent days, Cloud computing turns out to be exceptionally important which gives very accessible storage space on the cloud. Data deduplication is a most prominent data compression technique. This strategy is used to reduce the duplicity of information. To maintain capacity and decrease storage space in cloud storage information deduplication is used. Taking care of address following difficulties, this paper makes an endeavor to formalize the thought of secure and productive cloud storage. In this way, the component of openness and accessibility of the information. Text files are used as input. In the first module, the record is first processed for finding the essential features from the text files. For distinguishing the same sort of information these elements are utilized in this manner-containing comparative. This space based ordering of records is performed in advance. By Tiger, hash generation algorithm produces a key that will go through the 3DES algorithm for encoding. The results show the strategies proposed playing out a beneficial operation on contrasting and the customary system.
Keywords: Cloud Storage, Cryptographic Data, Data Management, Data Deduplication, Hash Tree

I. INTRODUCTION

This Image Cloud computing technology not only has contributed to various applications and influenced the operation of original networks but also allowed the Internet to exist in different corners with different devices. Nevertheless, with the appearance of numerous devices and data, data access and nodes management and control of cloud network become the significant issues to be emphasized because the efficiency of control methods enormously affects the performance and quality of cloud network.

A. Cloud Storage

It can be defined as "store data online in the cloud ". It provides reliability, openness, quick deployment and so on. It has four types of storage - public, private and hybrid [Figure 1].

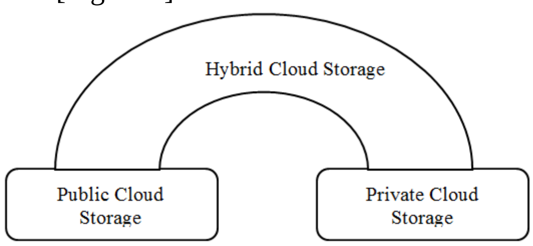

Figure 1. Cloud Deployment Model

1) Public Cloud Storage: In this, unstructured data stored in global centers. Customers have to buy for the storage area based on data per user. The cloud storage provider completely deals with the undertaking's public cloud storage.
2) Private Cloud Storage: The client who needs more control over their information mostly utilizes private cloud. Private cloud storage helps settle the potential for security and execution concerns while yet offering the upsides of cloud storage.
3) Hybrid Cloud Storage: It is a mix of public and private cloud storage. Where some common information lives in the endeavor private cloud. While other information is put away and available from a public cloud storage provider.

In this paper, we deal with the inherent safety exposures of symmetric encryption and advocate cloud deduplication. Which preserves the combined benefits of deduplication and symmetric encryption. To summarize our contributions:

- It assures block level deduplication and facts confidentiality simultaneously as coping with weaknesses raises through symmetric encryption.
- Data deduplication reduced hardware backup and storage cost.
- It increases storage efficiency.
- It preserves confidentiality and privateness even towards doubtlessly malicious cloud storage.
- Cloud deduplication works transparently with present cloud storage providers, for this reason, It's far completely likeminded with well-known storage.

Now a day cloud computing provider turning into very famous Cloud gives a higher way of storage with a green price. One primary hassle with the cloud is to control a large quantity of data That grant us to control information deduplication technique is used. Deduplication has many benefits it has a few safety issues. This motivates us to endorse a model which manage the security troubles of deduplication and provide authorized deduplication on the cloud server. In this research work, we developed a security structure for data privacy preservation of cloud data storage to make access to the data file in a secure manner in the large public cloud environment.

B. Data Deduplication

Whenever a user stores the data, multiple replicas created at several places i.e. the similar file stored at the several places by the different users it increases storage complexity. Data deduplication technique consider a prominent part in cloud computing. This is the technique which eleminates the duplicate data in the cloud. Thus, if the same file is uploaded by the user that is already stored then, the cloud provider will add the user to the owner list then he/she can access that file anytime. Deduplication can remove the storage complexity. The data deduplication and encryption are two different technologies. The after effect of encryption is to secure the information in the encoded form on the off chance that information is not encoded by clients the information is not secured by a cloud storage supplier. It has two levels -

i. File Level Deduplication: The elimination of duplicate data of the duplicate file happens in file level deduplication. Which is also usually known as single instance storage (sis).

ii. Block Level Deduplication: Block level deduplication removes the blocks of data that occurs in non-identical files. Block level deduplication is very popular; it gives more space than single instance storage file level deduplication.

II. LITERATURE SURVEY

Jin Li, et al., [10] developed an encryption technique to data encryption prior to outsourcing. In hybrid cloud, architecture duplicate check performs in which the file tokens are generated because of private cloud server. In this, every file attached with a unique file token. The secret key will not issue to the user directly. It will be managed kept because of private cloud server it means, it prevent the privilege key. To get file token the user has to request to a private cloud server after checking user identify the private cloud server issue the corresponding file token to the user. The duplicate check accomplished at the public cloud before file encoding. It will create minimal overhead compared to normal operation.

S. A. Maindakar and Dr. M z Shaikh [3] provides the secure efficient cloud storage system. They proposed a user side deduplication for storing and sharing information by the hybrid cloud. It uses proof of ownership pow of files also along with two factor authentication. One time password for security. This results overhead is very limited in realistic environment. Nesrine Kaaniche and Maryline Laurent [11] proposed and implement on Openstack. This is another customer side deduplication plot for putting away and sharing outsources information safely by via. Public cloud. It gives two level of access control check. It utilizes symmetric encryption for enciphering the information asymmetric encryption for the metadata documents. Each customer side uses a for every information key to encode the information. Enter incorporating access rights in a metadata document approves client can decode that information. It will be scrambled by just with his private key. Other than this arrangement is additionally appeared to be impervious to unapproved access to information.

Nimgire Reshma, et al., [2] proposed deduplication to secure the data by providing efficient privilege of users. Duplicate check done in hybrid cloud architecture. In this, duplicate check file tokens create because of private cloud server using secret key. The minimal overhead as compared to convergent encryption network transfer.

Pasquale Puzio, et al., [12] proposed Cloudedup is a comfortable and green storage service, which assures duplication at the level of blocks and reagent encryption records privacy at the same time. They used block stage deduplication instead of file stage deduplication. There is some trouble on the block degree when it comes to key control Extra layers of encryption are brought by means of the server and optional HSM so that they consist of a brand new factor for the minimum overhead. This solution may be easily applied to existing and complete strategies.

III. RELATED WORK

Although Managing data securely is of prior importance in cloud computing. The security includes the cryptographic techniques to implement with the data on the cloud. Therefore for the feature of accessibility and data availability i.e. anytime anywhere we can access the data providing a new solution which must be provided by the cloud.

A. Difference Between SHA1/SHA and MD5

- SHA/SHA1: It's key length is 160bits/20-byte digest. It's speed is slower than MD5 but it is more secure and stronger against brute force attack.
- MD5: It has 128 bits/ 16-byte digest key length. It is faster than SHA1 but less secure as compared to SHA1. It is cheaper to compute.

B. Cryptographic Technique Comparison

TABLE I
COMPARISON OF 3DES AND RSA

Factors	Triple DES(Triple Data Encryption Standard)	RSA(Rivest-Shamir-Adleman)
Key Length	56 bits	2048 bits
Cipher Type	Symmetric Block Cipher	Asymmetric Block Cipher
Time	Faster	Works slower
Security	Secure	Less secure

C. Comparison between Hashing and Indexing

- Indexing is an easy way of sorting some statistics on more than one field. Developing an index on a subject in a desk creates any other information structure which holds the field cost and pointer to the document it relates to. It substantially reduces run time computation with simple operations.

- Hash is a form of an index: it can be used to locate a document primarily based on a key. However it would not preserve any order of statistics. Based totally on hash, one cannot iterate to the succeeding or previous element.

D. Text Feature Extraction Techniques

i. Bag-of- words (BOW): From this algorithm, we frequently count a word happens in a document/file.

ii. Term frequency/inverse document frequency (TF-IDF): The method frequently emphasizes words on a given file or a document, while deemphasizing the repeated words in many documents at the similar time then TF-IDF is calculated.

IV. PROBLEM AND SOLUTION

The proposed work is aimed to provide the data deduplication approach for cloud's storage issues. In this context, a model is reported in this section which deals with the data to identify the duplicate data to remove from storage. In addition, it also now includes the data in storage which is found as a duplicate. The inverted index used that may help to locate the individual data in the cloud server. Additionally, the data searches are made easy for access (availability). Alongside we used the hash tree that helped for locating the similar contents based on their contents and also help to remove duplicity on the server.

Cloud server which is used for storing a large amount of data. That data can be any is any size and type such as image audio video text and others In this work the text file is required for work and demonstration of a system working and their functional importance Thus a provision is made by which the user chooses a text file from their system and upload it to the cloud storage. TF/IDF is a technique which used for removing extra words like special characters like separators (",", ".", ";", "!", "@") and stop keywords. After this for Domain Identification the Indexing technique used. It is assumed that some quantity of information and their options area unit existing on the server. Thus, this file primarily based options are compared with the present list features. The most matched extracted feature is considered. After characteristic the domain of information, it's thought of the file is treated because of the domain. After this process data encrypted for storing on the server. For data encryption, a merge algorithm using two different algorithms are generated. The original file which is given as input to the system is encrypted in this phase. During the encryption process first, the hash key is generated using the tiger hash generation algorithm. That produces 512 bits of key block but the 3DES is not accepting such length of key thus the last bits of the tiger hash algorithm is discarded. Remaining 168 bits are used with the 3DES algorithm to encrypt the file. The encrypted file is used in next phase of the process. The encrypted file is used in this phase than the data is split in a fixed size of blocks. In this process, the 512 bit of data block is created. That is not fixed it depends on the designer to create their own size. Each block of the file is not utilized with this phase of data processing. Here the individual block is treated with the SHA1 to generate the hash for the block data. The tree uses binary manner for mounting each block of data. Each block of data is mounted on the tree using the binary manner. Additionally, before constructing the tree, the block hashes are compared with the existing binary tree leaf nodes for finding the duplicate data. If the data is found the in any leaf node the tree is not constructed further and only a mapping for that file is created. After validating the file availability the file is stored in the server space and for searching the similar files on the storage the tree is used for making fast access of data.

A. Algorithm Used:

In this section we used two types of algorithm:

1) For File Uploading
2) For File Downloading

For File Uploading:

Begin
- Step-1 Read file
- Step-2 Duplication checks by cloud server
- Step-3 Sends the responses i.e. file exist or not
- Step-4 If File does not exist
  - 4.1 File Uploaded successfully
- Step-5 If File already exist
  - 5.1 Display "File already exist"

End

For File Downloading:

Begin
- Step-1 Read File
- Step-2 Duplication checks by cloud server
- Step-3 Sends the responses i.e. file exist or not
- Step-4 4 If File does not exist
  - 4.1 Display "File does not exist"
- Step-5 If File exist
  - 5.1 File Downloaded Successfully

End

V. CONCLUSION

The proposed Cryptography technique is a symmetric key encryption for data retrieval from end user access to secure data Transmission is initiated in a dispersed environment where the single storage is a secure way of authenticity and storage or hosting services. In addition of that for preventing the unauthorized access to the system a strong user authentication technique using the normal credential and multi replica of different uploaded data files are stored on cloud server and have been updating data files at the same place. Multi replica based data blocks are ensured that data integrity between two parties is secure in such a way that public cloud data are accessible throughout the session. Furthermore, for securing the data in storage and entrusted network 3 DES used to performing encryption and decryption and performing document indexing for hash tree generation. The proposed solution reduces storage complexity and increase availability of data.

VI.REFERENCES

[1] Junbeaom Hur, et al., "Secure data deduplication with dynamic ownership management in cloud storage

(Extended Abstract)" IEEE International Conference on Data Engineering. San Diego, CA, USA, 69-70,2017.

[2] Nimgire Reshma , et al., "Deduplication & secure authorized data using hybrid cloud". Imperial Journal of Interdisciplinary Research, 2(6), 415-419, 2016.

[3] Sumedha A Telkar and Dr. MZ Shaikh "Secured and efficient cloud storage data deduplication system" International Journal of Advanced Research in Computer and communication Engineering, 5(1), 301-304, 2016.

[4] Xin Yao, et al., "A Secure Hierarchical Deduplication System in Cloud Storage" IEEE/ACM 24th International Symposium on Quality of Service (IWQoS). Beijing, China, 1-10, 2016.

[5] Junbeaom Hur, et al., "Secure data deduplication with dynamic ownership management in cloud storage" IEEE International Conference on Data Engineering. San Diego, CA, USA, 3113 – 3125, 2016.

[6] Jingwei Li, et al., (2015) "Secure Auditing and Deduplication Data in Cloud" IEEE Transactions on Computers, 65(8), 2386 – 2396.

[7] Zuhair S. Al-Sagar1, et al., "Optimizing the cloud storage by data deduplication : A Study" International Research Journal of Engineering and technology (IRJET), 2(9), 2524-2527, 2015.

[8] Jadapalli Nandini and Ramireddy Navateja Reddy "Implementation of hybrid cloud approach for secure authorized deduplication" International Research Journal of Engineering and technology (IRJET), 2(3), 1297-1306, 2015.

[9] Jin Li, et al., "Secure deduplication with efficient and reliable convergent key management" IEEE Transactions on Parallel and Distributed Systems, 25(6), 1615 – 1625, 2014.

[10] Jin Li, et al., "Hybrid cloud approach for secure authorized deduplication" IEEE Transactions on Parallel and Distributed Systems, 26(5), 1206-1216, 2014.

[11] Nesrine Kaaniche and Maryline Laurent "A secure client side deduplication scheme in cloud storage environments" 6th International Conference on New Technologies, Mobility and Security (NTMS). Dubai, United Arab Emirates, 1-7, 2014.

[12] Pasquale Puzio, et al., "ClouDedup: Secure deduplication with encrypted data for cloud storage" IEEE 5th International Conference on Cloud Computing Technology and Science (CloudCom). Bristol, UK, 363-370, 2013.

[13] Explain Dedupication. from http:// www. webopedia. Com /term /d/data_deduplication.html.

[14] How deduplication works from http:// www.computer world.com /article / 2474479/data-center/data-deduplication-in-the-cloud-explained--part-one.html.

[15] Difference between Hashing and indexing from https://www. quora.com /what-are- the-major-differences –between -hashing -and-indexing.

[16] Difference between MD-5 and SHA-1 from http://lnxsysadm.blogspot.in/2010/12/what-is-difference-between -md5-and-sha.html.

[17] What is Triple DES from https:// www.tutorials point.com /cryptography /triple_ des.htm.