Data Deduplication in Client Side Using Hash Technique

P. Nirmala
Asst prof/CSE Department. Sri Muthukumaran Institute of Technology, Chennai-69.

S. Murugan@Prakasam
Asst prof/CSE Department. Sri Muthukumaran Institute of Technology, Chennai-69.

M. Janaki Raman
CSE Department, Sri Muthukumaran Institute of Technology, Chennai-69.

R. S. Arun Murugan
CSE Department, Sri Muthukumaran Institute of Technology, Chennai-69.

C. B. Lallit Kumar
CSE Department, Sri Muthukumaran Institute of Technology, Chennai-69.

Abstract – Data deduplication is one of the techniques which used to solve the repetition of data.The deduplication techniques are usually used in the cloud server for reducing the space of the cloud storage.To prevent the unauthorized use of data accessing and create duplicate data on cloud the encryption technique to encrypt the data before stored on cloud server.It consists two types of deduplications methods are available one is file name based and another one is file content based in the first method, we are not possible to process because same filename having different content and second method takes more time for content based search. To overcome the problem we are going to implement side data deduplication system, before upload the file to server. We first stores the file hash code in client side, Before uploading file the client machine it convert the file to hash code and checks temporary client chunks, if the hash code is same the file will not process to main server else if the hash code is different the files will upload to server.We make the initial attempt to officially address the problem of authorized data deduplication method. Different from traditional deduplication systems,in additional we are using the secret sharing scheme and shows that it incurs small encoding and decoding overhead compared to the network transmission overhead in the regular upload/download operations.

Index Terms – Cloud storage, Data deduplication, Proof of ownership, Dekey, Hybrid cloud.

1. INTRODUCTION

By Increasing quantity of data which make One of the serious issues in cloud storage services . In 2020 the amount of data storage growing about 40 trillion gigabytes,it can occupies lots of space in cloud storage to reduce it deduplication technique is used.Using Data deduplication is one of main data compression method for removing duplicate copies data in server, it is used to save bandwidth in server storage as well as to reduce the amount of storage space.To avoid few copies of data in cloud storage the deduplication system is used to reduce into one copy of original data.In hash technique data can be allocated in the complete file as data block or fine-grained data block or variable size data block.The deduplication is used to save maintenance cost like Dropbox,Gmail,Drive.,Mozy, and Memopal,etc..Anyhow, deduplication, while improving storage and bandwidth effectiveness, it is not compatible with traditional encryption.Particularly, traditional encryption produces their own keys to different user due to this it creates a huge amount of keys unnecessarily.Deduplication is impossible in traditional encryption because it can create similar copies of data to different users it will lead to different ciphertexts.Data privacy can be achieved by using Convergent encryption.Convergent encryption provides a possible option to implement deduplication process. Using cryptographic hash value that encrypts/decrypts can be done by using a convergent key process. Users keep the keys and send the ciphertext to the cloud and it is used to perform a key generation as well as data encryption.Identical convergent key and identical ciphertext are used in the same data copies for encryption. Convergent keys that allow0 the encryption process to convert plain text to cipertext and vice versa for decryption process.

To know how convergent encryption can be realized, we believe a baseline approach that equipment convergent encryption depends on this layered approach. That is, the original data copy is first and foremost encrypted with a convergent key resulting by the data copy itself and the convergent key is then encrypted a master key that will be kept locally and securely by the user.The encrypted convergent keys

are stored, along with the equivalent encrypted data copies in cloud storage.The process of master key can be used to get well the encrypted keys and hence the encrypted files. In this way, each user only needs to keep the master key and the metadata about the outsourced data.

Two critical issues can be created in baseline approach. The first one, increasing the number of users it will create a huge number of keys ineffective manner.Particularly, each user must associate an encrypted convergent key with each block of files is encrypted data copies,after it will restore the data copies. Whatever may be each and every user going to access the same copy of data that's why each user need own convergent key so that no other user can access the other's file. The total size of Dropbox is 8GB.To reduce duplicate copy of data in Dropbox deduplication technique is used.The number of users increases as soon as the number of keys also increases, then it will be led to the massive storage cost, as users must be owed for storing the bulky number of keys in the cloud it will affect pay-as-go model.

According to the second model, the baseline approach is untrustworthy, as it requires each user to keep his own master key. Incase master key is unfortunately lost, then the user data cannot get back if it is compromised by attackers, then the user data will be leaked.

To make this approach effective and reliably they are using new construction called Dekey, which is used to perform convergent key management on both user and cloud storage.

Our idea is to apply deduplication to the convergent keys and leverage secret sharing techniques.Particularly, we construct secret shares for the convergent keys and distribute them across multiple independent key servers.First user can upload the data then the System allows only one owner of the file. In case any other user uploads the same file, then the tag will indicate that the file already exists.This method is also used for faster recoveries of data.

To recover data copies, a user must access a small amount of key servers through authentication and obtain the secret shares to renovate the convergent keys. In this development it will be reliable and reduces storage space and it can prevent from attackers.

To overcome this problem a new construction key is build called Dekey. Dekey provides efficient and reliable through convergent key deduplication File-level and block-level both will support in Dekey.

The proposed security model demonstrates the Dekey analysis in secure to be specified as proposed security model.Dekey produces a limited number of keys in a secure manner.We apply Dekey by using a Ramp secret sharing system that gives the key to adapt different trustworthiness and privacy levels. Dekey occurs limited overhead in ordinary upload/download operations in easy manner.

2. RELATED WORK

To improve security in cloud computing deduplication process is used.The cloud storage is used to reduce the storage size of the tag integrity check for dedulpication system.To make a security system we can upgrade the system by transforming the predictable message into unpredictable message. Check whether it is related to proposed system

The message should be check whether it is predictable or unpredictable. The Related work follows below the procedures for processing.

2.1 Symmetric Encryption

A variety of cryptographic solutions have been proposed in the literature survey and it is used to keep the privacy of outsourced data. In traditional encryption every user encrypts data with an individual secret key. The robustness of key management can continue by utilizing the threshold secret sharing. Whatever may be, then on top of the studies do not think about the deduplication. The same data copies can be encrypted with own master key by dissimilar user.Due to this process it will lead to dissimilar cipertext and it will make dedupliaction not possible. That's why we are using Convergent Encryption. These are the processes of Symmetric encryption.

2.2 Convergent Encryption

Data privacy in de-duplication can be achieved by using convergent key.It is used for space-efficient secure outsourced storage in message locked encryption.There are also several implementations of convergent implementations of different convergent encryption variants for secure deduplication. It is recognized that some marketable cloud storage providers,they are organizing convergent encryption.Before using convergent key it will make a huge number of keys unnecessarily.Convergent key is mainly used to reduce the number of keys and security purpose.

2.3 Proof of Ownership

Proofs of ownership were proposed by Halevi et al for deduplication process, a file can be identified whether the file owner's or not. Such that a client can also confirm to the cloud storage server.Merkle Hash Tree proposed a Several Proof of ownership constructions depends on the client-side deduplication, which contain the bordered leakage setting.By proposing one more process as efficient and reliable of a file to select random bit-positions for the file proof.

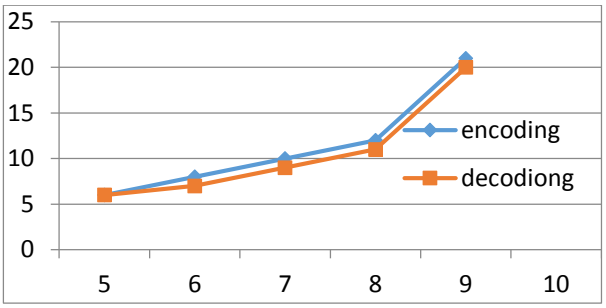


Fig 2: Impact of number of encryption with cloud server providers n on encoding/decoding times, where r ¼ 2 and n _ k ¼ 2

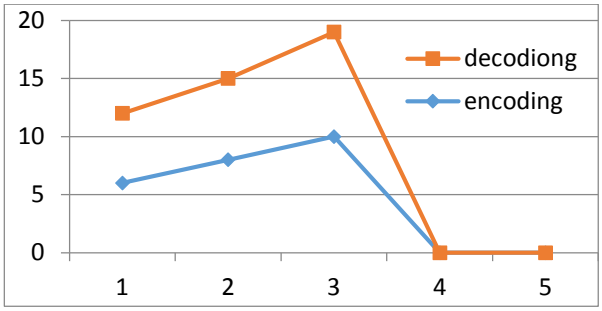


Fig no 3: Impact of reliability level n _ k on encoding/decoding times, where n ¼ 6 and r ¼ 2.

3. SECURE DEDUPLICATION

In this system, we represent a baseline approach to realizing convergent encryption in deduplication, and discuss the restrictions of the baseline approach in key management. To end this, we represent our construction Dekey, which belongs to ease the key management overhead and provide default tolerance guarantees for key management, while making the necessary protection properties of secure deduplication.

The baseline approach involves only the user and the storage-cloud server provider(Schedules caste Sub-Plan)Its idea is that each user has all his data copies encrypted with the corresponding convergent keys, which are then further encrypted by an independent master key. The encrypted convergent keys are outsourced to the storage-cloud server provider, while the master key is securely maintained by the user.

The details of the baseline approach are elaborated as follows. Based on the following system setup we are following here.

3.1 System Setup

The following setup includes two types of sub plan:

Step 1:They are started as :

1)A symmetric encryption system with the primitive functions and it plays a vital role for security

2) A convergent encryption system with the primitive functions.

3) A Proof of ownership algorithm for the file and a Proof of ownership algorithm for the block is used here.

Step 2: The storage-cloud server provider initializes into two types of storage systems: one is rapid storage system which is used to show duplicate files in tags.Another one is file storage system which is used for storing both encrypted convergent keys and encrypted data copies.

3.2. File Upload

User uploads a file.foremost, it performs file level deduplication as follows.

Step 1:Once an input file has been received, the user computes the input send to the tag file.

Step 2:After receiving tag If the file is already exist the tag as shown the file already exist otherwise no file is exist.

Step 3:If the file receives them “No file duplicated” it jumps to Step 5 procedure to go on block level duplication or answer has been file as been duplicated after finishing it,then it will check for proof of ownership in a file.Then it will show ownership of a file.

Step 4:If Proof of ownership is approved, the storage-cloud server provider just proceeds a file pointer to the user.

Step 5:The master key and input file depends on the following procedure:1) Segment file into chunks of data 2) for each chunk computes block tag 3) Send the chunk of data tags.

Step 6:Depends upon getting block tags they computes a block signal in the following manner:By using i for each file,if the file is already exists it will match to i=1 otherwise the file is not exists it will assume as i=1 which means there is no file is exists.

Step 7:After receiving the signal about the file, it will ensures about the file and it will follows following procedues: for each i,i=1the user runs Proof of ownership of a file is known.Afterwards the users do not need to upload the file.

Step 8:Each and every blocks, the user always computes the encrypted convergent keys with the convergent master key and master key.

Step 9:The user uploads the different blocks,i=0 all encrypted convergent keys and tag to the storage-cloud server provider, which then stores them in the file storage system.

3.3 File Download

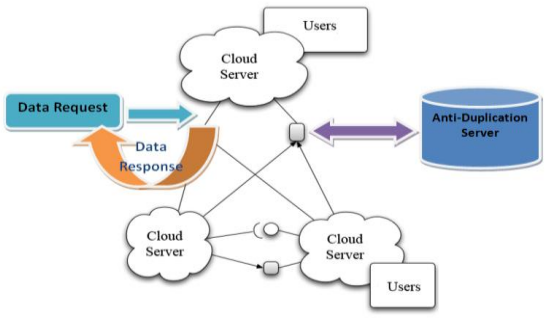
If the user needs to download a file . It Foremost sends a demand and the file name for the storage-cloud server provider and performs the following steps.

Step 1:Depends upon the receiving requests and file name, the storage-cloud server provider will make sure whether the user is qualified to download or not . Incase failed, the storage-cloud server provider send a returns the terminate signal to the user which specify the download failure. Otherwise storage-cloud server provider returns the corresponding ciphertexts and the encrypted convergent keys to the user.Then user can download the file.For security purpose ciphertext are used.

Step 2:Depends upon receiving the encrypted data from the storage-cloud server provider,to recover each convergent key using master key technique.Then it uses Key to make progress the original block. Then the user can download the original file.

3.4. System Model

We initially prepare a data outsourcing model used by Dekey. There are three entities, namely: the user, the storage cloud service provider gives the key management cloud service provider (encryption with cloud server provider), as elaborated in the below architecture diagram.



User: A user is an entity that needs to outsource data storage to the storage-cloud server provider and access the data later on. To save the upload bandwidth, the user only uploads unique data but does not upload any duplicate data,which may be owned by the same user or different users.

They provide the data outsourcing services and stores data to users.To reduce the storage cost, the storage-cloud server provider reduce the storage of the same copies of data using deduplication technique and keeps only different data.

A encryption with cloud server provider maintains convergent keys for users, and which provides users with minimum the amount of storage and computation of services to facilitate key management.Designed for fault tolerance of key management, we think about a encryption process in cloud server providers, each being an individual entity. All convergent key are distributed across multiple encryption in cloud service.

In this process, we submit a data duplicate to be either a whole file or a smaller-size block, and this tends to two types of deduplication: 1) file-level deduplication, which reduce the storage of any identical files, and 2) block-level deduplication, which divides a file into smaller fixed-size or variable-size blocks and reduce the storage of any identical blocks. By using fixed-size blocks that make it simple for the computations of block limitations, when using variable size blocks provides improved deduplication very effective. We organize our deduplication process in both file and block levels.Mainly to upload a file, a user first and foremost performs the file level duplicate verification. If the file is a duplicate, then all its blocks must be duplicated as well; otherwise, the user further performs the block-level duplicate check and identifies the unique blocks to be uploaded. Each data copy (i.e., a file or a block) is associated with a tag for the duplicate check (see Section 2). Each and every copy of data and tags will be stored in the cloud storage.

4. CONCLUSION

We suggest Dekey, an efficient and reliable convergent key administration system for secure deduplication. Dekey distributes convergent key shares across multiple key servers and applies deduplication among convergent keys itself, although preserving semantic security of convergent keys and privacy of outsourced data.For security purpose we are using private from a user and public key from a data provider.By using private and public key we can download a file that is called Hybrid cloud.It is shown that an authorized repeated copy of check method skill less overhead comparing convergent encryption and data transfer.

REFERENCES

[1] Boga Venkatesh,Anamika Sharma,Gaurav Desai,Dadaram Jadhav"Secure Authorised Deduplication by Using Hybrid Cloud Approach"B.E Students, Dept. of CSE, Trinity College of Engineering, Pune.

[2] P.Gokulraj,K.Kiruthika Devi,"Deduplication Avoidance Using Convergent Key Management in Cloud",Nandha College of Technology, Erode.

[3] N.O.Agarwal, Prof Mr. S.S.Kulkarni,"Secure Deduplication And Data Security With Efficient And Reliable CEKM",Department of Information Technology PRMIT&R, Badnera

[4] Nikhil O. Agrawal1, Prof.S.S.Kulkarni,"Secure Deduplication and Data Security with Efficient and Reliable Convergent Key Management",Information Technology, PRMIT&R, Badnera.

[5] Bhushan Choudhary, Amit Dravid,"A Study on Authorized Deduplication Techniques in Cloud Computing".

[6] Wee Keong Ng,Yonggang Wen,Huafei Zhu,"Private Data Deduplication Protocols in Cloud Storage".

[7] Shweta Pochhi, Vanita Babanne,"A Survey on Secure and Authorized Data Deduplication",Computer Engineering Department, RMD Sinhgad School of Engineering, Pune University, Pune.

[8] P. Anderson and L. Zhang. Fast and secure laptop backups with encrypted de-duplication. In Proc. of USENIX LISA, 2010.

[9] M. Bellare, S. Keelveedhi, and T. Ristenpart. Dupless: Server raided encryption for deduplicated storage. In USENIX Security Symposium, 2013.

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