International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2013): 4.438

# Secure Deduplication in Cloud Backup Services

**Nidhi Panpaliya[1], Prachi Sorte[2]**

[1, 2]Department of Information Technology, RMD Sinhgad School of Engineering, Pune, India

**Abstract:** ***Data deduplication is the technique which is used for avoiding or removing the duplication of the data, and it is frequently used in cloud storage to reduce the storage space of the data and the bandwidth. To perform the secure deduplication in the cloud it's a challenge. For provide the data security over the cloud convergent encryption technique is widely adopted. Convergent encryption technique is useful for efficiently and reliably manage the large number of convergent keys. The primary challenge is to correctly address the problem of achieving efficient and reliable key management in secure deduplication. Secondary data resource raises security and privacy concerns. Trusted third-party cloud service providers in the proposed system provide the confidentiality of data, reliability checking and also the access control mechanisms by number of internal and external attacks. As Deduplication improves the storage space, bandwidth efficiency but it is conflicting with the convergent encryption technique. The convergent encryption technique requires the different users to encrypt their data with their data with their respective key. As the same data copies of different user will confirm the method to individual cipher texts and making deduplication checking of data unfeasible. Convergent encryption provides an adequate option to implement data confidentiality while realizing deduplication. Convergent encryption is the technique which encrypts and decrypts the data copy with convergent key and which is calculated by cryptographic hash value of the content of the data copy itself. In key generation and data encryption technique users holds the key and send the cipher text to the cloud service provider. Encryption is the technique which deterministic and the identical data copies will create the similar convergent key and the same cipher text.***

**Keywords:** Cloud backup, data deduplication, Convergent encryption, deduplication efficiency.

## 1. Introduction

Nowadays, the personal computing systems like desktops, laptops, tablets, smart phones have become crucial platforms for numerous users, increasing the importance of data on these devices. To avoid data loss due to system failure, automatic deletion of data, or device theft/loss, individuals have improved their use of data protection and recovery tools in the personal computing devices. Because of the virtually infinite storage resources that are available on demand and charged according to usage of user, the cloud storage services (e.g., Amazon S3 and Google Storage) take considerable economic advantages to both cloud providers and cloud users. As shown in Figure 1, the data backup for personal storage has emerged to typically attractive application for outsourcing to cloud storage providers because users can manage data much more easily without having to be bothered about maintaining the backup infrastructure. This is feasible because the centralized cloud management has created effectiveness and cost variation point, and the cloud offers simple offsite storage for disaster recovery, which is always a critical concern for data backup.

Data backup for personal storage in the cloud storage environment implies a geographic division between the client and the service provider. Cloud storage that is usually bridged by wide area networks (WANs), data deduplication is an useful data compression approach that exploits data redundancy, divides the huge data objects into smaller parts, which are called as chunks. These chunks (i.e., typically a cryptographic hash of the chunk data), replace the frequent chunks with their fingerprints once chunk fingerprint index find, and only transfers or stores the unique chunks for the use of communication or storage efficiency.

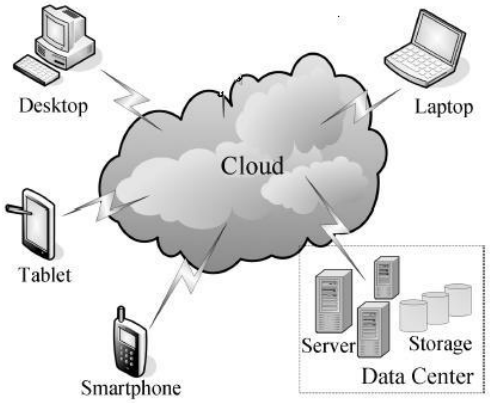



**Figure 1:** Cloud backup platform for personal computing devices.

There are some basics concepts for understanding which are as follows-

### Data Deduplication

Data deduplication is one of the most recent technologies in data storage, because it enables user to save money on data storage costs and the bandwidth costs to reallocate the data when replicating it to offsite for Disaster Recovery. This is big news for cloud provider, because if you store a smaller amount, you need a lesser amount of hardware. If user can deduplicate what user store, user can better exploit existing storage space, which helps to save money. The deduplication process removes the blocks that are not unique. The process consists of four steps such as: Split the input data into chunks. Calculate a hash value of all the block of data, consume these values to find out if another block of the similar data has earlier been stored. Replace duplicate data with a reference to the object earlier in the database.

**Convergent Encryption**

Convergent encryption scheme also known as content hash keying. This is a cryptosystem which produces the identical ciphertext from the same plaintext files. It has applications in cloud computing to remove duplicate files from storage without the provider having right to use to the encryption keys. Convergent encryption is open to an authorization of a file attack in which an attacker can capably confirm whether a target possesses a specific file by encrypting an unencrypted, or plain-text, version and then compare the output with files overcome by the target. This attack poses a problem for a user storing information which is publicly available or already held by the attacker- for example: Banned books or files that cause copyright violation. An argument could be made that a validation of a file attack is easily rendered unsuccessful by simply adding a unique portion of data such as a few arbitrary characters to the plain text before encryption; which causes the uploaded file to be unique and therefore results in a unique encrypted file. There are several implementations of convergent encryption scheme where the plain-text is broken down into blocks, based on content of the file then Each block independently performs convergent encryption which may by mistake overcome attempts at making the file unique by adding bytes at the beginning or end.

**Key Management**

The management of cryptographic keys in a cryptosystem is done by key management. This includes dealing with the generation, exchange, storage, utilize, and substitute of keys. It includes cryptographic key servers, user procedures, and other significant protocols. Key management scheme concerns keys at the user level, also between users or systems. This is in distinction to the key scheduling. Key scheduling on average refers to the internal behavior of key material within the procedure of a cipher. Proper key management is significant to the security of a cryptosystem. In practice it is possibly the most difficult feature of cryptography because it includes user training, organizational and departmental relations, and management between all of these elements.

## 2. Literature Survey

The traditional encryption requires different users to encrypt their data with their own keys. Thus, similar data copies of different users will lead to different cipher texts, making deduplication impossible. data outsourcing raises security and privacy concerns. The third-party cloud providers to properly implement confidentiality, reliability checking, and access control mechanisms beside any insider and outsider attacks.

In existing system they are using standard encryption scheme for identify duplicate blocks of data in cloud storage. In Cloud Storage, standard encryption of the same files produces same key and same cipher text. So Data deduplication of encrypted data is impossible. When user misplaced the key, there was impossible to restore the original content of the file. It is compromise by attackers. The existing encryption algorithm does not maintain the key management scheme.

A new cryptographic primitive, Message-Locked Encryption (MLE)[10], where the key under which encryption and decryption are performed is itself derived from the message. MLE provides a way to accomplish secure deduplication (space-efficient secure outsourced storage), An objective is targeted by numerous cloud-storage providers. It provides definitions both for privacy and for a form of reliability that call tag consistency. Which is based on this organization, it make both realistic and hypothetical contributions. On the realistic side, it provide ROM security analyses of a natural family of MLE schemes that includes deployed schemes. On the hypothetical side the challenge is standard model solutions, and here make relations with deterministic encryption, hash functions protected on correlated inputs and the sample-then-extract paradigm to send schemes under different assumptions and for different classes of message sources. MLE shows the primitive of both realistic and hypothetical attention.

In the traditional storage file systems and storage hardware, every layers contains different kinds of information about the data they handle and such information in one layer is usually not available to any other layers. Code sign for storage and application is probable to optimize deduplication based storage system [1], when the lower-level storage layer has broad knowledge about the data structures and their access characteristics in the higher-level application layer.

Deduplication approach reduces the storage capacity needed to store the data or to transfer the data on network. In cloud backup data storage resources are available on demand which helps to reduce the network space by breaking up incoming stream of data into small segments. To identify such segments block index technique [7] is used. Data deduplication techniques for data reduction are the most effective behavior to promote data storage efficiency by deleting data redundancy. Data reduction technique includes data compression, delta encoding and deduplication.

Data compression eliminates redundancies contained by data objects to characterize original information using fewer bits. This can be any lossy or lossless. Lossless compression reduces bits via identifying and eliminating statistical redundancy in data. The LZ compression methods [2] is most accepted algorithms for lossless storage. It is universal lossless data compression algorithm and It is simple to implement and probable for very high throughput in network implementation.

Lossless compression technique reduces bits by identifying slightly important information and removing it. It gives an equivalent substitution between information loss and the size reduction. In some popular applications, as images, audio and video, several loss of information is acceptable. Data compression [4] only achieves a partial data reduction ratio due to its intra object data reduction nature.

Chunk-based storage system utilizes the file similarity instead of chunk locality [6], Index reside in RAM and index kept on

disk. Extreme Binning exploits file similarity instead of locality to make only one disk access for chunk lookup per file as a replacement for of per chunk, thus alleviating the disk bottleneck problem. The new data structure in application-aware deduplication[3] is an application-aware index structure can widely mitigate the disk index lookup bottleneck by dividing a central index into many independent small indices to optimize lookup performance.

Data de-duplication is performed through hybrid cloud architecture which does not eliminate the redundant data completely. For the data security purpose when the data is stored over the cloud only, and if some data loss occurs it is unable to recover the data. The convergent encryption technique used for encrypting the data is inefficient. The same privilege key is used by the user for storage and retrieval of data for every time. The same privilege key is easily predictable by the hackers or intruders.

## 3. Working

The Original Data block is selected to out sourced into the cloud service provider. The File can be already exists in cloud storage or block of file can be already exists in cloud storage. The File or Block of file is selected to upload into cloud service provider and check whether the file or block is already exists.

**Hash Key Generation**
The Hash Key is generated according to the content of the file. The tag and hash key is derived from content of the file independently. Key generation algorithm that maps a data copy M to a convergent key K. key generation algorithm that generates using security parameter. The purpose of generating Hash key is to encrypt the data block with hash key.

**Encryption and Compression**
Convergent encryption provides data confidentiality in de duplication. A user (or data owner) derives a convergent key from each original data copy and encrypts the data copy with the convergent key. The user derives a tag for the data copy which will be used to detect duplicates. Assume that the tag accuracy properties hold, i.e., if two data copies are the identical, then their tags are the same. To identify duplicates, the user initially sends the tag to the server side to verify if the identical copy has been previously stored. Note that both the convergent key and the tag are separately derived and the tag cannot be used to assume the convergent key and compromise data confidentiality. Compression is helpful because it helps reduce resource usage and data storage space or transmission ability. The process of reducing the size of a data file is referred to as data compression in the context of data transmission.

**File Uploading:**
User uploads a file F. First, it performs file level De duplication. On input file F, the user computes and sends the file tag. Upon receiving, the S-CSP checks whether there exists the same tag on the S-CSP. If S-CSP replies the user with a response file duplicate, orno file duplicate otherwise. If the user receives the response no file duplicate, then it jumps to proceed with block-level deduplication. If the response is file duplicate, then the user runs PoW with the S-CSP to prove that it actually owns the same file F that is stored on the S-CSP.

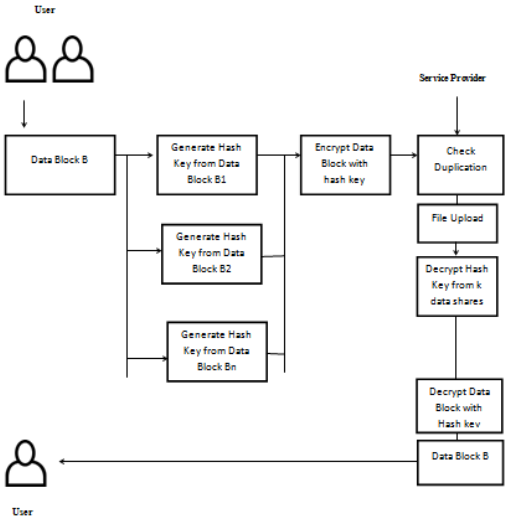



**Figure 2:** Architecture Diagram

**De Duplication**
Upon receiving the S-CSP checks whether there exists the similar tag on the S-CSP. If so, the S-CSP reply user with a response file duplicate, or no file duplicate. If the user receives the response ‘‘no file duplicate” then it jumps to proceed with block-level deduplication. If the response is “file duplicate” then the user runs PoW with the S-CSP to prove that it actually owns the same file F that is stored on the S-CSP. If PoW file F is passed, the S-CSP simply proceeds a file pointer of file F to the user, and no more information will be uploaded. If PoW file F fails, the S-CSP terminates the upload operation. After that user then performs block-level deduplication to additional remove any redundant blocks.

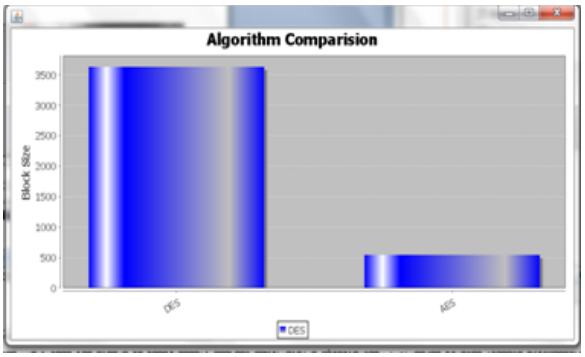
**File Downloading**
Suppose a user needs to download a file F. It initially sends a request and the file name to the S-CSP and performs the steps which are as follows- S1: Upon getting the request and file name, the S-CSP will check whether the user is suitable to download F. If user failed, then S-CSP would sends an abort signal to the user to be a sign of the download failure. Otherwise, the S-CSP returns the related cipher texts as fCig and the encrypted convergent keys fCKig to the user S2. S2: Upon getting the encrypted data from the S-CSP, the user primary uses their master key to recover all convergent key K to recover the original block Bi Decrypt CEðKi;CiÞ . Lastly, the user can get the original file F.

**Decompression and Decryption**
In order to use a compressed file, user have to primary decompressed it. The software used to decompress depends on how the file was compressed in the initial place. To decompress a .zip file user need software, such as WinZip. User downloads the file from the cloud service provider using hash key. Primary all user decrypt the hash key from k secret shares then user decrypt the data block with hash key.

**Convergent Encryption:**
Convergent encryption provides a feasible option to implement data confidentiality while realizing deduplication. It encrypts/decrypts a data replica with a convergent key, which is ensuing the computing of cryptographic hash value of the content of the data copy itself. After key generation and data encryption, users keep the keys and send the ciphertext to the cloud. While encryption is deterministic, the same data copies will generate the same convergent key as well as the same ciphertext. This allows the cloud to achieve deduplication on the ciphertexts. The ciphertexts can just be decrypted by the corresponding data owners with their convergent key. Thus the efficiently and reliably handle huge convergent keys, while achieving secure deduplication. Proposed system proposes a new structure called Dekey, which provides effectiveness and consistency guarantees for convergent key management on both user and cloud storage sides. The necessary idea is to concern deduplication to the convergent keys and force secret sharing techniques. Particularly, the secret shares for the convergent keys and distribute them across several independent key servers. Only the initial user who uploads the data is crucial to compute and distribute such secret shares, while all subsequent users who own the matching data replica not include compute and store these shares again. To improve data copies, a user must access a minimum number of key servers through authentication and obtain the secret shares to reconstruct the convergent keys.

## 4. Experimental Analysis

When user uploading the data and encrypting the data with encryption algorithm. Here proposed system compares the two encryption algorithm such as DES and AES on the basis of block size. As DES had 64bits block size and AES has 128bits block size. So the no. of blocks require to send over the network in DES are greater than that of AES. From these analysis the AES is more efficient than that of DES.




## 5. Conclusion

Propose an efficient and reliable convergent key management scheme for secure deduplication. Dekey applies deduplication with convergent keys and distributes convergent key shares across several key servers, preserve semantic security of convergent keys and privacy of outsourced data. Execute Dekey using the Ramp secret sharing scheme and show that it incurs small encoding/decoding overhead compared to the network transmission overhead in the regular upload/download operations.

## References

[1] Yinjin Fu, Hong Jiang, "Application-Aware Local-Global Source Deduplication for Cloud Backup Services of Personal Storage," IEEE Transactions On Parallel And Distributed Systems, VOL. 25, NO. 5, MAY 2014.
[2] Yin-Jin Fu, Nong Xiao, "Application-Aware Client-Side Data Reduction and Encryption of Personal Data in Cloud Backup Services," journal of computer science and technology 28(6):1012-1024 Nov. 2013.
[3] J. Malhotra, P. Ghyare, "A Novel Way of Deduplication Approach for Cloud Backup Services Using Block Index Caching Technique," IJAREEIE Vol.3, Issue 7, July 2014.
[4] A. ElShimi, R. Kalach, A. Kumar, J. Li, A. Oltean, and S. Sengupta, "Primary Data Deduplication Large Scale Study and System Design," in Proc. USENIX ATC, 2012, pp. 285-296.
[5] Y. Fu, H. Jiang, N. Xiao, L. Tian, and F. Liu, AA-Dedupe: "An Application-Aware Source Deduplication Approach for Cloud Backup Services in the Personal Computing Environment,"" in Proc. 13th IEEE Int"l Conf. CLUSTER Comput., 2011, pp. 112- 120.
[6] K. Eshghi, H. Khuern Tang, "A Framework for Analyzing and Improving Content-Based Chunking Algorithm," Hewlett-Packard Laboratories palo Alto, CA Feb 25,2005.
[7] D. Bhagwat, K. Eshghi, "Extreme Binning: Scalable, Parallel Deduplication For Chunk-based File Backup," 17th IEEE/ACM International Symposium on Modelling, Analysis and Simulation of Computer and Telecommunication Systems (MASCOTS" 2009), London, UK, September 2009.
[8] J. Li, X. Chen, M. Li, "Secure Deduplication With Efficient And Reliable Convergent Key Management," Ieee Transactions On Parallel And Distributed Systems, Vol. 25, No. 6, June 2014
[9] A. Katiyar and J. Weissman, \ ViDeDup: An Application-Aware Framework for Video De-Duplication," in Proc. 3rd USENIX Workshop Hot-Storage File Syst., 2011, pp. 31-35
[10] M. Bellare, S. Keelveedhi, and T. Ristenpart, ''Message-LockedEncryption and Secure Deduplication,'' in Proc. IACR CryptologyePrint Archive, 2012, pp. 296-3122012:631.