Secure Cloud Using Secure Data Deduplication Scheme

Supriya More[1], Sharmila Gaikwad[2]

[1]P.G. Student, Department of Computer Engineering, Rajiv Gandhi Institute of Technology, Mumbai, India
[2]Assistant Professor, Department of Computer Engineering, Rajiv Gandhi Institute of Technology, Mumbai, India

---***---

Abstract- *The cloud computing is metaphor for the Internet hence recent years have witnessed the trend of leveraging cloud based services for large scale content storage, processing, and distribution. Information deduplication is one of essential information pressure strategies for dispensing with copy duplicates of rehashing information, and has been generally utilized as a part of distributed storage to diminish the measure of storage room and spare transfer speed. Data Deduplication procedure identifies recurring data forms and reduces them to a single instance to save volume in the storage area network. To better ensure secure deduplication in cloud, this study paper discus secure deduplication for image, text and video.*

Keywords: Deduplication, hybrid cloud, image deduplication, text deduplication, video deduplication.

1. INTRODUCTION

Data deduplication is one of the critical information thickness methods for disposing of copy duplicates of recapping information, and it has been broadly utilized as a part of distributed storage to diminish the measure of storage area and spare data transfer capacity.

Definition 1: Data Deduplication is a technique which eliminates redundant data by storing only a single copy of each file, it reduces the space and the bandwidth requirements of data storage services like cloud, Data warehouse. It provides major saving in backup environments by using data compression and data deduplication. It is the most impactful storage technology used by following companies

1. In April 2008, IBM acquired Diligent
2. In July 2009, EMC acquired Data Domain
3. In July 2010, Dell acquired Ocarina

Many more company provides cloud-based storage such as Dropbox, Google Drive and Mozy can save money on capacity costs by means of deduplication [2], it means when two customers transfer the same record, the administration recognizes this and stores just a solitary duplicate. All of this are unique in relation to customary deduplication frameworks, the differential benefits of companies or data owner or data user are additionally considered in copy check other than the information itself and access only unique information which it want from cloud. This research demonstrates a few new deduplication developments supporting approved duplication check in a public as well as private and hybrid cloud storage. As mention in [3], to measure data duplication expression (1) is used for calculation of data deduplication rate.

The data duplication rate as:

$$\frac{\text{Projected Storage Size-Actual Storage Size}}{\text{Projected Storage Size}} \text{X100} \qquad (1)$$

Where the projected storage size is the sum of all data to backup, actual storage size is the actual amount of storage mandatory due to data de-duplication. For instance, if we need to store a total amount of 100 GB data from two machines A and B, and only 70 GB is needed for actual storage, then the data duplication rate for A and B is 30%.

2. MOTIVATION

With the arrival of cloud computing and its digital storage services, the growth of digital content has become uncontrollable at both enterprise and individual levels. According to the EMC Digital Universe Study (Gantz and Reinsel 2010), the global data supply had already reached 2.8 trillion Giga Bytes (GB) in 2012, with the expectation that volumes of data are projected to reach about 5247GB per person by 2020 [16]. Due to this explosive growth of digital data, there is a clear demand from Cloud Service Provider's (CSP) for more cost effective use of their storage and network bandwidth for data transfer. Previous deduplication systems cannot support secure deduplication in an authorized deduplication framework, every customer is issued a set of priority at the time of system initialization.

Targets behind secure deduplication are as per the following

1. Billing nature of cloud services

a. Pay As You Go: User needs to pay charges as per disk space utilized by him. If duplicate copies of file exist then user need to pay for duplicate file.

b. Duplicate file upload also increase bandwidth utilization, so it degrades network performance.

c. User need to afford higher cost for large bandwidth uses.

2. Access to Authorize Users

a. In cloud computing environment, same file could be shared to many users. Hence, only authorized user should have control of accessing shared file.

Authorized users should get download access to shared files in his access domain.

3. Confidentiality

a. Cloud service providers are the third party service providers. So, it's not secure to store confidential contents as it is on cloud.

b. To maintain confidentiality there is need to implement encryption and decryption scheme.

4. Indexing & Retrieval:

a. As deduplication avoiding duplicate data storage, document retrieval will be more efficient as index takes smaller space than files itself.

3. DEDUPLICTION STRATEGIES

This section discuss about two types of deduplication strategies first is file level deduplication and block level deduplication.

3.1 File level deduplication

File level deduplication is also known as Record level deduplication, as the name proposes, is constantly performed over a solitary document. Recognizable proof of same hash estimation of at least two records verifies that the documents are comparable [13]. It does not break the files into smaller chunks but rather uses entire file as chunks. As shown in Fig 1.This method only eliminates duplicate files and keep on single instance of file.

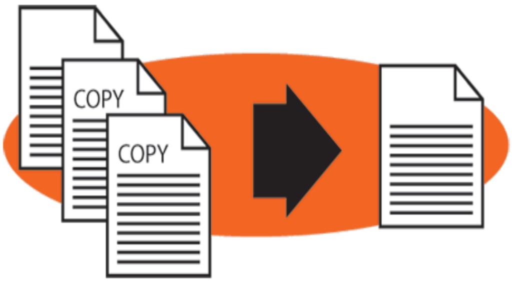

Fig 1: File Level Deduplication

i. Advantages of File level deduplication

(a.) Indexing performance

(b.) Low CPU usage

ii. Disadvantage of File level deduplication:

(a.) It cannot be applied for the large files with changing data.

3.2 Block Level deduplication

Block level deduplication is also known as square level deduplication is performed over pieces. Firstly, it isolates the records into pieces and stores only a solitary duplicate of every square. Settled measured pieces or variable sized lumps can be utilized with square level deduplication [13]. Fig 2 shows block level deduplication where deduplication is based on piece of file. There are two types of block level deduplication.

3.2.1 Fixed Size Chunking

Instead of using entire file as smallest unit it breaks the file into equally sized chunks, if a large file is changed then only the changed chunks must be reindexed and transferred to the backup.

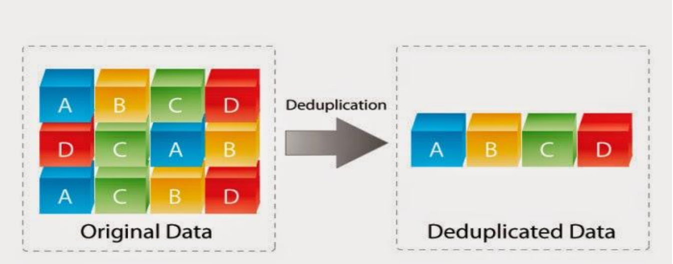

Fig 2: Block Level Deduplication

i. Disadvantage of Fixed size chunking:

It fails to detect redundant data if some bytes are files into equally sized chunks be re-indexed and transferred to the backup location inserted or deleted from the file because chunk boundaries are determined by offset rather than by content.

3.2.2 Variable Size Chunked

It defines breakpoints. This is usually done by fixed size overlapping sliding. At every offset of a file, the contents of the sliding becomes true window are analysed and a finger print f is calculated. If f satisfies the break condition, a new break point has been found and new chunk is created.

4. LITERATURE SURVEY

With the appearance of distributed computing, secure information deduplication has pulled in much consideration as from research group.

1) DupLESS: Server-Aided Encryption for Deduplicated Storage

Creator Mihir Bellere [1] designed Duplicate less Encryption for Simple Storage (DupLESS).Here client encrypt there data using key server (KS). Authentication of client is also done by key server and storage service are separate entity KS is inaccessible to attacker.

2) A Hybrid Cloud Approach for Secure Authorized Deduplication

The creators Jin Li, Yan Kit Li, Xiao Feng Chen, Patrick P. C. Lee and Wenjing Lou introduced that, several new deduplication constructions supporting authorized duplicate check in hybrid cloud architecture. As a proof of concept, introduced a prototype to system for authorized duplicate check scheme. It shows that system authorized duplicate check scheme incurs minimal overhead compared to normal operations [2].

3) Fast and Secure Laptop Backups with Encrypted Deduplication

The paper composed by creator Paul Anderson [3] about scrambled deduplication, an algorithm which takes advantage of the data which is common between users to increase the speed of backups, and reduce the storage requirements. This algorithm supports client end per user encryption which is necessary for confidential personal data. Files are copied into the backup store as file objects using the convergent encryption process. Directories are stored as directory objects, these are simply files which contain the normal directory Meta data for the children, and the encryption/hash keys for each child. It also supports a unique feature which allows immediate detection of common sub trees, avoiding the need to query the backup system for every file.

4) A Secure Video Deduplication Scheme in Cloud Storage Environments using H.264 Compression

The creator Fatema Rashid, Ali Miri, Isaac Woungang discussed in [4], a scheme is proposed that achieves a secure video deduplication in cloud storage environments. Its design consists of embedding a partial convergent encryption along with a unique signature generation scheme into a H.264 video compression scheme. The partial convergent encryption scheme is meant to ensure that the proposed scheme is secured against a semi-honest Cloud Service Provider (CSP) and the unique signature generation scheme is meant to enable a classification of the encrypted compressed video data in such a way that the deduplication can be efficiently performed on them.

5) Secure Deduplication with Efficient and Reliable Convergent Key Management

Authors Jin Li, Xiaofeng Chen, Mingqiang Li, Jingwei Li, Patrick P.C. Lee, and Wenjing Lou, formally address the problem of achieving efficient and reliable key management. They introduced a baseline approach, here user holds an independent master key for encrypting the convergent keys. Dekey is a proposed method by which users do not need to manage any keys on their own but instead securely distribute the convergent key shares across multiple servers [5].

6) Proof of ownership in remote storage system

Creators Shai Halevi, Danny Harnik, Benny Pinkas, and Alexandra Shulman-Peleg, identify attacks. This attack exploit client side deduplication. It allow an assailant to gain access to arbitrary size files of other users based on very small hash signatures of these files. The hash signature is known to attacker. Hence, assailant can convince the storage service and the server lets the assailant download the entire file. To overcome such attacks, Creators introduce the notion of proof of ownership, which lets a client efficiently prove to a server that that the client holds a file, rather than just some short information about it. Creators assigned the approach of proof of ownership, under rigorous security definitions, and rigorous ability requirements of Petabyte scale storage systems [6].

7) RevDedup: A Reverse Deduplication Storage System Optimized for Reads to Latest Backup

The creator Chun-Ho Ng and Patrick P. C. Lee discussed about RevDedup, is a deduplication system that optimizes reads to the latest backups of virtual machine (VM) images using reverse deduplication. RevDedup discard duplicates from old data. It is done by shifting dissolution to old data while keeping the layout of new data as sequential as possible. It evaluates RevDedup prototype using a 12 week span of real world VM image snapshots of 160 users. RevDedup achieves high deduplication efficiency, high backup throughput, and high read throughput [7].

8) Twin Clouds: An Architecture for Secure Cloud Computing

The creator Sven Bugiel, Stefan, Ahmad-Reza Sadeghi, and Thomas Schneider introduced architecture for secure outsourcing of data and arbitrary computations to an untrusted commodity cloud. Here, the user communicates with a trusted cloud. It encrypts and verifies the data stored. Creators split the computations such that the trusted cloud is mostly used for security critical operations in the less time critical setup phase. Queries to the outsourced data are processed in parallel by the fast commodity cloud on encrypted data [8].

9) Private Data Deduplication Protocols in Cloud Storage

The creator Wee Keong Ng, Yonggang Wen and Huafei Zhu introduced private data deduplication protocol. A private data deduplication protocol allows a client who holds a private data and proves to a server by summary string of the data. It's client responsibility that data should not revealing further information to the server. A construction of private deduplication protocols based on the standard cryptographic assumptions is then presented and analysed [9].

10) A Secure Client Side Deduplication Scheme in Cloud Storage Environments

The creator Nesrine Kaaniche, Maryline Laurent introduce, a new client side deduplication design for firmly storing and sharing outsourced data via the public cloud is known as

Open Stack. Initially, it ensures better confidentiality against unauthorized users. Every client computes per data key to encrypt the data that is store in the cloud. Here, this data access is managed by the data owner. In short, Swift is a cloud based storage system, which stores data and allows write, read, and delete operations on them [10].

11) Venti: a new approach to archival storage

The creator Sean Quinlan and Sean Dorward describes a network storage system, called Venti. In this, unique hashes of a block's contents acts as the block identifier for read and write operations. This approach enforces a write once policy. It prevented accidental or malicious destruction of data. The redundant copies of a block can be coalesced, reducing the consumption of storage and simplifying the implementation of clients. Venti is a building block for constructing a variety of storage applications such as logical backup, physical backup and snapshot file systems [11].

12) A Secure Cloud Backup System with Assured Deletion and Version control

The creator Arthur Rahumed, Henry C. H. Chen, Yang Tang, Patrick P. C. Lee, and John C. S. Lui shows that, Fade Version, a secure cloud backup system that serves as a security layer on top of today's cloud storage services. Fade Version pursue the standard version controlled backup design. It disposed the storage of redundant data across different versions of backups. On top of this, Fade Version applies cryptographic protection to data backups. Creators implement a proof-of-concept prototype of Fade Version and conduct empirical evaluation atop Amazon S3. The main goal is to make both version control and assured deletion compatible with each other in a single design [12].

13) A Survey Paper on Removal of Data Duplication in a Hybrid Cloud

The creator Prakash Gapat, Snehal Khillare, Akshay Khiste, and Rohini Pise have shown difference between traditional encryption algorithm and Convergent Encryption Algorithm. For the preservation of delicate data convergent encryption technique is used and then data is stored on cloud storage, In order to make system more secure, the different privileges to users are again considered while checking duplicate content. But the problem which occurs in this approach is that even if some contents of both files are different, it stores them as two different files which lead to reduction in cloud storage space. The solution for this problem is to perform apply technique for deduplication (block level).In this approach the file which is to be stored on cloud storage is divided into number of different blocks based on contents and deduplication is performed on these blocks [13].

14) A Hybrid Cloud Approach for Secure Authorized Deduplication

The creator Prof. N.B. Kadu, Mr. Amit Tickoo, Mr.Saurabh I. Patil, Mr. Nilesh B. Bhagat, and Mr. Ganesh B. Divte introduced that, new deduplication constructions supporting authorized duplicate check in a hybrid cloud architecture. As a proof of concept, creators implement a prototype of proposed authorized duplicate check scheme and conduct testbed experiments using prototype. Creators conclude that authorized duplicate check scheme incurs minimal overhead compared to convergent encryption and network transfer [14].

5. SECURE DEDUPLICATION SCHEME FOR DATA FORMATS

This section discuss secure deduplication scheme for various data format like Image, text and video.

5.1 Secure Deduplication of Image

As mention in [15], Image deduplication contain four methods

i. Feature Extraction: Image deduplication needs a mechanism to compare image content [15, 17].

ii. High-dimension Indexing: In large scale image retrieval, because of the influence of "curse of dimensionality", the traditional indexing technologies fall sharply when facing high dimensional data. So, effectively organize index structure to improve the processing capabilities of high dimensional.

iii. Accuracy Optimization: Although the grey block feature can effectively identify duplicate images, but it is a coarse comparison based on block units, and the image details can't be well handled.

iv. Centroid Selection: In image deduplication, the collection of duplicate images can be regarded as a cluster. The highest perceptual quality representative image is called as centroid image. A centroid image is regarded as the center point in a cluster. Other images in the cluster, which are created logical points to the centroid image, can be derived from the centroid image by using the standard image transformations like downscaling, down-quality, storage format conversion, etc.Secure deduplication of images in the cloud involves three components, namely, an image compression scheme, a partial encryption scheme, and a hashing scheme as shown in Fig 3. Typically, on the user's side, the user will process the image by applying image compression, partial encryption, and will calculate the hash signature of the image, in that order. Next, he/she will only send the hash signature to the Cloud Service Provider (CSP). Image hash is based on Set Partitioning in Hierarchical Tree (SPIHT) compression algorithm. The SPIHT algorithm is based on the fact that there is a correlation between the coefficients that are in different levels of the hierarchy pyramid of the underlying structure. It maintains this information in the zero trees by grouping insignificant coefficients together. Typically, each 2x2 block of coefficients at the root level of this tree structure corresponds to further trees of coefficients. Basically, the SPIHT algorithm can be in three phases, namely, initialization, sorting, and refinement phases. On the CSP side, the CSP will compare the received image hash against all the signatures already present in the cloud storage [16]. If a match is not found, the CSP will instruct the user to upload the image. Otherwise, the CSP will update the image metadata and then will deduplicated the image by saving only a single, unique copy.

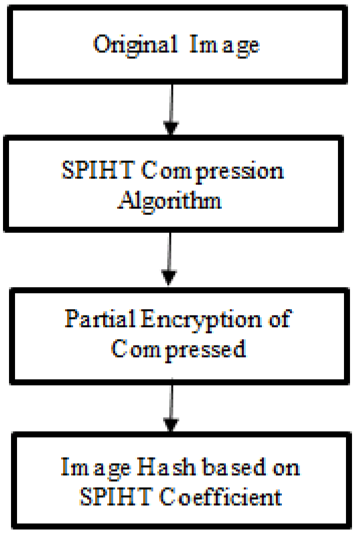
Original Image
SPIHT Compression Algorithm
Partial Encryption of Compressed
Image Hash based on SPIHT Coefficient

Fig 3: Image Deduplication Process

Table 1 shows difference between text file deduplication and Image deduplication. The main difference is in text file deduplication it store only single text file but while image deduplication first selection of centroid is required [15].

Table 1. Difference between text deduplication and Image deduplication

	Text Deduplication	Image Deduplication
1	Read Text File	Read Image File
2	Data Partition	Image Pre-process
3	Hash Computation	Feature Extraction
4	Index Lookup: an exact matching	Index Lookup: an approximate matching
5	Accuracy Optimization: comparison byte by byte	Accuracy Optimization: compare number of same element
6	Store one copy	Centroid selection and store centroid image

5.2 Secure Deduplication Scheme for video

The detailed description of the Video deduplication scheme as mention in [4] is as follows.

1. H.264 Video Compression:

i.The prediction step: The coded video sequence generated by the H.264 algorithm is made of a sequence of coded pictures. Each picture is divided into fixed size macro blocks. Macro blocks are the basic building blocks in the H.264 algorithm.

ii.The transform step: This includes the transform, scaling, and quantization sub steps. In the H.264 algorithm, an integer transform such as the 4×4 Discrete Cosine Transform (DCT) is used.

iii.The entropy coding: In the H.264 algorithm, two entropy coding are supported named as context adaptive variable length coding (CAVLC) and the context adaptive binary arithmetic coding (CABAC). The CABAC has better coding efficiency compared to the CAVLC.

2. Signature Generation from the Compressed Videos

The signature generation is carried out in the compressed domain, and the signatures are generated from the information produced in the transform domain of the H.264 compression algorithm. The content dependent robust bits are extracted from the macro blocks and are further used as the signature for authenticating the compressed video.

3. Selective Encryption of the Compressed Videos

The encryption process starts by first generating the compressed bit stream for the video. First, the user classifies the video into six different categories, namely high, medium and low intensity motion (for complex texture) and high, medium and low intensity motion (for non-complex texture) by utilizing the information generated in the intra prediction mode, Discrete Cosine Transform (DCT) coefficients, and motion vectors. The convergent encryption is employed to derive the key for partial encryption from the content of the compressed video rather than getting the key chosen by the users individually. Therefore, for the same content, the same key will be generated without the users knowing each other. Thus, different users will have the same key as well as the same encrypted videos, irrespective of the knowledge of each other keys. This will make it easier for the CSP to compare the encrypted parts of the videos and perform deduplication in case of duplicated videos without actually decrypting or decompressing the video data.

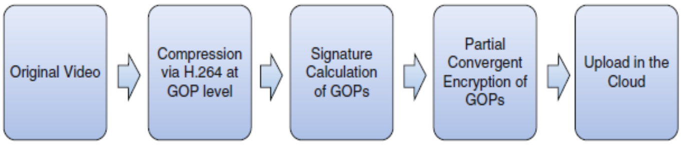
Original Video
Compression via H.264 at GOP level
Signature Calculation of GOPs
Partial Convergent Encryption of GOPs
Upload in the Cloud

Fig 4: Secure Deduplication of Video data format

Secure deduplication of videos in the cloud storage involves three components: H.264 video compression scheme, signature generation from the compressed videos, and selective encryption of the compressed videos as shown in Fig.5. First, the user compresses the original video using the H.264 compression algorithm. Second, he/she calculates the signatures based on the Group of Picture (GOP) from the output bit stream. Third, he/she encrypt the important parts of DCT coefficients and motion vectors according to the type of the videos. After these processing steps on the original video, it will be uploaded in the cloud storage. The CSP will then check for the identical GOPS with the help of the signatures and encrypted data. If identical GOPs are detected, the CSP will delete the new data and update the metadata for this particular data already in the cloud storage. In this way, the CSP will save huge space by performing cross-user video deduplication in the cloud storage [4].

6. HYBRID CLOUDE FOR SECURE DEDUPLICATION

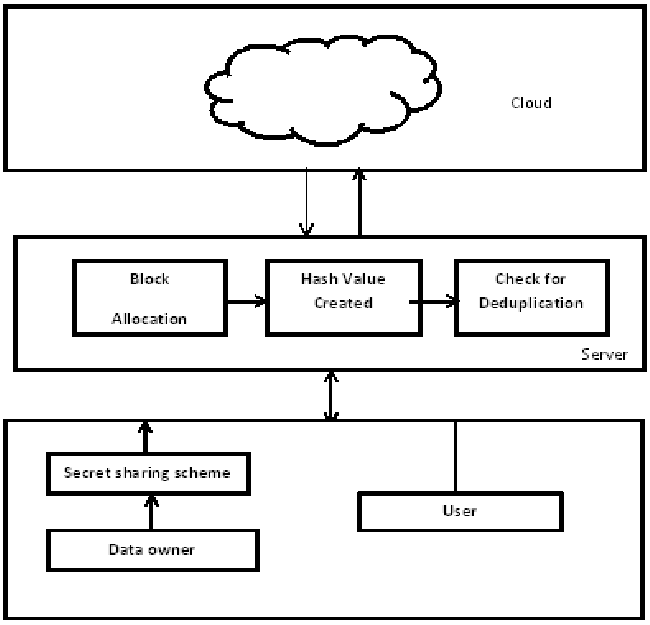

Fig 5: Architecture of Hybrid Cloud for Secure Deduplication

The architecture shown in Fig.5 includes three entities viz. the user, the storage cloud service provider (S-CSP) and server. The task of each is as given below:

Cloud Service Provider: This is an element that gives an information stockpiling service out in the open cloud. The S-CSP as shown in [2], 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.

Data Users: A user is an object that wants to outsource data storage from the Cloud service provider and access the data later. In a storage system supporting deduplication, the user only uploads unique data but does not upload any identical data to save the upload bandwidth, which may be owned by the same user or different users. In that cloud service provider maintained security by using security algorithm, each document is secured with the Convergent key encryption and Symmetric key encryption [2].

Server*:* This is a middleware of cloud service provider and data user, it accomplish main functioning, and it perform all security related and it also eliminate duplicate data by executing various algorithm and technique.

To determine security in these model following criteria should be consider as shown in [2].

Authorized copy check: Authorized client can utilize his/her individual private keys to produce question for certain document and the benefits he/she claimed with the assistance of private cloud, while the general population cloud performs copy check specifically and tells the client if there is any copy. For the security of record token, two viewpoints are characterized as un-produce capacity and in-recognize capacity of document token. The subtle elements are given underneath.

Enforceability of record token/copy check token: Unauthorized clients without suitable benefits or record ought to be kept from getting or creating the document tokens for copy check of any document put away at the S-CSP. In distinguishability of document token/copy check token, it requires that any client without questioning the private cloud server for some record token, user cannot get any valuable data from the token, which incorporates the document data or the benefit data. Data Confidentiality is also maintain [2].

7. ADVANTAGES

Following advantages of data deduplication are based on [18]

1. Administration of the regularly expanding volume of information is done.

2. Reduced storage allocation and efficient volume replication- Only unique data is written to disk hence reduce storage allocation.

3. Effectively increase network bandwidth- No duplicate copies need to be transmitted over the network if deduplication is takes place at the source.

4. Fast Recoveries ensure that line-of business process continue unimpeded. This will helpful in disaster management.

5. Using authorized deduplicated check, convergent encryption, signature security is maintained while image, text, video data deduplication hence cloud computing become secure.

8. CONCLUSION

In this paper, different deduplication strategies like file level and block level deduplication was studied. Also discussed secure deduplication methods for data formats like image deduplication, text deduplication, video deduplication. Finally, discussed a few new deduplication developments supporting approved copy check in hybrid cloud, in which the copy check tokens of documents are produced by the private cloud server with private keys.

REFERENCES

[1] M. Bellare, S. Keelveedhi, and T. Ristenpart., "DupLESS: Server aided encryption for deduplicated storage", USENIX Security Symposium, 2013.

[2] Jin Li, Yan Kit Li, Xiaofeng Chen, Patrick P. C. Lee, Wenjing Lou.", Hybrid Cloud Approach for Secure Authorized Deduplication" TPDS.2014.2318320, IEEE, 2014.

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

[4] Fatema Rashid, Ali Miri, Isaac Woungang "A Secure Video Deduplication Scheme in Cloud Storage Environments using H.264 Compression", First International Conference on Big Data Computing Service and Applications, IEEE, March 2015.

[5] J. Li, X. Chen, M. Li, J. Li, P. Lee, and W. Lou. "Secure deduplication with efficient and reliable convergent key management". In IEEE Transactions on Parallel and Distributed Systems, 2013.

[6] S. Halevi, D. Harnik, B.Pinkas and A.Shulman-Peleg. "Proofs of ownership in remote storage systems". In Y. Chen, G. Danezis, and V. Shmatikov, editors, ACM Conference on Computer and Communications Security, pages 491–500. ACM, 2011.

[7] Ng and P. Lee. "RevDedup: A reverse deduplication storage system optimized for reads to latest backups". In Proc. of APSYS, Apr 2013.

[8] S. Bugiel, S. Nurnberger, A. Sadeghi, and T. Schneider. "Twin clouds: An architecture for secure cloud computing". In Workshop on Cryptography and Security in Clouds (WCSC 2011), 2011.

[9] W. K. Ng, Y. Wen, and H. Zhu. "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.

[10] Nesrine Kaaniche, Maryline Laurent. "A Secure Client Side Deduplication Scheme in Cloud Storage Environments", Proceedings of the FAST 2002 Conference on File and Storage Technologies, Monterey, California, USA January 28-30, 2002.

[11] S. Quinlan and S. Dorward." Venti: A new approach to archival storage". In Proc. USENIX FAST, Jan 2002.

[12] 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 3rd International Workshop on Security in Cloud Computing, 2011.

[13] Prakash Gapat, Snehal Khillare, Akshay Khiste, Mrs. Rohini Pise, "A survey paper on Removal of Data Duplication in a Hybrid Cloud". IEEE Computer, 395-0072, Jan-2016.

[14] N.B. Kadu, Amit Tickoo, Saurabh I. Patil, Nilesh B. Bhagat, Ganesh B. Divte "A Hybrid Cloud Approach for Secure Authorized Deduplication", International Journal of Scientific and Research Publications, Volume 5, Issue 4, ISSN 2250-3153, April 2015

[15] Ming Chen, Shupeng Wang and Liang Tian "A High-precision Duplicate Image Deduplication Approach" JOURNAL OF COMPUTERS, VOL. 8, NO. 11, NOVEMBER, 2013.

[16] Fatema Rashid and Ali Miri, "Deduplication Practices for Multimedia Data in the Cloud", S. Srinivasan (ed.), Guide to Big Data Applications, Studies in Big Data 26, DOI 10.1007/978-3-319-53817-4_10, 2018.

[17] Tzay-Yeu Wen,"Large Scale Image Deduplication", PDF available online http://vision.stanford.edu/teaching/cs231a_autumn1213_internal/project/final/writeup/nondistributable/Wen_Paper.pdf

[18] Venencia D'costa," What are the real benefits of data deduplication in Cloud?", blog available online on http://blog.webwerks.in/cloud-hosting-blog/what-are-the-real-benefits-of-data-deduplication-in-cloud