International Journal of Scientific Engineering and Research (IJSER)
ISSN (Online): 2347-3878 (UGC Approved, Sr. No. 48096)
Index Copernicus Value (2015): 62.86 | Impact Factor (2015): 3.791

Implementation of Cloud Deduplication Mechanism Using Sentence Based Chunking On Hybrid Cloud Architecture

Arti Walzade[1], Vidhate S.P[2]

[1]Pune University, Vishwabharti Academy College of Engineering, Ahmednagar, Maharashtra India

[2] Professor, Pune University, Vishwabharti Academy College of Engineering, Ahmednagar, Maharashtra India

Abstract: ***Now a day's cloud technology are being continuously used in IT field as well as in other to keep the data, cloud re-copying is also a service which has to be focused on. As the cloud service has improved huge focus in last few years. Cloud storage massive management has become important. As the cloud computing services are rapidly being used in recent days for storage and other purposes, cloud deduplication is also such service which has to be focused on. Data deduplication is a technique for reducing the amount of storage space an organization needs to save its data. In most institutes, the storage systems contain duplicate copies of many pieces of data. Deduplication eliminates these extra copies by saving just one copy of the data and replacing the other copies with pointers that lead back to the original copy. For backup and disaster recovery applications, most companies and organizations used deduplication but it can be used to free up space in primary storage as well. To avoid this duplication of data and to maintain the confidentiality in the cloud Hybrid cloud concept is used. For safety and to secure the confidentiality of sensitive data while supporting deduplication, the convergent encryption technique has been proposed to encrypt the data before out sourcing. To better protect data security, this paper makes the first attempt to formally address the problem of authorized data deduplication using content level duplication checking. This paper surveying various works previously done in the area of cloud services and therefore, re-copying of data over cloud storages has still scope of improvement. Inspection on the papers or researches emphasizes various deduplication idea and the ways they are differ from each other for efficient deduplication. Thus as there are many ways of deduplication in clouds, an efficient technique is to be search out having less drawbacks and more outcome.***

Keywords: re-copying, hybrid storage on cloud, authorization, data security, privilege, deduplication, credential, hybrid cloud

1. Introduction

Private, community or public clouds are the compositions of hybrid cloud, offering the benefits of multiple deployment models.

Current time is distributed computing time. Distributed computing has massive variety of degree in data sharing in current period. Distributed computing is give accurate measure of virtual environment concealing the stage and working frameworks of the client. Client use the assets for exchanging information. It may, client need to pay by the process of utilization of assets of cloud. Client can transfer the vast sum data on cloud and exchanged information to a large number of clients. Cloud suppliers are offer diverse administrations, for e.g., framework as an administration, stage as an administration, and so forth. Client not has to buy the assets. As the data is get exchanged by the client might be it is basic notification to deal with this regularly expanding information on the cloud. To make well information administration in the distributed computing. We use duplication technique, which is the best technique in cloud. This technique is turning out to be more moderation for information DE duplication. This system sends the information over the system required little measure of information. This technique has application in information administration and organizing. Information duplication is the procedure of decreasing copy file Also it is the best pressure system for the information DE duplication. This system has application in information administration and organizing. Rather than keeping excess duplicate file of the same information DE duplication just keep unique duplicate and give just references of the first duplicate to the repetitive information. The process of checking the duplication process is two; one is document level duplication check and second is piece content level duplication check. In the document level duplication technique check is expel the same name record from the capacity and square level DE duplication are evacuated the copy pieces. DE duplication techniques need of the some security system. In the conventional system client need to encode his own particular information.

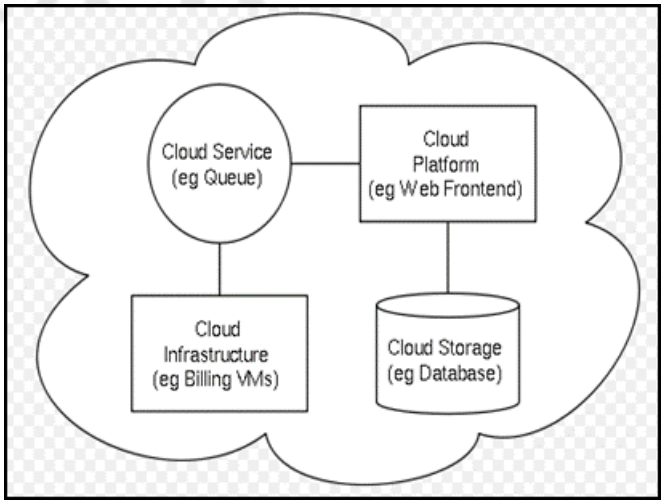


Figure 1: Cloud architecture and services

To maintain a security from the unapproved information DE duplication focalized information DE duplication is proposed to uphold the data privacy while checking the information

duplication. The cloud giving various administrations as attended in the above figure, for example, stage, administrations, base as an administration, and database as an administration.

In this part we are utilizing as a part of distributed storage as an administration. We are utilizing client accreditations to check the confirmation of the client. In that cases cloud is available two sort of cloud such private cloud and open cloud. In private cloud store the client accreditation and in the open cloud client information present out. Open cloud and private cloud are available in the half and half cloud structural engineering. When any client forward solicitation to people in general cloud to get to the data he have to present his data to the private cloud then private cloud will give a record token and client can get the notifications to the document lives on the general population cloud. We have utilized a half and half cloud construction modeling as a part of proposed. We have to need to mind the file name in record information duplication and information DE duplication is checked at the square level. On the other hand, client needs to recover his information or download the information record he have to download both of the document from the cloud server this will prompts perform the operation on the same record this abuses the security of the distributed storage.

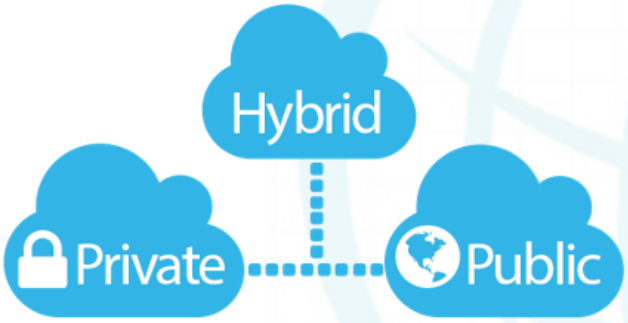


Figure 2: Hybrid Cloud Architecture.

2. Literature Survey

A. DupLESS: Server-Aided Encryption for Deduplicated Storage

Showing the e.g., Mozy, Dropbox and others perform deduplication to spare space by just putting away one duplicate of every document or file transferred. Should customers routinely scramble their documents, be that as it may, funds are lost. Message-bolted encryption (the most unmistakable appearance of which is concurrent encryption) certify this strain. In any case it is intrinsically subject to savage power assaults that can recoup records falling into a known set. In DupLESS, customers encode under message-based keys acquired from a key-server by means of an absent PRF convention. It secures customers to store scrambled information with a current administration, have the administration perform deduplication for their advantage, but then accomplishes solid privacy ensures. We demonstrate that encryption for deduplicated stockpiling can accomplish execution and space reserve funds near that of consuming the stockpiling administration with plaintext information [1].

B. Fast and Secure Laptop Backups with Encrypted De-duplication

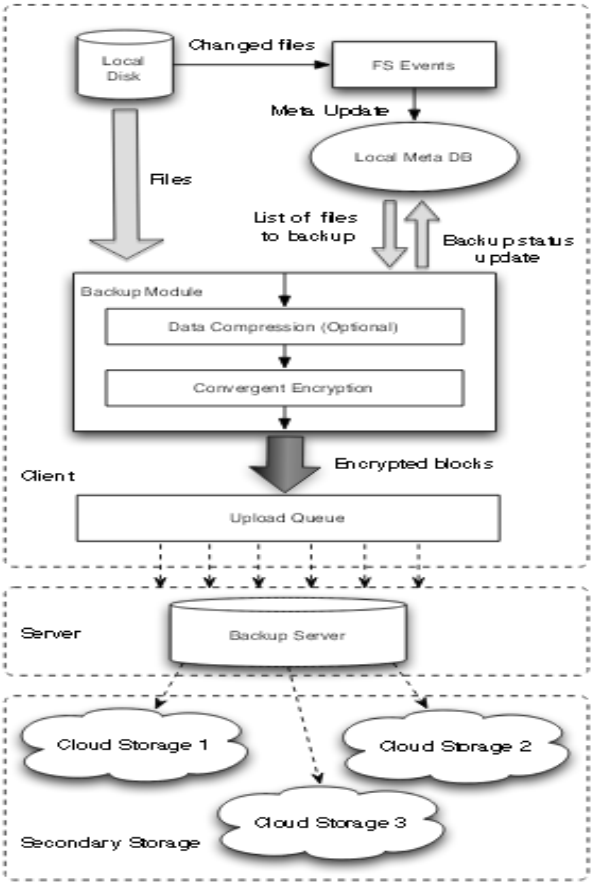


Figure 3: System diagram

Different types or individual data now store extensive amounts of individual and corporate information on tablets or home PCs.

By doing this type of work it is helpless against burglary or equipment disappointment. Ordinary ideal arrangements are not appropriate to this environment, and reinforcement administrations are every now and again deficient. This paper depicts a calculation which exploits the information which is basic between clients to build the pace of reinforcements, and decrease the capacity necessities. This calculation bolsters customer end per-client encryption which is essential for classified individual information. It like-wise underpins a one of a kind element which permits prompt location of normal sub trees, dodging the need to question the reinforcement framework for each document. It means the same data uses by different users have take large space and reduce the performance of your PC. We portray a model usage of this calculation for Apple Operating System X, and present an investigation of the potential viability, utilizing genuine information acquired from an arrangement of ordinary clients. At last, we talk about the utilization of this model in conjunction with remote distributed storage, and present an investigation of the common place cost reserve funds [2].

C. Secure Deduplication with Efficient and Reliable Convergent Key Management

Deduplication is a system for taking out copy duplicates of information, and has been broadly utilized as a part of distributed storage to decrease storage space and transfer data transfer capacity. Promising as it perhaps, an emerging test is to perform secure deduplication in distributed storage. In this paper, Albeit joined encryption has been widely received for secure deduplication, a basic problem of making focalized encryption down to earth is to productively and dependably deal with an immense number of united keys. This paper makes the first endeavor to formally notify the issue of accomplishing effective and dependable key administration in secure deduplication. Firstly we introduce a pattern approach in which every client holds an autonomous expert key for scrambling the aim keys and outsourcing them to the cloud.

On the second way, such a standard key administration plan produces a tremendous number of keys with the expanding number of obliges clients and clients to dedicatedly secure the expert keys. To this end, we propose Dekey , another development in which clients don't have to deal with any keys all alone however rather safely circulate or transfer the united key shares over different servers. Security examination exhibits that Dekey is secure as far as the definitions determined in proposed security model.

D. Proofs of Ownership in Remote Storage Systems.

Distributed storage frameworks are turning out to be progressively prominent. A promising innovation that holds their expense down is de-duplication, which stores just a solitary duplicate of rehashing information. Customer side deduplication endeavours to recognize deduplication opportunities as of now at the customer and save the transmission capacity of transferring duplicates of the existing documents or files to the server. After that process we looks assaults that endeavour customer side de-duplication, permitting an aggressor to access self-assertive size records of different clients in view of a few hash marks of these documents.

All the more particularly, an aggressor who knows the hash mark of a record can persuade the capacity advantage that it possesses that document, henceforth the server lets the assailant download the whole record. (In parallel to our work, a subset of these assaults was as of late presented in the wild regarding the Dropbox record synchronization administration) To overcome of this problem, we present the thought of verifications of-possession (PoWs), which lets a customer effectively present to a server that that the customer holds a document, as opposed to simply some short data about it. We formalize the concept of evidence of-proprietorship, under thorough security definitions, and thorough productivity prerequisites of Petabyte scale stockpiling frameworks. We then present arrangements in view of particular encodings and Merkle trees, and investigate their security. We actualized one variation of the plan. Our execution estimations show that the plan causes just a few overhead contrasted with guileless customer side deduplication [4.]

E. Private Data Deduplication Protocols in Cloud Storage

Other proposed system call private information deduplication convention, a deduplication system for private information stockpiling is presented and formalized. Naturally, a private information deduplication convention allow a customer who holds a private information demonstrates to a server who have a synopsis string of the information that he/she is the proprietor of that information without uncovering additional data to the server. Our idea can be seen as a supplement of the cutting edge open information deduplication conventions of Halevi et al. The security of private information deduplication conventions is formalized in the recreation based system in the connection of two-gathering calculations.

A development of private deduplication conventions in view of the standard cryptographic suspicions is then introduced and examined. We demonstrate that the proposed private information deduplication convention is provably secure accepting that the basic hash capacity is crash flexible, the discrete logarithm is hard and the eradication coding calculation can deletion up to α-division of the bits in the vicinity of malignant enemies in the vicinity of vindictive foes.

To the best our insight this is the first deduplication convention for private information stockpiling [6].

3. Proposed System

In the proposed system we are doing duplication check in authenticated way. For the file duplication check proof of ownership is also set at the time of file upload the proof is added with the files this proof will decide the access privilege to the file. It is decide who can perform duplication check of the file. User is needed to submit his/her file and proof of ownership of the file before sending the request to for the duplicate check Request to the cloud. When there is file on the cloud and also privileges of the user only that time to approved the duplicate check request.

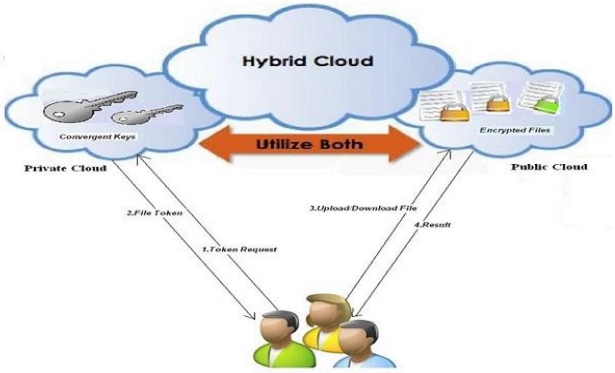


Figure 4: Overview of the system

Above fig.3 shows the proposed system architecture which comprises of public cloud, private cloud and user. Proposed system architecture contains only one public cloud and one is private cloud. All data of user is contains in public cloud

such as files. And private cloud consists of user credentials. User for every transaction with the public cloud need to take token from the private cloud. If the user's credentials stored at the public cloud and private cloud are get matched then user can have assess for the duplicate check. Following operations are need to be done in the authenticate duplicate check.

A. Encryption of File:

We are using secrete key resides at the private cloud to encrypt the user data and this key is used to convert plain text to cipher text and again for the decryption of the user data. To encrypt and decrypt we have used three basic functions as follows:

1) Key GenSE: It is generate the secrete file by using security parameter. In this k is the key generation algorithm.
2) EncSE (k, M): In this we have generated a cipher text using formulae M is the text message and k is the secrete key.
3) DecSE (k, C): In this we have to generate plain text using C is the cipher text and k is the encryption key.

B. Confidential Encryption of data:

This ensures the data confidentiality in the duplication. User derives a convergent key from each original data and encrypt the data copy with the generated convergent key. User also add the tag for the data so that the tag will helps to find the duplicate data. By using converged key generation algorithm to encrypt the user data. This will ensures the security, authority and ownership of the data.

Figure 5: Confidential data encryption

C. Proof of Data

When file upload and download user need to provide proof of the data. User need to submit his/her convergent key which was generated at the time of file upload. To generate the hash value of the data we have used MD5 massage digest version 5 algorithm to generate the hash value of the user data. If there is any small change in data occur the hash value of that data get changed.

4. Algorithm

Chunking per sentence:

Step 1: Upload a file
Step 2: Read the file into buffer reader.
Step 3: for all the sentences in the buffer reader
Step 4: Read each line till first full Stop is detected.
Step 5: Consider the data till first full stop as a chunk
Step 6: While all the sentences are read select data till every full stop as a chunk
Step 7: Check if the currently read chunks are available in the cloud data base.
Step 8: If yes the chunk is duplicate increase
a. The count for duplicate chunk
b. Else the chunk is original keep the chunk count as it is.

Block Level Algorithm (10 Sentences)

Step 1: Upload a file
Step 2: Read the file into buffer reader.
Step 3: for all the sentences in the buffer reader
Step 4: Read each line till 10th full Stop is detected.
Step 5: Consider the data till 10th full stop as a chunk
Step 6: While all the sentences are read select data till every 10th full stop as a chunk.
Step 7: Check if the currently read chunks are available in the cloud data base.
Step 8: If yes the chunk is duplicate increase
a. The count for duplicate chunk.
b. Else the chunk is original keep the chunk count as it is.

5. Mathematical Model

Set theory:

- S = {R,T,P,H,D}
- R = Registration of the user with specifying size on cloud.
- T = token generated and forwarded to user through mail for activation.
- P- User Privileges
- H- Hash function calculation.
- D = Matching contents of user uploaded data with existing database
- R = {r0,r1}
- Where ,
- r0- Provide information to the registration authority.
- r1- Registration authority validate the information.
- r2- user get cloud id and user id.
- r0 → t1
- T = {t1,t2}
- t1- token gives to user through mail.
- t2- get privilege to user.
- D = {d0,d1,d2,d3}
- Where,
- d0- get the data file name and key
- d1- Generate hash function and encrypt file.
- d2- heck matching content by entering upload button.
- t2 →d3
- d3- download/update/upload file by providing token/key or any other details.

Graphical Analysis:

File Size:

To evaluate the effect of file size to the time spent on different steps, I upload 100 unique files of particular file size and record the time break down. Using the unique files

enables us to evaluate the worst-case scenario where I have to upload all file data. The average time of the steps from test sets of different file size are plotted in Figure 5. The time spent on downloading, encryption, upload increases linearly with the file size, since these operations involve the actual file data and incur file I/O with the whole file. In contrast, other steps such as duplicate check and token generation only use the file metadata for computation and therefore the time spent remains constant. With the file size increasing from 10MB to 120MB.

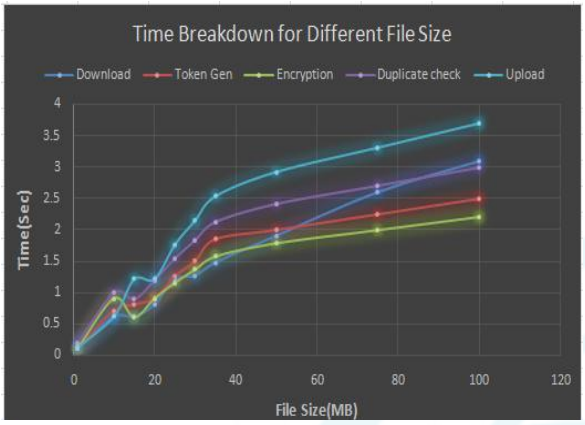


Figure 4.1: Time Breakdown for different Fie size

Number of Stored Files:
To evaluate the effect of number of stored files in the system, I upload different number of unique size files and record the breakdown for every file upload. From Figure 5, every step remains constant along the time. Token checking is done with a hash table and a linear search would be carried out in case of collision.

6. Results

The proposed system should re-address user from uploading duplicate data on cloud. Data stored on cloud must be in secure encrypted format. Malicious user not able to upload or download data on cloud. The user who has proof of ownership only that user can modify data.

User Login

Figure 5.4: User Login

User Activation

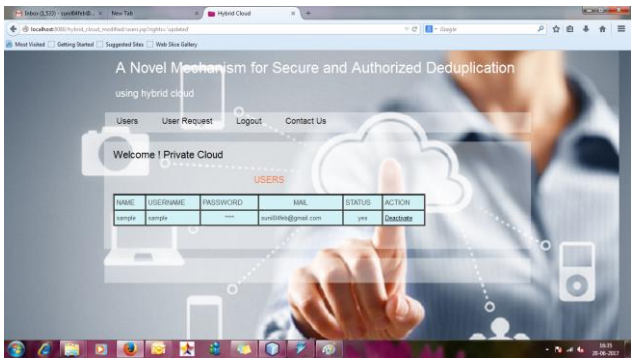
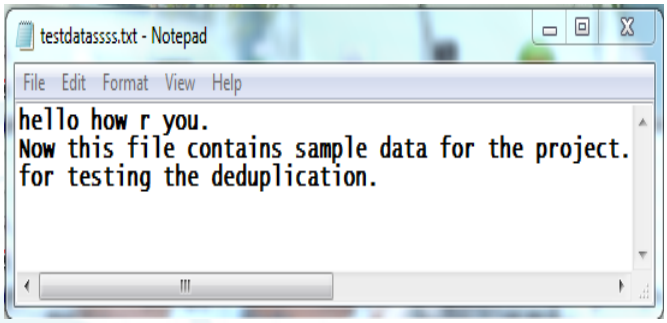
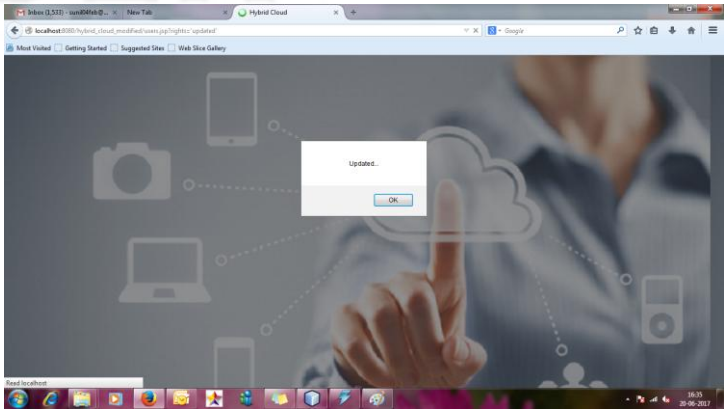


Figure 5.5: User Activation



File to be uploaded for deduplication check.

File Uploading and Updating



Percentage of duplicate data found

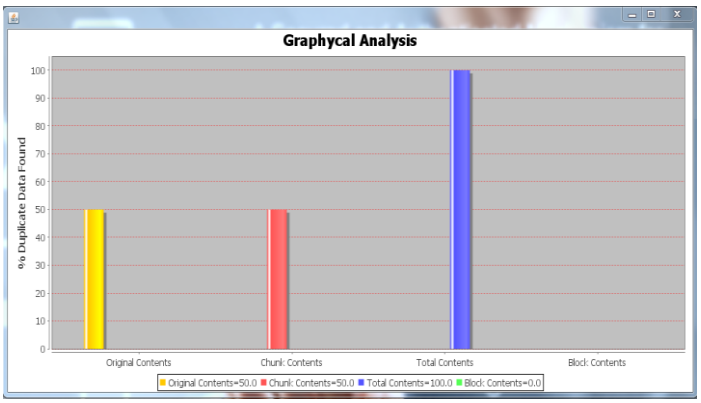


Figure 5.7: Percentage analysis of duplicate data found

7. Conclusion

The allover system provided reason that our proposed framework information DE duplication of record is done approves way and safely. In this we have additionally proposed new duplication check system which produce the token for the private document. We have settled more basic piece of the cloud information stockpiling which is just endured by diverse systems. Proposed routines guarantee the information duplication safely.

References

[1] M. Bellare, S. Keelveedhi, and T. Ristenpart. Dupless: Serveraided encryption for deduplicated storage. In USENIX Security Symposium, 2013.
[2] P. Anderson and L. Zhang. Fast and secure laptop backups with encrypted de-duplication. In Proc. of USENIX LISA, 2010.
[3] J. Li, X. Chen, M. Li, J. Li, P. Lee, andW. Lou. Secure deduplication with efficient and reliable convergent key management. In IEEE Transactions on Parallel and Distributed Systems, 2013.
[4] 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.
[5] C. Ng and P. Lee. Revdedup: A reverse deduplication storage system optimized for reads to latest backups. In Proc. of APSYS, Apr 2013.
[6] 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.