Efficient method of Secure Authorized Data Deduplication at Small Block Level Approach in Cloud Storage

Komal Kasana*[1], Dr. S.K Yadav*[2]

1. M.Tech(CSE), Research Scholar
2. Associate Professor

Dept. Of Computer Science and IT, SHUATS, Allahabad, India

1.komalkasana.cs0022@gmail.com

2.sanajay.yadav@shuats.edu.in

Abstract—**In this paper, we proposed a new serverside deduplication scheme for mixed data. It is applying data deduplication on small block level and if data block are unique data in that case first it will get encrypted then uploaded to cloud storage. Moreover, the proposed system guarantees data uprightness against any name anomaly attack. In this way, security is enhanced in the proposed system. For deduplication hash code comparison in proposed system introduced a new technique called authentication of multi-level block signature. This technique provides a mechanism which is used to reduce comparison time of hash code in data base. The adequacy examination comes to fruition show that the proposed contrive is for all intents and purposes as capable as the existing system, while the additional computational overhead is unimportant.**

Keyword: ***Deduplication, Ownership, Cloud Storage, multi-level block signature, hash code, authentication***

1. INTRODUCTION

Cloud storage is one of the digital storage solutions that utilizes multiple servers (typically spread across multiple locations) which ensure to safely store files such as site backups etc. The data is stored on dedicated servers, which provides unlimited accessibility wherever an internet connection is available, along with an increase in backup file security which ensures files not to be hacked. To reduce resource consumption, many cloud storage services, such as Dropbox, Wuala, Mozy, and Google Drive, employ a deduplication technique, where the cloud server stores only a single copy of redundant data and provides links to the copy instead of storing other actual copies of that data, regardless of how many clients ask to store the data [1][3][4][5][6]. The savings are significant, and reportedly, business applications can achieve disk and bandwidth savings of more than 90%.

Deduplication is a technique that ensures only one unique instance of data is retained on storage media, such as disk, flash or tape. Excess information squares are supplanted with a pointer to the one of kind information duplicate. The system is utilized to enhance stockpiling use and can likewise be connected to arrange information exchanges to diminish the quantity of bytes that must be sent. In the deduplication procedure, novel lumps of information, or byte designs, are distinguished and put away amid a procedure of investigation. Information deduplication is a particular information pressure system for disposing of copy duplicates of re-hashing information. Related and to some degree synonymous terms are savvy (information) pressure and single-example (information) stockpiling.

However, from a security perspective, the shared usage of users' data raises a new challenge. In cloud storage services, deduplication technology is commonly used to reduce the space and bandwidth requirements of services by eliminating redundant data and storing only a single copy of them. Deduplication is most effective when multiple users outsource the same data to the cloud storage, but it raises issues relating to security and ownership. Proof-of-ownership[5] schemes allow any owner of the same data to prove to the cloud storage server that he owns the data in a robust way. However, many users are likely to encrypt their data before outsourcing them to the cloud storage to preserve privacy, but this hampers deduplication because of the randomization property of encryption [1].

Cloud computing provides scalable, low-cost, and location-independent online services ranging from simple backup services to cloud storage infrastructures [2]. The fast growth of data volumes stored in the cloud storage has led to an increased demand for techniques for saving disk space and network bandwidth.

Some of its benefits are as under:-

- Generate data tags before uploading as well as audit the integrity of data having been stored in cloud.
- It Decreases size of occupation of data in the cloud which is one of the critical challenge of cloud storage services is the management of the ever increasing volume of data
- It provides the security of the stored data by AES encryption algorithm.

- It maintains the privacy of the user by deduplicating the encrypted data.
- A user can check the stored data's integrity in the proposed scheme by auditing the encrypted data.

2. LITERATURE SURVEY

On the basis of extensive literature survey related to the data deduplication with dynamic ownership management in cloud storage has been taken into consideration in this section.

M. Bellare et al.(2013). Studied the problem of providing secure outsourced storage that both supports deduplication and resists brute-force attacks. They designed a system, DupLESS, that combines a CE-type base MLE scheme with the ability to obtain message-derived keys with the help of a key server (KS) shared amongst a group of clients. The clients interact with the KS by a protocol for oblivious PRFs, ensuring that the KS can cryptographically mix in secret material to the per-message keys while learning nothing about files stored by clients. The mechanisms ensure that DupLESS provides strong security against external attacks which compromise the SS and communication channels (nothing is leaked beyond file lengths, equality, and access patterns), and that the security of DupLESS gracefully degrades in the face of comprised systems. The low performance overhead results in part from optimizing the client-to-KS OPRF protocol, and also from ensuring DupLESS uses a low number of interactions with the SS. they showed that DupLESS is easy to deploy: it can work transparently on top of any SS implementing a simple storage interface, as shown by the prototype for Dropbox and Google Drive.

N. Baracaldo et al. (2014). Represented that Cloud reckoning has developed as exceptionally useful for organizations that hopes to decrease their expenses, send new applications quickly or that do not have any need to stay up their own process framework. In any case, late info ruptures in clear distributed storage suppliers have created customers be increasingly troubled concerning the classification of their (outsourced) info.

There are things wherever client info was conferred to and spilled by cloud provider representatives that had physical access to the capability medium, and moreover wherever cloud clients accessed alternative customer's info within the wake of getting been distributed physical warehousing quality already allotted to a different customer e.g., then alternative client had worn out its distributed storage membership.

J. Li et al.(2015). Proposed new deduplication system, a hybrid cloud architecture to solve the problem. The private keys for privileges will not be issued to users directly, which will be kept and managed by the private cloud server instead. In this way, the users cannot share these private keys of privileges in this proposed construction, which means that it can prevent the privilege key sharing among users in the above straightforward construction. To get a file token, the user needs to send a request to the private cloud server. The intuition of this construction can be described as follows.

- To perform the duplicate check for some file, the user needs to get the file token from the private cloud server. The private cloud server will also check the user's identity before issuing the corresponding file token to the user. The authorized duplicate check for this file can be performed by the user with the public cloud before uploading this file. Based on the results of duplicate check, the user either uploads this file or runs PoW.

M.Bellare et al.(2015). Provided a deduplication scheme secure for messages that are both correlated and parameter dependent. Additionally, our scheme is standard-model, not ROM. The key new ingredient is interaction. In the proposed scheme, upload and download are interactive protocols between the client and server. To specify and analyse these protocols, we define a new primitive, interactive MLE or iMLE. This system provides a syntax and definitions of security, then specify and prove correct the proposed protocols. iMLE turns out to be interesting in its own right and yields some other benefits. It provides the first secure deduplication scheme that permits incremental updates. This means that if a client's message changes only a little, for example due to an edit to a file, then, rather than create and upload an entirely new cipher text, she can update the existing one with communication cost proportional only to the distance between the new and old plaintexts. This is beneficial because communication is a significant fraction of the operating expenditure in outsourced storage services. For example, transferring one gigabyte to the server costs as much storing one gigabyte for a month or longer in popular storage services.

Junbeom Hur et al.(2016) has proposed a secure deduplication scheme for encrypted data that has dynamic ownership management capability. Its construction is based partially on a randomized convergent encryption scheme in order to randomize the encrypted data, which renders the proposed scheme secure against the chosen-plaintext attack while still allowing deduplication over the data. The proposed scheme is further integrated into the re-encryption protocol for owner revocation.

3. Problem Statements

Now days everybody using cloud services, storing data, sharing data with others and people knows that cloud is a third party resource so many concern are there like data security, privacy, data ownership, space for storage, bandwidth, data deduplication etc. Major issues are data security, data deduplication, dynamic data ownership which is not as per our expectation, we need more enhancement in these concern. As we analyse many existing work in the field of data security, dynamic data ownership. Many existing plan using deduplication but they are not addressing deduplication at small block level.

4. Contributions

We propose a data deduplication scheme over dynamically distributed data in cloud storage. The proposed system ensures that individual secure access to the common data is possible in cloud storage, which is believed to be the most basic test for successful and secure appropriated big data storage organizations in nature where ownership changes effectively. In existing system deduplication checked on file level and no security for data but the proposed system introduced the small block level deduplication, enhance data security and provide dynamic data ownership in cloud storage.

5. SYSTEM ANALYSIS

5.1. Existing System

When a user uploads data that already exist in the cloud storage, the user should be deterred from accessing the data that were stored before he obtained the ownership by uploading it (backward secrecy).Existing system are handling data deduplication at file level. The dynamic ownership changes may occur very frequently in a practical cloud system, and thus, it should be properly managed in order to avoid the security degradation of the cloud service. Most of the existing schemes have been proposed in order to perform a POW process in an efficient and robust manner, since the hash of the file, which is treated as a "proof" for the entire file, is vulnerable to being leaked to outside adversaries because of its relatively small size. A data owner uploads data that do not already exist in the cloud storage, he is called an initial uploaded; if the data already exist, called a subsequent uploaded since this implies that other owners may have uploaded the same data previously, he is called a subsequent uploader as shown in fig 5.1[1].

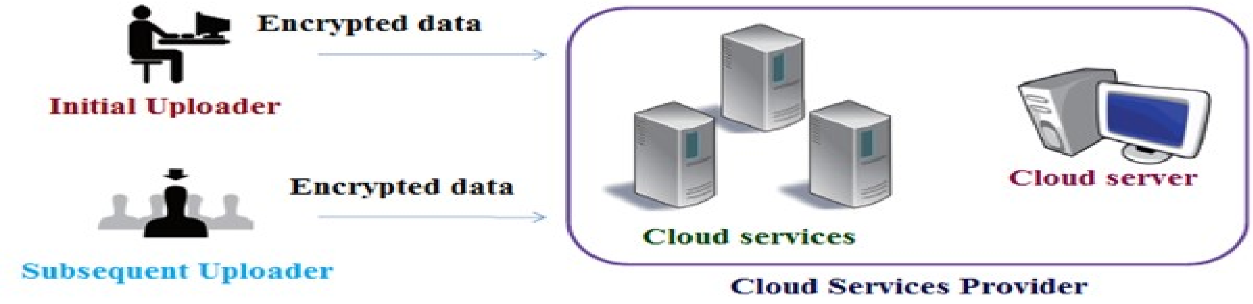

Fig. 5.1 Architecture of Existing System

5.2. Proposed System

In Proposed system we propose a new server-side deduplication plan for mixed data. It empowers the cloud server to control access to outsourced data despite when the ownership changes intensely by manhandling randomized joined encryption and secure ownership pack key scattering, a deduplication service over encoded data.

5.3. Advantages of the Proposed System

- Generate data tags before uploading as well as audit the integrity of data having been stored in cloud.
- Enables secure deduplication through introducing a Proof of Ownership protocol and preventing the leakage of side channel information in data deduplication.
- Integrity auditing and secure deduplication directly on encrypted data.

6. FLOW CHART OF DATA DEDUPLICATION

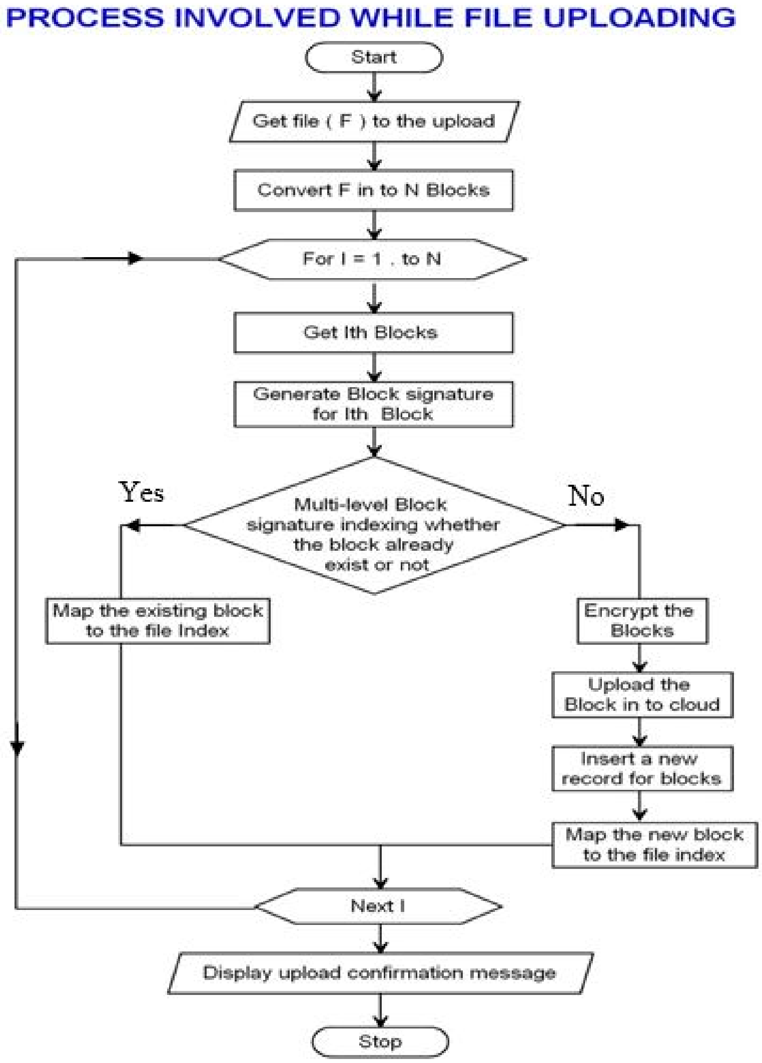
PROCESS INVOLVED WHILE FILE UPLOADING
Start
Get file (F) to the upload
Convert F in to N Blocks
For I = 1 . to N
Get Ith Blocks
Generate Block signature for Ith Block
Multi-level Block signature indexing whether the block already exist or not
Yes
No
Map the existing block to the file Index
Encrypt the Blocks
Upload the Block in to cloud
Insert a new record for blocks
Map the new block to the file index
Next I
Display upload confirmation message
Stop

Fig.6.1. Flow Chart for Upload Process

While uploading the file, in first step the file is broken in small blocks based on given block size after that hash code get generated for all blocks, while generating hash code it will check whether it is new block of data or duplicate block of data based on hash code if hash code matched with existing hash code means it is duplicate block of data and if it is not matching means it is new data, all new block of data we will encrypt using AES encryption then we will upload to the cloud drive.

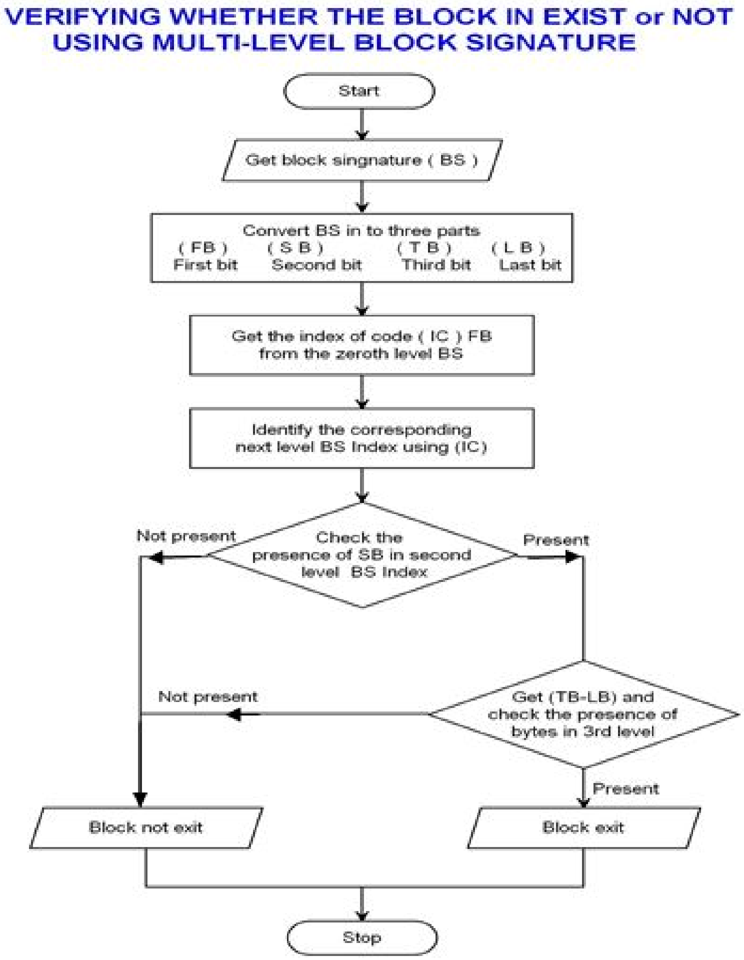
VERIFYING WHETHER THE BLOCK IN EXIST or NOT
USING MULTI-LEVEL BLOCK SIGNATURE
Start
Get block singnature (BS)
Convert BS in to three parts
(FB) First bit
(S B) Second bit
(T B) Third bit
(L B) Last bit
Get the index of code (IC) FB from the zeroth level BS
Identify the corresponding next level BS Index using (IC)
Check the presence of SB in second level BS Index
Not present
Present
Get (TB-LB) and check the presence of bytes in 3rd level
Not present
Present
Block not exit
Block exit
Stop

Fig.6.2. Flow Chart for Multi-Level Block Signature

Proposed system is providing security to the data using AES encryption as mention in uploading file flow chart figure 6.1. For deduplication detection in small block level it is using concept of Multi-level block signature which improving performance of our proposed system shown in figure 6.2.

While hash code comparison in proposed system introduced a new technique called authentication of multi-level block signature. This technique provides a mechanism which is used to reduce comparison time of hash code in data base. Here it breaks hash code in three part, first part having two digit, second part two digit and last part having remaining digit, while comparison first it will check two digit if it match then it will check another two bit if it also match then it will compare remaining bit.

7. RESULT

When a new file is stored on existing system then it stores the file as new in cloud storage. If the same file is uploaded then the server implements deduplication scheme in which it stores the single instance of the file. But when the same file is modified then it lacks in deduplicating the data because a minor change within the file will change its hash tag and the modified file will be accepted as a new file on server in existing system.

But proposed system fixes this issue by using block level deduplication approach. When a new file is uploaded, the file is fragmented among different blocks of equal sizes. The hash code of each block is generated. Each hash tag is compared with the stored hash tag in order to check whether it is new data or not. If it is new then it will be encrypted and then finally uploaded to the cloud storage. But if server finds the data already in the cloud then it will save only single instance of the file on the cloud storage. File is stored in blocks in cloud storage. If a minor modification is done within the file then a new block with its hash code get generated. This new encrypted block is stored in cloud. While downloading blocks from cloud drive it will decrypt block content and then re downloading all the blocks, it will merge all the blocks, to make a single file.as shown in table 7.1 and fig 7.2, in the long run the block level deduplication technique is more efficient in comparison with the existing system where the deduplication is done on file level.

Table 7.1. Comparison of Space Optimization in Cloud

File Name	File Size (kB)	Existing System Space occupied (kB)	Proposed System Space occupied (kB)
File-1.txt	2780/1024=2.71	3	4200/1024=4.10
File-2.txt	2808/1024=2.73	3	2100/1024=2.05
File-3.txt	2803/1024=2.73	3	700/1024=0.68
File-4.txt	3355/1024=3.27	4	1400/1024=1.36
File-5.txt	3356/1024=3.27	4	700/1024=0.68
Total size(kB)	14.71	21	8.87

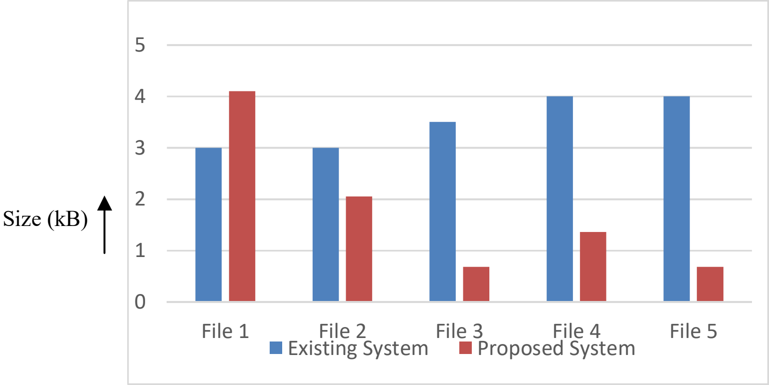

Fig. 7.2 Comparison of Space Optimization

8. Conclusion

The proposed system provides a secure and authorized block level data deduplication scheme over dynamically distributed data in cloud storage and ensure individual secure access to the common data which is possible in cloud storage, which is believed to be the challenge for successful and secure appropriated big data storage organizations in nature where ownership changes effectively. It provides secured and authorized deduplication scheme at small block level which helps in saving space in cloud

Future Enhancement

Further work should be done to reduce comparison time while deduplication of data, because when data block will increase it will take more time to compare with new block hash tag.For data security server should use Advanced Encryption Techniques. Like AES and DES

REFERENCES

.

[1] **Junbeom Hur, Dongyoung Koo, Youngjoo Shin, and Kyungtae Kang.(2016). "***Secure Data Deduplication with Dynamic Ownership Management in Cloud Storage"* IEEE Transactions on Knowledge and Data Engineering.

[2] **L. Mingqiang, C. Qin, P.P.C. Lee. (2015).** *"CDStore: Toward Reliable, Secure, and Cost-Efficient Cloud Storage via Convergent Dispersal Dispersal,"* Proc. USENIX Annual Technical Conference, pp. 120.

[3] **M. Bellare, S. Keelveedhi, T. Ristenpart. (2013).** *"DupLESS: Serveraided encryption for deduplicated storage,"* Proc. USENIX Security Symposium.

[4] **M. Bellare, S. Keelveedhi. (2015).** *"Interactive message-locked encryption and secure deduplication,"* Proc. PKC 2015, pp. 516–538

[5] **N. Baracaldo, E. Androulaki, J. Glider, A. Sorniotti. (2014).** *"Reconciling end-to-end confidentiality and data reduction in cloud storage,"* Proc. ACM Workshop on Cloud Computing Security, pp. 21–32.

[6] **Patil Shweta , Pawar Mohini , Nikam Pratima , Jondhale Sonali , Prof V.K.Wani. (2016).** *" Data Deduplication In Hybrid Cloud With Secured Authorization,"* e-ISSN No.:2349-9745.