International Research Journal of Advanced Engineering and Science
ISSN (Online): 2455-9024

A Study of Deduplication in Cloud Storage

Guljar P. Shaikh
Computer Engineering, Pune University, Pune, Maharashtra, India-411046

Abstract*— *Now days, in this fast and digital world, Storing of data is important for singles and also for the large organization. As the amount of data generated by individuals or by organization is huge and storing of that data is major concern. Sometimes duplicate data are also being stored which is not memory sufficient or wastage of space. Thus, Storage optimization techniques are needed to store large amount of data over large storage areas like cloud. Data Deduplication is one the optimization technique that reduces duplicate data to be stored. Data being stored on large storage areas or in cloud is in encrypted format for security concern.

Keywords*— *Deduplication, cloud storage, convergent encryption, optimization techniques.

I. INTRODUCTION

Now days with the growing of population and use of technology, it leading to many problems for storing data [1]. The technology growth is increasing the amount of storage or communication and technique devices. The Cloud offers apparently infinite resources to store data. Due to availability of cloud storage, large amount of data get saved on cloud & get retrieved by users by using privileges and keys whenever desired. One of the major intimidations of cloud is to preserve the escalating number of data. To make data management more scalable in cloud, deduplication is well-known and effective system. Data Deduplication is enthusiastic compression method. In this it reduces copies of duplicate data. Deduplication method is also used for less storage space utilization and this can applied on network for transferring data to cut down the volume that is to be transmitted. Its better to delete same copies rather than maintaining it and wasting space.

And Deduplication does the same thing i.e. eliminating copied data and keeping only unique data. By Using a Cloud, we can achieve better results of Deduplication in terms of security and storage. Cloud offer us many advantages like scalability, consistency, cost savings & deployment with the improved control, superior. Convergent Encryption gives confidentiality on data. Data Encryption and data decryption accomplished by the key derived from fil. Key is generated with hash function. After the key generated and data is encrypted, owners protect those keys then send cipher text to cloud. To keep data safe and to maintain distance from attackers or unauthorized people, the term is introduces named proof of ownership i.e. POW. This protocol is essential for confirmation that the user certainly owns the identical file when a replica is found. After confirmation, users having similar data, will be given "pointer". this pointer is generated for not to upload same copy of data over cloud. With the help of that pointer a user can download the file which is encrypted from server. And that file is decrypted by respective keys only. Convergent Encryption will be acting as process of authorization of cloud to execute Deduplication on the texts and PoW check for not permitted user to have access to file.

II. BACKGROUND

1. Deduplication:

Deduplication is essentially a compression technique for deleting redundant data. Deduplication is process done before the storing data onto memory or space. [1] Deduplication can be classified as file level deduplication and block level deduplication based on granularity [4].

a) *File level Deduplication*: This deduplcation takes into the entire file, thus even small update or append makes the file different from previous version of it and thereby reducing deduplication ratio.[5]
b) *Block level Deduplication*: in this deduplication the data chunks are considered for deduplication [5].

Deduplication can further classified based on location of deduplication i.e., as client side deduplication and as source side deduplication.

Client side deduplication ensures that bandwidth is saved as it only send hash value of checked file to the server in case of duplication exists [11].

Data Deduplication is majorly used in many applications like metadata management, backup, primary or secondary storage, etc. for storage optimization.

2. Convergent Encryption:

Convergent encryption is an encryption way which supports the data Deduplication [1]. By convergent encryption, the key named encryption key is created in combination of hash values. So applying this technique to identical plaintexts would create same cipher text, and this helps in performing Deduplication further [11].

3. Proof of Ownership:

Data Deduplication is works by processing the cryptographic hash function on to data and by using this hash value we can determine the alike data [1], [2]. When duplicate copy is found then new data is not uploaded over storage but the pointer to that file ownership is generated and updated so wastage of space and bandwidth is not done.

At the time of client side Deduplication, the hash values of data or files are computed at client side first and then send to server side for duplicate check. [3] An offender, who gains access to the hash value of the data which is not authorized to them, may claim Deduplication of file and thereby gaining access to the file. To prevent such an attack, a Proof of Ownership (PoW) has been proposed. This PoW works as an interactional algorithm between two parties - a prover and verifier to prove the ownership of the file [7].

Verifier: this computes a short value of data

Guljar P. Shaikh, "A study of deduplication in cloud storage," *International Research Journal of Advanced Engineering and Science*, Volume 3, Issue 1, pp. 105-106, 2018.

Prover: this need to compute short value of M and send it to verifier for claiming ownership.

Data Deduplication Flow Chart:

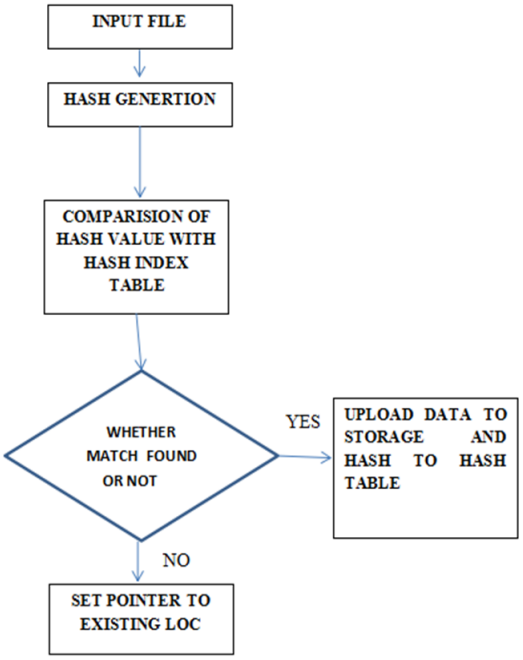

Fig. 1. Flow of data De-Duplication.

Practical Applications

Data deduplication aid to accomplish aims of optimizing data and scaling of storage capacity. It also offers man of applicable ways to the cloud service providers to attain these goals.
This includes following some of ways:

Capacity Optimization:

Data Deduplication cut down the storage space used for storing that data. This also achieves greater storage ratio. This method uses compression and variable-size blocking, which leads to achieve great optimization ratios [6].

Performance and Scalability:

Data Deduplication can process multiple volumes of files and data at the same time without affecting other workloads on the server.

Data Integrity and Reliability:

The integrity of data is achieved by the process of Data Deduplication is maintained [3]. To ensure this integrity this method uses consistency, check-sum and validation on data. The Deduplication also ensure about the data recovery in case of data corruption [6].

Bandwidth Efficiency:

The data encryption is performed on client side only and based on hash value comparison sending only unique data to server saves more bandwidth. This results into faster downloading of files at times and minimizes bandwidth utilization.

III. Conclusion

Data Deduplication is one of the well-known available technique in cloud storage which is used for saving of bandwidth and utilizing more storage Space. But, in some cases this method i.e. the data deduplication is not that feasible with data which is in format of encrypted since, different key encryption convert same data into different formats.

In this paper we have mentioned some of methods of Deduplication where Deduplication methods are carried out on encrypted data in cloud storage or any other large storage area. Most of simple method of this works on basis of the convergent encryption. This is simple and also more compatible with encrypted data as well.

For more security concern a strategy needs to be developed which will intensify storage optimization without go across on encryption method; by giving Deduplication proficiency in data storage servers where the available data is encrypted.

References

[1] G. P. Shaikh, S. D. Chaudhary, P. Paygude, and D. Bhattacharyya, "Achieving secure deduplication by using private cloud and public cloud," *International Journal of Security and Its Applications*, vol. 10, no. 5, pp. 17-26, 2016.
[2] G. P. Shaikh, "De-duplication with authorization in hybrid cloud approach for security," *International Journal of Computer Sciences and Engineering*, vol. 4, special issue 4, pp. 1-4, 2016.
[3] G. P. Shaikh, Prof. S. D. Chaudhary, and Prof. P. S. Paygude, "Achieving data confidentiality by usage of hybrid cloud and deduplication," *International Journal of Computer Science and Mobile Computing*, vol. 5, issue 7, pp. 245–252, 2016.
[4] G. Shaikh, "A survey on deduplication strategies and storage systems," *IOSR Journal of Computer Engineering (IOSR-JCE)*, pp. 85-90, 2015.
[5] A. Ka, A. Ganesha, and Sunitha C, "A study on deduplication techniques over encrypted data," *Procedia Computer Science*, vol. 87, pp. 38–43, 2016.
[6] D. T. Meyer and W. J. Bolosky, "A study of practical deduplication," *ACM Transactions on Storage (TOS)*, vol. 7, issue 4, pp. 14, 2012.
[7] Halevi, D. Harnik, B. Pinkas, and A. Shulman-Peleg, "Proofs of ownership in remote storage systems," *Proc. ACM Conf. Comput. Commun. Security*, pp. 491–500, 2011.
[8] J. R. Douceur, A. Adya, W. J. Bolosky, D. Simon, and M. Theimer. "Reclaiming space from duplicate files in a serverless distributed file system," *IEEE Proceedings 22nd International Conference on Distributed Computing Systems*, pp. 617-624, 2002.
[9] P. Anderson and L. Zhang, "Fast and secure laptop backups with encrypted de-duplication," *Proc. 24th Int. Conf. Large Installation Syst. Admin*., pp. 29-40, 2010.
[10] B. Choudhary and A. Dravid "A study on authorized deduplication techniques in cloud computing," *International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)*, vol. 3, issue 12, pp. 4191-4194, 2014.
[11] Yingdan Shang and Huiba Li, "Data deduplication in cloud computing systems," *International Workshop on Cloud Computing and Information Security (CCIS)*, pp. 483-486, 2013.

Guljar P. Shaikh, "A study of deduplication in cloud storage," *International Research Journal of Advanced Engineering and Science*, Volume 3, Issue 1, pp. 105-106, 2018.