

Secure Updatable Data Deduplication in Cloud

Dr.B.Tirapathi Reddy¹, K.Lasya Priya², G.Ramya Sri³, P.Nikitha⁴ and A.Laxmi Samvidha⁵

¹⁻⁵ Department of CSE, KLEF, Guntur, India

Email: tirapathireddyb@kluniversity.in, {lasyapriya2000, ramyagollapudi1701, nikipunnani123, samvidha.ankem05}@gmail.com

Abstract—With the constant usage of Storage techniques in daily lives, it has been important to establish a system which properly manages the storage space. Almost all the storage space of cloud is occupied by the duplicate data which have been gathered or downloaded from various resources and saved in the system. So it increases storage space. Backing up of data from the cloud will increase redundant data which is uploaded into the cloud earlier. Due to the increase in the amount of redundant data, proper utilization of the storage resources and bandwidth is not feasible. Once a user uploads the data into the cloud there is no guarantee for security of the data. In this paper deduplication is applicable to lessen the redundant data and updating of users data without any security issues and data loss using secure encrypted algorithms.

Index Terms— Data Deduplication, Cloud computing, Client side, Storage, Bandwidth and SHA-256.

I. INTRODUCTION

Storage of cloud is a model of data storage where data will be stored in pools. The physical storage will have multiple servers in distinct locations. Ref. [2] Storing of our data in cloud will be done through servers. When we use cloud for storage purposes, we intend to store the same file more than once even may be with different names. So, almost all the cloud is occupied with repetitive data. Thus, the redundant data is occupying more disk space. There is a chance where different users upload the same file more than once. So there will be a huge wastage in disk space. To overcome this problem, there is only one Solution which is Data Deduplication or dedupe, in short, Data Deduplication is a specialized technique to eliminate duplicate copies of data. Ref. [1] This Deduplication frees up a lot of storage when applied on large volume of data.

Data Deduplication can be done in different ways based on different conditions:

- Based on Location
- Based on Granularity
- Based on Ownership

A. Based on Location

In this, Data Deduplication can occur at two levels : deduplication can be done either at source side/client side and target level/ Server side. Source level de-duplication eliminates redundant blocks before transmitting data to the server or backup storage. Ref. [3] In this, there is no need of additional hardware Source level Deduplication reduces network bandwidth and upload costs.

In Target based deduplication, deduplication is done after transmitting the file to the server or the cloud. Ref. [3] This type of deduplication increases costs even though it has an advantage of performance when compared to source deduplication.

B. Based on Granularity

In this, Data Deduplication can be done mainly in two ways: File level Deduplication and Block Level Deduplication. In the file level Deduplication, the whole file is being checked with the existing files using a hash value. Ref. [3] If the file is unique, it is stored else the same file is used. In case of Block Level Data Deduplication, the whole file is broken into chunks and each chunk is checked for deduplication. Each chunk of data is processed using a hash algorithm and hash values are stored. If the file is updated only that chunk or block is updated, rest all blocks are not updated. This behavior makes block level deduplication more effective. In the Chunking process, we can divide the blocks using fixed size or variable size. Ref. [8] Dynamic Sized Chunking is more effective than Fixed Sized Chunking as the file is dynamically broken into variable sized blocks.

C. Based on Ownership

In this type of Data Deduplication, there are two kinds: One user and Cross user.

In One user, the Data or the files are deduplicated for the same user. There may be chance where a user uploads a same content with different filenames. So, it can also be eliminated in this type of deduplication. But in case of Cross User, the files are checked for deduplication among different users. Suppose two or more users can upload a same file to the cloud which can be eliminated for improving storage utilization and reducing the bandwidth. Ref. [6] In this case, the user gets the file as uploaded but internally only one file will be there in which the user's id is tagged in the owners list of that file.

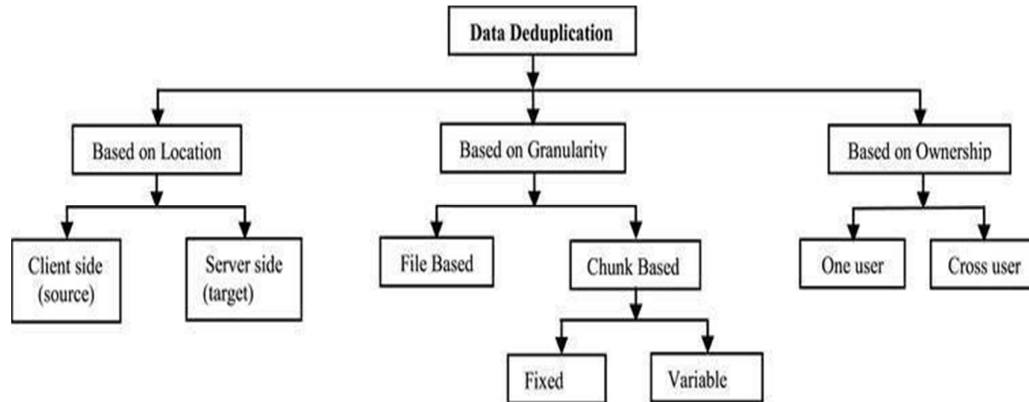


Figure 1. Types of Data Deduplication

Ref. [4] Since everything got improved and reduction of on premises servers has been seen, this technique will be used for every cloud provider in order to keep the cloud storage free from redundant files. Since a basic idea about the topic is done. The following deals with advantages of proposed system and implementation.

II. LITERATURE REVIEW

Little Research has been conducted on the Deduplication techniques. This survey helps to understand about data deduplication a, different types of data deduplication and efficient solution for it. Ref. [5] Hyungjune proposed a privacy preserving and updatable data deduplication technique. Ref. [10] Pasquale Puzio proposed a secured deduplication technique with encrypted data for cloud storage. In that, he proposed a secured architecture in which additional layers of encryption are added on top of convergent algorithm. Ref. [9] Mihir Bellare proposes DupLess Server aided encryption for deduplicated storage for cloud storage service provider. Ref. [9] Bellare et al. proposed formal definitions of message-locked encryption. Ref. [18] B.Tirapathi Reddy at proposed a system which mainly concentrates on proof of ownership of the file. In this paper, a cuckoo filter is used which consists of hash functions to calculate the file hash value. If a duplicate copy is being uploaded, the server challenges the user to prove the ownership by asking the token value of that file.

III. PROPOSED SYSTEM

In this paper, we performed Data Deduplication on the client side. The goal of the system is to guarantee data confidentiality while applying data deduplication. To overcome the weakness of convergent encryption

mechanism, a server component is included in the architecture. It performs additional encryption to protect all the files uploaded by the user. For every user, a token is generated upon registering and the token is sent to the user. Only after Activation of the user, the user can be able to upload or download the file. In this system, we even added the updating option where a user can update a file in the cloud and updated file is stored in the cloud. This system mainly focuses on the client-side deduplication and the updatable Data Deduplication mechanism. Before saving the file, the content is being checked by the system whether a copy is present in the cloud or not. If the same file or the same content with different file name exists, the file is not uploaded to the cloud. If it is a new file, the file is uploaded to the cloud. If same file is uploaded by more than one user into the cloud, then user will get an alert message called uploaded but duplicate copy will not be uploaded into the cloud. Instead, the owners of the existing file will be updated.

IV. SYSTEM MODEL

The proposed System contains the following components:

A. User

User is any entity that wants to upload file to the external storage area. User can upload, download and update a file in the cloud.

B. Private Cloud

Here Private Cloud authenticates the user after registering. It provides an additional security to the user. It provides a unique token to every user.

C. Cloud Service Provider

CSP here refers to Admin of the Cloud. Admins have every detail what a user have done. Ref. [10] Admin cannot view the files, but the admin has the access of every detail the user has done in terms of uploading, downloading and updating.

V. PROTOCOL

This section describes the process of uploading, downloading and updating.

First, a user U needs to register in the system. Then, a request is sent to private cloud to accept or reject the cloud access to that user U. After the Acceptance of Private Cloud, the Token is sent to the User for accessing cloud.

A. Upload

When a User wants to upload a file F, user clicks on the file to be uploaded and clicks on upload Button. Then, the content in that file is checked with the existing contents by calculating a unique hash value for the file using Hashing algorithm called SHA-256. Then the encrypted file will be stored into the cloud along with the hash value of that file.

B. Download

When a User U wants to download a particular file, user proceeds by clicking on the Download option of that respective file, that corresponding file is downloaded after the decryption process.

C. Update

When a User U wants to update a particular File, user clicks on the respective file to update. Then the content is displayed on the screen so that user can update accordingly. Later, by clicking on update option, the file update its content, updating time, size etc.. When a file is updated, the respective file hash value is updated accordingly with the content.

D. Encryption Algorithm

In this mechanism, to encrypt and decrypt the content using a hash value, we have used AES Algorithm. AES (Advanced Encryption Standard) Algorithm is a standard Cipher Algorithm in which it takes 128 bits plain text and converts it into cipher text using keys of 128,192,256 bits. It is considered very secure. AES is also known as Rijndael established by National Institute of Standards (NIST), U.S. AES is widely used in both software and hardware as it has built-in key length flexibility and key management is also easy in this algorithm. This algorithm has its own structure for encryption and decryption on sensitive data and is applied in hardware and

software all over the world. It will be immensely difficult to hackers to have an idea about the actual data while encrypting through AES algorithm.

E. SHA Algorithm

SHA is a hash function which generates a hash value. In cryptography for a given data, SHA (Secure Hash Algorithm) is a hash function which generates a unique value. Nobody can predict the original content from the hash value. No two texts will have the same hash value. Even updating a small character, the SHA-algorithm will produce a new hash value.

VI. ADVANTAGES

A. Cost Effective

As less amount of storage is required when deduplication is done. Cost is reduced significantly. As less amount of data is transferred, Disaster Recovery is improved.

B. Effective use of network bandwidth:

Ref. [14] If the data deduplication is done on the client side, there is no need to transfer redundant data to the cloud.

C. Reduced Storage Allocation

Deduplication can reduce storage needs by up to 80% for files. This helps in storing data to disk cost efficiently, taking advantage of its speed.

D. Efficient Volume Replication

Since, only unique data is written to cloud, only those blocks need to be replicated. This can reduce traffic for replicating data by 90% depending on the application.

E. Faster recovery and continuity

By eliminating redundant data from the equation, deduplication enables backups to be recovered much faster. In doing so, it minimizes downtime and helps keep business continuity plans on track.

VII. DISADVANTAGES

There will be no disadvantages for using Data deduplication. Instead, we must choose the right technique along with secure encryption algorithms. There may be a risk of data corruption if something went wrong. So, it is required to create backups of the data whenever deduplication technique is used. In simple terms, the appropriate deduplication technique should be used whenever and wherever required.

VIII. CONCLUSION

In this paper, a client-side deduplication technique and updating of users data in a secured way is proposed in the cloud storage. The proposed system removes the addition of redundant files to the cloud. The deduplication approach is used in this paper will reduce redundant files, lower the usage of storage capacity and overcoming the bandwidth issues by maintaining security. Here, the whole process of deduplication will process in an encrypted way so that no data loss will happen during the transfer and retrieval of data. The efficiency of this system can be seen in terms of usage of bandwidth as it reduces the overhead that must be uploaded to the cloud.

ACKNOWLEDGMENT

Our sincere thanks to Dr. B.Tirapathi Reddy for his outstanding support and esteemed guidance for the successful completion of the project work. We express our gratitude to KLEF Management for providing us with adequate facilities. We appreciate all the team members for their cooperation and support in the completion of our work.

REFERENCES

- [1] https://en.wikipedia.org/wiki/Data_deduplication
- [2] https://en.wikipedia.org/wiki/Cloud_storage
- [3] <https://searchstorage.techtarget.com/definition/data-deduplication>

- [4] <https://www.computerworld.com/article/2474479/data-deduplication-in-the-cloud-explained--part-one.html>
- [5] Hyungjune Shin*, Dongyoung Koo†, Youngjoo Shin‡, and Junbeom Hur “Privacy-preserving and Updatable Block-level Data Deduplication in Cloud Storage Services” In IEEE 11th International Conference on Cloud Computing, 2018
- [6] A Study on Authorized Deduplication Techniques in Cloud Computing, Int. J. Adv. Res. Computer Engg. Technol. (IJARCET), 3(12) (2014).
- [7] J. R. Douceur, A. Adya, W. J. Bolosky, P. Simon and M. Theimer. “Reclaiming space from duplicate files in a serverless distributed file system,” In Proceedings of the 22nd International Conference on Distributed Computing Systems (ICDCS’02), pp. 617-624, 2002
- [8] Lauren Whitehouse. “The pros and cons of file-level vs. block-level data deduplication technology[Online],” Available: <http://searchdatabackup.techtarget.com/tip/Thepros-and-cons-of-file-level-vs-block-level-data-deduplication-technology>, accessed Dec. 11, 2017.
- [9] M. Bellare, S. Keelveedhi, and T. Ristenpart. “Message-locked encryption and secure deduplication,” In: Johansson T., Nguyen P.Q. (eds) Advances in Cryptology - EUROCRYPT 2013, vol 7881, pp. 296-312, 2013.
- [10] P. Puzio, R. Molva, et al. “Clouddedup: Secure deduplication with encrypted data for cloud storage,” In IEEE 5th International Conference on Cloud Computing Technology and Science (CloudCom’13), Dec 2013.
- [11] Storer, Mark W., et al. “Secure data deduplication,” In Proceedings of the 4th ACM International Workshop on Storage Security and Survivability (StorageSS’08), pp. 1-10, 2008
- [12] R. SHOBANA*, K. SHANTHA SHALINI, S. LEELAVATHY and V. SRIDEVI “DE-DUPLICATION OF DATA IN CLOUD” in 2016
- [13] Pasquale Puzio, ClouDedup: Secure Deduplication with Encrypted Data for Cloud Storage, <http://elasticsecurity.com/2013/12/10/clouddedup-securededuplication>
- [14] <https://www.webwerks.in/blogs/what-are-real-benefits-data-deduplication-cloud>
- [15] Yongjun Zhao and Sherman S. M. Chow. “Updatable blocklevel message-locked encryption,” In Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security (ASIA CCS’17), pp. 449-460, 2017.
- [16] B. Tirapathi Reddy, P. S. Kiran, T. Priyanandan, C. V. Chowdary and B. J. Aditya, "Block Level Data-Deduplication and Security Using Convergent Encryption to Offer Proof of Verification," 2020 4th International Conference on Trends in Electronics and Informatics (ICOEI)(48184), Tirunelveli, India, 2020, pp. 428-434, doi:10.1109/ICOEI48184.2020.9143055
- [17] B.Tirapathi Reddy, K.V. Padmasree " Secure Data Deduplication by Using Convergent Key Technique" International Journal Of Engineering And Technology(UAE),2018, 7 (2.32) (2018) 459-461.
- [18] B.Tirapathi Reddy, Dr.M.V.P.Chandra Sekhara Rao, “Privacy Preserving Proof of ownership for data in cloud storage Systems”, International Journal Of Engineering and Technology(UAE),2018, vol.7(2.8) pp 13-17.