P-ISSN: 2204-1990; E-ISSN: 1323-6903
DOI: 10.47750/cibg.2021.27.03.235

DSDA: A DESIRABLE AND SECURE DEDUPLICATION APPROACH FOR OUTSOURCED DATA IN CLOUD ENVIRONMENT

Esther Daniel[1], N.Susila[2], S.Durga[3]
[1,3]Department of Computer Science and Engineering, Karunya Institute of Technology and Sciences, Coimbatore, India
[2]Department of Information Technology, Sri Krishna College of Engineering and Technology, Coimbatore, India
E-mail: estherdaniell@gmail.com[1],
susila@skcet.ac.in[2],durga.sivan@gmail.com[3]

ABSTRACT: With ever-increasing and voluminous growth of data the enterprise cloud storage servers is critical since they provide efficient and desirable services to the end-users taking advantage of the cloud offerings in terms of providing services such as availability, reduced bandwidth, computational and storage cost. One of the solutions to provide low-cost storage services is Data Deduplication. This enables to save enormous storage space of the storage providers and thus reducing the cost. This paper proposes a deduplication technique termed Desirable and Secure Deduplication Approach for cloud storage environment (DSDA) eliminating duplicate data files and images securely. This approach combines convergent based encryption with a cuckoo filter to speed up the similarity search query and thus improves the performance while preserving the security. The experimental inference of this proposed approach exhibits reduced storage and communication cost.

KEYWORDS: Deduplication, Cloud Storage, Communication cost, secure deduplication.

I. INTRODUCTION

In the rise of the internet, there‟s a massive increase in digital data such as images, files, videos being produced, and stored in enormous quantities. In this context, the remote storage service providers namely cloud service providers (CSP) are popularly used as they provide efficient and convenient storage for the clients or end-users exploiting the benefits and cost savings of the cloud storage technologies in terms of transmission, computational, bandwidth and storage costs. According to an estimate [1], the amount of data produced across the globe doubles every year and has to be retained and archived efficiently for further use. One best technology to satisfy this necessity is cloud storage with service delivered over the Internet. It is forecasted that in 2020,more than 2 billion cloud users worldwide will be storing their data in the third party controlled remote datacentres which can be accessed without any additional encumbrances on the users' devices. A few of the most popular cloud storage service providers are Amazon S3, Dropbox, iCloud, OneDrive.According to Antony Adshead [2], every individual user has only 25% of unique data and the remaining 75% of the total data has similar common data contents linked to several other users. Added to that approximately on an average a business organization has two to five copies of a file stored across the devices and few of those enterprises have even roughly ten copies of the files. Grounded on these facts it is essential to implement data deduplication methods that enable the CSPs to efficiently manage and store a unique copy of the data without any duplications thus saving storage costs. Deduplication is a method of removing redundant blocks of data and lessening the usage of storage space. This maximizes the usage of cloud storage as the data transferred over the internet is reduced thus saving the bandwidth and storage requirements. Deduplication techniques can be broadly classified as in Fig 1.Deduplication is a process of the following steps: A single unique copy of the file blocks is to be stored.

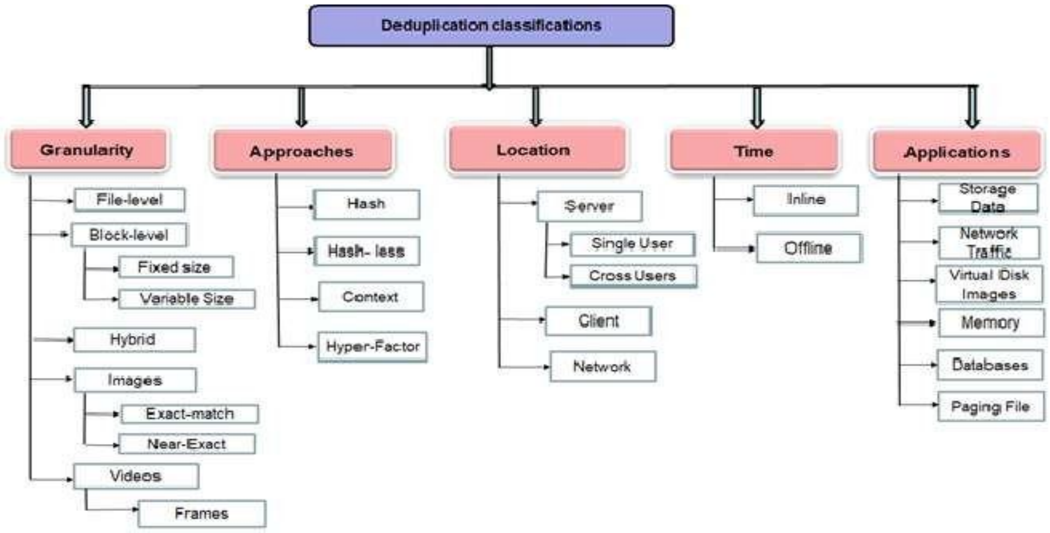

Fig 1: Categorization of Data Deduplication

If a user tries to store a duplicate file block on a storage server then the cloud storage service provider creates a link pointing to the original data instead of storing a duplicate copy. Deduplication is based on the two-level split-ups namely block and file-level deduplication. When a single copy of a file as a whole is stored then it is called file-level deduplication. The file is divided as blocks of data and stored. The block of each file is checked for uniqueness and if and duplication found the block is duplicate is discarded and only a link to that block is created. This way of storing the unique copy of the block is called as block-level deduplication. The hash value for each block is used to detect the duplication of the blocks. By comparing the blocks hash values the unique copy of the block can be retained [3-5].

The rest of the paper is organized asa review of existing literature in Section 2. The deduplication DSDA architecture and algorithm are explained in Section 3. Section 4 analyses the performance results of the system and Section 5 concludes the paper.

II. RELATED WORKS

The various literature emphasizing secure deduplication was studied. The DupLess system architecture by Bellareet al.[4] delivers a secure deduplication storage resisting attacks such as bruteforce.A set of clients of a company encrypts their information to be stored by secret keys managed by the Key Server(KS).This system uses convergent key encryption with a 128 bit key for SHA256.The storage space is reduced and the performance is improved in storing the encrypted data.However,a single point of failure at the key server causes a huge loss.

P-ISSN: 2204-1990; E-ISSN: 1323-6903
DOI: 10.47750/cibg.2021.27.03.235

Kaaniche et al.,[5]emphasized common security challenges in the cloud environment, the author had proposed the open stack shift, a client-side deduplication scheme. The access to the data is maintained by the clients and the keys will be distributed for the authorized users. The disadvantage is that the computation overhead for the encryption and decryption occurs at the client-side. S-POW Protocol, S-POW1 Protocol proposed by Pietro et al.[6] utilizes the data structure of hash map F that maps finite strings to 4- tuples of pointers „ptr" ,an array „res" consisting responses and indexes.Random access of the disks is required to fetch the data. In s-POW3 particularly, the server receives a digest from the client to be used as a key for lookup access. This scheme was a secure and better performance on the client side I/O. The security mechanisms of the proposed system relied on information retrieval rather than the computational assumptions. Malicious users can be able to get involved in attacks. In this scheme proposed by Keong ng et al.[7], the main objective is to overcome the difficulty of the encrypted string. The processes that were carried out in this proposed scheme are Private data deduplication protocols, Erasure Coding algorithm. Based on the output string of the original data the commitment scheme is implemented in this deterministic algorithm.The Merkle tree construction with the commitment of blocks at leaf nodes and the overall output of the commitments is the root of a file F. The advantage is that the PoW can be done without the file based on the encryption technique. However, this scheme doesn"t support public data deduplication. Halevi et al.[8], incorporated Proof of Ownership(PoW) for proving the file ownership. This scheme had used the Merkle hash tree for the creation of proofs for challenges. A Merkle tree delivers a concise "commitment" to the buffer allowing it to open and verify the single block of the buffer instead of a whole buffer.Better Deduplication was possible by using the proposed method. But the disadvantage was the computing Merkle tree was expensive. The overall running time was calculated by using the following formula $C_{time}+S_{time}+N_{time}$ and the performance was thus calculated. Yang et al.[9], the objective is to solve the problem of cost increases insecure systems. The authors had developed aa efficient scheme to support multiple client-side user deduplication checks over an encrypted file. The spotchecking scheme enables the client to check only a part of the original file selected randomly instead of the complete file.The disadvantage is that it is not a provably secure scheme but it can reduce the client-side burden. Li Wei et al.[10], introduced SecCloud for auditing purposes. This scheme includes both the concepts of POW and POR. SecCloud includes the initialization phase, storage phase,computation phase, Commitment, Verification, and Batch auditing. This scheme provides secure data storage, privacy but the communication costs were not noticed. The two main requirements for the proposed cloud storage scheme are Data integrity and storage efficiency [11]. This includes KeyGen, Setup,Challenge. The signature is checked for validity and a decision is taken whether to proceed or stop with access. The challenge raised by the user to the cloud server enables to check the data authentication tags and the files for deduplication.Thusthe integrity of data as well as eliminating duplicate copy are achieved.Jingwei Li et al.[12], proposed two schemas they are SecCloud, SecCloud+. SecCloud overcomes the need for integrity auditing of files and deduplication of data. In SecCloud+, the integrity auditing of data and deduplication are done on encrypted data. So SecCloud+ enables the guarantee of file confidentiality. The disadvantage is the communication cost get increases based on the file size and the proposed scheme cannot maintain the constant cost.

P-ISSN: 2204-1990; E-ISSN: 1323-6903
DOI: 10.47750/cibg.2021.27.03.235

III. DSDA ALGORITHM

The objective of the DSDA protocol in Fig 2. is to reduce the overload of storage services and minimizing the usage of storage space. The storage space can be preserved by eliminating the extra copies of the existing files in the storage servers. To preserve the privacy and security of the files stored by the users the files are stored in an encrypted format. The file is protected employing secure key exchange so no intruders or attackers can decrypt the file. The keys are maintained safe so that only the members who possess the key will be able to access the files. To improve the security for every file-download dual level authentication is executed thus improving the security along with saving the storage space on the cloud. To store a single unique copy of a file or a part of a file owned by multiple users the DSDA algorithm uses hash-based encryption functions. Based on the service level agreements between the cloud storage service provider and the client the storage servers will have data backup for the availability of data and thus data loss can be avoided. The security of the data is maintained by encryption techniques. The ThirdParty Auditor(TPA) enables the check for the data loss or misbehaviors if any on behalf of the users.A deduplication storage server can store both data and metadata. In a distributed storage environment, there is a specific independent server that stores all metadata, and the data are stored in a sequence of object store storage racks. The critical factors influencing the similarities checking in the deduplication server are file size, file type, file popularity, and user privileges. The deduplication can be done at the source i.e. at the client-side or at the target destination which is the storage servers. The client-side deduplication checks will save huge bandwidth costs as it avoids unnecessary data transmission from the data owner. The server-side deduplication check allows cross-level checking with other user"s data and thus saving 70% to 80% of storage space by avoiding keeping an extra copy of the redundant data. TheCuckoo filters enable to speed up the query transmission and retrieval process from the membership data set.

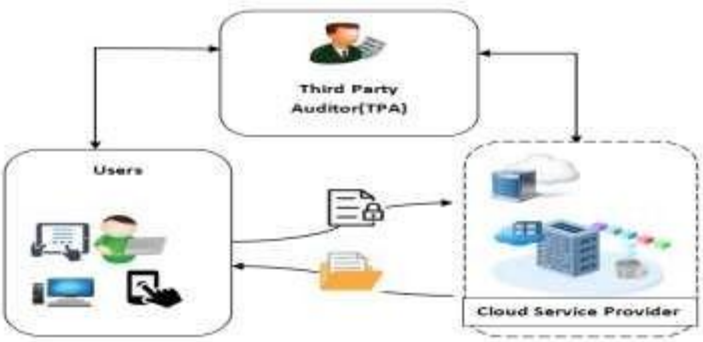

Fig 2. Deduplication Storage System Architecture

Duplicate files will be identified by matching the hash values generated for each block for each file. If the hash values remain the same then the existence of the duplicate file is confirmed otherwise, there will be no presence of duplicate files in the storage.

3.1 Algorithm for deduplication

Input:

Step 1: *The client will choose the file to be stored.*

Step 2: *At the client side,*

The data file will be divided into blocks.

Individual block of data is considered as a node of the tree.

The hash values for each data block or node will be computed this acts as a key for encryption Then the encryption using the DES algorithm is done and a file tag will be generated for the file. The File tags are stored at the TPA for lookup of all the hash table values or fingerprints

Cuckoo filter holds the fingerprints and enables them to check whether the data block is a unique copy to be stored.

Every block has the block id||hash value||relative index||location||Timestamp

Step 3: *At the TPA side,*

The client will now upload the file id along with the hash values to the TPA.

The TPA will maintain an index table for storing the hashes for the blocks and enables quick look-up for similarity and also checks malicious attacks on the servers

If the hash values match with each other,then the TPA will notify the client about the duplicate file. Otherwise, the uploading process will become successful.

Step 4: *At the server-side,*

Verify the duplicate files by comparing the stored hash values with the newly generate hash values. If the hash does not match then the file is not duplicate, so the file will be saved in the database.

Load the blocks and the node list to the server

Otherwise, the file will not get stored in the database. Remove the duplicated chunks

Step 5:For Image data,

Convert the image file into a matrix form.

Resize the image to dimension (128 x 128) or (256 x 256) to get common dimension for all images. Find the correlation coefficient for both images.

If the correlation is greater than 0.70 than print the images are nearly the same. Else print images are different.

The filters enable to speed up the search query results and also helps in achieving high throughput. The TPA stores and look-up the fingerprint hashes of data blocks in a file. In the image recognition process, we often need to find the similarity between the two images that are between the test image and its equivalent training database image. The metrics, coefficient of correlation gives the degree of correlation between any of the two images.

Correlation Coefficient is calculate using the following formula:

$$r = \frac{\sum_m \sum_n (A_{mn} - \overline{A})(B_{mn} - \overline{B})}{\sqrt{\left(\sum_m \sum_n (A_{mn} - \overline{A})^2\right)\left(\sum_m \sum_n (B_{mn} - \overline{B})^2\right)}}$$

where $\overline{A}$ = mean2(A), and $\overline{B}$ = mean2(B).

…………1

The correlation coefficient has value r = 1 if the two input images are same and equally correlated, r = 0 if the images are fully uncorrelated,
r = -1 if the images are inversely correlated.

This sample metric CoC in eq 1. is used for image recognition, which is often used as image quality measures. Correlation is the techniquefor generatingaconceivable outcome where there exists a linear solution wasthat a linear connectionbetween two measured quantities. A correlation coefficient which determines the similarity between the given two images.

IV. PERFORMANCE ANALYSIS

The performance is evaluated based on the computation cost taken at the client, TPA side, and the server. And also, the time is taken for uploading based on the different file sizes. The time cost for the hash value generation for the data blocks of different file sizes. By considering these mentioned performance metrics, the performance of the approach is evaluated and the result is analyzed with the existing protocols.Figure 3 shows the time taken for the upload process (i.e., Block separation+ Encryption+ Hash values) respect to the file size. The time taken for computation is almost equal to the ordinary upload process. Due to the insertion of the deduplication process at the client-side, there is no certain overhead given to the client. As the TPA is undertaking the verification process for the deduplication, the client side computation is also decreased. The storage space required is also minimized due to the avoidance of the repeated copies of the same file. The storage space usage minimization leads to less storage cost. Thus the proposed work is providing both proper uses of space in storage and the integrity of files by undergoing simplified steps.Fig.4 shows the time taken for the download process. As in DSDA, some extra features are added during the download process to improve the security of the storage, the usual process should not be exceeded in the time taken for the client to download. In such a case, the above graph is analyzed for several types of files download. The time taken is not so much exceeded than the usual download process.

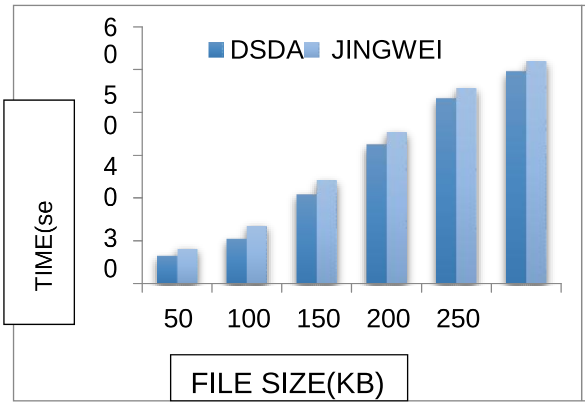

Fig.3Time taken for File Upload

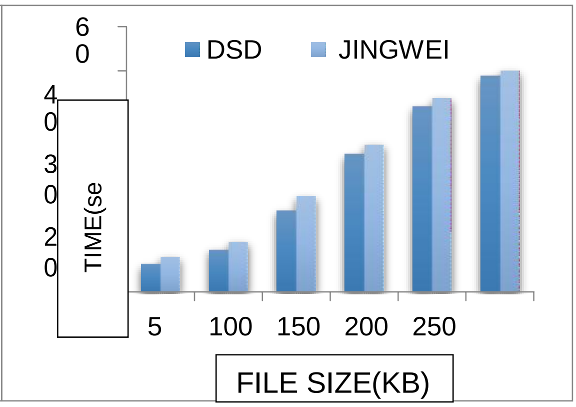

Fig. 4 Time taken for File Download

The comparison made between the proposed work and the existing work. The overload at the client side and the computation cost is minimized for about 20% in the proposed work.The overload at the client side and the computation cost is a little bit minimized for about 15% in the proposed work.

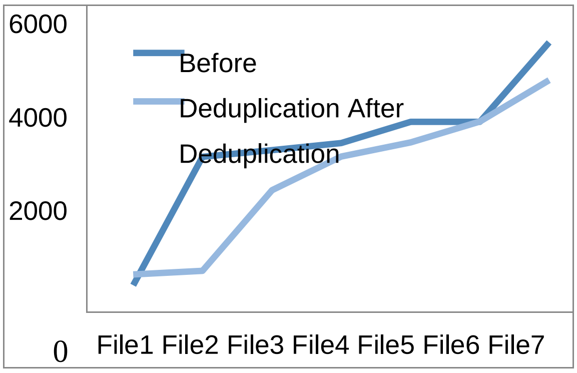

Figure 5 Deduplication Progress

Fig. 5 shows the deduplication progress for each files stored in the server. Storage consumption is noticed for different types of files. The proposed work saves the storage space based on the file's duplication contents.

V. CONCLUSION

In this paper, we propose a DSDA scheme that uses convergent based encryption with cuckoo filters to securely find the similarity of the data. The prominent feature of this scheme is reduced computation cost and storage cost. Deduplication techniques balanced against user‟s privacy concerns is highly cumbersome. The combined

P-ISSN: 2204-1990; E-ISSN: 1323-6903
DOI: 10.47750/cibg.2021.27.03.235

encryption and filter based DSDA ensure the users that their data has been saved in the cloud without tampering or manipulations.

VI. REFERENCES

[1]. https://www.statista.com/topics/3150/data-storage/ accessed on 23/4/18 [2]. https://www.computerweekly.com/feature/A-guide- to-data- de-duplication

[3]. Shin, Youngjoo, Dongyoung Koo, and JunbeomHur. "A survey of secure data deduplication schemes for cloud storage systems." *ACM Computing Surveys (CSUR)* 49, no. 4 (2017): 74.

[4]. S. Keelveedhi, M. Bellare, and T. Ristenpart, "Dupless: Serveraided encryption for deduplicated storage," in Proceedings of the 22Nd USENIX Conference on Security, ser. SEC"13. Washington, D.C.: USENIX Association, 2013, pp. 179–194.

[5]. NesrineKaaniche, Maryline Laurent " A Secure Client Side Deduplication Scheme Cloud Storage Environments" 6TH International Conference On New Technologies,Mobility And Security Year 2014

[6]. R. Di Pietro and A. Sorniotti, "Boosting efficiency and security in proof of ownership for deduplication," in Proceedings of the 7th ACM Symposium on Information, Computer and Communications Security, ser. ASIACCS "12. New York, NY, USA: ACM, 2012, pp. 81–82.

[7]. W. K. Ng, Y. Wen, and H. Zhu, "Private data deduplication protocols in cloud storage," in Proceedings of the 27th Annual ACM Symposium on Applied Computing, ser. SAC "12. New York, NY, USA: ACM, 2012, pp. 441–446.

[8]. S. Halevi, D. Harnik, B. Pinkas, and A. Shulman-Peleg, "Proofs of ownership in remote storage systems," in Proceedings of the 18th ACM Conference on Computer and Communications Security. ACM, 2011, pp. 491–500.

[9]. Yang, Chao, Jianfeng Ma, and Jian Ren. "Provable Ownership of Encrypted Files in De-Duplication Cloud Storage." Ad Hoc & Sensor Wireless Networks26.1-4 (2015): 43-72.

[10]. L. Wei, H. Zhu, Z. Cao, W. Jia, A. Vasilakos, Seccloud: bridging secure storage and computation in cloud, in: 30th International Conference on Distributed Computing Systems Workshops (IEEE ICDCSW 2010), Genova, Italy, June 21–25, 2010

[11]. J. Yuan and S. Yu, "Secure and constant cost public cloud storage auditing with deduplication," in IEEE Conference on Communications and Network Security (CNS), 2013, pp. 145–153.

[12]. JingweiLi,Jin Li, DongqingXie and Zhang Cai," Secure Auditing and Deduplicating Data in cloud", 0018-9340 (c) 2015 IEEE Transactons on Computers.

[13].

P-ISSN: 2204-1990; E-ISSN: 1323-6903
DOI: 10.47750/cibg.2021.27.03.235

AUTHORS PROFILE

Esther Daniel is working as an Assistant Professor in the Department of Computer Science and Engineering at Karunya Institute of Technology and Sciences, India. She has obtained her Bachelor of Engineering from Bharathiar University and a Master of Engineering from Karunya University. She completed her Ph.D. from Anna University. Her area of interests includes computer networking, cloud computing, machine learning, and information security. She has published papers in reputed national and international journals and conferences. She is acting as a Reviewer for several international journals such as Elsevier Journal of Computers & Electrical Engineering, IEEE Transactions on computer, Science China and Inderscience publications.

N.Susila is currently working as Professor and Head of Information Technology Department, Sri Krishna College of Engineering and Technology with a total experience of 19 yrs. She has completed her Bachelor's Degree in Computer Science and Engineering from PeriyarManiammai College of Technology for Women, Bharathidasan University in the year 2001. She has completed her Master's Degree in Computer Science and Engineering from Sathyabama Institute of Science and Technology in the year 2005 as a 9th Rank Holder. She has completed her Ph.D. in the field of Information and Communication Engineering in the area of Cloud Computing in the year 2017. She has published various papers in International Journals and Conferences. She is acting as a Reviewer for several international journals like Expert Systems, IGI Global, Wiley, etc. She has authored a good number of books and book chapters.

P-ISSN: 2204-1990; E-ISSN: 1323-6903
DOI: 10.47750/cibg.2021.27.03.235

S.Durga is currently working as Assistant Professor, Department of Computer Science and Engineering at Karunya Institute of Tech. and Sciences, India. She received the degree of BTech with Distinction in Information Technology. She received her ME with distinction in Network and Internet Engineering. She completed her Ph.D. in IT. Her research interest includes computer networks, IoT, Machine learning, and mobile cloud computing. She has a good number of publications in peer-reviewed journals and conferences. She is a reviewer in the Elsevier Journal of Computers & Electrical Engineering and Inderscience Journal of Cloud Computing.