JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p. 5382-5393
https://publishoa.com
ISSN: 1309-3452

Systematic Review of Data Deduplication in Cloud Storage and its Challenges

K. E. Narayana[1], Dr. K. Jayashree[2]

[1] Assistant Professor, Department of Computer Science and Engineering ,Rajalakshmi Engineering College, Chennai, India, narayanake@gmail.com

[2] Professor, Department of Artificial Intelligence and Data Science, Panimalar Engineering College, Chennai, India,
k.jayashri@gmail.com

Received: 2022 March 15; **Revised:** 2022 April 20; **Accepted:** 2022 May 10

Abstract

In recent years, the data storage system plays important role in the technology called big data management and cloud storage systems. Duplication of the data in the cloud system is a key problem, which decreases the data storage efficiency in the system. To avoid this problem, the data deduplication technology helps to improve the efficiency of the data storage systems. The data deduplication shrinks the need for storage by removing and avoiding uploads of duplicate data in the particular cloud environment. The integrity and confidentiality of the data are essential factors of the data deduplication technologies. As well as, the security measure of the data storage systems needs to be considered while building the data deduplication technologies. Because most common data encryption algorithms cannot be applied while performing the data deduplication process. This paper describes the process behind the data deduplication and analyzes various data deduplication techniques in terms of security threats.

Keywords: Cloud Storage System, Confidentiality, Data Deduplication, Encryption Algorithms, and Integrity.

1. Introduction

In recent times, cloud computing is termed “computing as a utility” due to its flexibility and scalability. Cloud computing has a huge amount of benefits such as omnipresent access to the network, individual payments, resource provisioning demand, autonomous access for location, and quantifying resources for end-users or cloud customers. The data storage also known as storage-a-service provided by cloud technology is useful for individuals and also to government and Big tech companies who do not have adequate storage space [1]. The fast improvement of cloud computing and data storage technologies causes a huge impact on the usage of data processing because the cloud servers provide fast computing and efficient storage anytime and anywhere for cloud users [2]. The cloud system provides quantifiable storage space to avoid large data extraction [3]. Performing data deduplication in the cloud is beneficial for Cloud Service Providers(CSPs). Uploading the same data in a cloud system can lead to a decrease in storage efficacy. To overcome this problem, data deduplication is used to reduce the storage space [4]. The EMC authority conducted the survey in cloud storage, reviewed about 75 percent of the uploaded files as duplicated data in the cloud [5].

Data deduplication techniques are generally divided into two parts, which are Server-side deduplication and Client-side deduplication techniques. Server-side deduplication refers to the deduplication operation examining the uploaded data from the client and removing the duplicate files. Client-side deduplication means the deduplication operation performs before the file is uploaded by the cloud user [6]. Client-side deduplication is great, saves the bandwidth [7].In reality, many users can upload similar data to the cloud system, this permits CSP can accomplish data deduplication for all their end-users to reduce the storage cost and space. Data deduplication proves that higher cost savings can achieve, it saves a minimum of 50% of the cloud storage costs on standard files, and a maximum of up to 95% on backup systems [8], because the cloud retains the only distinct copy of the duplicated data and creates a restore path link to the particular file for the cloud end users [9]. Virtual machines are targeted by the attacker to steal the confidential information through side channel attack[10]. The benefits and opportunities of cloud storage are numerous, and it relieves data owners of the burden of scalable storage administration and maintenance (DOs). [10].

The objectives and contribution of the research work is discussed below:

1. The data deduplication is done by using three different kind such as storage based data deduplication, level based data deduplication and type based deduplication. The performance of the data deduplication is measured by time consumption of the system.
2. To safeguard the confidentiality and integrity of data, convergent encryption (CE) and Proof of Ownership (PoW) are utilized. Several other approaches have been investigated to address client security

concerns, including Provable Data Possession (PDP), Proof of Retrievability (POR), secure keyword search, DupLESS, Proof of Storage with Deduplication (PSD), Dekey, Message-Locked Encryption, Attribute-Based Encryption (ABE), and Identity Based Encryption (IBE).

3. The incompatibility between classical encryption and deduplication is addressed. Unlike previous research, which assumes that all files require equal security, this paper examines a strategy that secures data based on its popularity.

2. Taxonomy of data deduplication in cloud storage

The deduplication techniques can be classified into the following types: Primary and secondary deduplication depending on storage i.e., Type 1. Source and target deduplication based on type i.e., Type 2. In-line and post-processing deduplication based on processing time i.e., Type 3. Local and global-level deduplication based on level i.e., Type 4. The 4 factors plays an important role in the taxonomy of data deduplication which are type, storage, time, and level-based deduplication. Figure. 1. illustrate the graphical representation of data deduplication taxonomy.

These are further classified as follows:

- Global/distribution and local deduplication is a level based method
- Post-process and In-line deduplication is timing based method
- Target and source-level based deduplication is type method
- Primary and secondary storage depending on deduplication method.

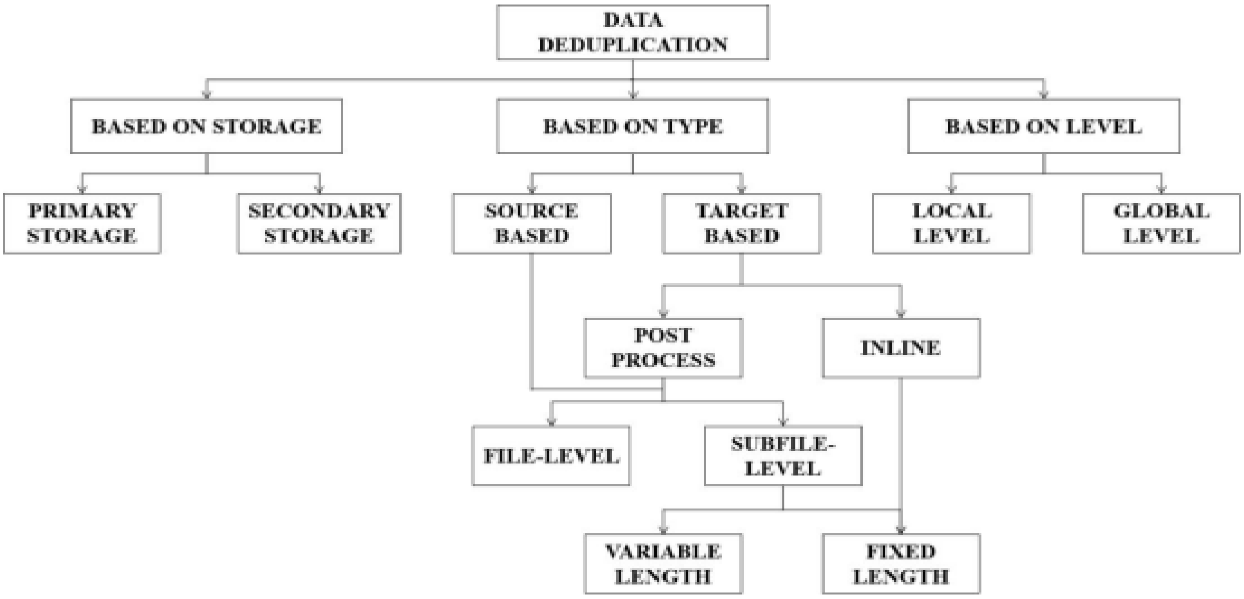

Fig.1. Taxonomy of data deduplication techniques

Yuan, *et al.* [11], proposed an authorized data deduplication system to perform a re-encryption algorithm depending on the convergent all-or-nothing convert, abbreviated as CAONT and bloom filter for sampling the random bit. The developed deduplication method efficiently reduces the system overhead because the data proprietors were only re-encrypt a small slice of the package over the CAONT, as an alternative of re-encrypting the whole set of a package. The data owners employed a bloom filter-based location selection mechanism, which provides a safe deduplication system by rejecting cloud users who are unable to get sensitive data from the data owner. The stub reserved attack affects the rekeying-aware encrypted deduplication (REED) technique.

Wang, *et al.* [12], proposed a decentralized fair payment protocol in a cloud deduplication storage system and make a reasonable payment using ethereal smart contract technology. At the same time, this system performs the functions of monitoring, data tracking, and decentralized fair payment. The fair payment protocol reduces the cloud storage cost and a load of the server-side, by using the deduplication technology. The advantage of the decentralization of block chain technology was permitting direct contacts without the contribution of trustworthy third parties. The block chain-based decentralized system has no complete structure which is the important disadvantage of this approach.

Fan, *et al.* [13], established a secure deduplication system depending on the Trusted Execution Environment (TEE). The deduplication only performed for the correct privilege cloud users because this scheme assigned the privilege set for each cloud user. The scheme can resist selected plain text attacks and chosen-cipher text attacks because the scheme enhances encryption with users' privileges and relies on TEE to deliver secure key supervision.TEE was used to replace third-party servers. Furthermore, the user may determine the acceptable charge of the search cost without using a third party, even if the cloud was hostile, by employing the block chain approach and the hash function.

Saharan, *et al.* [14], developed a QuickDedup for virtual machines(VMs) storage deduplication and in a cloud server. The QuickDedup algorithm obtained better results in the VM images, and efficiently saves the metadata storage. The proposed method was more effective than a traditional system in storage spaces and communication bandwidth .QuickDedup reduces the hash calculations in some candidate blocks. The possibility of a hash solution was developed based on the hashes used in the block-level deduplication. In absence of additional in-built verification, false positives cause loss of data integrity.

Ebinazer, *et al.* [15], developed a data deduplication technique to improve security by using Bloom filter(BF) along with radix tire (RT) model (SDD-RT-BF). The developed SDD-RT-BF system contains three important stages such as authentication type deduplication, valid data ownership, and role key updating. To prevent leakage of information, the convergent encryption algorithm is applied and plays re-encryption is performed for authorized data deduplication resourcefully. The RT structure is established to plan the roles and keys relationship to handle the authorized request by the key management center. The implementation of data updating was performed by BF and also it's capable of increasing the system efficiency and fast searching. However, when the data was large over the cloud due to enormous data the time probabilistic approach may lead to giving false-positive results.

3. Comparative analysis

Table1: The comparative analysis of the data deduplication techniques.

Si. No	**Author(s)**	**Method**	**Advantage**	**Limitations**	**Performance metrics**
1	Nayak and Tripathy (2019) [1]	Server-based deduplication method for cloud storage applied to check supports across the key servers.	The Multiple key servers are used to deal with a particular point failure and the process of key generation distributes the load amid the key servers. Individuals of security data do not work with handy using deduplication method or security challenges leading to many	Tag inconsistency attack (The breach in the data of the integrity). The developed method is vulnerable to security attacks such as brute-force attacks. This model has some attacks due to the convergent encryption being deterministic and key less.	Computation time for key servers is 2×10^4 ms 2×10^4 ms, computation time for files is 11×10^4 ms 11×10^4 ms.

			corporate of the legal documents.		
2	Xiong,Zhang and Tang, Liu and Yao (2019)[2]	Convergent encryption based on secure role re-encryption. Privacy data leakage prevents to use of role re-encryption algorithms in the cloud.	Expenditure reduces the management. The secure hash function resisted the collision attack. Data deduplication can be 83% reduced up to backup systems and memory systems for 68%. The specific file can access the corresponding privilege only can the user perform the deduplication in the cloud.	The shortcoming of the re-encryption scheme was the decentralized structure of the system, which is not complete. The file size increases the computational cost of generating file tokens.	Comparing the computational cost with different methods. The computational cost of the proposed system challenge generation is 3000ms.
3	Premkamal, Pasupuleti and Singh, Alphonse(2020) [3].	For massive data storage on cloud servers, EABACC-SD (advanced attribute-based access management with secure	The EABAC-SD scheme proved that data consistency resists duplicate attack faking.	Ownership data not verified; Data protection is essential when the same data is uploaded by multiple users. The data ownership management does not provide; it creates	Compared the time with different file sizes. The computation time for 100MB is 1.5s

		deduplication) was developed. Dynamic ownership management achieves the EABAC-SD scheme using the group key.		false data which may chance to uploaded by the ownership revoked owner. This method suffers from computational and communication overhead during duplications and process encryption,	and 1GB is 9s.
4	Liang, Yan and Deng (2020) [4]	In this method, the scheme of client-controlled deduplication for payoff structure and unified discount analyze the feasibilities and this structure discount under for individuals. Extensive tests using a real-world dataset have been conducted to illustrate the usefulness of the suggested mechanism.	The free-riding behaviours overcome the privacy issues. It helps to how owner's data analyze and strategies choose for holder's data based on the utility function.	C- DEDU needs a data Owner to stay online. So they have to pay relatively more for duplication than the data holder. This program will not be able to retrieve user data using adequate incentives. C-DEDU cannot be deployed smoothly .The data owners are difficult to guarantee online all the time, the model delay service avoid is too complicated.	Calculate the Duplication percentage. The developed method deduplication in different time generation percentage is 95%.
5	Bai, Yu, and	The first duplicate data	The tag consistency problem can solve the	If the data is accidentally modified	When it comes to auditing

	Gao(2020) [5]	integrity supporting a scheme of modification of the ownership. The integrity of the cloud services is ensured, and dynamic access control over the data is supported.	scheme of symmetric encryption by integrity check an additional phase. The model problem can be solved for data leakage with high probability. The computation owner-side overhead efficiently reduces using a method of integrity auditing and overhead storage.	or deleted by its user then cannot retrieve successfully which means deleted data no longer be successfully retrieved.	time, single and batch auditing are calculated. The auditing time per task for single auditing is 330ms. The auditing time per task for batch auditing is 500ms.
6	Wu, Li, and Wang, Ding (2019) [6]	With public cloud auditing, anonymity is preserved when deduplication cloud storage is used (CPDA). The unique copy in the deduplication cloud system storage is supported by public integrity auditing. The secure authentication tag deduplication	The server storage saves the client-side deduplication, costs communication, and bandwidth network, which both benefits for Clients and cloud servers.	The deduplication on data encrypted The problem is tough. Due to malicious attacks, the method can leak the static hash value so this leads to vulnerability for data owners	Comparing the costs of communication in the auditing phase and storage phase. The communication cost in the auditing phase of the proposed system is 100KB. The computational time of the proposed system is

		realizes the method of CPDA.			10000ms.
7	Wang, Wang and Song ,Zhang (2019) [7]	The key-sharing technique based on the secure deduplication proof of ownership The possessing users only can claim the data to retrieve the convergent key (CK).	The storage and maintenance of many CKs are also dangerous. Dekey to is called a new construction method that solves the problem of CK management. The assumption of the CDH is to resolve the issues like invisible computational in cluster CDH. Method of two-phase deduplication data was used to solve the issues of convergent encryption.	The CE is not secure for semantically and brute-force attacks defined from an offline point of view when the selected data from a set Of predictable. Communication overhead and collusion attacks cannot be resisted.	Comparing the computation time and storage. The computation overhead of the proposed system is 32s.
8	Li, Xu, and Zhang (2019) [8]	Methods of client-side encrypted data deduplication (CSED) were efficient and secure. A dedicated key is introduced in the CSED server in MLE	Deduplication reduced the storage overhead significantly. For cloud storage systems, Message Locked Encryption reduces both communication and computing overhead. The end user-side deduplication	The developed method was vulnerable to illegal attacks and content distribution. Where the data distributed can adversary to other users via the detecting without a cloud server.	Calculates the number of blocks/chunks and computing time for that. The computational time of 500 blocks/chunks is 570ms.

		keys generation which resists brute-force attacks.	approach was highly efficient in terms of communication overhead.		

4. Problem statement

- In the primary storage system, the data deduplication is performed by using the fixed block method which decreases the performance of the system. Random inheritance of primary data leads to creating many chunk data at various locations which increases re-assembly times.
- Energy management is one of the important tasks for every data processing operation in cloud computing. The consumption of quite an amount of energy produces lots of radiation. The VMs consolidate dynamically and use efficient storage as few of the methods used will decrease the usage of energy. Data deduplication can help to increase storage usage by deleting duplicate copies from memory space.
- In deduplication, the post-process and inline deduplication is based on the time of the operation. There is no pre-data storage space in offline or post-process deduplication, which ensures better performance of the system. If the storage of the cloud system is close to it’s full volume, then the duplicate data in the storage causes a huge problem. In-line deduplication does not need big storage space because it does not save duplicate information. Moreover, this may degrade the performance of the storage because the computation takes time.

5. Conclusion

In recent decades, the data traffic and data duplication on the data storage systems are causing a huge problem in cloud storage systems. To solve this problem, data deduplication comes into the picture by planning network traffic and addressing the duplicate data in the storage systems. The main goal of the Data deduplication technologies is to build the system to minimize the price of the cloud storage system. The architecture of the cloud environment is built by the remote servers which can leak the data to another node by the internal or external features. Data leakage and tempering are avoided to solve the two important factors which are integrity and confidentiality. The integrity of the cloud storage system has been solved by removing the duplicate data in the system and it's called data deduplication. And, confidentiality is achieved by encrypting the data which we are going to store in the cloud systems. In this review paper, we analyzed various kinds of Data deduplication technologies that are proposed in recent years. From the analysis, those recent methodologies are performing well in terms of omitting duplicate information in the storing system. But, need some modification to expand the security factors of the cloud system. The security threats are increasing day by day which leads to increases in the security in data deduplication methodologies. In future work, the research scholars need to work to improve the security factor in data duplication technologies.

References

[1] Kayak, S.K. and Tripathy, S., SEDS: secure and efficient server-aided data deduplication scheme for cloud storage. International Journal of Information Security, 19(2), pp.229-240.(2020).

[2] Xiong, J., Zhang, Y., Tang, S., Liu, X. and Yao, Z., Secure encrypted data with authorized deduplication in cloud. IEEE Access, 7, pp.75090-75104. (2019).

[3] Premkamal, P.K., Pasupuleti, S.K., Singh, A.K. and Alphonse, P.J.A., Enhanced attribute based access control with secure deduplication for big data storage in cloud. Peer-to-Peer Networking and Applications, 14(1), pp.102-120.(2021).

[4] Liang, X., Yan, Z. and Deng, R.H. Game theoretical study on client-controlled cloud data deduplication. Computers & Security, 91, p.101730.(2020).

[5] Bai, J., Yu, J. and Gao, X., Secure auditing and deduplication for encrypted cloud data supporting ownership modification. Soft Computing, 24(16), pp.12197-12214.(2020).

[6] Wu, J., Li, Y., Wang, T. and Ding, Y., CPDA: A confidentiality-preserving deduplication cloud storage with public cloud auditing. IEEE Access, 7, pp.160482-160497.(2019).

[7] Wang, L.,Wang, B.,Song, W. and Zhang, Z.,A key-sharing based secure deduplication scheme in cloud storage. Information Sciences, 504, pp.48-60. (2019).

[8] Li, S., Xu, C. and Zhang, Y., CSED: Client-side encrypted deduplication scheme based on proofs of ownership for cloud storage. Journal of Information Security and Applications, 46, pp.250-258. (2019).

[9] Shen, W., Su, Y. and Hao, R., Lightweight cloud storage auditing with Deduplication supporting strong privacy

protection. IEEE Access, 8, pp.44359-44372.(2020).

[10] Daniel, E. and Vasanthi, N.A., LDAP: a lightweight deduplication and auditing protocol for secure data storage in cloud environment. Cluster Computing, 22(1), pp.1247-1258. (2019).

[11] Yuan, H., Chen, X., Li, J., Jiang, T., Wang, J. and Deng, R., Secure cloud data deduplication with efficient re-encryption. IEEE Transactions on Services Computing.(2019).

[12] Wang, S., Wang, Y. and Zhang, Y., Blockchain-based fair payment protocol for deduplication cloud storage system. IEEE Access, 7, pp.127652-127668.(2019).

[13] Fan, Y., Lin, X., Liang, W., Tan, G. and Nanda, P., A secure privacy preserving deduplication scheme for cloud computing. Future Generation Computer Systems, 101, pp.127-135.(2019).

[14] Saharan, S., Somani, G., Gupta, G., Verma, R., Gaur, M.S. and Buyya, R., QuickDedup: Efficient VM deduplication in cloud computing environments. Journal of Parallel and Distributed Computing, 139, pp.18-31.(2020).

[15] Ebinazer, S.E. and Savarimuthu, N., An efficient secure data deduplication method using radix trie with bloom filter (SDD-RT-BF) in cloud environment. Peer-to-Peer Networking and Applications, pp.1-9.(2020).