International Journal of Engineering Technology and Management Science
Website: ijetms.in Issue: 2 Volume No.9 March - April – 2025
DOI:10.46647/ijetms.2025.v09i02.110 ISSN: 2581-4621

Secure and Efficient Data Deduplication in Joint Cloud Storage

Poluka Venkata Hymasree[1], CH. Sri Lakshmi Prasanna[2]
[1]*MCA Student, Dr.K.V.Subba Reddy Institute of Technology, Kurnool, Andhra Pradesh, India*
[2]*Assistant Professor, Dr.K.V.Subba Reddy Institute of Technology, Kurnool, Andhra Pradesh, India*

Abstract
Data deduplication can efficiently eliminate data redundancies in cloud storage and reduce the bandwidth requirement of users. However, most previous schemes depending on the help of a trusted key server (KS) are vulnerable and limited because they suffer from revealing information, poor resistance to attacks, great computational overhead, etc. In particular, if the trusted KS fails, the whole system stops working, i.e., single-point-of-failure. In this paper, we propose a Secure and Efficient data Deduplication scheme (named SED) in a Joint Cloud storage system which provides the global services via collaboration with various clouds. SED also supports dynamic data update and sharing without the help of the trusted KS. Moreover, SED can overcome the single-point-of-failure that commonly occurs in the classic cloud storage system. According to the theoretical analyses, our SED ensures the semantic security in the random oracle model and has strong anti-attack ability such as the brute-force attack resistance and the collusion attack resistance. Besides, SED can effectively eliminate data redundancies with low computational complexity and communication and storage overhead. The efficiency and functionality of SED improves the usability in client-side. Finally, the comparing results show that the performance of our scheme is superior to that of the existing schemes.

INTRODUCTION

WITH the rapid development of cloud storage, more and more individuals and enterprises tend to outsource their sensitive data to remote cloud service providers in a pay-per-use manner [1], [2], [3], [4], [5]. According to the study from Internet Data Center (IDC) sponsored by Dell EMC, the digital universe is doubling in size every two years and the volume of the universe data is expected to reach 44 zettabytes (ZB) or 44 trillion gigabytes (GB) in 2020 (more than 5200 gigabytes for each man, woman, and child) [6]. However, the growth of data puts heavy pressures on cloud service providers. To cope with it, a straightforward method is to require cloud service providers continuously increasing the capacity of storage space, so as to meet users' requirements for high-quality storage services.

However, cloud service providers may store plentiful and repetitive data (such as movies, music and genome data), which inevitably incurs a mass of redundant storage and backup space, consequently to cost a vast amount of computing and management overhead during its whole life cycle. To solve this problem, Bolosky *et al*. first proposed the technique of data de duplication [7], which decreases the redundant storage space and bandwidth by eliminating duplicate copies and only storing one copy of them. Nowadays, data deduplication techniques have been widely deployed by cloud service providers, such as Drop box [8], Google Drive [9] and Memopal [10]. Researches [11], [12] have shown that most of the genome data ($\geq 83\%$) and disk ($\geq 90\%$) of business applications can be reduced by exploiting data de duplication technique. While the technique of data de duplication has plentiful advantages, there are still some security challenges that need to be addressed. In particular, the cloud service providers are often assumed to be not fully trusted, which may try to infer and analyze the outsourced data [13], [14], [15], [16]. To protect the confidentiality of their sensitive data, cloud users generally encrypt their data before outsourcing them to cloud service providers. However, different users encrypt the same data with their private keys, which leads the same data to output different cipher texts, and makes the function of data de duplication unachievable. Douceur

DOI:10.46647/ijetms.2025.v09i02.110 ISSN: 2581-4621

et al. [17] proposed the first feasible solution to protect the confidentiality of data and achieve de duplication on cipher texts. However, the cloud user encrypts sensitive data with a convergent key, which is derived from the hash value of the data and unchanged. It will lead the revoked cloud user to access the sensitive data through the reserved convergent key.

User revocation is a severe security problem in the cloud environment. We take the case in genome research as an example to illustrate this point. Considering the enormous volume of genome datasets, genome researchers tend to use the cloud to store the genome data [18]. Google Genomics [19] and Amazon [20] have deployed specific platforms for managing and analyzing genome data. However, some sensitive genome data produced by disease sequencing projects must be protected. For example, when a researcher is no longer a member of the genome project, he will be prohibited from accessing the genome datasets. This problem has been addressed by using techniques such as re-encryption and group key distribution [21], [22], [23]. By using symmetric encryption (such as AES-128 or AES-256) to re-encrypt the sensitive data and distribute group key for group users, those schemes can support user joining and user revocation. Although the re-encryption scheme uses a new encryption key to encrypt the entire message to protect the data confidentiality, it will result in a waste of excessive computation overhead. William *et al*. [24] provided evidence that, in situations involving even a minimal amount of policy dynamism, the cryptographic enforcement of access controls is likely to carry prohibitive costs.

Recently, Li *et al*. [25], [26] proposed a rekeying-aware encrypted deduplication storage system (REED), which supports a lightweight re-encryption. Instead of re-encrypting the entire package, data owners are only required to reencrypt a small part of it through the CAONT, thereby effectively reducing the computation overhead of the system. However, we point that the REED is vulnerable to the stub-reserved attack, which will be described in Section 3.2. In short, if a revoked cloud user keeps the last bytes of a package as *stub* package, he can use the reserved *stub* package and *trimmed* package (downloaded from the cloud service providers) to recover the plaintext. Therefore, existing proposed schemes cannot well support secure dynamic ownership management of cloud users and efficient re-encryption.

Existing System

Convergent encryption is one of the main approaches to ensure the security of data in deduplication, which can protect the outsourced data against the untrusted and malicious CSPs. Bellare et al: [10] formalized a primitive as message-locked encryption (MLE) scheme. Then, some variants [12], [13] were proposed based on the work of Bellare. However, these MLE-based schemes were facing many potential risks because the keys used to encrypt files are derived from the files themselves. Abadi et al: [11] designed a full randomized scheme and a deterministic encrypted scheme for bounded message distributions based on the non-degenerate efficiently computable bilinear map. Li et al: [15] presented a scheme to achieve reliable key management in deduplication. Then, Jiang et al: [14] showed a secure scheme for deduplication with the randomized tag. However, they did not mention the requirements of data updating.

Later, in [19], Hur et al: considered the dynamic ownership management for secure deduplication and each client needs to store the keys located on the path of a binary
tree of all keys to achieve deduplication. Li et al: [15], [32] proposed secure deduplication with key management based on secret sharing schemes. After that, Shin et al: [20] designed a decentralized server-aided encryption for deduplication. But it needed multiple interactions between users and KS, which gave the attackers opportunities to get useful information from the communication [33]. In [34], Miao et al: proposed a secure deduplication for multi server- aided. To achieve fine-grained data deduplication,

Xia et al. [3] designed a Fast and efficient content-defined chunking approach. Zhao et al. [35] proposed deduplication scheme based a Docker registry architecture, which deduplicates layers for space saving and reduces layer restore overhead. In terms of security, the common attacks for recent deduplication schemes were discussed in [18], [36]. Beside, some emerging technologies such as

DOI:10.46647/ijetms.2025.v09i02.110 ISSN: 2581-4621

blockchain [22], [37], [38] are used for deduplication in various applications, which has be focused on by researchers recently.

Disadvantages

➢There is no Intra-deduplication which just considers that the data owner has outsourced his/her data by the same KS. It is more efficient for the backup system which is not in an existing system.

➢There is no technique which inter-deduplication that considers the data outsourced by different data owners through multiple KSs.

Proposed System

In this paper, we propose a secure and efficient data deduplication scheme SED without the help of the trusted KS in the JointCloud storage system. Some sub-algorithms of our SED are inspired by the fully randomized tag generation algorithm [11] which helps with duplicates detection and protects the outsourced data against the collusion attacks. Different from the previous deduplication schemes, our SED ensures that the ciphertext and the tag can satisfy semantic security. Any adversary cannot get any useful information from the tag and ciphertext. Moreover, our SED is the first scheme that supports data update and data sharing securely.

In our SED, we design the encryption algorithm that supports data deduplication, update, and sharing. To the best of our knowledge, SED is the first scheme considering the case that data owner shares his/her outsourced data to permitted users. Especially, a master encryption key is generated by the collaboration of the participating CSPs. It ensures the flexibility and security of key generation. The data access control based on authentication of SED also helps with the implementation of data update and sharing operations. Then, SED combines the intra-deduplication and inter-deduplication techniques to eliminate duplicates in the JointCloud system, which improves the efficiency of data deduplication. After that, the theoretical analyses indicate that the SED has superior performance such as data confidentiality, data integrity, strong attacks and collusion resistances, and functionality. In order to evaluate the complexity experimentally, SED is implemented and simulated based on Crypto++ [23], GNU [24], and PBC [25] librariesin Ubuntu. The evaluation results show that SED is efficient and costs low computational overhead.

Advantages

➢The encryption algorithm and the tag generation algorithm of the proposed SED ensure the semantic security. Moreover, SED can resist the typical attacks such as the brute-force attack, tampering attack, and collusion attack.

➢The SED implements secure deduplication without the help of the trusted key server. It also supports data updating and sharing cross-clouds. Furthermore, SED solves the single-point-of-failure issue and improves the scalability of the classic deduplication scheme.

➢We conduct the experiments of SED scheme by invoking cryptographic libraries, verifying that our SED is efficient and has low computational overhead. We specially simulate inter-deduplication and intra-deduplication and the results demonstrate that they can improve the efficiency of duplicates detection.

Literature Survey

1. “Oruta: Privacy reserving Public Auditing for Shared Data in the Cloud,” AUTHORS: B. Wang, B. Li, and H. Li, It is common practice to share data with multiple users in addition to storing it in the cloud using cloud data services. Sadly, there are hardware/software failures and human errors that raise questions about the reliability of cloud data. Without having to download all of the data from the cloud server, both data owners and public verifiers can now effectively audit the integrity of cloud data through a number of mechanisms. Be that as it may, public examining on the respectability of imparted information to these current systems will definitely uncover secret data character protection to public verifiers. Public auditing of shared cloud-based data is made possible by a novel privacy-preserving mechanism that we propose in this paper. In particular, we make use of ring signatures to generate the verification metadata needed to check that shared data is correct.

DOI:10.46647/ijetms.2025.v09i02.110 ISSN: 2581-4621

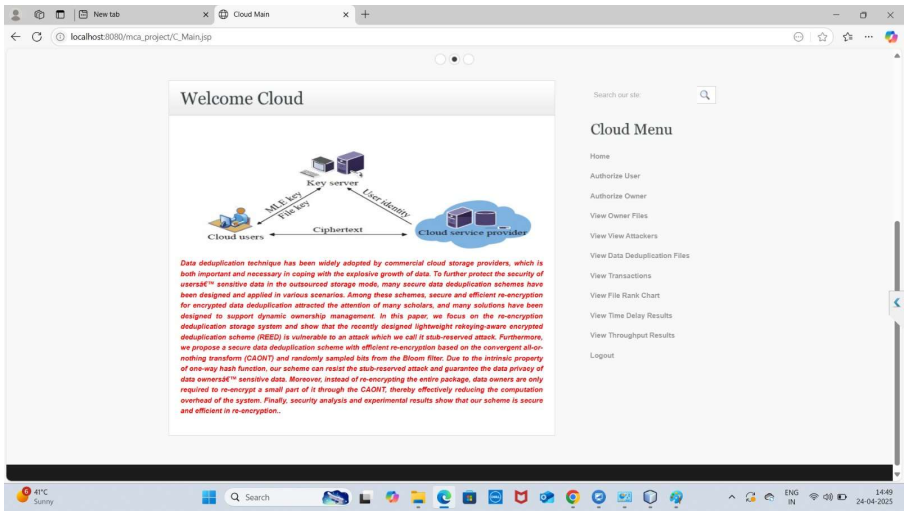
Public verifiers are able to effectively verify shared data integrity without retrieving the entire file because our mechanism keeps the identity of the signer on each block of shared data private. In addition, rather than verifying each auditing step individually, our mechanism can perform multiple auditing tasks simultaneously. Our exploratory outcomes show the adequacy and proficiency of our component when evaluating shared information trustworthiness.

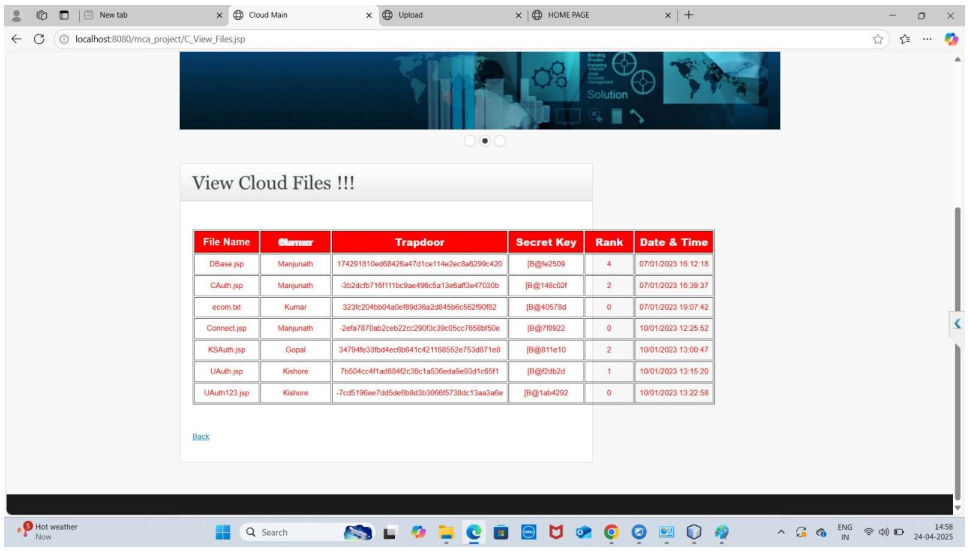
2. "Security Difficulties for the Public Cloud," Creators: K. Ren, C. Wang, and Q. Wang, In this discussion, I will initially examine various squeezing security challenges in Distributed computing, including information administration reappropriating security and secure calculation rethinking. The security of cloud-based data storage will then be my primary focus. One of the basic services is cloud storage, which lets people outsource their data to the cloud for its attractive benefits. However, significant security concerns regarding the correctness of the storage arise because the owners no longer have physical possession of the outsourced data. As a result, it becomes crucial and challenging to enable secure storage auditing in the cloud environment using novel methods. In this discussion, I will introduce our new exploration endeavors towards capacity rethinking security in distributed computing and portray both our specialized methodologies and security and execution assessments.

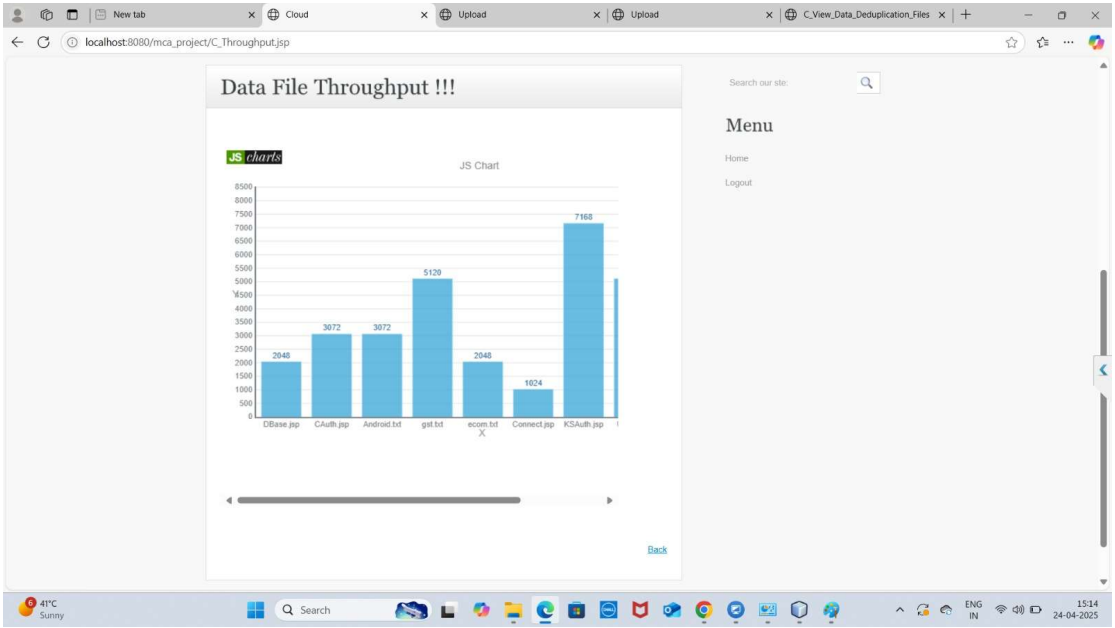
3. "Protection Saving Public Examining for Information Stockpiling Security in Distributed computing," Creators: C. Wang, Q. Wang, K. Ren, and W. Lou. Cloud computing is the long-awaited vision of computing as a utility in which users can remotely store their data in the cloud and use high-quality applications and services that are available on demand from a shared pool of configurable computing resources. Users may be spared the burden of maintaining and storing local data through data outsourcing. In any case, the way that clients never again have actual ownership of the conceivably enormous size of reevaluated information makes the information honesty security in Distributed computing an exceptionally difficult and possibly considerable errand, particularly for clients with compelled processing assets and capacities. As a result, enabling public auditing for cloud data storage security is crucial so that users can rely on an outside auditor to verify the accuracy of outsourced data when necessary. The following two fundamental requirements must be satisfied before introducing an efficient third party auditor (TPA) in a secure manner: 1) TPA ought to be able to effectively audit cloud data storage without requiring a local copy of the data and without putting the cloud user under any additional online burden; 2) The third-party auditing procedure should not introduce any new risks to the privacy of user data. To create a privacy-preserving public cloud data auditing system that satisfies all of the aforementioned requirements, we use and uniquely combine random masking and the public key-based homomorphic authenticator in this paper. To help productive treatment of numerous inspecting errands, we further investigate the strategy of bilinear total mark to expand our primary outcome into a multi-client setting, where TPA can play out various reviewing undertakings all the while. Provably secure and highly effective, the proposed strategies are demonstrated by extensive security and performance analysis.

International Journal of Engineering Technology and Management Science
Website: ijetms.in Issue: 2 Volume No.9 March - April – 2025
DOI:10.46647/ijetms.2025.v09i02.110 ISSN: 2581-4621

Results







DOI:10.46647/ijetms.2025.v09i02.110 ISSN: 2581-4621

Conclusion

In this paper, we propose a Bloom filter-based location selection method and a secure data de duplication scheme with efficient re-encryption. Owing to the inherent property of one-way hash function, our scheme is secure against the stub-reserved attack and guarantees the data privacy of the data owners' sensitive data. In addition, instead of re-encrypting the entire package, data owners are only required to re-encrypt a small part of it through the CAONT, which saves excessive computation over head. We also prove that our scheme can achieve the desired security goals and provide detailed simulation tests. The experimental results show that our scheme is efficient in re-encryption.

References

[1] X. Chen, J. Li, J. Weng, J. Ma, and W. Lou, "Verifiable computation over large database with incremental updates," *IEEE Trans. Computers*, vol. 65, no. 10, pp. 3184–3195, 2016.
[2] M. Gerla, J.Weng, and G. Pau, "Pics-on-wheels: Photo surveillance in the vehicular cloud," *International Conference on Computing, Networking and Communications*, pp. 1123–1127, 2013.
[3] X. Chen, J. Li, J. Ma, Q. Tang, and W. Lou, "New algorithms for secure outsourcing of modular exponentiations," *IEEE Trans. Parallel Distrib. Syst.*, vol. 25, no. 9, pp. 2386–2396, 2014.
[4] H. Yuan, X. Chen, T. Jiang, X. Zhang, Z. Yan, and Y. Xiang, "Dedupdum: Secure and scalable data deduplication with dynamic user management," *Inf. Sci.*, vol. 456, pp. 159–173, 2018.
[5] H. Huang, X. Chen, Q. Wu, X. Huang, and J. Shen, "Bitcoinbased fair payments for outsourcing computations of fog devices," *Future Generation Comp. Syst.*, vol. 78, pp. 850–858, 2018.
[6] IDC. (2014) The digital universe of opportunities : Rich data and the increasing value of the internet of things. [Online]. Available: https://www.emc.com/leadership/digitaluniverse/ 2014iview/index.htm
[7] W. J. Bolosky, S. Corbin, D. Goebel, and J. R. Douceur, "Single instance storage in windows 2000," in *Conference on Usenix Windows Systems Symposium*, 2000.
[8] Dropbox. (2007). [Online]. Available: http://www.dropbox.com
[9] GoogleDrive. (2012). [Online]. Available: http://drive.google.com
[10] Memopal. (2018). [Online]. Available: http://www.memopal.com
[11] Netapp. (2008) Netapp deduplication helps duke institute for genome sciences and policy reduce storage requirements for genomic information by 83 percent. [Online]. Available: http://www.netapp.com
[12] M. Dutch, "Understanding data deduplication ratios," in *SNIA Data Management Forum*, 2008, pp. 1–13.
[13] T. Jiang, X. Chen, J. Li, D. S. Wong, J. Ma, and J. K. Liu, "TIMER: secure and reliable cloud storage against data re-outsourcing," *Information Security Practice and Experience - 10th International Conference*,
pp. 346–358, 2014.
[14] X. Chen, B. Lee, and K. Kim, "Receipt-free electronic auction schemes using homomorphic encryption," *Information Security and Cryptology - ICISC 2003, 6th International Conference, Seoul, Korea, November 27-28, 2003, Revised Papers*, pp. 259–273, 2003.
[15] J. Wang, X. Chen, J. Li, K. Kluczniak, and M. Kutylowski, "Trdup: enhancing secure data deduplication with user traceability in cloud computing," *IJWGS*, vol. 13, no. 3, pp. 270–289, 2017.
[16] X. Zhang, X. Chen, J.Wang, Z. Zhan, and J. Li, "Verifiable privacypreserving single-layer perceptron training scheme in cloud computing," *Soft Comput.*, vol. 22, no. 23, pp. 7719–7732, 2018.
[17] 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 *ICDCS*, 2002, pp. 617–624.
[18] L. D. Stein, "The case for cloud computing in genome informatics," *Genome Biology*, vol. 11, no. 5, pp. 207–207, 2010.
[19] GoogleGenomics. (2018). [Online]. Available: https:// cloud.google.com/genomics/

DOI:10.46647/ijetms.2025.v09i02.110 ISSN: 2581-4621

[20] Amazon. (2018). [Online]. Available: https://aws.amazon.com/
[21] J. Hur, D. Koo, Y. Shin, and K. Kang, "Secure data deduplication with dynamic ownership management in cloud storage," *IEEE Trans. Knowl. Data Eng.*, vol. 28, no. 11, pp. 3113–3125, 2016.
[22] J. Li, X. Chen, X. Huang, S. Tang, Y. Xiang, M. M. Hassan, and A. Alelaiwi, "Secure distributed deduplication systems with improved reliability," *IEEE Trans. Computers*, vol. 64, no. 12, pp.3569–3579, 2015.
[23] T. Jiang, X. Chen, Q. Wu, J. Ma, W. Susilo, and W. Lou, "Secure and efficient cloud data deduplication with randomized tag," *IEEE Trans. Information Forensics and Security*, vol. 12, no. 3, pp. 532–543, 2017.
[24] W. C. G. III, A. Shull, S. Myers, and A. J. Lee, "On the practicality of cryptographically enforcing dynamic access control policies in the cloud," *IEEE Symposium on Security and Privacy*, pp. 819–838, 2016.
[25] J. Li, C. Qin, P. P. C. Lee, and J. Li, "Rekeying for encrypted deduplication storage," in *46th Annual IEEE/IFIP International Conference on Dependable Systems and Networks*, 2016, pp. 618–629.