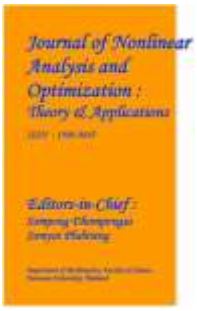
Journal of **N**onlinear **A**nalysis and **O**ptimization
Vol. 16, Issue. 1: 2025
ISSN : **1906-9685**



CLOUD STORAGE OPTIMIZATION: A FRAMEWORK FOR DETECTING AND ELIMINATING DATA REDUNDANCY

B. AMARNATH REDDY[1], CH.VINAY [2]

[1]Assistant Professor, Dept. of MCA, QIS College of Engineering and Technology, Ongole, Andhra Pradesh.

[2]PG Scholar, Dept. of MCA, QIS College of Engineering and Technology, Ongole, Andhra Pradesh.

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.

Index Terms – Data Deduplication, Cloud Computing, Access Control, Storage Management, hybrid cloud.

I. INTRODUCTION

CLOUD storage is a platform to provide large scale data storage and service access at a “pay-as-you-go” fashion. However, a lot of redundant data in cloud storage has seriously wasted and occupied storage resources. Data deduplication is an effective technology to detect and eliminate redundant data. After that, only a single copy of the data is uploaded and stored. Thus, the data deduplication technology can reduce the bandwidth requirement of client-side and improve the space utilization efficiency of server-side. Currently, it has widely used in various cloud computing services to improve user experience and save storage space.

The classic data deduplication scheme and its variants where the framework consists of a key server (KS), a cloud storage provider (CSPs), and users, ensure the security depending on the trusted KS. What is worse, these classic schemes may suffer from the single-point-of-failure and “platform lock-in” issues. If the trusted KS fails, the cloud storage system stops working and data outsourcing protocols cannot be implemented. Recently, a new model of cloud computing, called as Joint- Cloud computing system , has been designed to solve the above-mentioned issues well. The network architecture of JointCloud consists of users and multiple CSPs providing various services. These clouds collaborate together without the trusted KS and the users can connect with any one of them to get computing services. Obviously, JointCloud can provide efficient cross-cloud services and satisfy the requirements of globalized cooperative cloud services by the multilateral collaboration among various clouds. Moreover, it can be built in the decentralized system. JointCloud computing has drawn a lot of attention from both academia and industry.

A massive data breach incidents affecting billions of personal data are far common. Therefore, the outsourced data are usually asked to be encrypted to ensure data confidentiality in cloud storage systems. However, it is difficult to detect and delete the duplicated copies in the ciphertext domain. Because the ciphertexts of the same plaintext encrypted by different users using traditional encryption algorithms are different. In order to implement encrypted deduplication, convergent encryption (CE) and its variants have been proposed. In these schemes, data are encrypted using the keys derived from the data themselves. That is, the secret key is deterministic and the scheme is vulnerable.

Specifically, there are some security vulnerabilities, e.g., 1) the tag reveals the hash value of plaintext, which is vulnerable to the chosen-plaintext attack; the ciphertext dissatisfies the semantic security; 3) the predictable plaintext cannot resist the brute-force attack; 4) users have to bear a great computational burden to protect their data against malicious attackers, etc. The one type of solution improves the security of outsourced data by limiting the manners of data accessibility and availability in the classic system model. That is, they use key management and access control techniques to manage encryption keys and ensure outsourced data access security.

II. LITERATURE SURVEY

A. A survey of indexing techniques for scalable record linkage and deduplication

Record linkage is the process of matching records from several databases that refer to the same entities. When applied on a single database, this process is known as deduplication. Increasingly, matched data are becoming important in many application areas, because they can contain information that is not available otherwise, or that is too costly to acquire. Removing duplicate records in a single database is a crucial step in the data cleaning process, because duplicates can severely influence the outcomes of any subsequent data processing or data mining. With the increasing size of today's databases, the complexity of the matching process becomes one of the major challenges for record linkage and deduplication. In recent years, various indexing techniques have been developed for record linkage and deduplication. They are aimed at reducing the number of record pairs to be compared in the matching process by removing obvious nonmatching pairs, while at the same time maintaining high matching quality. This paper presents a survey of 12 variations of 6 indexing techniques.

B. Coordinate memory deduplication and partition for improving performance in cloud computing

Both limited main memory size and memory interference are considered as the major bottlenecks in virtualization environments. Memory deduplication, detecting pages with same content and being shared into one single copy, reduces memory requirements; memory partition, allocating unique colors for each virtual machine according to page color, reduces memory interference among

virtual machines to improve performance. In this paper, we propose a coordinate memory deduplication and partition approach named CMDP to reduce memory requirement and interference simultaneously for improving performance in virtualization. Moreover, CMDP adopts a lightweight page behavior-based memory deduplication approach named BMD to reduce futile page comparison overhead meanwhile to detect page sharing opportunities efficiently. And a virtual machine based memory partition called VMMP is added into CMDP to reduce interference among virtual machines. According to page color, VMMP allocates unique page colors to applications, virtual machines and hypervisor.

C. The design of fast content-defined chunking for data deduplication based storage systems

Content-Defined Chunking (CDC) has been playing a key role in data deduplication systems recently due to its high redundancy detection ability. However, existing CDC-based approaches introduce heavy CPU overhead because they declare the chunk cut-points by computing and judging the rolling hashes of the data stream byte by byte. In this article, we propose FastCDC, a Fast and efficient Content-Defined Chunking approach, for data deduplication-based storage systems. The key idea behind FastCDC is the combined use of five key techniques, namely, gear based fast rolling hash, simplifying and enhancing the Gear hash judgment, skipping sub-minimum chunk cut-points, normalizing the chunk-size distribution in a small specified region to address the problem of the decreased deduplication ratio stemming from the cut-point skipping, and last but not least, rolling two bytes each time to further speed up CDC.

III. PROPOSED SYSTEM

The overview of our proposed system is shown in the below figure.

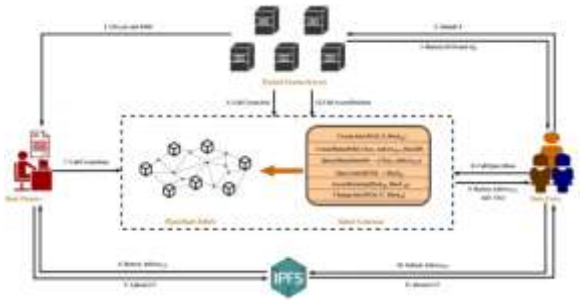


Fig. 1: System Overview

Implementation Modules

Cloud service provider (CSPs).

- It is an entity that has significant resources and enough computational power to provide distributed cloud computing services and execute protocols.

Owner

- The owner who has the original data. Especially, the owner first uploading the data into CSPs is called as initial uploader and the users subsequently uploading the same data into CSPs are called as subsequent uploader

Users

- The non- owner who does not possess original data but wants to get the plaintext of the outsourced data from CSPs

IV. RESULTS

Fig.2: Cloud Login

Fig.3: Data Owner Login

Fig.4: File Rank

Fig.5: File Upload Time Delays

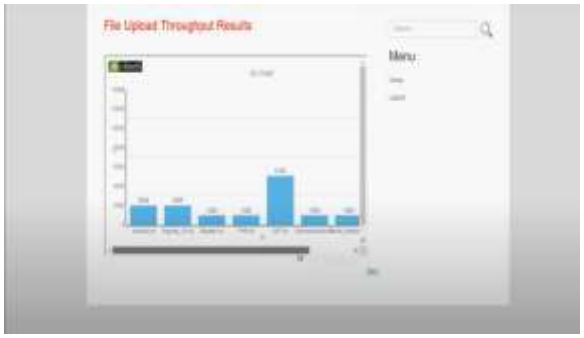


Fig.6: File Upload Time Delays Results

V. CONCLUSION

In this project, we have designed a secure and efficient scheme SED for data deduplication without the help of the trusted KS. The proposed SED has reduced the communication and computation overhead of client-side and improved efficiency based

on the CDH problem in the JointCloud storage system. Its concise algorithms of encrypting and generating tag satisfy the semantic security and the tag consistency (including security and validity), respectively. Moreover, SED improves the scalability and solves the single-point-of-failure of KS in the classic cloud storage system. SED has strong capacity against typical attacks such as the brute-force attack and the collusion between malicious CSPs and unauthorized users. Besides, SED supports dynamic supports data operations, including deletion, modification, and sharing, which improves the functionality and usability. To the best of our knowledge, SED is the first scheme considering the case that data owner shares his/her outsourced data to the permitted users. According to the theoretical and experimental analyses, our SED is secure and has low computation, communication, and storage complexity. From the comparison with the previous scheme, our SED is more secure, efficient, and functional.

REFERENCES

[1] P. Christen, “A survey of indexing techniques for scalable record linkage and deduplication,” IEEE Transactions on Knowledge and Data Engineering, vol. 24, no. 9, pp. 1537–1555, 2012.

[2] G. Jia, G. Han, J. J. P. C. Rodrigues, J. Lloret, andW. Li, “Coordinate memory deduplication and partition for improving performance in cloud computing,” IEEE Transactions on Cloud Computing, vol. 7, no. 2, pp. 357–368, 2019.

[3] M. A. Shaik, A. Fatima, M. Parveen, A. Soumya Rani, A. Mohammad and A. Rahim, "Dual-Model Approach for Lung Disease Classification Using Convolutional Neural Networks and Support Vector Machines," 2024 International Conference on Integrated Intelligence and Communication Systems (ICIICS), Kalaburagi, India, 2024, pp. 1-6, doi: 10.1109/ICIICS63763.2024.10860090.

[4] J. Li, J. Li, D. Xie, and Z. Cai, “Secure auditing and deduplicating data in cloud,” IEEE Transactions on Computers, vol. 65, no. 8, pp. 2386–2396, 2016.

[5] L. Liu, Y. Zhang, and X. Li, “Keyd: Secure key-deduplication with identity-based broadcast encryption,” IEEE Transactions on Cloud Computing, pp. 1–1, 2018.

[6] M. Ali Shaik, P. Sairam, P. Rahul, T. Sadhvik and P. Nithin, "Titanic Survival Predictor Using Machine Learning Algorithms," 2024 4th Asian Conference on Innovation in Technology (ASIANCON), Pimari Chinchwad, India, 2024, pp. 1-6, doi: 10.1109/ASIANCON62057.2024.10837935.

[7] Y. Zheng, X. Yuan, X.Wang, J. Jiang, C.Wang, and X. Gui, "Toward encrypted cloud media center with secure deduplication," IEEE Transactions on Multimedia, vol. 19, no. 2, pp. 251–265, 2017.

[8] H. Wang, P. Shi, and Y. Zhang, "Jointcloud: A cross-cloud cooperation architecture for integrated internet service customization," in 2017 IEEE 37th International Conference on Distributed Computing Systems, 2017, pp. 1846–1855.

[9] M. A. Shaik and E. Ravithreyini, "Enhanced BreastNet Architecture and Comparison with State-of-the-Art Models," 2024 5th International Conference on Data Intelligence and Cognitive Informatics (ICDICI), Tirunelveli, India, 2024, pp. 604-608, doi: 10.1109/ICDICI62993.2024.10810949.

[10] M. Bellare, S. Keelveedhi, and T. Ristenpart, "Message-locked encryption and secure deduplication," in International Conference on the Theory and Applications of Cryptographic Techniques, 2013, pp. 296–312.

AUTHORS Profile

Mr. B. Amarnath Reddy is an Assistant Professor in the Department of Master of Computer Applications at QIS College of Engineering and Technology, Ongole, Andhra Pradesh. He earned his M.Tech from Vellore Institute of Technology (VIT), Vellore. His research interests include Machine Learning, Programming Languages. He is committed to advancing research and fostering innovation while mentoring students to excel in both academic and professional pursuits.

Mr. Ch. Vinay has received his Degree in B.Sc Chemistry from Acharya Nagarjuna University 2022 and pursuing MCA degree in Computer Science at Qis College of Engineering and Technology affiliated to JNTUK in 2023-2025.