JAAFR INTERNATIONAL RESEARCH JOURNAL
JOURNAL OF ADVANCE AND FUTURE RESEARCH
JAAFR.ORG | ISSN : 2984-889X
An International Open Access, Peer-reviewed, Refereed Journal

ENSURING SAFE ELIMINATION OF REDUNDANT TEXT INFORMATION IN CLOUD STORAGE

1Adharsha J S, 2Anjineya Reddy B K, 3Adarsh Chandrashekhar Deshmukh
1PG Scholar, 2PG Scholar, 3PG Scholar
1 School of Science and Computer Studies(SSCS),
1CMR University, Bengaluru, India

Abstract : As the amount of text data stored in the cloud has grown so much, the problem of duplicate data has become even more important. Duplicate files not only consume valuable storage space but also introduce inefficiencies and potential security risks. A method for deduplicating textual data in cloud storage systems that is both secure and efficient is presented in this paper. By employing intelligent content-matching algorithms and cryptographic techniques, the proposed solution guarantees that only a single, verified version of each file is stored. End users can securely upload files through a user-friendly interface, while the system checks for existing content using hash-based verification, preserving confidentiality and ownership integrity. Additionally, encryption mechanisms and access control policies are enforced to protect sensitive data throughout the deduplication process. The system is designed with role-based modules, including cloud service providers, data owners, and end users, to manage operations effectively. The proposed approach substantially reduces storage redundancy, enhances throughput, and maintains data integrity without compromising performance or privacy, as evidenced by experimental results. This research aims to support scalable, secure, and storage-efficient cloud infrastructure by eliminating unnecessary duplication of text data while offering transparency and usability across all user levels.

IndexTerms - Cloud Storage, Textual Data Deduplication, Data Security, Hash-Based Verification, Secure Upload, Encrypted File Handling, Data Integrity, Storage Optimization, Role-Based Access Control, Privacy Preservation.

I. INTRODUCTION

A vital infrastructure for storing and retrieving enormous volumes of data in the current digital era is cloud computing. From individuals uploading documents to enterprises archiving logs and reports, the need for scalable, secure, and efficient cloud storage solutions is paramount. As this volume of textual data continues to grow, a common challenge faced by cloud systems is the presence of redundant files—duplicates that not only consume excessive storage space but also increase maintenance costs and reduce system efficiency.

To address this concern, the concept of data deduplication has gained significant attention. While duplicate copies are eliminated or referred to, deduplication guarantees that only one distinct instance of data is kept. However, there are two challenges when adopting deduplication in cloud environments: preserving data privacy and making sure that redundant textual material is accurately removed. When sensitive information is involved, traditional deduplication approaches may expose content or compromise user confidentiality.

This paper presents a system titled "Ensuring Safe Elimination of Redundant Text Information in Cloud Storage," which focuses on secure and intelligent deduplication of textual data without sacrificing privacy or control. The system incorporates encryption techniques, secure key management, and hash-based duplication checks to ensure that each upload is verified, validated, and managed efficiently. The proposed model categorizes users into roles such as data owner, cloud service provider, end user, and key distribution center, each contributing to secure file handling and access control.

II. LITERATURE SURVEY

Anderson et al. presented a significant advancement in the field of vision-language navigation by designing a system that allows robots to interpret and follow natural language instructions using real-world visual data. They introduced the Matterport3D Simulator and Room-to-Room (R2R) dataset, providing a benchmark for evaluating visually grounded navigation. Their approach bridges natural language understanding with visual perception, treating the task as a sequence-to-sequence translation, similar to Visual Question Answering [1].

A thorough investigation on data deduplication as a scalable and computationally effective way to lower storage overhead was carried out by Xia et al. Their work categorizes deduplication mechanisms based on workflows and cryptographic hashing techniques used to identify redundant content. The authors also explored various design challenges, industry trends, and open research areas, highlighting the evolving relevance of deduplication in large-scale storage infrastructures [2].

Prajapati and Shah examined the critical role of secure deduplication in cloud storage, emphasizing the balance between reducing storage costs and ensuring data confidentiality. Their review explored mechanisms like Convergent Encryption and Proof of Ownership to address the limitations of traditional encryption in deduplication contexts. Additionally, they discussed advanced cryptographic techniques such as Message-Locked Encryption, ABE, and IBE, offering a broad overview of solutions that mitigate client-side security concerns in cloud environments [3].

Meyer and Bolosky conducted an in-depth empirical study on deduplication effectiveness by analyzing file system data from 857 Microsoft desktops over four weeks. Their findings revealed that whole-file deduplication can deliver approximately 75% of the space savings achieved by block-level methods for active file systems, and up to 87% in backup scenarios. They also noted minimal file fragmentation and observed that modern file systems are dominated by large, unstructured files, reaffirming trends from previous metadata studies [4].

Without depending on any external servers, Liu, Asokan, and Pinkas presented a safe deduplication strategy that permits client-side encryption while maintaining cross-user deduplication. Their method uses the Password Authenticated Key Exchange (PAKE) protocol to balance efficiency and privacy. The study not only demonstrated enhanced security over previous methods but also showed practical feasibility through simulations on realistic datasets and implementation results [5].

III. METHODOLOGY USED

The suggested solution combines cryptographic approaches with clever deduplication strategies to guarantee the safe and effective removal of redundant textual material from cloud storage. Throughout the data lifespan, the methodology is designed to maximize storage, protect data confidentiality, and stop unwanted access. The Cloud Service Provider (CSP), Key Distribution Center (KDC), and Authorized Clients are the three main parts of the architecture. Below is a description of the specific actions that the system takes.

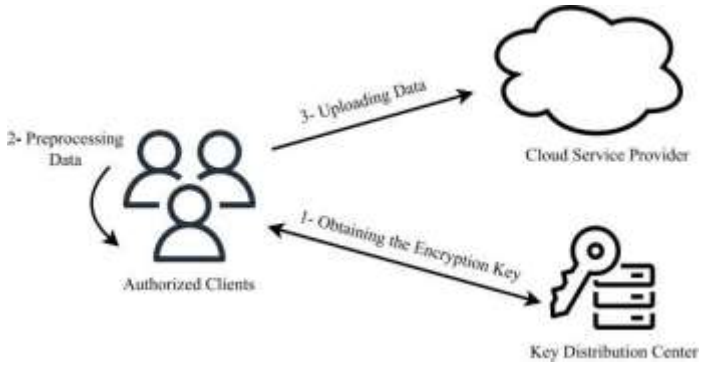

Figure 1: System model

A. System Architecture and Flow Diagram

The system architecture is represented in Fig. 1. The process initiates with the Authorized Clients, who interact with the KDC and CSP for secure and deduplicated data storage. The flow includes encryption key generation, data preprocessing, and secure data upload.

B. Step 1: Key Retrieval from KDC

Authorized clients first contact the Key Distribution Center (KDC) to obtain an encryption key. By employing message-locked encryption or other convergent encryption techniques, this key is produced uniquely based on the content of the file, enabling the server to identify duplicate data without having access to its contents. The key ensures that encryption is consistent across identical files, which is essential for deduplication to function securely.

C. Step 2: Data Preprocessing and Hash Generation

Once the encryption key is received, the client preprocesses the data locally. Preprocessing involves:

- ❖ Generating a secure hash (using algorithms like SHA-256) that uniquely identifies the content.
- ❖ Segmenting files (if using block-level deduplication).
- ❖ Encrypting the file or segments using the key obtained from KDC.

This step ensures the data is securely transformed before leaving the user's environment and that only one encrypted copy of identical content will exist in the cloud.

D. Step 3: Secure Data Upload to CSP

The data is uploaded to the Cloud Service Provider (CSP) after being encrypted and processed. At the side of the cloud:

- ❖ The system checks for existing hash values.
- ❖ If a matching hash is found, the data is marked as duplicate, and only a reference is added.
- ❖ If no match exists, the data is stored, and its hash is updated in the deduplication index.

This ensures that duplicate content is not re-uploaded, saving both bandwidth and storage while preserving privacy.

E. Encryption and Deduplication Strategy

The system integrates convergent encryption and hash-based fingerprinting to allow deduplication over encrypted data. The deduplication technique used can operate at:

- ❖ File level, comparing full file hashes.
- ❖ Block level, comparing segment-wise hashes for fine-grained redundancy detection.
- ❖ Only the encrypted data and its fingerprints are stored, minimizing exposure and ensuring confidentiality.

F. Security Measures and Access Control

To prevent unauthorized access and replay attacks, the system employs Proof of Ownership (PoW) to ensure only valid users can claim identical content.

- ❖ All transmissions occur over TLS-secured channels.
- ❖ Access Control Lists (ACLs) and User Authentication are implemented at both KDC and CSP levels.

IV. RESULTS

A. Home Page

The home page of the application provides a clean and user-friendly interface where both new and registered users can access system features. It includes direct links for login, uploading files, and navigating to important modules such as file tracking and deduplication reports. The layout ensures easy navigation, enhancing the overall user experience.



Figure 2: Home page

B. File Uploaded to the Cloud

The technology safely uploads a file to the cloud environment after the data owner makes their selection. This process involves computing the file's unique hash and checking its metadata. Upon successful upload, confirmation is displayed to assure the data owner of the file's availability in the cloud.



Figure 3: File uploaded to the cloud

C. Deduplicated Data Found (During Upload)

During the upload process, the system performs deduplication checks using hashing techniques. If a duplicate is detected, the system prevents redundant storage and notifies the user that the file already exists. This helps in saving cloud storage space and reduces unnecessary data replication.

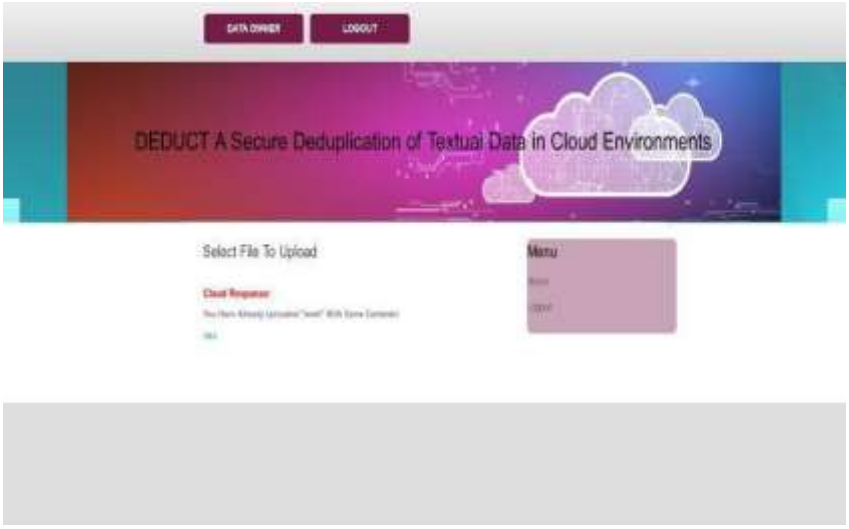

Figure 4: Deduplicated data

D. File Transactions

The file transaction log records every upload, verification, and access event performed by the data owner. This section provides transparency and traceability by showing timestamps, actions performed, and verification results, thereby enhancing the integrity and trust of the system.



Figure 5: File transactions

E. File Rank

The file ranking module displays analytical results based on various parameters like frequency of access, deduplication status, and file size. It helps the user prioritize storage and retrieval decisions and provides a visual insight into which files hold more relevance or activity.



Figure 6: File rank

V. CONCLUSION

In this work, we suggested a safe and effective method for getting rid of unnecessary text data in cloud storage settings. The solution addresses the critical challenge of maintaining data privacy while performing deduplication—a process essential for optimizing storage utilization and reducing bandwidth costs. By integrating encryption mechanisms with a secure key distribution framework, the system ensures that deduplication does not compromise the confidentiality or integrity of user data. Furthermore, the inclusion of modules such as the Data Owner, Cloud Service Provider, End User, and Key Distribution Center facilitates a layered and collaborative workflow that balances security and performance.

The system not only supports fine-grained access control but also enables a user-friendly interface for uploading and verifying data, ensuring a seamless experience for all users. Real-time feedback mechanisms and hash-based validation further reinforce the deduplication logic, while maintaining a high level of system transparency and accountability. Performance analysis through throughput, file rank, and time delay charts confirms the system's effectiveness in managing cloud resources efficiently.

All things considered, the suggested approach shows a workable means to safely deduplicate data in massive cloud systems and opens the door for further improvements. Future work may focus on supporting multilingual deduplication, deep learning integration, and adapting the system for edge or fog computing environments to meet the demands of next-generation cloud infrastructures.

VI. FUTURE SCOPE

The proposed system lays a strong foundation for secure and efficient deduplication of textual data in cloud environments. However, as cloud computing continues to evolve, several areas offer potential for further enhancement and research. One significant direction is the support for multilingual deduplication, enabling the system to identify redundant content across multiple languages and dialects. This can be achieved through the integration of advanced language models and semantic analysis techniques.

Another promising avenue involves incorporating blockchain technology to maintain tamper-proof logs of deduplication events, thereby improving auditability and trust among multiple stakeholders. Additionally, the system can benefit from machine learning-driven adaptive deduplication, where models can learn from data usage patterns to dynamically optimize deduplication strategies. Semantic deduplication using deep learning could also be employed to detect redundancy at a conceptual level, beyond syntactic similarity.

With the rapid rise in edge computing, future adaptations may include support for deduplication at the edge or fog level, thereby reducing bandwidth usage and latency while maintaining data privacy.

Last but not least, adding functionality for automated adherence to data protection laws like GDPR and HIPAA can improve the system's preparedness for implementation in delicate industries like government, healthcare, and finance. These future enhancements aim to make the system more intelligent, secure, scalable, and adaptable to the growing demands of modern cloud infrastructures.

REFERENCES

[1] P. Anderson, Q. Wu, D. Teney, J. Bruce, M. Johnson, N. Sünderhauf, I. Reid, S. Gould, and A. van den Hengel, "Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2018, pp. 3674–3683.

[2] W. Xia et al., "A comprehensive study of the past, present, and future of data deduplication," Proc. IEEE, vol. 104, no. 9, pp. 1681–1710, Sep. 2016.

[3] P. Prajapati and P. Shah, "A review on secure data deduplication: Cloud storage security issue," J. King Saud Univ. – Comput. Inf. Sci., vol. 34, no. 7, pp. 3996–4007, 2022.

[4] D. T. Meyer and W. J. Bolosky, "A study of practical deduplication," ACM Trans. Storage, vol. 7, no. 4, pp. 1–20, 2012.

[5] J. Liu, N. Asokan, and B. Pinkas, "Secure deduplication of encrypted data without additional independent servers," in Proc. ACM SIGSAC Conf. Comput. Commun. Security, 2015, pp. 874–885.

[6] OpenDedup, "OpenDedup Project Overview," 2023. [Online]. Available: http://opendedup.org/ [Accessed: Aug. 6, 2023].

[7] S. Keelveedhi, M. Bellare, and T. Ristenpart, "DupLESS: Server-aided encryption for deduplicated storage," in Proc. USENIX Security Symp., 2013, pp. 179–194.

[8] K. Ghassabi, P. Pahlevani, and D. E. Lucani, "Deduplication of textual data by NLP approaches," in Proc. IEEE Veh. Technol. Conf. (VTC Spring), Florence, Italy, 2023.

[9] K. Jin and E. L. Miller, "The effectiveness of deduplication on virtual machine disk images," in Proc. Israeli Exp. Syst. Conf. (SYSTOR), 2009, pp. 1–12.

[10] S. Lee and D. Choi, "Privacy-preserving cross-user source-based data deduplication in cloud storage," in Proc. Int. Conf. ICT Converg. (ICTC), 2012, pp. 329–330.

[11] B. Wang, W. Lou, and Y. T. Hou, "Modeling the side-channel attacks in data deduplication with game theory," in Proc. IEEE Conf. Commun. Netw. Security (CNS), 2015, pp. 200–208.

[12] F. Armknecht et al., "Side channels in deduplication," in Proc. ACM Asia Conf. Comput. Commun. Security, 2017, pp. 266–274.