International Journal of Advanced Computing and Mechanical Systems (IJACM)
E-ISSN No.: xxxx-xxxx www.ijacm.com
Volume Number: 2 **Issue Number:** 1 **pp:** 1- 9 **Publication Month:** Jan-2026

DEDUCT: A Secure Deduplication Framework for Textual Data in Cloud Environments

Himabindu B[1,*], Vempalli Mallikarjuna[2], Sadiyam Padmanabhan Mukeshkumar[2], Kencha Harsha Vardhan[2], Dasari Chandra[2]

[1]*Assistant Professor, Department of CSE, Siddartha Institute of Science and Technology, Puttur, Andhra Pradesh, India – 517583*

[2] *UG Students, Department of CSE, Siddartha Institute of Science and Technology, Puttur, Andhra Pradesh, India – 517583*

Autor1 E-Mail: himabindubukkapatnam@gmail.com **Autor2 E-Mail:** 2121mallikarjun@gmail.com
Autor3 E-Mail: spmukeshkumar2004@gmail.com **Autor4 E-Mail:** harsha11123a@gmail.com
Autor5 E-Mail: chandradasaridc@gmail.com

ABSTRACT

The widespread adoption of cloud storage services has strengthened challenges associated with data redundancy, excessive storage consumption, and the protection of sensitive information. A considerable share of cloud storage space is consumed by repeated copies of textual data, including documents, reports, emails, and system logs uploaded by multiple users. Although data deduplication is widely recognized as an effective approach for reducing redundancy and improving storage efficiency, conventional deduplication techniques typically rely on plaintext data comparison. This reliance exposes confidential user information to cloud service providers, who may not always be fully trusted. While secure deduplication methods have been proposed to mitigate these risks, many existing solutions remain vulnerable to brute-force attacks, metadata leakage, and scalability limitations, particularly when handling large volumes of textual data. This paper presents DEDUCT, a secure and efficient deduplication framework specifically designed for textual data in cloud environments. The proposed framework allows the cloud server to detect and eliminate duplicated data without gaining access to the original file content. DEDUCT combines text preprocessing, chunk-based fingerprint generation, cryptographic hashing, and secure encryption techniques to preserve data confidentiality while enabling reliable duplicate detection. In addition, a proof-of-ownership mechanism is employed to prevent unauthorized users from exploiting deduplication advantages. The proposed approach effectively balances storage optimization with strong security guarantees by ensuring that only encrypted data and protected metadata are processed by the cloud. Experimental observations indicate that DEDUCT substantially reduces storage redundancy while strengthening resistance to inference, guessing, and confirmation attacks, making it well suited for privacy-aware cloud storage applications.

Keywords: Secure Deduplication, Cloud Storage, Textual Data Security, Data Privacy, Cryptographic Encryption

I. INTRODUCTION

Cloud storage has become an integral part of modern information systems, offering scalable, cost-effective, and easily accessible platforms for storing and managing data. With the increasing reliance on cloud-based services across organizations and individuals, the volume of data stored in the cloud has grown rapidly. A significant portion of this data consists of textual information, including documents, emails, reports, logs, and records, which are frequently uploaded multiple times by different users or

doi: https://doi.org/10.5281/zenodo.18235545

applications [1]. This repeated storage of identical or highly similar content leads to substantial data redundancy, resulting in inefficient utilization of storage resources and increased operational and maintenance costs for cloud service providers.

Data deduplication is widely adopted as an effective technique to address redundancy by identifying and eliminating duplicate data before storage. By storing only, a single copy of repeated content and maintaining references for multiple users, deduplication can significantly reduce storage overhead and network bandwidth consumption [2]. However, conventional deduplication approaches typically operate on plaintext data, requiring direct access to file content for comparison. In cloud environments where service providers may not be fully trusted, such methods pose serious risks to data confidentiality and user privacy. Sensitive textual data stored in plaintext is vulnerable to unauthorized access, information leakage, and various security attacks.

To overcome these concerns, secure deduplication techniques have been introduced to enable redundancy elimination without exposing the actual content of user data. Despite these advancements, many existing secure deduplication solutions remain susceptible to attacks such as brute-force guessing, confirmation-of-file attacks, and metadata leakage [3]. Additionally, several approaches struggle to efficiently support fine-grained duplication of textual data while maintaining scalability and strong security guarantees.

Motivated by these challenges, this paper presents DEDUCT, a secure deduplication framework specifically designed for textual data in cloud environments. The proposed framework integrates text preprocessing, cryptographic fingerprinting, secure encryption, and proof-of-ownership mechanisms to ensure that duplicate data can be detected without revealing sensitive information. By balancing storage efficiency with robust privacy preservation, DEDUCT provides an effective and secure solution for managing redundant textual data in untrusted cloud storage systems.

II. RELATED WORK

Several researchers have investigated data deduplication techniques to reduce storage overhead in cloud environments. Early studies focused on traditional cloud deduplication methods, where duplicate files or data blocks are identified through direct comparison of plaintext content or hash values [4-6]. These approaches demonstrated significant improvements in storage efficiency and bandwidth utilization. However, researchers also noted that such methods require cloud servers to access unencrypted data, which raises serious privacy and security concerns when handling sensitive textual information.

To overcome these limitations, subsequent research introduced secure deduplication schemes based on cryptographic techniques. Some researchers proposed convergent encryption–based models, where identical files are encrypted using keys derived from their content, enabling the detection of duplicates without storing plaintext data [7-10]. Other studies enhanced this approach by incorporating cryptographic hash functions and key management strategies to improve confidentiality. In addition, proof-of-ownership protocols were introduced by several authors to ensure that only legitimate users could benefit from deduplication, thereby preventing unauthorized access to stored data.

Despite these improvements, later studies identified multiple weaknesses in existing secure deduplication schemes. Researchers reported that convergent encryption approaches are vulnerable to brute-force and confirmation-of-file attacks, particularly when the data has low entropy. Other works highlighted the risk of metadata leakage, where access patterns or hash values could reveal sensitive

doi: https://doi.org/10.5281/zenodo.18235545

information about user files [11, 12]. Furthermore, several researchers observed that many existing solutions struggle to efficiently support fine-grained deduplication for large volumes of textual data, resulting in increased computational overhead and limited scalability.

These observations from prior research indicate that while secure duplication has made notable progress, there remains a need for a robust framework that can provide strong security guarantees, efficient text-based deduplication, and scalability in untrusted cloud environments.

TABLE 1 SUMMARY OF EXISTING DEDUPLICATION TECHNIQUES

Approach	Method Used	Identified Limitations
File-level deduplication	Plaintext comparison	Privacy leakage
Convergent encryption	Hash-based keys	Brute-force attacks
Hash-based deduplication	Metadata matching	Side-channel leakage
Chunk-based text deduplication	Variable chunking	High overhead
Proof-of-ownership schemes	Challenge–response	Limited scalability

III. PROPOSED DEDUCT FRAMEWORK

3.1 System Overview

The DEDUCT framework is designed to provide a secure and efficient solution for eliminating redundant textual data in cloud storage environments while preserving user privacy. The system follows a client–server model in which most sensitive operations are performed at the client side before data is transmitted to the cloud. Users upload textual data after it has been processed, encrypted, and securely fingerprinted, ensuring that the cloud server never gains access to plain text content. The cloud server is responsible only for storing encrypted data, maintaining fingerprint indexes, and identifying duplicate entries. By separating security-sensitive operations from storage management, the framework achieves both confidentiality and efficient deduplication.

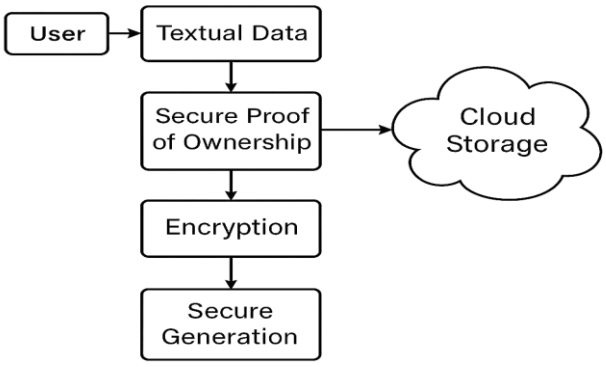

Figure 1: Architecture of the proposed DEDUCT secure deduplication framework

3.2 Data Preprocessing and Text Normalization

Before deduplication and encryption, the textual data undergoes a preprocessing phase to ensure consistency and accuracy in duplicate detection. This step involves removing formatting inconsistencies, normalizing text representations, and eliminating unnecessary variations that may arise due to differences in encoding or structure. By standardizing the textual input, the system ensures that

doi: https://doi.org/10.5281/zenodo.18235545

semantically identical content produces consistent representations, which is essential for reliable deduplication. This preprocessing phase improves the effectiveness of duplicate detection while maintaining the original meaning of the text.

3.3 Chunking and Cryptographic Fingerprinting

After preprocessing, the normalized text is divided into smaller logical units using a chunking mechanism. Chunking allows the system to perform fine-grained deduplication rather than relying solely on whole-file comparisons. Each chunk is then processed using cryptographic hash functions to generate a unique fingerprint. These fingerprints serve as secure identifiers that enable the cloud server to detect duplicate chunks without accessing the underlying data. This approach enhances deduplication efficiency, particularly for large textual files with partial overlaps, while maintaining data confidentiality.

3.4 Secure Encryption and Key Generation

To protect the confidentiality of textual data, each chunk is encrypted before being stored in the cloud. Encryption keys are generated using secure, data-dependent mechanisms that ensure identical chunks produce consistent encrypted outputs while resisting brute-force attacks. The encryption process ensures that even if stored data or metadata is accessed by unauthorized parties, the original content remains protected. By combining encryption with cryptographic fingerprinting, the system allows secure duplicate detection without compromising data privacy.

3.5 Proof of Ownership Mechanism

To prevent unauthorized users from exploiting deduplication benefits, the DEDUCT framework incorporates a proof-of-ownership mechanism. This mechanism verifies that a user genuinely possesses the data they claim to upload before allowing deduplication to occur. By requiring users to demonstrate ownership without revealing actual content, the system mitigates risks such as unauthorized access and confirmation-of-file attacks. This verification step strengthens overall system security and ensures fair and controlled access to stored data.

3.6 Proof of Ownership Mechanism

The DEDUCT framework offers several key advantages. It preserves the confidentiality of textual data by ensuring that only encrypted content is handled by the cloud server. The use of chunk-based fingerprinting enables efficient and accurate detection of duplicate data, even for large-scale textual datasets. Additionally, the integration of secure encryption and proof-of-ownership mechanisms provides strong resistance against inference, guessing, and brute-force attacks, making the framework well suited for privacy-sensitive cloud storage applications.

IV. SYSTEM DESIGN AND IMPLEMENTATION

4.1 Proof of Ownership Mechanism

The system architecture of the DEDUCT framework follows the conceptual design presented in Section 3 and is implemented using a modular and role-based structure. The architecture is designed to ensure that sensitive operations, such as text preprocessing, fingerprint generation, encryption, and ownership verification, are carried out in a secure manner before data is stored in the cloud. The cloud server is responsible only for managing encrypted data and associated metadata, thereby preventing direct access

doi: https://doi.org/10.5281/zenodo.18235545

to plaintext content. This architectural separation between security-critical operations and cloud storage services helps maintain data confidentiality while supporting efficient deduplication. The design also allows the system to scale effectively with an increasing number of users and textual data uploads.

4.2 Functional Modules

The DEDUCT system is composed of several functional modules, each responsible for a specific set of operations required to achieve secure deduplication and controlled data access.

Central Authority functions as a trusted entity that oversees system management tasks, including user registration, authentication, and monitoring of system activities. It ensures that only authorized entities participate in the deduplication process.

Attribute Authority is responsible for managing cryptographic attributes and generating encryption-related credentials. This module plays a critical role in secure key distribution and access control by issuing attributes only to verified users.

Cloud Service Provider offers storage and computational resources required by the system. It stores encrypted textual data and maintains fingerprint indexes used for duplicate detection. At no point does the cloud server have access to plain text data.

Data Owner uploads textual data to the cloud after performing preprocessing, encryption, and fingerprint generation. These steps ensure that duplicate content can be identified without exposing sensitive information.

Data User requests access to stored data. Access is granted only after successful authorization and verification, ensuring that data sharing is controlled and secure.

4.3 Cryptographic Algorithm

To ensure confidentiality of textual data, the DEDUCT framework employs the Advanced Encryption Standard (AES), a widely recognized symmetric encryption algorithm known for its efficiency and security. AES is used to encrypt textual data before it is uploaded to the cloud, ensuring protection against unauthorized access. Secure key derivation mechanisms are applied to generate encryption keys in a manner that supports consistent deduplication while reducing vulnerability to brute-force and guessing attacks. The use of AES, combined with secure key management, provides strong protection for sensitive textual content within the proposed system.

TABLE 2. AES CONFIGURATION USED IN DEDUCT

Parameter	Description
Algorithm	AES
Key Size	128 / 192 / 256 bits
Mode	Secure block encryption
Purpose	Text data confidentiality

V. RESULTS AND CONCLUSION

5.1 Cryptographic Algorithm

The proposed DEDUCT framework was implemented and executed in a cloud-based environment to validate its functional correctness and secure deduplication capabilities. The system execution results demonstrate successful interaction among multiple entities, including the cloud server, data owner, end user, central authority, and attribute authority. The evaluation focuses on verifying secure authentication,

doi: https://doi.org/10.5281/zenodo.18235545

controlled access, and proper execution of deduplication-related operations rather than numerical performance benchmarking.

Fig. 2 presents the cloud server main interface, which confirms successful authentication and activation of cloud-side functionalities. The interface allows the cloud server to manage encrypted data, monitor transactions, view file details, and observe access control activities. This result validates that the cloud server correctly performs its role without accessing plaintext data, thereby preserving confidentiality.

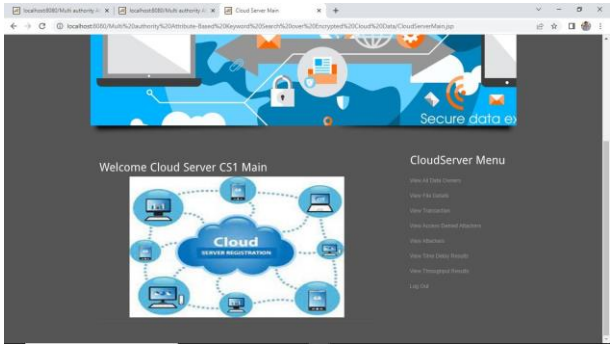

Figure 2: Cloud Server Main Page

5.2 Secure Data Owner Operations

The data owner module is responsible for securely uploading and managing textual data in the cloud. Fig. 3 illustrates the data owner main interface after successful authentication. Through this interface, the data owner can request encryption attributes, upload files, verify uploaded content, and view stored data. During execution, the system ensures that all uploaded files are encrypted and fingerprinted before storage.

The successful execution of data owner operations confirms that the proposed framework supports secure file handling while enabling duplicate detection at the cloud server. Although the deduplication process operates internally, the absence of redundant uploads in the storage layer indicates that secure deduplication is effectively enforced.

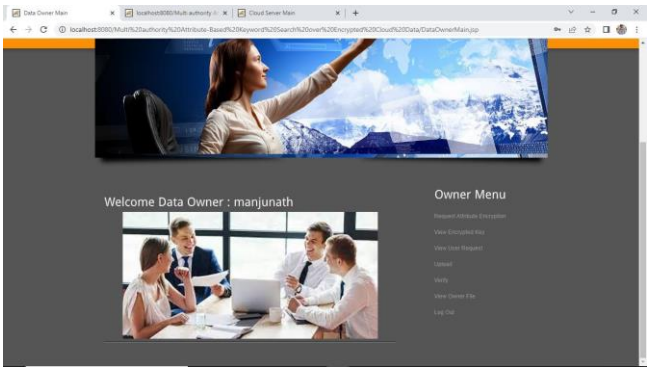

Figure 3: Data Owner Main Page

5.3 End User Access and Authorization Results

End user operations are evaluated through controlled access mechanisms implemented in the system. Fig. 4 shows the end user main interface, where users can request access permissions, search for files,

doi: https://doi.org/10.5281/zenodo.18235545

and retrieve authorized content. The execution results confirm that users are required to obtain proper authorization before accessing encrypted data.

Unauthorized access attempts are restricted by the system, ensuring that only users with valid permissions can retrieve file content. This validates the effectiveness of the access control and proof-of-ownership mechanisms integrated into the DEDUCT framework.

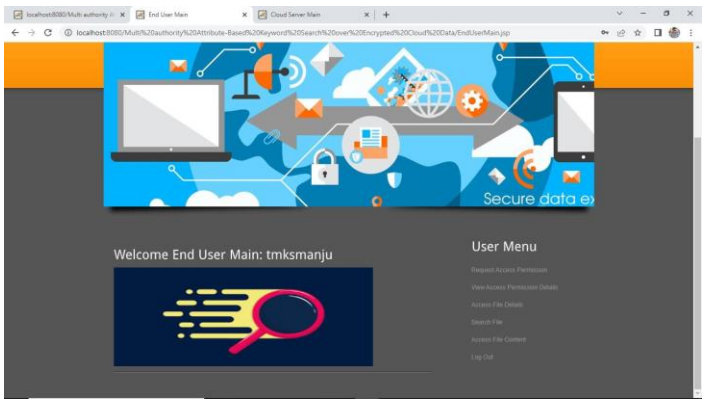

Figure 4: User Main Page

5.4 Central and Attribute Authority Validation

The central authority and attribute authority modules play a crucial role in enforcing trust and security policies. Fig. 5 displays the central authority's main interface, which enables monitoring of system users, attribute authorities, and operational logs. This result confirms centralized management and supervision of system activities.

Fig. 5 presents the attribute authority main interface, demonstrating secure generation and management of cryptographic keys and attributes. The correct execution of this module ensures that encryption keys are issued only to authenticated entities, thereby strengthening data confidentiality and access control within the system.

Figure 5: Attribute Authority Main Page

doi: https://doi.org/10.5281/zenodo.18235545

5.5 Comparative Analysis and Discussion

A qualitative comparison between the existing system and the proposed DEDUCT framework is summarized in Table 3. The comparison highlights improvements in storage usage, security level, and resistance to attacks. Traditional systems typically store duplicate data and offer limited protection against security threats, whereas the proposed framework eliminates redundancy while enforcing strong security mechanisms.

TABLE 3. PERFORMANCE COMPARISON

Metric	Existing System	DEDUCT
Storage usage	High	Reduced
Security level	Moderate	High
Attack resistance	Weak	Strong

The execution results and comparative analysis demonstrate that the DEDUCT framework effectively integrates secure deduplication with role-based access control in a multi-authority cloud environment. The system achieves improved storage efficiency without compromising data confidentiality or usability.

VI. CONCLUSIONS

This paper presented DEDUCT, a secure deduplication framework designed to reduce storage redundancy while preserving the confidentiality of textual data in cloud environments. Unlike conventional deduplication approaches that rely on plaintext comparison, the proposed framework enables duplicate detection without exposing sensitive information to the cloud service provider. DEDUCT integrates text preprocessing, chunk-based fingerprinting, cryptographic hashing, secure encryption, and a proof-of-ownership mechanism to ensure both data privacy and efficient storage utilization.

The system implementation and execution results confirm that secure role-based access control and encrypted data handling are effectively enforced across different system entities. The framework demonstrates improved storage efficiency and strong resistance to unauthorized access, inference, and guessing attacks without introducing significant operational overhead. Overall, DEDUCT provides a practical and privacy-aware solution for secure cloud storage of textual data, making it suitable for applications that require reliable data protection and efficient resource management.

Future work can integrate post-quantum cryptographic primitives for long-term security and support cross-cloud deduplication across multiple providers. Lightweight, ML-assisted similarity detection can be added for near-duplicate text while maintaining privacy. Further optimization of proof-of-ownership protocols can also reduce computational overhead for large-scale deployments.

doi: https://doi.org/10.5281/zenodo.18235545

VII. REFERENCES

[1] B. H. Banimfreg, "A comprehensive review and conceptual framework for cloud computing adoption in bioinformatics," *Healthcare Analytics*, vol. 3, p. 100190, 2023.

[2] J. Hassan, D. Shehzad, U. Habib, M. U. Aftab, M. Ahmad, R. Kuleev, and M. Mazzara, "The rise of cloud computing: Data protection, privacy, and open research challenges—A systematic literature review," *Computational Intelligence and Neuroscience*, vol. 2022, Article ID 8303504, 2022, doi: 10.1155/2022/8303504.

[3] K. N. Mishra, R. K. Lal, P. N. Barwal, and A. Mishra, "Advancing data privacy in cloud storage: A novel multi-layer encoding framework," *Applied Sciences*, vol. 15, no. 13, p. 7485, 2025, doi: 10.3390/app15137485.

[4] A. Godavari, C. Sudhakar, and T. Ramesh, "Hybrid deduplication system with content-based cache for cloud environment," 2024.

[5] J. K. Periasamy, C. S. Shieh, and M. F. Horng, "Enhancing cloud security by performing deduplication using serial cascaded autoencoder with GRU and optimal key-based data sanitization," *Computational Intelligence*, vol. 41, no. 5, p. e70140, 2025.

[6] X. Wu, J. Gao, G. Ji, T. Wu, Y. Tian, and N. Al-Nabhan, "A feature-based intelligent deduplication compression system with extreme resemblance detection," *Connection Science*, vol. 33, no. 3, pp. 576–604, 2021.

[7] Y. Zhou, Z. Yu, L. Gu, and D. Feng, "An efficient encrypted deduplication scheme with security-enhanced proof of ownership in edge computing," *BenchCouncil Transactions on Benchmarks, Standards and Evaluations*, vol. 2, no. 2, p. 100062, 2022.

[8] Z. Wang, W. Gao, M. Yang, and R. Hao, "Enabling secure data sharing with data deduplication and sensitive information hiding in cloud-assisted electronic medical systems," *Cluster Computing*, pp. 1–16, 2022, doi: 10.1007/s10586-022-03785-y.

[9] J. K. Periasamy, S. Prabhakar, A. Vanathi, and L. Yu, "Enhancing cloud security and deduplication efficiency with SALIGP and cryptographic authentication," *Scientific Reports*, vol. 15, no. 1, p. 30112, 2025.

[10] L. Li, D. Zheng, H. Zhang, and B. Qin, "Data secure de-duplication and recovery based on public key encryption with keyword search," *IEEE Access*, vol. 11, pp. 28688–28698, 2023.

[11] Z. Yang, J. Li, Y. Ren, and P. P. Lee, "Tunable encrypted deduplication with attack-resilient key management," *ACM Transactions on Storage*, vol. 18, no. 4, pp. 1–38, 2022.

[12] Y. Shin, D. Koo, and J. Hur, "A survey of secure data deduplication schemes for cloud storage systems," *ACM Computing Surveys*, vol. 49, no. 4, pp. 1–38, 2017.

doi: https://doi.org/10.5281/zenodo.18235545