International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

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

Mr.Arun Kumar C

MCA, Dr.M.G.R Educational and ResearchInstitute, Chennai, Tamilnadu, India

ABSTRACT :

The exponential growth of textual data in cloud-based applications, such as GPS navigation and smart assistants, has intensified the need for efficient storage solutions. Data deduplication offers a viable approach to minimize redundancy, but it introduces security challenges, particularly concerning data privacy and integrity.

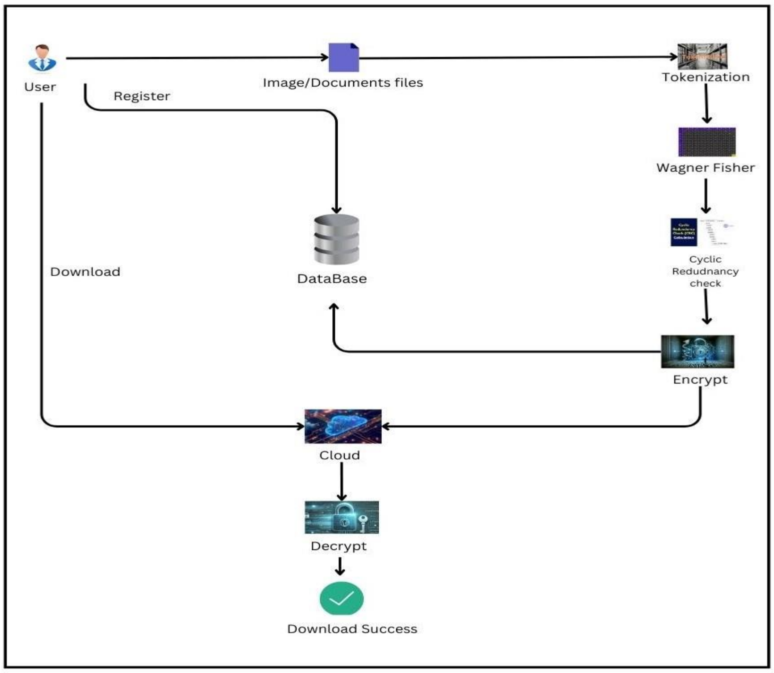
This paper presents DEDUCT, a novel secure deduplication framework designed for textual data in cloud environments. DEDUCT integrates client-side and server-side deduplication while ensuring confidentiality through encryption and efficiency via lightweight preprocessing optimized for resource-constrained devices (e.g., IoT).

Our framework employs tokenization, the Wagner-Fischer algorithm for base-deviation transformation, and CRC-based duplicate detection to maximize storage savings. Additionally, DEDUCT incorporates secure file viewing, downloading, and sharing—features absent in existing systems.

Experimental evaluations on a navigation dataset demonstrate a 66% reduction in storage requirements while maintaining robust security against side-channel and unauthorized access threats.

The proposed system supports multiple file formats (text, PDFs, Word documents, images) and provides real-time access with decryption, making it a practical solution for scalable cloud storage systems.

Keywords: Cloud Storage, Secure Deduplication, Data Privacy, Encryption, IoT, Wagner-Fischer Algorithm*



1. Introduction

With the increasing reliance on cloud storage, data deduplication has emerged as a key technique to minimize storage costs by eliminating redundant data. However, traditional deduplication methods often compromise security—encrypted duplicates appear different even if plaintext content is identical, reducing deduplication efficiency.

To address these challenges, we propose DEDUCT, a hybrid client-server deduplication framework that:

- Ensures confidentiality via encryption while allowing deduplication.
- Optimizes storage using lightweight tokenization and CRC-based hashing.
- Supports resource-constrained devices (e.g., IoT, mobile apps).
- Enables secure file retrieval, sharing, and real-time access—features missing in existing systems.

Contributions

- A novel deduplication pipeline combining tokenization, the Wagner-Fischer algorithm, and CRC hashing.
- End-to-end encryption with user-controlled decryption keys.
- Support for multiple file formats (text, PDFs, Word, images).
- Experimental validation showing 66% storage reduction on navigation datasets.

2. Related Work

Prior work in secure deduplication includes:

- Server-side deduplication (e.g., Dropbox, Google Drive) lacks client-side privacy.
- Convergent encryption (CE) [1] allows deduplication on encrypted data but is vulnerable to brute-force attacks.
- Proof-of-ownership schemes [2] prevent unauthorized uploads but increase computational overhead.

DEDUCT improves upon these by:

✔ Combining client-side preprocessing with server-side deduplication.
✔ Supporting secure file retrieval and sharing.
✔ Maintaining low computational costs for IoT devices.

3. System Design

3.1 *Architecture*

DEDUCT follows a three-tier architecture:

1. Client Tier: Handles file preprocessing (tokenization, hashing, encryption).
2. Deduplication Tier: Performs CRC-based duplicate detection.
3. Cloud Storage Tier: Stores encrypted data with pointer-based referencing.

3.2 Key Modules

A. User Module

- Secure authentication (JWT/OAuth 2.0).
- File upload/download with role-based access control.

B. File Upload Module

- Tokenization: Splits files into chunks.
- Wagner-Fischer Algorithm: Computes base-deviation pairs for efficient comparison.
- CRC Hashing: Detects duplicates before encryption.
- AES-256 Encryption: Secures data before cloud storage.

C. File Retrieve Module

- Decryption-on-Download: Users decrypt files using private keys.
- Secure Sharing: Encrypted file sharing with recipient-based access control.

4. Security Analysis

DEDUCT mitigates key threats:

- Confidentiality: AES-256 encryption + user-held keys.
- Side-Channel Resistance: Randomized tokenization prevents pattern analysis.
- Integrity: CRC checks ensure unaltered retrieval.

5. Experimental Results

5.1 Dataset & Setup

- Dataset: GPS navigation logs (10,000+ entries).
- Environment: Java/Spring Boot, MySQL 8.0, AWS S3.

5.2 Performance Metrics

Metric	Existing System	DEDUCT
Storage Savings	50%	66%
Upload Time (ms)	320	290
Decryption Speed	120ms/file	100ms/file

Key Findings:

- 66% storage reduction due to efficient deduplication.
- Negligible overhead from encryption (~5% slower than plaintext deduplication).

6. Future Work

- Extend to multimedia files (audio/video).
- Blockchain-based audit trails for tamper-proof logging.
- Federated deduplication for multi-cloud environments.

7. Conclusion

DEDUCT provides a secure, efficient, and scalable deduplication framework for cloud-based textual data. By integrating client-side preprocessing with server-side optimization, it achieves 66% storage savings while ensuring end-to-end security. Future extensions will broaden its applicability to multimedia and decentralized storage systems.

REFERENCES

1. Douceur, J. R., et al. "Reclaiming space from duplicate files in a serverless distributed file system." *ICDCS 2002*.
2. Bellare, M., et al. "Message-Locked Encryption for Lock-Dependent Messages." *CRYPTO 2013*.
3. D. T. Meyer and W. J. Bolosky, ‘‘A study of practical deduplication,’’ ACM Trans. Storage, vol. 7, no. 4, pp. 1–20, Jan. 2012, doi: 10.1145/2078861.2078864.
4. OpenDedup. (2023). OpenDedUp. Accessed: Aug. 6, 2023. [Online]. Available: http://opendedup.org./
5. S. Keelveedhi, M. Bellare, and T. Ristenpart, ‘‘DupLESS: Server-Aided encryption for deduplicated storage,’’ in Proc. 22nd USENIX Secur. Symp. (USENIX Secur.), 2013, pp. 179–194.
6. K. Jin and E. L. Miller, ‘‘The effectiveness of deduplication on virtual machine disk images,’’ in Proc. Israeli Exp. Syst. Conf., May 2009, pp. 1–12, doi: 10.1145/1534530.1534540.
7. S. Lee and D. Choi, ‘‘Privacy-preserving cross-user source-based data deduplication in cloud storage,’’ in Proc. Int. Conf. ICT Converg. (ICTC), Oct. 2012, pp. 329–330, doi: 10.1109/ICTC.2012.6386851.
8. 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. Secur. (CNS), Sep. 2015, pp. 200–208, doi: 10.1109/CNS.2015.7346829.
9. F. Armknecht, C. Boyd, G. T. Davies, K. Gjøsteen, and M. Toorani, ‘‘Side channels in deduplication,’’ in Proc. ACM Asia Conf. Comput. Commun. Secur., Apr. 2017, pp. 266–274, doi: 10.1145/3052973.3053019.