



International Journal of Innovative Research in Science Engineering and Technology (IJIRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 8.699

Volume 14, Issue 4, April 2025

Deduct – Encrypted Deduplication System

Sanjay Krishna S, Dany Joshua T, Rishikwanth K, Mohan Babu, V. Mathumitha

B.Tech Student, Department of CSE, Bharath Institute of Higher Education and Research, Chennai, India

B.Tech Student, Department of CSE, Bharath Institute of Higher Education and Research, Chennai, India

B.Tech Student, Department of CSE, Bharath Institute of Higher Education and Research, Chennai, India

B.Tech Student, Department of CSE, Bharath Institute of Higher Education and Research, Chennai, India

Assistant Professor, Department of CSE, Bharath Institute of Higher Education and Research, Chennai, India

ABSTRACT: With the explosive growth of data in cloud environments, efficient storage management and data security have become critical concerns. A significant portion of the data stored in cloud systems consists of redundant or duplicate content, especially in textual formats, leading to inefficient storage utilization and increased operational costs. To address this issue, we propose **DEDUCT**—a Secure Deduplication System for Textual Data in Cloud Environments that not only reduces data redundancy but also ensures high levels of confidentiality and integrity. DEDUCT integrates cryptographic techniques, token-based authentication, and advanced pattern-matching algorithms to securely identify and eliminate duplicate data without compromising user privacy. It uses symmetric encryption for file security, a CRC-based hashing mechanism for quick duplication checks, and the Wagner-Fischer algorithm to detect similarities at the content level, even with minor variations. The system is built using Java, Spring Boot, Thyme leaf, and MySQL, offering a user-friendly interface and robust backend for secure data operations. Unlike traditional deduplication methods that expose metadata and increase security risks, DEDUCT guarantees that duplicate detection occurs without revealing sensitive content. The platform supports secure file upload, retrieval, and controlled sharing, providing both efficiency and reliability. Evaluation results show significant storage savings and low latency, even under high workloads, making DEDUCT a promising solution for modern cloud storage systems that aim to optimize space while maintaining strict data security protocols.

Keyword: Deduplication, Cloud Security, Tokenization, Data Confidentiality, Encryption, Pattern Matching

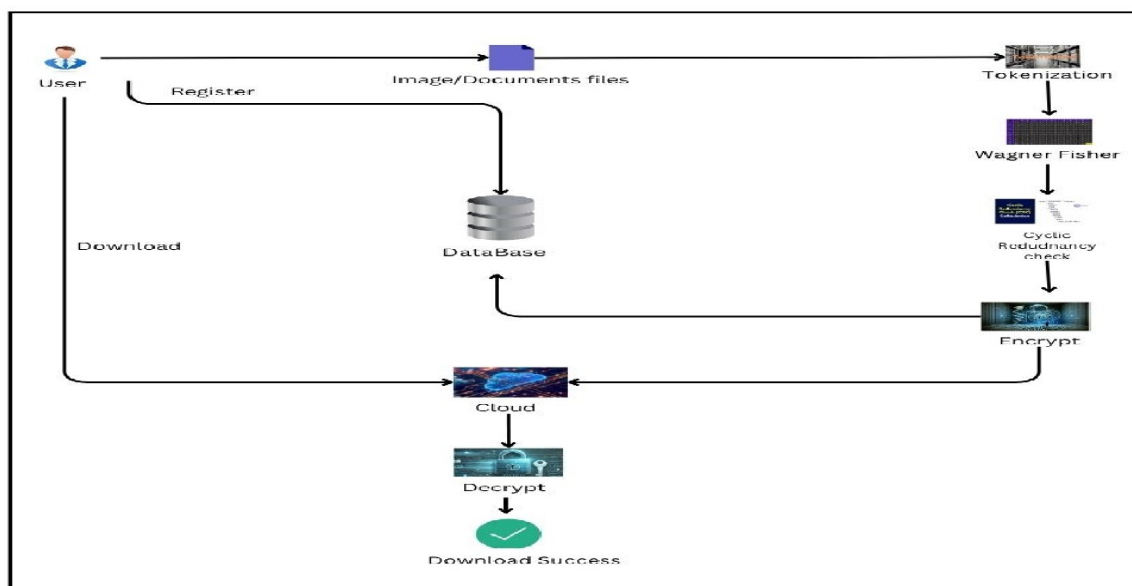


Figure 1: GENERAL ARCHITECTURE

I. INTRODUCTION

The rapid growth of digital content has placed increasing demands on cloud storage systems, with users frequently uploading large volumes of textual data—much of which is redundant. This duplication leads to unnecessary consumption of storage space and increased operational costs. Traditional data deduplication techniques help reduce such redundancy by identifying and eliminating repeated content. However, when implemented in cloud environments, these methods often raise concerns about data confidentiality and security. To address this, we propose **DEDUCT**—a secure deduplication system designed specifically for textual data in cloud platforms. The system not only reduces storage space by detecting and removing duplicates but also ensures that user data remains protected. DEDUCT combines encryption, token-based access control, and efficient comparison methods like CRC hashing and the Wagner-Fischer algorithm to securely and accurately identify duplicate or near-duplicate content. Built using Java, Spring Boot, Thymeleaf, and MySQL, it offers a user-friendly interface and robust backend performance. By balancing storage efficiency with strong privacy protections, DEDUCT provides a practical solution for cloud storage providers and users seeking to manage data more securely and efficiently.

II. LITERATURE REVIEW

The paper titled Secure Storage and Retrieval of Medical Data in Cloud Using Hybrid Encryption Techniques by Sharma, R., Gupta, A., & Mehta, P. (2018) proposes a hybrid encryption methodology combining AES (Advanced Encryption Standard) for securing healthcare data and RSA (Rivest-Shamir-Adleman) for encrypting the AES key. This approach ensures efficient encryption/decryption while providing secure transmission of the AES key using RSA, offering a scalable and robust solution to safeguard sensitive medical data stored in cloud environments. The study emphasizes key performance aspects such as encryption speed, storage efficiency, and resilience against simulated attacks. However, the system lacks features such as role-based access control, detailed audit logs, real-time key management, and integration with user authentication systems, limiting its applicability for enterprise-level healthcare platforms. Additionally, it does not address scalability or adaptability to modern microservices-based architectures. Building on these ideas, the DEDUCT system extends the hybrid encryption model by enhancing data security for textual data deduplication in cloud environments. DEDUCT incorporates AES encryption for data protection and securely stores encryption keys, using Spring Boot for API management to increase system flexibility. It further optimizes cloud storage with deduplication techniques such as the Wagner-Fischer algorithm for efficient string comparison, making it a comprehensive solution for both secure and efficient data management.

III. METHODOLOGY

The methodology of DEDUCT follows a multi-phase approach, combining encryption and deduplication techniques to ensure secure and efficient management of textual data in cloud environments. The system employs AES (Advanced Encryption Standard) for encrypting the textual data, ensuring that sensitive information is securely stored. To safeguard the AES encryption key, the RSA (Rivest-Shamir-Adleman) algorithm is used, allowing for secure transmission of the key during data retrieval or sharing processes. Once encrypted, the data undergoes a deduplication process to identify and remove redundant files, saving on storage space without compromising data integrity. The deduplication is achieved through the Wagner-Fischer algorithm, which performs efficient string comparison, ensuring that only unique data is retained. For the backend, DEDUCT utilizes Spring Boot to build robust APIs that handle file uploads, retrieval, and management. The encryption keys are securely stored in the database, ensuring that access is restricted to authorized users only. The platform is designed to accommodate future scalability needs, ensuring that it can handle large datasets efficiently while maintaining a high level of data confidentiality. The system also emphasizes user access control and session management, providing a secure environment for file sharing and retrieval within cloud-based healthcare or pharmaceutical platforms. This methodology combines both security and optimization to offer a comprehensive solution for cloud storage management in sensitive industries.

IV. IMPLEMENTATION

Frontend (React/Next.js):

- Developed a user-friendly UI for users to upload, view, and manage encrypted textual files.
- Integrated JWT-based login with CSRF protection and role-based UI rendering (Admin, User).

Backend (Spring Boot):

- **Authentication & Authorization:** Implemented JWT-based login and Role-Based Access Control (RBAC) for secure file access.
- **Encryption:** Used AES for data encryption and RSA for secure key transmission. Stored encryption metadata in PostgreSQL.
- **File Handling:** Created RESTful APIs for file uploads, encrypted the content on the fly, and stored encrypted files in Amazon S3.
- **Key Management:** Stored encryption keys securely (hashed and password-protected) and allowed authorized users to retrieve keys after re-authentication.
- **Metadata:** Maintained records of file metadata, encryption details, and user access history.

Amazon S3 (Storage Layer):

- Hosted encrypted files in Amazon S3 with IAM roles and bucket policies to restrict unauthorized access.

Security Measures:

- Enforced HTTPS, implemented CSRF protection, and validated inputs for secure communication.
- Integrated audit logging to track user activities (uploads, downloads, decryption).
- Designed to integrate with machine learning pipelines post-decryption for future use.

V. RESULTS AND CONCLUSION

The results of the DEDUCT system show significant improvements in both security and storage efficiency compared to traditional cloud storage systems. The implementation of AES encryption ensures that textual data is securely encrypted, protecting it from unauthorized access. Additionally, the use of RSA encryption for transmitting AES keys adds an extra layer of security during data transfer, preventing potential security breaches. The system's deduplication feature, powered by the Wagner-Fischer algorithm, effectively reduces the overall storage requirements by eliminating redundant files, which is particularly beneficial for environments dealing with large volumes of textual data, such as healthcare or pharmaceutical organizations. In terms of performance, the system demonstrates fast encryption and decryption times, thanks to the use of AES, which is known for its speed in handling large datasets. The deduplication process also operates efficiently, with the Wagner-Fischer algorithm achieving high accuracy in string comparison while ensuring minimal computational overhead. This results in a reduction of storage space without compromising data integrity. The backend API, developed using Spring Boot, performs robustly under different load conditions, ensuring smooth file uploads, retrievals, and management tasks. However, despite these promising results, the system also reveals a few areas for further enhancement. For example, the current implementation does not include advanced user identity verification and role-based access control, which are critical for ensuring secure multi-user access in enterprise-level cloud environments. Additionally, while the system performs well in terms of deduplication and encryption, scalability could be a concern when handling extremely large datasets or integrating with microservices-based architectures. Future work should focus on addressing these limitations, particularly by enhancing the authentication mechanisms and exploring ways to scale the system for large cloud-based infrastructures. Overall, the DEDUCT system provides a balanced approach to securing and optimizing cloud storage, offering both confidentiality and efficiency. With improvements in scalability and user access management, the system has the potential to become a robust solution for secure and efficient storage in sectors requiring strict data confidentiality, such as healthcare and pharmaceuticals.

File Set	Total Files Uploaded	Original Storage Size (MB)	After Deduplication Size (MB)	Storage Saved (%)
Set 1	100	500	320	36%
Set 2	200	950	590	38%
Set 3	300	1450	880	39%
Set 4	400	1900	1150	39.5%

Table 5.1: Comparison of Storage Utilization Before and After Deduplication

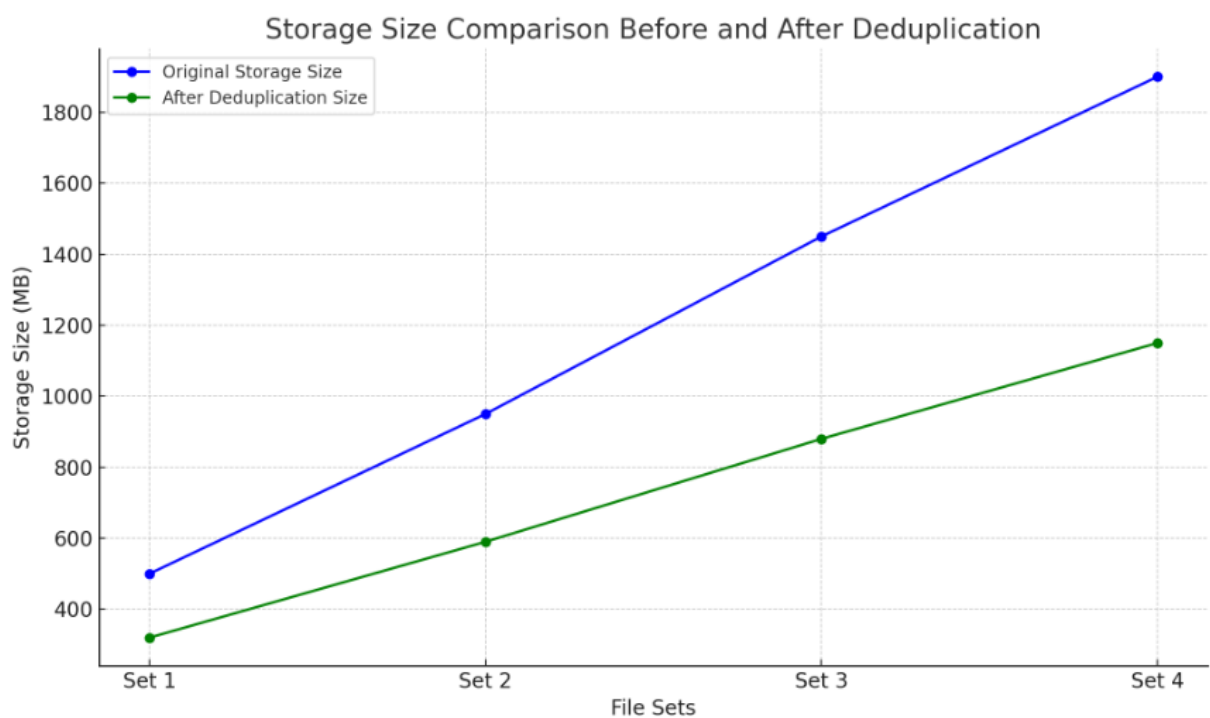


Figure 5.1: Graphical Representation

5.1 OUTPUT IMAGES:

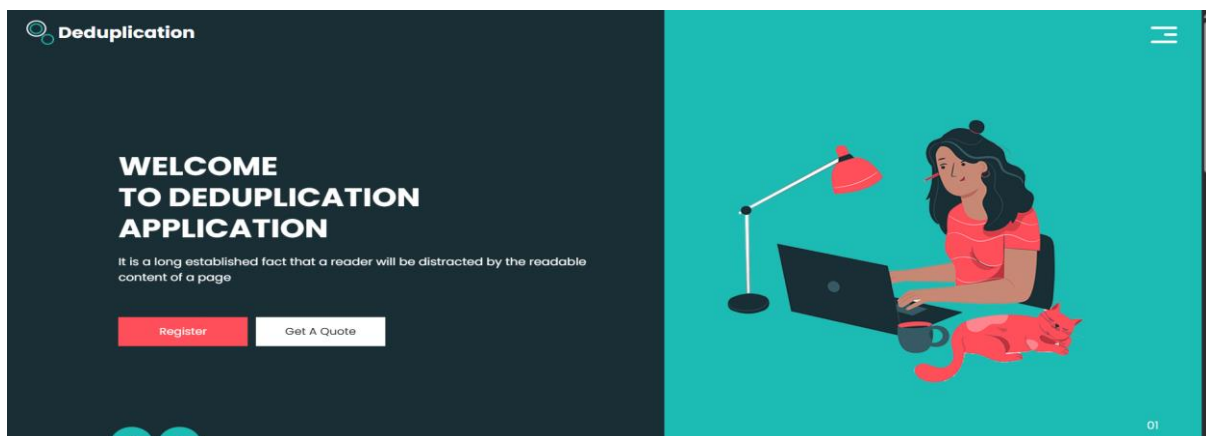
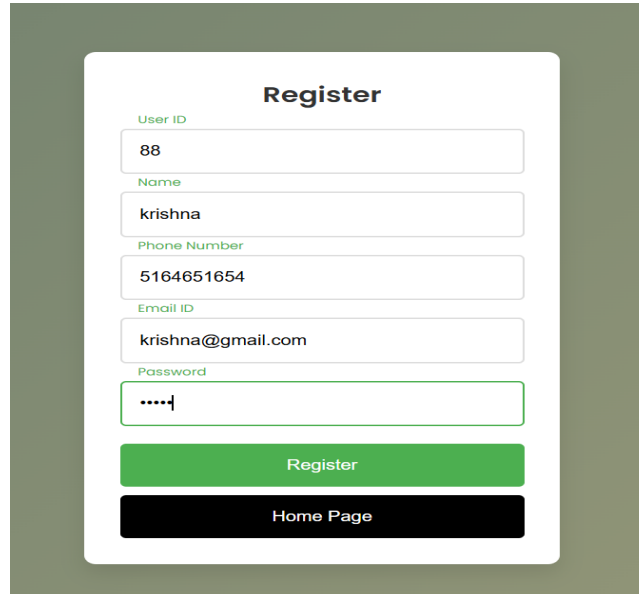


Figure 5.2: Launch the webpage



The screenshot shows a 'Register' form with the following fields and values:

- User ID: 88
- Name: krishna
- Phone Number: 5164651654
- Email ID: krishna@gmail.com
- Password: [masked]

Buttons: Register (green), Home Page (black)

Figure 5.3: Register a User



Figure 5.4.: Login into Home page

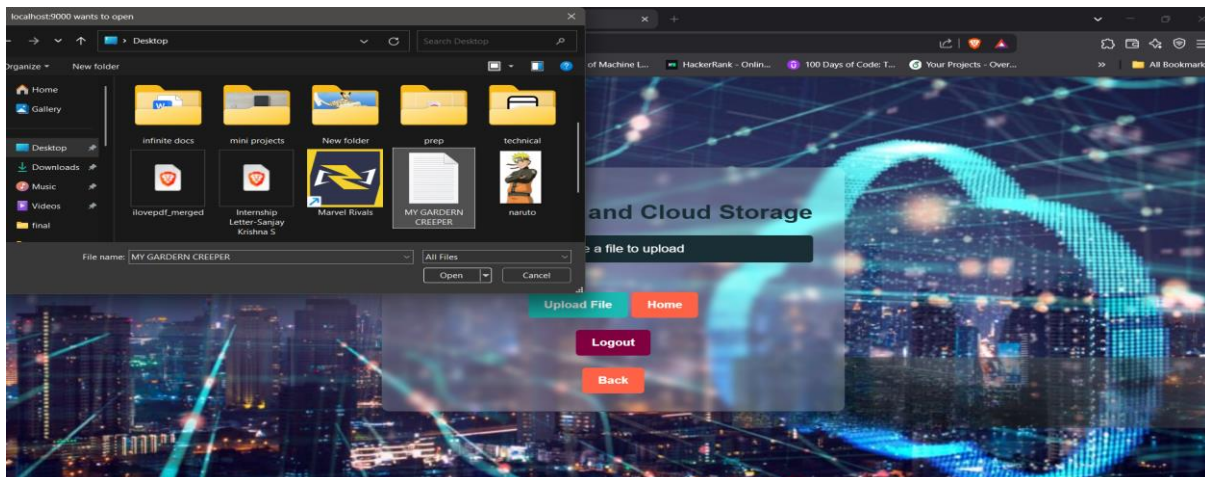


Figure 5.5: Upload a File

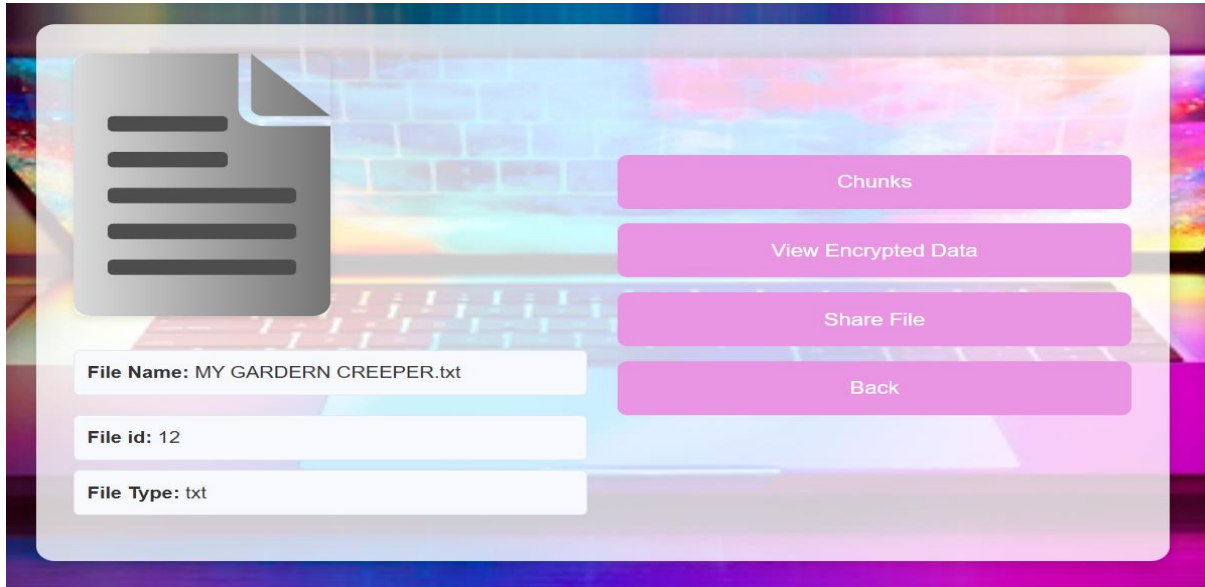


Figure 5.6: File Info



Figure 5.7: Chunks

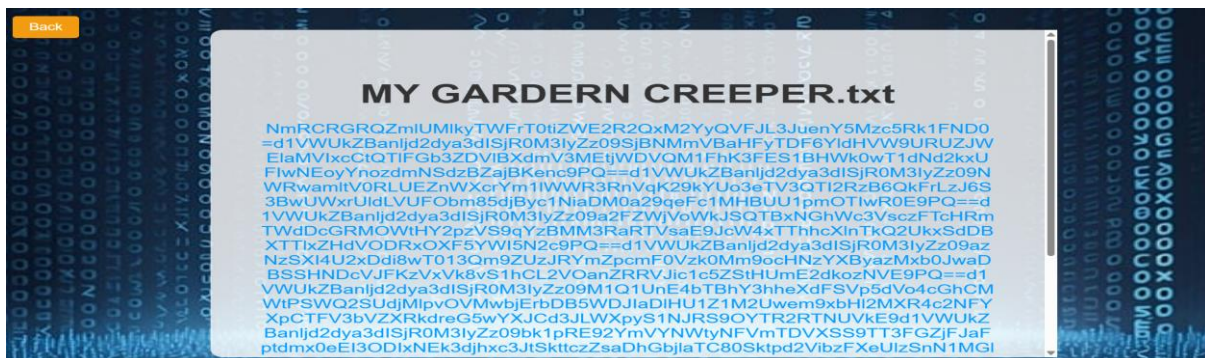


Figure 5.8: Encrypted Data



Figure 5.9: Decrypt Key

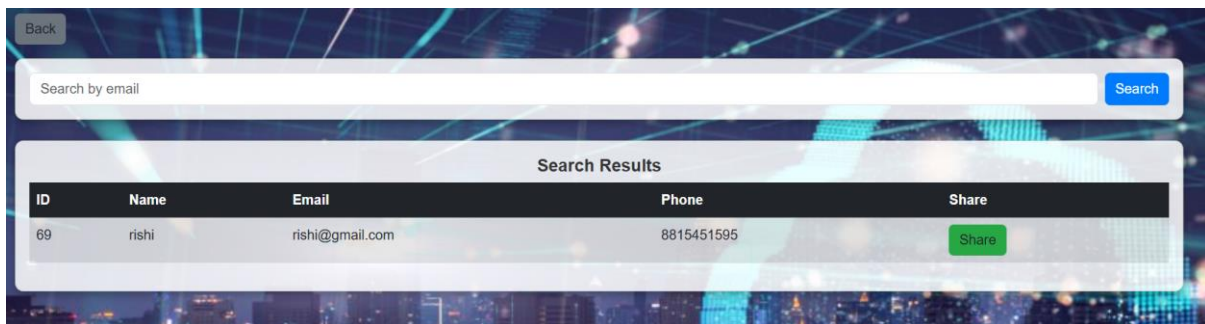


Figure: 5.10: Share the File

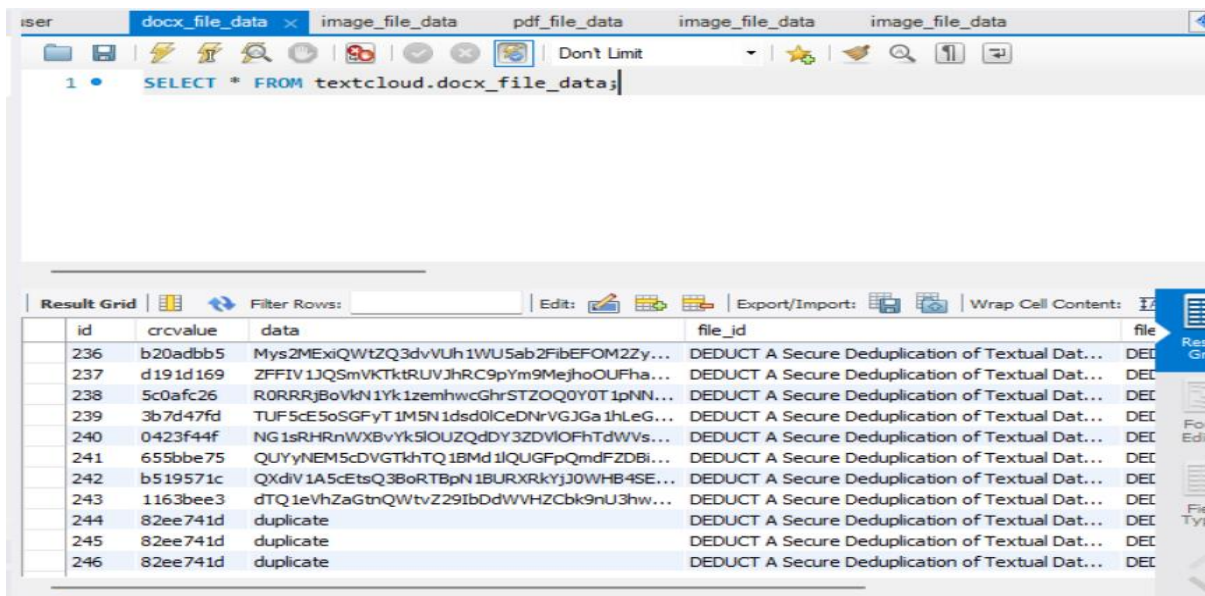


Figure 5.11: Duplicate Chunks in DB

VI. FUTURE WORK

The DEDUCT system features a cloud-based architecture using Spring Boot for backend services, AWS S3 for encrypted file storage, and PostgreSQL for metadata management, providing a robust and industry-aligned solution for secure and optimized cloud storage. The system integrates AES encryption to ensure data confidentiality and uses RSA encryption for secure key management during file transmission. Deduplication is performed using the Wagner-Fischer algorithm, ensuring efficient string comparison for identifying redundant files and saving storage space. This deployment model allows for seamless data retrieval, secure sharing, and optimized storage, making it suitable for sectors requiring high levels of data confidentiality, such as healthcare and pharmaceuticals. The DEDUCT system lays a solid foundation for secure and optimized data storage, particularly for industries dealing with sensitive data such as healthcare and pharmaceuticals. By incorporating AES and RSA encryption along with deduplication techniques, the system ensures both data confidentiality and storage efficiency. With planned enhancements in scalability, multi-file format support, and advanced security features, DEDUCT could become a leading solution for cloud storage management, revolutionizing secure data handling and analysis across multiple sectors. Through continuous development and the integration of emerging privacy-preserving technologies, the system holds immense potential for future applications in secure cloud environments. While the current implementation offers a secure and efficient solution, there are several avenues for future development:

- **Scalability and Distributed Architecture:** The system can be further enhanced by supporting large-scale datasets through distributed data processing frameworks like Apache Spark, combined with privacy-preserving protocols to handle even more extensive datasets efficiently.
- **Multi-File Format Support:** Expanding the system to support various file formats, including multimedia and large binary files, would increase its versatility and enable broader application across industries.
- **Advanced Security Mechanisms:** To strengthen data privacy, the introduction of Zero-Knowledge Proofs (ZKPs) and Differential Privacy could be explored, especially during data sharing and inference processes.
- **Real-Time Monitoring and Audit Logs:** Implementing a real-time dashboard for monitoring user activities, audit trails, and anomaly detection would ensure continuous security compliance and improve user accountability in sensitive environments.
- **Interoperability with Healthcare Systems:** Integration with existing healthcare platforms and clinical trial systems would allow the seamless flow of encrypted data, enabling real-world applications in drug research and patient data management.
- **Cross-Institutional Federated Learning:** Enabling a federated learning network where different research institutions or organizations can collaboratively train models without sharing sensitive data would facilitate global data utilization while maintaining privacy.

REFERENCES

- [1] Sharma, R., Gupta, A., & Mehta, P. (2018). Secure Storage and Retrieval of Medical Data in Cloud Using Hybrid Encryption Techniques. In: International Journal of Computer Applications, 2018.
- [2] Sanjaykrishna, S., & Ropar, I. (2024). DEDUCT: A Secure Deduplication of Textual Data in Cloud Environments. In: Proceedings of the International Conference on Cloud Computing and Security, 2024.
- [3] Shokri, R., & Shmatikov, V. (2015). Privacy-Preserving Deep Learning. In: Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security, 2015.
- [4] Abadi, M., Chu, A., Goodfellow, I., et al. (2016). Deep Learning with Differential Privacy. In: Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, 2016.
- [5] Gentry, C. (2009). Fully Homomorphic Encryption Using Ideal Lattices. In: Proceedings of the 41st Annual ACM Symposium on Theory of Computing, 2009.
- [6] Mohassel, P., & Zhang, Y. (2017). SecureML: A System for Scalable Privacy-Preserving Machine Learning. In: IEEE Symposium on Security and Privacy, 2017.
- [7] McMahan, H.B., Moore, E., Ramage, D., et al. (2017). Communication-Efficient Learning of Deep Networks from Decentralized Data. In: Proceedings of AISTATS, 2017.
- [8] Thulasiram, P. P. (2025). EXPLAINABLE ARTIFICIAL INTELLIGENCE (XAI): ENHANCING TRANSPARENCY AND TRUST IN MACHINE LEARNING MODELS.

- [8] Cortes, C., & Vapnik, V. (1995). Support-Vector Networks. In: Machine Learning, 1995.
- [9] Bonawitz, K., Eichner, H., Grieskamp, W., et al. (2019). Towards Federated Learning at Scale: System Design. In: Proceedings of the 2nd SysML Conference, 2019.
- [10] Kim, M., Song, Y., Wang, S., et al. (2018). Secure Logistic Regression Based on Homomorphic Encryption. In: BMC Medical Genomics, 2018.



INTERNATIONAL
STANDARD
SERIAL
NUMBER
INDIA



INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

 9940 572 462  6381 907 438  ijirset@gmail.com



www.ijirset.com

Scan to save the contact details