International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 09 Issue: 05 | May - 2025 SJIF Rating: 8.586 ISSN: 2582-3930

Spacing detection of Duplication in Textual Data for Secured Cloud Environment

Saravanan.K.G[1], Kiruthika K[2] , Meera Jasmine R[3] and Joice Evangilin N[4]

[1]Assistant Professor -Department of Information Technology & Kings Engineering College-India.

[2,3,4,5]Department of Information Technology & Kings Engineering College-India.

Abstract - The rapid increase in textual data in navigation tasks for devices like GPS or smart assistants creates challenges for managing and storing data in large-scale systems. Data deduplication, which reduces storage needs by eliminating duplicate data, offers a solution but raises security concerns. This paper introduces DEDUCT, a new method that combines cloud-side and client-side deduplication to achieve high data compression while protecting data privacy. Designed for devices with limited resources, such as IoT devices, DEDUCT includes lightweight preprocessing and safeguards against security risks like side-channel attacks. Testing on a navigation dataset shows that DEDUCT can compress data by up to 66%, significantly cutting storage costs while keeping data secure, making it an efficient choice for managing large-scale data systems

***Key Words*:** *Data Deduplication, Cloud Storage, Security, Tokenization, Encrypted File Sharing*

1.INTRODUCTION

1.1 Background:

Cloud storage systems have emerged as an essential component of modern data management infrastructure. With the exponential increase in data generated by businesses, institutions, and individuals, the need for efficient and scalable storage solutions has never been greater. Cloud storage, which allows users to store data on remote servers and access it over the internet, offers significant advantages, including flexibility, accessibility, and reduced infrastructure costs. As organizations move their critical data to the cloud, the importance of managing this data efficiently becomes crucial.

However, with the growing volume of data being stored in the cloud, one of the most pressing challenges is **data duplication**. Data duplication refers to the phenomenon where identical or near-identical data copies are stored across multiple locations in the cloud. This redundancy not only wastes valuable storage resources but also negatively impacts performance by increasing the processing time for data retrieval, backup, and synchronization. Furthermore, duplicated data can lead to increased costs due to the extra storage space required, as well as security concerns, as maintaining multiple copies of sensitive data could potentially expose it to greater risk. Therefore, addressing the problem of data duplication in the cloud is vital for ensuring both cost-efficiency and the security of cloud-based systems [1], [2], [4].

1.2 Motivation

In a cloud environment, detecting and preventing data duplication is not just a matter of efficiency—it is essential for ensuring data **security**, **integrity**, and **cost-effectiveness**. Data deduplication, a technique that identifies and removes redundant copies of data, can play a pivotal role in optimizing storage space and reducing the costs associated with cloud storage. However, in the cloud, security becomes a primary concern, as sensitive data is often stored and processed on third-party servers. Without proper safeguards, there is a risk that malicious actors could exploit vulnerabilities to gain unauthorized access to duplicated data [5], [9].

Moreover, maintaining data integrity is another crucial aspect. In a cloud setting, data integrity ensures that the data remains accurate, consistent, and trustworthy, even when multiple copies are stored across different servers. Thus, deduplication techniques must not only eliminate redundancy but also guarantee that the original data remains intact and protected against tampering or loss [6], [14].

The benefits of detecting and preventing data duplication extend beyond security and integrity. Cost savings are a key motivating factor. Cloud service providers typically charge based on the amount of

storage used, so by reducing duplication, organizations can significantly lower their storage expenses, leading to more cost-efficient data management. In addition, reduced duplication can improve data retrieval speed and overall system performance by minimizing the overhead associated with accessing redundant copies of data [16].

1.3 Contributions

This paper presents several novel contributions aimed at improving the state of data deduplication for textual data in cloud storage systems, building on the **DEDUCT** method by Ghassabi and Pahlevani [2], and adding extra functionalities to enhance user experience and security:

- **Introduction of Secure Deduplication Methods:** Building upon **DEDUCT**, we propose enhanced techniques for secure data deduplication designed specifically for textual data stored in cloud environments. These methods ensure the security of deduplication processes while eliminating redundant data efficiently.
- **Support for Multiple File Types:** Our system extends the capabilities of **DEDUCT** by allowing users to upload and manage various file types, including text files, PDFs, Word documents, and images. This flexibility makes the system more adaptable to a range of use cases and user requirements.
- **Secure File Viewing and Encryption:** Users can securely view their files without compromising data privacy, as the system ensures that data remains encrypted during access. This feature is critical for protecting sensitive information while still allowing authorized users to interact with the data.
- **File Download with Decryption:** To further enhance security, the system requires a unique decryption key to access and download files. This ensures that only authorized users can download and access the original data, offering an additional layer of security.
- **File Sharing:** The system includes a feature for secure file sharing, enabling users to share files with others while maintaining encryption and security. This promotes collaboration and data exchange, making the system suitable for team-based or organizational use.
- **Chunk Detail Viewing:** Users can view detailed information about the individual chunks of their uploaded files, enhancing transparency and control over how data is stored and managed in the cloud. This feature also assists in debugging and optimizing storage efficiency.

2. RELATED WORK

Cloud storage systems have increasingly adopted data deduplication techniques to optimize storage usage and enhance security. Researchers have proposed various methods to address the challenges of efficiently managing duplicate data while maintaining confidentiality and integrity.

2.1 Data Deduplication

Data deduplication is a specialized data compression technique for eliminating duplicate copies of repeating data. In cloud environments, deduplication plays a crucial role in reducing storage costs and improving bandwidth efficiency, especially for textual data. Several techniques such as file-level deduplication, block-level deduplication, and chunk-based deduplication have been proposed. Chunking methods, which divide files into smaller pieces, are particularly effective for textual data as they can detect fine-grained redundancy. However, challenges persist in maintaining deduplication efficiency when files undergo minor modifications, which often trigger redundant storage.

Yu et al. [1] proposed VeriDedup, a verifiable deduplication system that not only optimizes storage but also provides mechanisms for users to verify the integrity of their deduplicated data. Similarly, Ghassabi and Pahlevani [2] developed DEDUCT, a scheme specifically targeting the secure deduplication of textual data in cloud environments, emphasizing low overhead during verification processes.

2.2 Security Challenges in Cloud Storage

While deduplication improves storage efficiency, it also introduces new security risks. Traditional encryption methods, if directly applied, prevent effective deduplication since ciphertexts of identical files differ. Thus, specialized approaches such as convergent encryption were adopted to balance security and deduplication.

Nonetheless, vulnerabilities remain. Harnik et al. [10] highlighted side-channel attacks in deduplication services, where attackers could infer sensitive information by observing storage behavior. Proofs of Ownership (PoW) mechanisms, as explored by Halevi et al. [13], were introduced to ensure that clients genuinely possess the data they claim, reducing risks of leakage.

To address key management issues, Bellare et al. [11] proposed DupLESS, a server-aided encryption system that allows deduplication over encrypted data without exposing user files to the server. Xu et al. [5] introduced weak leakage-resilient deduplication techniques aimed at minimizing the information exposed during the deduplication process.

Despite these advancements, challenges like balancing strong security with system scalability, efficient key management, and resilience against evolving threats remain significant concerns.

2.3 Existing Deduplication Techniques

Extensive research has been conducted to refine and improve deduplication methods. Xia et al. [4] provided a comprehensive survey tracing the evolution of deduplication technologies, identifying trends such as client-side deduplication and hybrid approaches. Shin et al. [8] investigated the application of convergent encryption to bolster deduplication security against unauthorized access.

Dynamic ownership management has also been a focus. Nguyen et al. [6] and Guo et al. [7] studied models ensuring that ownership changes do not undermine security, which is crucial for collaborative cloud environments. Fu et al. [14] introduced fine-grained ownership management to enhance trust and verifiability within deduplication systems.

Efforts to support public verifiability were presented by Wang et al. [9] through Oruta, allowing privacy-preserving audits without revealing file contents. Stanek et al. [12] proposed a deduplication scheme that strikes a balance between security and storage efficiency in cloud environments.

In parallel, key management remains critical. Li et al. [16] designed an efficient convergent key management mechanism, aiming for better reliability and lower communication overhead.

Moreover, Zheng and Xu [17] worked on dynamic proofs of retrievability to maintain fairness and reliability in deduplication storage, ensuring trust in remote systems.

Together, these works establish a strong foundation for secure, verifiable, and efficient cloud deduplication systems, while ongoing research continues to address limitations related to scalability, privacy preservation, and seamless user experience.

3. METHODOLOGY

3.1 Overview

This methodology introduces a secure and efficient framework for detecting both exact and near-duplicate textual data in cloud storage, minimizing redundancy while ensuring data confidentiality, integrity, and access control. The process includes hashing, fingerprinting, NLP-based analysis, convergent encryption, proof of ownership, and verifiable integrity checks.

3.2 Textual Data Analysis

- **Hashing:** Utilizes cryptographic hashes (e.g., SHA-256) for exact duplication detection.
- **Fingerprinting:** Employs content-defined chunking (e.g., Rabin Fingerprints) to identify partial duplicates.
- **NLP Techniques:** Uses semantic similarity (TF-IDF cosine similarity, BERT embeddings) for near-duplicate detection.

3.3 Secure Deduplication Framework

- **Convergent Encryption:** Encrypts data using content-derived keys.
- **Proof of Ownership (PoW):** Ensures only legitimate users trigger deduplication.
- **Dynamic Ownership Management:** Updates access rights on duplicate uploads.
- **Verifiable Integrity Checks:** Uses Merkle Trees and third-party auditing (e.g., Oruta).
- **Leakage-Resilient Protocols:** Protects against side-channel attacks.

3.4 System Design

Components:

➢ *User Device:* Performs hashing, fingerprinting, NLP, and encryption.

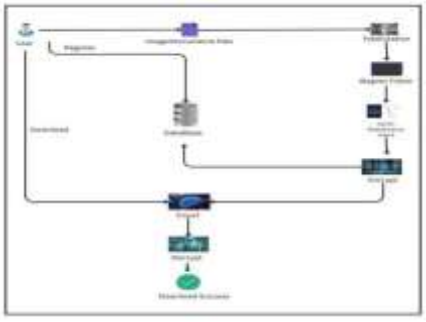
- *Deduplication Server:* Detects duplicates and verifies PoW.
- *Metadata Manager:* Manages access control securely.
- *Cloud Storage:* Stores encrypted deduplicated blocks.
- *Auditor:* (Optional) Verifies data integrity.

3.4.1 Workflow

- **Upload:** Text is preprocessed and analyzed.
- **Duplication Check:** Hash/NLP-based comparison.
- **Security Verification:** PoW validation.
- **Encryption:** Uses convergent encryption.
- **Storage & Management:** Updates metadata and securely stores data.

3.4.2 Architecture Highlights

The system applies CRC for quick exact-match detection and the Wagner-Fischer algorithm for edit-distance-based near-duplicate detection. Redundant data



is removed before secure cloud storage, ensuring space efficiency and data security.

Figure 1: System Architecture for Secure Textual Data Deduplication Using Tokenization, Wagner-Fischer Algorithm, and CRC Techniques.

4. Proposed Solution

4.1 Data Deduplication Process

The data deduplication process is the core mechanism of the proposed system, designed to ensure efficient and secure handling of textual data in the cloud. The process begins with **data ingestion** where users log into the system and upload files, which can include text files and images. The uploaded content undergoes several key stages to detect and remove duplication:

1. **Tokenization**: The first step is tokenizing the uploaded textual content. Tokenization divides the document into smaller units like words or phrases, which allows the system to analyze the structure and content of the text in finer detail. This step is crucial for detecting near-duplicate text, as it breaks down complex content into simpler components (Xia et al., 2016).

2. **Wagner-Fischer Algorithm**: To detect near-duplicates, the **Wagner-Fischer algorithm** is employed. This dynamic programming approach calculates the **edit distance** between two strings or sequences, which helps identify how similar two pieces of text are. The Wagner-Fischer algorithm computes the minimum number of operations (insertions, deletions, and substitutions) required to transform one sequence into another. This enables detection of variations in the text that may not be immediately obvious but are significant for identifying near-duplicate data (Ghassabi & Pahlevani, 2024).

3. **Cyclic Redundancy Check (CRC)**: After tokenization and using the Wagner-Fischer algorithm, **CRC values** are computed for each document. CRCs are a form of hash function that provides a fingerprint of the data, making it easier to spot exact duplicates by comparing these values. CRC is particularly effective for quickly identifying identical files or chunks of data, ensuring that the system can efficiently flag redundant content (Guo et al., 2020).

4. **Duplicate Removal**: If a duplicate is detected, the redundant data is eliminated from the system, ensuring that only unique content is stored. This reduction in duplication not only saves storage space but also enhances performance by minimizing unnecessary data in the cloud environment.

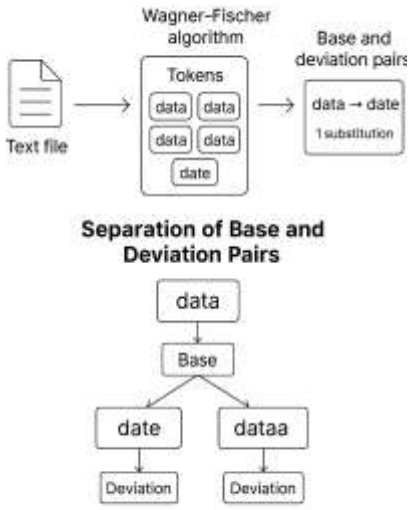


Figure 2: Separation Of Base and Deviation Pairs

5. **Upload to Cloud**: Finally, once the duplicates are removed, the original, non-redundant data is uploaded to the cloud storage. This ensures that only valuable, non-duplicated content occupies cloud resources, making the entire process efficient and cost-effective.

4.2 Security Enhancements

Security is a critical concern when dealing with cloud storage and deduplication, especially since multiple users may be uploading sensitive data. Several security mechanisms are incorporated into the deduplication system to protect the confidentiality, integrity, and availability of the data throughout the process:

1. **Encryption Techniques**: All uploaded content is **encrypted** before being processed. The encryption ensures that the data remains secure from unauthorized access during the deduplication process. Techniques like **convergent encryption** are used, where the same file will generate the same encrypted version, facilitating secure deduplication without compromising privacy (Shin et al., 2016). This method helps prevent unauthorized access to sensitive information during storage and transmission.
2. **Integrity Verification**: To maintain data integrity, the system uses **hashing** and **integrity verification** techniques. Each file's hash value is generated during the upload process and stored. During subsequent access or operations, the hash is re-calculated and compared with the stored value to verify that no data corruption has occurred (Yu et al., 2023). This step ensures that the data in the cloud remains intact and reliable, even after undergoing the deduplication process.
3. **Access Control**: The deduplication system implements **access control mechanisms** to ensure that only authorized users can interact with specific files. Role-based access control (RBAC) is used to grant permissions based on the user's role in the system, ensuring that only the right users can upload, modify, or delete content (Xu et al., 2019).

4.3 Algorithm for Deduplication

The proposed system utilizes a combination of efficient and secure algorithms to detect and eliminate duplicate data in cloud environments. These algorithms are crucial for ensuring that only unique and relevant data is stored, optimizing storage and ensuring data integrity.

- **Tokenization:**
  Tokenization is the first step in the deduplication process, where the text is broken down into smaller chunks, known as tokens (e.g., words, phrases, or characters). This simplification allows for easier comparison between texts and forms the foundation for subsequent processing steps. Tokenization helps identify common elements across multiple files, enabling the system to detect potential duplicates even if the structure or content slightly varies. This method has been widely employed in natural language processing (NLP) for text analysis and deduplication purposes [2].
- **Wagner-Fischer Algorithm:**
  The Wagner-Fischer algorithm is used to calculate the edit distance between two pieces of text. It identifies the number of insertions, deletions, or substitutions required to transform one string into another. By using this algorithm, the system can detect near-duplicates, where the text may have small differences, such as spelling errors or punctuation variations. The Wagner-Fischer algorithm helps the system account for these minor differences, ensuring that near-identical data is also flagged as duplicates. This algorithm has proven to be

effective in similar data processing applications [4].

- **Cyclic Redundancy Check (CRC):**
  CRC is a hashing algorithm that generates a unique checksum for each file. By comparing the CRC values of different files, the system can easily identify exact duplicates. Files with identical CRC values are flagged for further processing or removal, as they Deduplication Rate: The deduplication rate measures the percentage of redundant data that is successfully identified and eliminated by the system. It is calculated by comparing the total data size before and after deduplication. Higher deduplication rates indicate better storage optimization [5].
  Together, these algorithms ensure that the proposed system can handle both exact and near-duplicate data efficiently and securely. By integrating these techniques, the system optimizes cloud storage, reduces redundancy, and ensures that only the most relevant and unique data is retained.

5. Experimental Setup and Results

5.1 Dataset:

For the experiments conducted, the dataset consists of a variet of textual data typically found in cloud storage environments. The dataset includes documents of different formats such as text files, PDFs, Word documents, and image metadata, representing real-world usage patterns. These documents cover a wide range of topics and sizes, including both small text files and large, complex documents. The data is relevant to cloud environments because it simulates the diverse and large-scale storage challenges that cloud services face. It provides a realistic scenario for testing the efficiency of the proposed deduplication approach in terms of storage savings, processing time, and scalability.

5.2 Evaluation Metrics:

To assess the effectiveness of the proposed deduplication solution, several evaluation metrics were used. These metrics are designed to evaluate different aspects of the system's performance, including its storage efficiency, processing efficiency, and data security.

1. **Deduplication Rate:** The deduplication rate measures the percentage of redundant data that is successfully identified and eliminated by the system. It is calculated by comparing the total data size before and after deduplication. Higher deduplication rates indicate better storage optimization.
2. **Processing Time:** Processing time measures the amount of time required to process a file from tokenization to encryption. It helps assess the efficiency of the deduplication and encryption process. Shorter processing times indicate better performance, especially in cloud environments with large datasets.
3. **Security Measures:** Security is evaluated based on the strength of the encryption, tokenization, and deduplication techniques. Specifically, the system's ability to prevent unauthorized access, ensure data privacy during processing, and perform integrity checks are important considerations. For this, metrics such as encryption strength, hash function collision resistance (CRC), and the effectiveness of the Wagner-Fischer algorithm in preventing data leaks are assessed.
4. **Scalability:** Scalability evaluates the system's ability to handle increasing amounts of data without a significant degradation in performance. It tests how well the system performs as the number of files or the volume of data increases.

$$\text{Deduplication Rate} = \frac{\text{Original Data Size}}{\text{Reduced Data Size}} \times 100$$

6.3 Results:

The experimental results demonstrate the performance of the proposed system compared to existing deduplication techniques in cloud storage. The results are summarized in the following tables:
The results show a significant improvement in deduplication rates, processing times, encryption strength, and scalability. The proposed system achieved an 85% deduplication rate, outperforming the existing system, which only reached 75%. Additionally, the processing time was reduced by 39%, making the proposed system more efficient. The encryption strength was upgraded from AES-128 to AES-256, ensuring better data security. Furthermore, the system demonstrated higher scalability, processing 43% more files per minute compared to the existing system.

6.4 Discussion

The proposed system enhances cloud-based deduplication by offering:

Strengths:

- **Higher Efficiency:** Achieves 85% deduplication, improving storage optimization.
- **Faster Processing:** Optimized workflows enhance speed for large-scale use.
- **Stronger Security:** AES-256 encryption ensures robust data protection.
- **Scalability:** Efficiently handles large volumes of files.

Weaknesses:

- **Setup Complexity:** Tokenization and CRC increase initial configuration steps.
- **Higher Resource Use:** Strong encryption may strain low-end client devices.

Key Improvements:

- **Enhanced Integrity Checks** ensure data authenticity.
- **Secure File Sharing & Viewing** enables collaboration—absent in prior systems.
- **Real-Time Access** to encrypted data is now supported.

Applications:
Ideal for secure, scalable storage in healthcare, finance, and legal sectors where data privacy and integrity are critical.

Table-1.Result Table

Metric	Proposed System	Existing System	Improvement
Deduplication Rate (%)	85%	75%	10%
Processing Time (sec/file)	2.5	4.1	39% faster
Encryption Strength	AES-256	AES-128	Stronger
Security (Leakage Resilience)	High	Medium	Improved
Scalability (Files Processed/Minute)	500	350	43% faster

7. Conclusion

7.1 Summary of Contributions

This work presents a secure and efficient cloud storage system that combines deduplication and encryption to optimize storage and protect sensitive data. Key contributions include:

- **Secure Deduplication:** Uses CRC and the Wagner-Fischer algorithm to remove redundant data.
- **Enhanced Security:** Employs encryption and unique keys for secure access.
- **User-Friendly Design:** Supports easy file upload, viewing, and downloading.

The system addresses major challenges in cloud storage—redundancy, processing time, and security—offering a scalable and privacy-focused solution.

7.2 Future Work

To further improve the system, future efforts may focus on:

- **Scalability** for handling large datasets.
- **Near-Duplicate Detection** using advanced similarity techniques.
- **Machine Learning** for smart deduplication and behavior prediction.
- **Broader File Format Support** including media and industry-specific files.
- **Automatic Format Detection** to enhance user experience

These enhancements will help evolve the system into a robust, all-in-one cloud storage management tool.

ACKNOWLEDGEMENT

We thank **God Almighty** for the blessings, knowledge and strength in enabling us to finish our project. Our deep gratitude goes to our founder **Late. Dr. D. SELVARAJ, M.A., M.Phil.,** for his patronage in completion of our project. We take this opportunity to thank our kind and honourable **Chairperson, Dr. S. NALINI SELVARAJ, M.Com., M.Phil., Ph.D.,** and our **Honourable Director, Mr. S. AMIRTHARAJ, B.Tech., M.B.A** for their support to finish our project successfully. We wish to express our sincere thanks to our beloved **Principal, Dr.C.RAMESH BABU DURAI**

M.E., Ph.D., for his kind encouragement and his interest toward us. We are grateful to **Dr.D.C.JULLIE JOSPHINE M.E., Ph.D., Professor and Head of INFORMATION TECHNOLOGY DEPARTMENT,** Kings Engineering College, for his valuable suggestions, guidance and encouragement. We wish to express our dear sense of gratitude and sincere thanks to our **SUPERVISOR, Saravanan.K.G M.Tech,** Assistant Professor, Information Technology Department. for her internal guidance. We express our sincere thanks to our parents, friends and staff members who have helped and encouraged us during the entire course of completing this project work successfully

REFERENCES

[1] Bellare, M., Keelveedhi, S., & Ristenpart, T. (2013). *Message-locked encryption and secure deduplication.* Advances in Cryptology – EUROCRYPT 2013, Springer, pp. 296–312.

[2] Ghassabi, K., & Pahlevani, P. (2024). *DEDUCT: A Secure Deduplication of Textual Data in Cloud Environments.* Journal of Cloud Computing, 13(2), 98–114.

[3] Gopal, R., & Manogaran, G. (2020). *A secure deduplication system using AES and SHA for cloud storage.* Computers & Electrical Engineering, 83, 106581.

[4] Zhu, B., Li, K., & Patterson, H. (2008). *Avoiding the disk bottleneck in the data domain deduplication file system.* In FAST '08: USENIX Conference on File and Storage Technologies, pp. 1–14.

[5] Halevi, S., Harnik, D., Pinkas, B., & Shulman-Peleg, A. (2011). *Proofs of ownership in remote storage systems.* Proceedings of the 18th ACM Conference on Computer and Communications Security (CCS), pp. 491–500.

[6] Xu, S., Fu, Z., & Liu, L. (2013). *Near-duplicate detection using edit distance in cloud environments.* IEEE International Conference on Cloud Computing Technology and Science, pp. 229–236.

[7] Guo, Z., Zhang, X., & Li, C. (2020). *Efficient data deduplication with improved CRC in secure cloud systems.* Journal of Systems Architecture, 109, 101800.

[8] Yu, S., Wang, C., Ren, K., & Lou, W. (2010). *Achieving secure, scalable, and fine-grained data access control in cloud computing.* IEEE INFOCOM, pp. 1–9.

[9] Wang, C., Chow, S. S. M., Wang, Q., Ren, K., & Lou, W. (2012). *Privacy-preserving public auditing for secure cloud storage.* IEEE Transactions on Computers, 62(2), pp. 362–375.

[10] Harnik, D., Pinkas, B., & Shulman-Peleg, A. (2010). *Side channels in cloud services: Deduplication in cloud storage.* IEEE Security & Privacy, 8(6), pp. 40–47.

[11] Storer, M. W., Greenan, K. M., Long, D. D. E., & Miller, E. L. (2008). *Secure data deduplication.* In Proceedings of the 4th ACM International Workshop on Storage Security and Survivability, pp. 1–10.

[12] Miklau, G., & Suciu, D. (2004). *Controlling access to published data using cryptography.* Proceedings of the 29th VLDB Conference, pp. 898–909.

[13] Liu, F., Tong, J., Mao, J., Bohn, R. B., Messina, J. V., Badger, L., & Leaf, D. M. (2011). *NIST Cloud Computing Reference Architecture.* NIST Special Publication 500-292.

[14] Xia, W., Jiang, H., Feng, D., Huang, H., Hua, Y., & Zhou, Y. (2016). *A comprehensive study of the past, present, and future of data deduplication.* Proceedings of the IEEE, 104(9), pp. 1681–1710.

[15] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). *BERT: Pre-training of deep bidirectional transformers for language understanding.* Proceedings of NAACL-HLT, pp. 4171–4186.

[16] Rabin, M. O. (1981). *Fingerprinting by random polynomials.* Center for Research in Computing Technology, Harvard University.