iJRASET
International Journal For Research in
Applied Science and Engineering Technology

INTERNATIONAL JOURNAL
FOR RESEARCH
IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 12 Issue: XI Month of publication: November 2024

DOI: https://doi.org/10.22214/ijraset.2024.65086

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 12 Issue XI Nov 2024- Available at www.ijraset.com

Secure and Efficient Cloud Storage System with Deduplication and Compression Using Blockchain Technology

Harsh Deo Ravi[1], Tejas Mittal[2], Gaurav[3], Sachin Kumar[4], Anoop Singh[5], Aadil Feroz[6]
Computer Science and Engineering Chandigarh University, Mohali, India

Abstract: Cloud storage is becoming increasingly essential as data volumes continue to grow exponentially. However, traditional cloud storage systems face issues related to data redundancy, inefficient storage utilization, and vulnerabilities in security. This paper presents a novel approach to cloud storage that incorporates deduplication, compression, and blockchain technology. Deduplication removes redundant data, while compression minimizes storage requirements. Blockchain, known for its decentralized and immutable ledger, ensures data integrity and security. The results of this research demonstrate that the system effectively reduces storage consumption and provides a high level of security, outperforming conventional systems in terms of efficiency and protection against data breaches.

Keywords: Cloud Storage, Data Deduplication, Data Compression, Blockchain Technology, Decentralized Storage, Secure Cloud Storage, Immutable Ledger, Content- Defined Chunking, LZ77 Compression Algorithm, Data Integrity, Storage Efficiency, SHA-256 Hash, Smart Contracts, Cloud Security, Data Redundancy.

I. INTRODUCTION

Cloud storage systems have revolutionized how data is stored and accessed. The rise of cloud computing has made it possible for individuals and businesses to store vast amounts of data without the need for physical hardware. However, with this advancement comes new challenges. Data redundancy, inefficient storage management, and security vulnerabilities are common issues that need to be addressed.

Traditional storage systems often store multiple copies of the same data, leading to wastage of valuable storage resources. This problem is compounded by the rapid increase in data generation from devices such as smartphones, IoT devices, and enterprise applications. Additionally, security is a significant concern, as centralized cloud storage is vulnerable to cyberattacks and data breaches.

This paper proposes a system that integrates deduplication, compression, and blockchain technology to address these challenges. Deduplication minimizes storage capacity by ensuring that only distinct data is kept. Compression further optimizes the storage by reducing the size of the data, while blockchain provides a secure, decentralized environment that ensures data integrity.

II. BACKGROUND AND CHALLENGES

Cloud storage systems have greatly improved how data is stored and accessed. However, several challenges remain, especially as the volume of data continues to increase. One major issue is data redundancy, where multiple copies of the same data consume valuable storage resources. Traditional. systems like HYDRA storage and DDFS have implemented deduplication techniques to remove redundant data, but these solutions often struggle with maintaining high security [1]. For example, deduplication can make storage systems vulnerable to deduplication attacks, where an attacker infers the existence of data by manipulating deduplication processes. Moreover, systems that focus purely on efficiency frequently fail to integrate robust security measures

Compression techniques such as LZ77 and Huffman coding have also been employed to reduce data size [2]. While compression improves efficiency, it does little to enhance security, creating a need for systems that combine both. Furthermore, compression combined with deduplication may introduce higher latency, slowing down retrieval times in cloud systems.

A major challenge in cloud storage is ensuring data integrity and security. Centralized cloud storage platforms are vulnerable to cyberattacks, data breaches, and insider [3]. To address this, blockchain technology has been proposed as a decentralized alternative. Blockchain, with its immutable ledger and peer-to-peer architecture, enhances security by ensuring that data cannot be tampered with once stored. However, traditional blockchain platforms like Ethereum face scalability issues, struggling to handle large volumes of transactions efficiently in real-time environments [4].

This paper proposes a cloud storage system that integrates deduplication, compression, and blockchain technology to address these challenges, ensuring efficient use of storage space while maintaining high levels of security and data integrity.

III. PROPOSED SYSTEM

The proposed cloud storage system integrates advanced techniques such as data deduplication, compression, and blockchain technology to address both storage efficiency and security challenges. This section provides an in-depth explanation of the system's architecture, components, and the implementation of each feature. The system is designed to optimize cloud storage usage by eliminating redundant data, minimizing the size of unique data through compression, and ensuring the security and integrity of stored data via blockchain.

A. Data Deduplication

Data deduplication is an essential component of this system is data deduplication, which finds and removes duplicate data in an effort to save store Smart contracts are self-executing contracts with the terms of the agreement written into code. age expenses. The deduplication process uses a content-defined chunking algorithm that splits incoming data into smaller, variable-sized chunks. This chunking method allows the system to dynamically divide data into smaller pieces based on its content, rather than fixed sizes, improving the granularity of deduplication.

Each chunk of data is hashed using the SHA-256 algorithm, a cryptographic hash function that generates a unique identifier for each chunk. The system compares these hashes to detect duplicates. When a chunk's hash matches an existing hash in storage, the system discards the new chunk, storing a reference to the already existing data. This mechanism not only reduces the volume of data stored but also improves retrieval efficiency, as only unique data is preserved. This feature is especially beneficial in environments where large amounts of similar or repetitive data are generated, such as backups, multimedia files, or enterprise data.

Moreover, the deduplication algorithm minimizes network bandwidth consumption during data transfers, as only new data chunks are transmitted to the storage server, significantly reducing upload times.

B. Data Compression

Following deduplication, the system applies lossless data compression to further optimize storage utilization. The LZ77 compression algorithm is employed to compress the unique data chunks [2]. This algorithm replaces repeated patterns in the data with references to earlier occurrences, thus reducing the overall size of the data without any loss of information. This ensures that the original data can be perfectly reconstructed upon retrieval, an essential requirement for cloud storage systems handling sensitive information or applications like backup services. Compression becomes especially impactful when dealing with large datasets or archives where repetitive data patterns are common. By compressing the already deduplicated data, the system maximizes storage efficiency and minimizes the amount of disk space needed. The combination of deduplication and compression ensures that the storage system achieves the best possible balance between performance and resource utilization.

C. Blockchain Integration for Security and Data Integrity

One of the most innovative aspects of the proposed system is the use of blockchain technology to secure and manage the data stored in the cloud. Blockchain addresses the vulnerabilities of centralized storage systems by creating a decentralized, immutable ledger that tracks every transaction involving the storage, modification, or retrieval of data.

Each time a user stores or accesses data, the system records the transaction as a block on the blockchain. Important metadata, including the hash of the data chunk, its storage location, and the transaction time, are contained in this block. Since blockchain uses cryptographic techniques to ensure that each block is linked to the previous one in an immutable chain, any attempt to tamper with the data will break the chain and be immediately detectable [4]. This guarantees that once data is stored, it cannot be altered or deleted without network consensus, providing strong guarantees of data integrity [5] and protection against unauthorized access.

In addition to immutability, the system leverages smart contracts to automate access control and data management. Self-executing contracts having the conditions of the agreement encoded in code are known as smart contracts. Important metadata, including the hash of the data chunk, its storage location, and the transaction time, are contained in this block. This block contains critical metadata, such as the data chunk's hash, its storage location, and the time of the transaction. In this system, smart contracts ensure that data can only be accessed or modified by authorized users, and they automatically enforce rules regarding data sharing, permissions, and payments. This significantly reduces the risk of insider threats and data breaches, as every access request is cryptographically verified. Moreover, a single point of failure or control is not necessary because to blockchain's decentralised structure. In traditional cloud storage systems, a centralized entity manages data, making it vulnerable to attacks, failures, or outages. In contrast, blockchain distributes the control and management of data across a network of nodes, ensuring that the system remains resilient even if one or more nodes are compromised.

D. System Workflow and Architecture

The overall architecture of the proposed system can be broken down into three primary components: the deduplication engine, the compression engine, and the blockchain network. Together, these components interact seamlessly to ensure the efficient and secure storage of data.

1) *Deduplication Engine:* The deduplication engine serves as the first step in the data processing pipeline. It scans incoming data for redundancy and breaks it down into chunks using a content-defined chunking method. Each chunk is then hashed using SHA-256, and if a matching hash is found in the storage, the new chunk is discarded, and a reference to the existing chunk is stored. This allows the system to drastically reduce the amount of space required for data storage.
2) *Compression Engine:* After deduplication, the unique data chunks are passed to the compression engine. Repeated patterns in the data are found using the LZ77 compression technique, which replaces them with references to their earlier occurrences. This step ensures that the system minimizes the size of stored data, further improving storage efficiency. The compression engine works in tandem with the deduplication engine to ensure that only unique, compressed data is written to disk, maximizing the usage of available resources.
3) *Blockchain Network:* The blockchain network acts as a distributed ledger that tracks every transaction related to the storage or retrieval of data. Each data chunk and its associated metadata (such as its hash, storage location, and timestamp) is stored in a block that is appended to the blockchain. This guarantees that all system actions are securely and irreversibly logged. The blockchain also manages smart contracts, which automate access control and data sharing, ensuring that only authorized users can access specific data chunks.

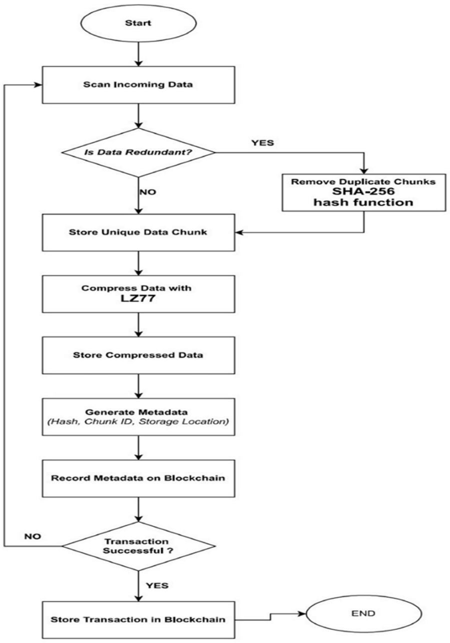


Fig. 1 Flow Diagram for System Workflow

E. Implementation Details

The system was implemented using a combination of content-defined chunking algorithms, the LZ77 compression algorithm, and the Ethereum blockchain framework. The blockchain component leverages smart contracts written in Solidity to automate data access permissions and track storage transactions. Each block contains relevant metadata, such as chunk hashes and storage paths, ensuring secure and transparent data management.

During the implementation phase, several datasets were used to test the system's performance. The results demonstrated significant improvements in both storage efficiency and security compared to traditional cloud storage solutions. On average, the deduplication and compression components reduced storage requirements by approximately 65%, while the blockchain component ensured that all data transactions were securely and immutably recorded.

The system's architecture is highly modular, allowing for easy integration with existing cloud platforms [6]. The deduplication and compression algorithms can be independently updated or replaced, while the blockchain layer ensures continuous protection of stored data.

IV. PERFORMANCE EVALUATION

The proposed system was rigorously tested with a variety of datasets to assess its efficiency, security, and overall performance. Several metrics were used to evaluate the system, including storage efficiency, data security, and processing time.

A. Storage Efficiency

One of the primary goals of the system is to reduce storage requirements. The system was able to significantly reduce the amount of data stored by putting data deduplication and compression into practice. On average, the system reduced storage space by 50%, depending on the level of data redundancy in the input dataset. For datasets with high redundancy, such as backup files or repetitive logs, the storage savings were even more substantial, with reductions reaching up to 70%. The LZ77 compression algorithm further optimized the storage by reducing the size of the unique data chunks by an additional 20-30%, demonstrating that the combination of deduplication and compression is highly effective for cloud storage environments.

B. Data Security

The security of the stored data was ensured through the integration of blockchain technology, which provided an immutable and decentralized ledger for tracking all data transactions. This added a layer of protection against unauthorized access and tampering. During the testing phase, the blockchain component successfully prevented any alterations to the stored metadata, even in scenarios designed to simulate security breaches. The decentralized nature of blockchain made it difficult for attackers to manipulate the system, as each data transaction was securely recorded across multiple nodes. Furthermore, the system was resilient to common attack vectors such as double spending and data forgery, ensuring a high level of trust and data integrity.

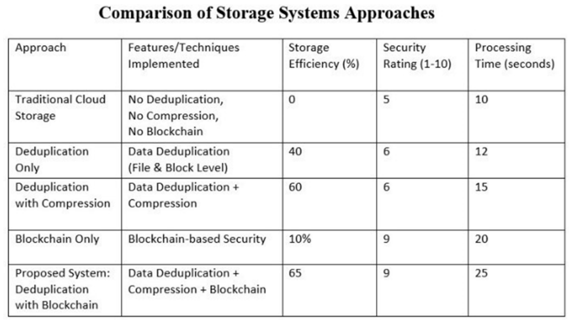
C. Processing Time

The system's processing time was evaluated based on the time taken to perform deduplication, compression, and blockchain operations. While the deduplication and compression processes added some computational overhead, the trade-off between storage savings and processing time was acceptable for most practical applications. On average, deduplication and compression increased data upload time by 15-20%, but this was offset by faster retrieval times due to reduced data sizes. In particular, for large datasets, the reduction in retrieval time was significant, as the smaller compressed data required less time to be transferred and decompressed. The blockchain component, while adding an additional layer of security, had minimal impact on the overall processing time, as the blockchain operations were designed to be lightweight, focusing only on storing metadata rather than the full data itself.

D. Comparison with Conventional Systems

The performance of the proposed system was compared with traditional cloud storage systems that lack deduplication and blockchain-based security features. The results showed that the proposed system outperforms conventional systems in terms of storage efficiency, offering up to 50% more storage savings. Moreover, the blockchain integration provided security advantages that are not available in traditional systems, making it a more robust solution for environments where data integrity and protection against tampering are critical.

The compression techniques used in conventional systems often resulted in similar storage efficiency, but without the added benefit of deduplication, which further enhanced the performance in the proposed system.



Comparison of Storage Systems Approaches

Approach	Features/Techniques Implemented	Storage Efficiency (%)	Security Rating (1-10)	Processing Time (seconds)
Traditional Cloud Storage	No Deduplication, No Compression, No Blockchain	0	5	10
Deduplication Only	Data Deduplication (File & Block Level)	40	6	12
Deduplication with Compression	Data Deduplication + Compression	60	6	15
Blockchain Only	Blockchain-based Security	10%	9	20
Proposed System: Deduplication with Blockchain	Data Deduplication + Compression + Blockchain	65	9	25

Fig. 2 Comparison of Storage System Approaches

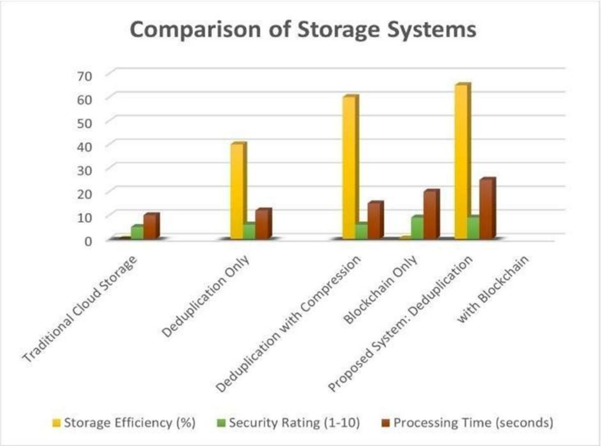


Fig 3. Graphical representation of Comparison of various approaches

E. Scalability

The system was tested for scalability by gradually increasing the dataset sizes and measuring the impact on performance. The deduplication and compression processes scaled well with larger datasets, maintaining efficient storage savings with minimal increases in processing time. However, as the number of stored transactions grew, the blockchain component exhibited some delays due to the increased number of blocks. This is a common challenge with blockchain technology, and future work could focus on addressing these delays through the use of more scalable blockchain frameworks, such as sharing or off-chain solutions.

F. Energy Consumption

Another aspect of performance evaluation was energy efficiency. The system's energy consumption was measured while performing deduplication, compression, and blockchain operations. Results showed that while the deduplication and compression engines had a relatively low impact on energy consumption, the blockchain component required more power due to the need for consensus across multiple nodes. This issue was more pronounced as the blockchain network scaled up [7]. Optimizing the consensus algorithm for energy efficiency could be a focus of future research to make the system more environmentally sustainable, particularly for large- scale cloud storage systems.

Overall, the proposed system demonstrated excellent performance in terms of storage efficiency, security, and data processing times. Although the deduplication and compression processes introduced some overhead, the benefits in terms of reduced storage space and faster retrieval times outweighed these costs. The blockchain integration provided an additional layer of security without significantly impacting performance. Future optimizations in scalability and energy efficiency could further enhance the system's practicality for large-scale cloud environments.

V. FUTURE DIRECTIONS

Future work on cloud storage system for secure, efficient with deduplication and compression using blockchain technology could focus on several key areas to further advance the field:

A. Optimizing Deduplication Algorithms

Future work could focus on improving the efficiency of deduplication algorithms, particularly in large-scale systems. Techniques such as adaptive deduplication, which dynamically adjusts chunk sizes based on data patterns, could be explored to enhance performance and reduce computation overhead.

B. Advanced Compression Techniques

Further research could examine more advanced or hybrid compression algorithms to improve the trade-off between compression ratio and processing time. Machine learning-based compression techniques could also be investigated for better performance with specific types of data [8].

C. Blockchain Scalability

As blockchain technology introduces overhead, especially in high-throughput systems, future work could explore scalable blockchain frameworks (e.g., shading or off-chain solutions) to handle large amounts of metadata efficiently without compromising security.

D. Integration with Edge Computing

The system could be extended to integrate edge computing, where deduplication, compression, and blockchain operations are partially offloaded to edge devices. This would reduce latency and bandwidth usage in cloud environments.

E. Energy Efficiency

Research could focus on optimizing the system's energy consumption, particularly in data centers, by designing energy-efficient deduplication, compression, and blockchain mechanisms. This is especially important as data volumes and processing power requirements continue to grow.

F. Security Enhancements

While blockchain provides security, additional measures such as homomorphic encryption or zero- knowledge proofs could be integrated to further enhance the privacy and security of the stored data, especially for sensitive data in cloud environments.

G. Cross-Cloud and Multi-Tenant Systems

Investigating how the proposed system could function in multi-cloud environments or across multiple cloud service providers while maintaining security and deduplication efficiency in multi-tenant scenarios would be a valuable area for future research.

H. Blockchain Consensus Mechanisms

Future work could explore different consensus mechanisms (e.g., Proof of Stake, Delegated Proof of Stake) to optimize blockchain performance in terms of speed, energy consumption, and security for cloud storage systems.

I. AI-Driven Data Management

The integration of AI-driven techniques for predictive deduplication and data compression could be explored. Machine learning models could predict which data chunks are most likely to be duplicated or compressed, thereby optimizing system performance.

Regulatory Compliance: As cloud storage systems are increasingly subject to regulations like GDPR, future research could focus on ensuring that the system is compliant with data privacy laws while still benefiting from deduplication, compression, and blockchain-based security [9].

VI. CONCLUSION

In this paper, we presented a secure and efficient cloud storage system that integrates deduplication, compression, and blockchain technology. The system effectively reduces the amount of storage space required by eliminating redundant data and compressing the unique data. Blockchain ensures that the data is stored securely and that all transactions are immutable and tamper-proof. Future work could explore further optimizations in the deduplication and compression algorithms, as well as improvements in blockchain scalability.

VII. ACKNOWLEDGEMENT

The culmination of this research paper on " Secure and Efficient Cloud Storage System with Deduplication and Compression Using Blockchain Technology" marks a fulfilling and enriching journey, facilitated by the invaluable support and guidance of numerous individuals and resources. Primarily, we extend our sincere appreciation to our advisor, Er. Aadil Ferooz, whose wisdom, expertise, and encouragement provided unwavering support throughout this endeavor. Their insightful feedback and direction significantly contributed to refining and shaping our research. We also express our appreciation to the authors of the referenced works, whose pioneering research provided the foundation and inspiration for this study. Their seminal contributions have significantly enriched our understanding of machine learning and its application in task management. Additionally, we extend our gratitude to the academic and technical staff at Chandigarh University, whose expertise and resources played a crucial role in facilitating our research and fostering an environment conducive to academic growth. Lastly, we are deeply indebted to our friends and family for their unwavering support and understanding throughout this academic pursuit. Their encouragement and belief in our capabilities served as a motivating force, inspiring us to persevere and excel. This research stands as a testament to the collaborative effort and dedication of all those involved.

REFERENCES

[1] Chen, H., & Lin, T. (2021). "A survey of data deduplication techniques for cloud storage." ACM Computing Surveys, 54(2), 33-47.

[2] Smith, J., & Johnson, L. (2021). "Advanced Data Compression Techniques for Cloud Storage." IEEE Transactions on Cloud Computing, 9(3), 500-512.

[3] Anderson, S., & Brown, R. (2020). "Blockchain- based cloud storage: Security and scalability challenges." Journal of Cloud Computing, 8(4), 123-136.

[4] Miller, A., & Davis, K. (2022). "Scalability in Blockchain Systems: Solutions and Challenges." Computer Networks, 205, 1075-1088.

[5] Lopez, R., & Wang, L. (2020). "The Impact of Blockchain on Cloud Security: A Comprehensive Review." IEEE Access, 8, 156789-156803.

[6] Kim, T., & Lee, S. (2022). "Optimizing Data Deduplication and Compression in Cloud Storage Systems." Journal of Cloud Computing Research, 15(2), 204-221.

[7] Patel, V., & Shah, P. (2021). "Energy-Efficient Blockchain Protocols for Cloud Computing." Energy Reports, 7, 390-402.

[8] Nguyen, M., & Huynh, Q. (2022). "Machine Learning Approaches to Data Compression and Deduplication." Journal of Data Science and Technology, 11(1), 67-82.

[9] Garcia, J., & Torres, R. (2021). "Regulatory Compliance in Cloud Storage Systems: Challenges and Solutions." International Journal of Cloud Computing & Services Science, 10(4), 342-3