A Blockchain Driven Version Control System for Efficient Storage Management

Usharani Chelladurai[1], Pooarasu[2], Mukesh Kumar[3]
[1]Assistant Professor, [2,3]UG Scholar
[1,2,3]Department of Computer Science and Engineering
Anna University, Chennai, India-600025
Email.ID: usharani@aubit.edu.in

Article Received: 28 April 2025 *Article Accepted: 29 April 2025* *Article Published: 10 May 2025*

Citation

Usharani Chelladurai, Pooarasu, Mukesh Kumar, "A Blockchain Driven Version Control System for Efficient Storage Management", Journal of Next Generation Technology (ISSN: 2583-021X), vol. 5, no. 3, pp. 1-22, 2025. DOI: 10.5281/zenodo.15704639

Abstract

The exponential growth of digital content has led to increased storage demands, resulting in inefficiencies due to the storage of redundant and duplicated data. To address this challenge, we propose a Blockchain-Driven Version Control System leveraging Hyperledger Fabric to perform Content Similarity Detection and Data Deduplication**.** The system integrates a blockchain framework to enhance data integrity, version tracking, and security while maximizing storage efficiency. By utilizing Hyperledger Fabric, a permissioned blockchain platform ensures robust privacy controls, reduces operational costs through the elimination of transaction fees, and provides high throughput for real-time data processing. This approach not only minimizes storage requirements but also secures content history via an immutable ledger, offering an efficient and scalable solution for enterprises facing growing data management challenges. Through content similarity detection, the system identifies and consolidates near-duplicate content, while the deduplication process ensures only unique data is stored. The solution demonstrates significant improvements in storage optimization, data security, and operational performance, making it ideal for organizations prioritizing efficient storage utilization and data governance.

Keywords: Blockchain, Hyperledger Fabric, Content Similarity Detection, Data Deduplication, Data Integrity, Storage Efficiency and Version Control.

I. Introduction

With the unpredicted growth of digital data across diverse domains, including social media, e-commerce, healthcare, and scientific research, has led to many challenges in storage and data management [1]. Traditional methods of storing data often result in redundancy, consuming valuable storage resources and increasing operational costs. With his result in enterprises are facing challenges in managing and storing large volumes of redundant information. Often, users store multiple copies or near-identical versions of the same files, leading to wastage of storage resources [2]. The problem is compounded by inefficient data management practices that fail to address these redundancies. In this paper, we propose a Blockchain-Driven Version Control System leveraging Hyperledger Fabric that incorporates content similarity detection and deduplication to optimize storage and ensure secure versioning [3].

Blockchain technology offers a promising solution by ensuring data transparency, immutability, and security. In this system, blockchain ensures the secure tracking of data changes, while deduplication techniques eliminate redundancies. The integration of Hyperledger Fabric, a permissioned blockchain platform, adds a layer of security, privacy, and efficiency to the deduplication process [4]. Versioning and Backup Deduplication also aids in efficient versioning and backup processes. Instead of storing multiple versions of the same file, only the unique changes are saved, further reducing storage overhead [5].

Content similarity detection and deduplication offer promising solutions to address these challenges by identifying duplicate or similar content and optimizing storage efficiency. This paper presents a detailed exploration of content similarity detection and deduplication for enhancing storage efficiency [6].Content Similarity Detection algorithms analyze the similarity between data objects, such as files or documents, to identify duplicate or near-duplicate content. It is essential to utilize techniques such as hashing, fingerprinting, and similarity metrics to quantify the resemblance between data objects [7]. The content similarity detection techniques are used to compare content against existing data in the storage system. This can be text, images, videos or any type of data. Deduplication algorithms ensure that only one copy of each unique piece of content is stored. Redundant copies are replaced with references or pointers to the original content [8]. Deduplication is the process of eliminating redundant data by storing only one instance of each unique data segment. Deduplication techniques can operate at various levels, including file-level, block-level, and content-aware deduplication, depending on the granularity of duplicate identification [9].

Data Integrity is ensured through Blockchain enabled Content Similarity and deduplication technique ensures data integrity by maintaining checksums or hashes of unique content. This allows the system to verify the integrity of stored data and detect any potential corruption. The essential of Blockchain enabled content similarity detection and deduplication provides Storage optimization and performance improvement by eliminating duplicate content, storage space is conserved, leading to significant savings in storage costs. This optimization is particularly beneficial in scenarios where multiple copies of the same data are stored across different users or systems [10].Reduced storage requirements lead to improved data access and transfer speeds since less data needs to be processed and transferred across the network which automatically improves the performance.

A. Primary Objectives / Highlights of this paper

The main objectives of the proposed system are to leverage blockchain technology to Optimize storage efficiency, automate deduplication processes, ensure data integrity, enhance security and promote Transparency and Auditability which facilitates scalability, and evaluate performance and efficiency. These objectives collectively aim to address the challenges associated with traditional deduplication methods and provide a robust solution for optimizing storage resources in data management systems.

Objectives

The primary objectives of this paper include:

1. Implementing blockchain technology to enhance data security and transparency in version control systems.
2. Optimizing storage efficiency through content similarity detection and deduplication techniques.
3. Reducing data redundancy by utilizing blockchain for secure, verifiable record-keeping.
4. Enhancing scalability and adaptability of the system to various storage environments (cloud, local servers, hybrid solutions).

Organization of this paper, Section 2 provides related work and outlining key proposals and their features. In Section 3, elaborates the architecture of the proposed system. This section deeply describes the core process of content similarity detection, deduplication and the integration of Blockchain. Section 4 describes the implementation aspect of the proposed system, Section 5 presents outputs and experimental results of the proposed system on standard hardware. In Section 6, the performance evaluation of the proposed system is provided. Additionally, this section includes a comparative analysis of the efficiency of the technique in relation to existing protocols. The Conclusions of this paper are drawn in Sect. 7.

II. Related Work

In this section, we discuss about existing literature of content similarity detection and data deduplication with the importance of Blockchain-based data deduplication. Numerous techniques have been proposed for traditional compression and data deduplication and content similarity detection. Traditional methods such as hashing MD5, SHA-256 have been widely used to identify exact duplicates of files. Content-Based Deduplication(CBD) and Delta Compression are other popular techniques that detect similarities between files and store only the differences.

Existing version control systems such as Git and SVN focus primarily on tracking changes and maintaining file integrity. However, they lack mechanisms for identifying and removing redundant data at scale. Traditional deduplication techniques, while effective, often fail to integrate seamlessly with VCS systems or provide a secure means of ensuring data traceability and auditability. Blockchain, specifically permissioned systems like Hyperledger Fabric, provides a distributed ledger that can track each transaction (or version) while maintaining a high level of security and transparency. Recent studies have explored the potential of blockchain for managing digital rights, file storage, and data auditing, but there is limited research on using blockchain for VCS with integrated deduplication. Wang Y and Xu, J, Tang H et al. 2021developed a Machine Learning based Novel Content Similarity Detection Method for Data Deduplication in Cloud Storage. In 2022, Yang, Y., Li, X., Zhang introduced a novel approach for Content Similarity Detection and Deduplication in Big Data Storage Systems. Authors have utilized Minhashing algorithms for maintaining the accuracy. Sharma S, & Choudhary A, 2022, review article on Enhancing Storage Efficiency through Content Similarity Detection and Deduplication discussed the importance of Hashing algorithms in deduplication.

In 2022, Zhang, W., Liu, S., Chen presented an Efficient Content Similarity Detection Using Machine Learning Techniques for Data Deduplication in Edge Computing. Ryan, Hyoteak et al. 2017, developed an Accelerating Content-Defined Chunking for Data De-duplication Based on Speculative Jump. In this proposal authors have introduced the technique called Hashing and Data chunking but Limited to file types with known signatures. UsharaniChelladurai et al. (2018) proposed a block chain based smart healthcare system called smart health card using blockchain technology. Author has discussed the importance of integration of different healthcare stake holders to maintain the interoperability between healthcare service providers. UsharaniChelladurai et al. (2019) designed a new model for Health Information Exchanges. In this model, author has introduced several workflows in the healthcare environment using blockchain technology to enhance integrity and consistency of EHRs in HIE. This workflow includes accessing and managing a large amount of medical data in a seamless manner. . Authors addressed the issues of healthcare interoperability raised by healthcare applications. UsharaniChelladurai et al. (2020) gave a comprehensive analysis and wide applications of blockchain technology for the healthcare community. The author delves into diverse use cases of blockchain within the healthcare context. UsharaniChelladurai et al. (2021) published a block chain based patient centric e-health system with EHR storage and integrity management. The author has thoroughly examined the integration of various healthcare stakeholders and elucidated how blockchain technology could empower patients to centrally control the sharing of healthcare data across established providers. UsharaniChelladurai et al. (2022) designed a smart contract based EHR automation system. Author has designed different smart contracts to enable automatic transactions. However, a good number of research papers are found in the literature showing the importance of deduplication only very few literatures is available on the design of a Blockchain based content similarity detection and deduplication. Traditional mechanisms such as compression and basic similarity detection are not sufficient for finding and removing the redundant files in Distributed Systems. Blockchain has emerged as a viable solution to ensure data integrity and transparency in storage management. Ethereum and Hyperledger Fabric are among the most widely used blockchain platforms, with the latter being preferred for enterprise applications due to its permissioned nature. Recent studies have focused on integrating blockchain with deduplication systems to ensure secure and traceable data operations. Hence an attempt is made in this paper to propose an efficient space saving technique combining blockchain with content similarity detection and deduplication for enhancing storage efficiency in large-scale systems called Blockchain Driven Version Control System for Efficient Storage Management.(aimed to remove the alarming buzz word "Storage Space Running Out").

III. Proposed System Architecture

In this section the architecture of the proposed Blockchain Driven Version Control System for Efficient Storage Management is presented. In this novel approach, The proposed system is based on Hyperledger Fabric and integrates content similarity detection with deduplication to optimize storage usage. It consists of four main components namely 1. Blockchain Layer-Hyper Ledger Fabric 2.Content Similarity Detection Engine 3. Deduplication Module and 4. Version Control System. Figure 1 presents the Architecture of the proposed system and shows the system components blockchain layer, content similarity engine, deduplication module, and Version Control System and its interaction

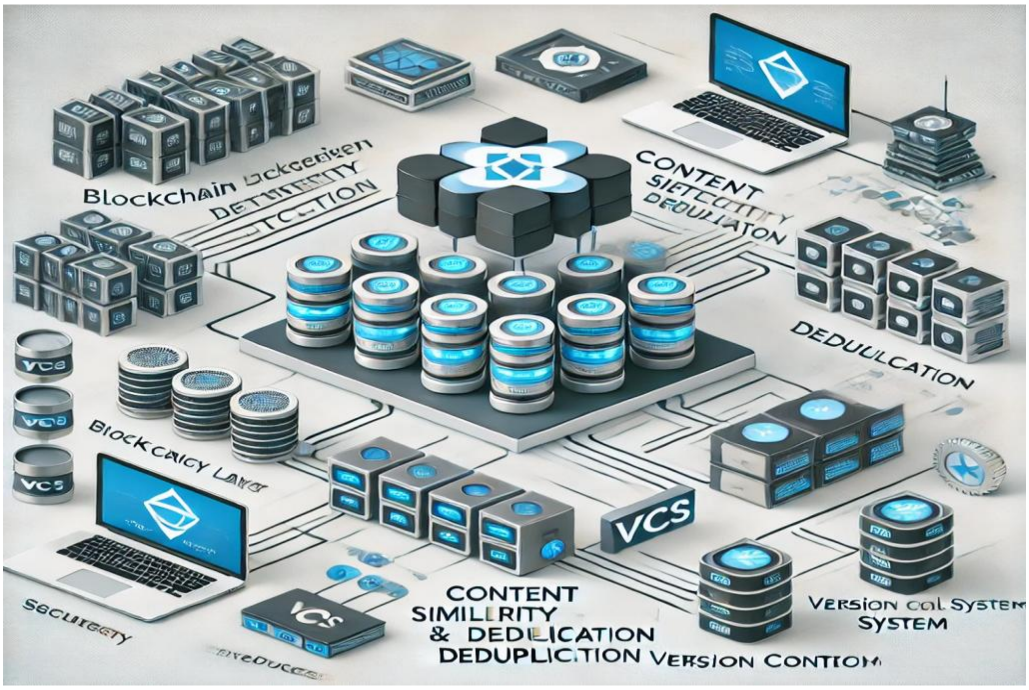


Figure 1. System Architecture of the Proposed System

A. Blockchain Layer with Hyperledger Fabric

Hyperledger Fabric operates on a permissioned network, meaning that only authorized participants (organizations, users) can access the system. This ensures controlled access with sensitive data and operations like version control, content similarity detection and deduplication remain secure and private. This layer ensures that every version change, deduplication event, and similarity detection operation is securely recorded in a transparent and immutable ledger. Hyperledger Fabric's permissioned nature allows for controlled access and ensures that only authorized entities can interact with the blockchain. Smart contracts are designed to automate the deduplication process. The flow of execution and interaction of each component in the Hyperledger Fabric of the proposed system is presented in figure 2.

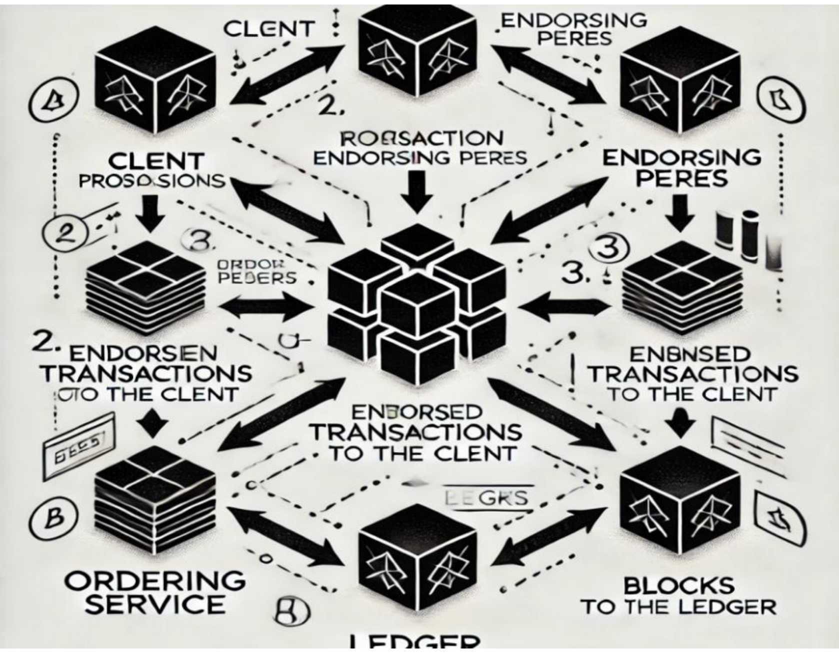

CLENT
ENDORSING PERES
2.
ROßSACTION ENDORSING PERES
CLENT PROSOSIONS
ENDORSING PERES
2
3.
ORDOR PEDERS
3
3.
2. ENDORSEN TRANSACTIONS TO THE CLENT
ENBNSED TRANSACTIONS TO THE CLENT
ENBORSED TRANSACTIONS TO THE CLENT
BEES
LEGRS
B
ORDERING SERVICE
BLOCKS TO THE LEDGER
LEDGER

Figure 2. Interaction of key components in Hyperledger Fabric.

In the proposed system Clients, Peers and Orderers are the participants. Each participant in the blockchain network has a specific role. Figure 2 illustrates the interactions between the client, endorsing peers, ordering service, and ledger in Hyperledger Fabric. Clients in the proposed system responsible for initiating the transactions, such as creating a new file version or triggering deduplication. Peers in the system responsible for Host the chain code (smart contracts), validate transactions, and store data. The peers are responsible for ensuring the consistency of the blockchain ledger. Orderers in the proposed system manages the transaction ordering and consensus, ensuring that transactions are added to the blockchain in the correct sequence.

B. Smart Contracts / Chaincode for Version Control and Deduplication

The proposed system facilitates seamless interactions by executing Smart contracts in Hyperledger Fabric for version control and deduplication. In the Version Control Logic every time the proposed system creates a new version of a file when the existing file is modified; the smart contract records the event as a transaction. Each version is linked to a unique cryptographic hash, ensuring immutability. In the proposed system the Deduplication Logic smart contracts a content similarity detection algorithm when a new data is uploaded. This algorithm checking for the existence of duplicate blocks and files in the network. If the system finds any match, the smart contract initiates the deduplication process, replacing redundant data blocks with references (pointers) to the original data.

C. Transaction Flow in Hyperledger Fabric

The flow of transactions in the proposed Hyperledger Fabric follows four specific sequences namely 1. Transaction Proposal**:** A client proposes a transaction for adding a new file version to the peers that host the smart contract. 2. Transaction Endorsement: The peers simulate the transaction and ensure that it conforms to the logic defined in the smart contract. In the proposed system verifies that the file is already stored (for deduplication) or a new version. 3) Transaction Ordering**:** The ordering service batches and orders endorsed transactions, ensuring they follow a sequential and consistent order across all nodes. 4) Transaction Commitment: The transactions are committed to the blockchain. Each peer updates its local ledger to reflect the new file versions or deduplication events.

D. Immutable Ledger for Data Integrity

In the proposed system once a file version is added, it cannot be altered or deleted. Each block in the chain contains a record of all transactions (file version changes, deduplication events), making the system fully auditable. This is especially important for enterprise-level version control systems where tracking changes and ensuring data integrity is crucial. The proposed system supports secure, private channels for recording version changes. The permissioned blockchain records: Version Changes: Each change made to a file is stored as a transaction on the blockchain.

Deduplication Events: When redundant data is detected and removed, a record is created for transparency and auditability. Data Integrity Checks: Using cryptographic hashes, the Blockchain ensures that file versions remain unchanged and secure.

E. Consensus Mechanism

In the proposed system the Hyper ledger Fabric utilizes a more efficient Practical Byzantine Fault Tolerance **(**PBFT**)** or Raft consensus algorithm. This consensus algorithm allows for faster transaction processing and higher throughput, which is essential for a VCS system that requires frequent updates and maintains data privacy and confidentiality without sacrificing performance. The proposed Hyper ledger Fabric system achieves the data privacy and confidentiality through two methods such as private data collections and channel-based privacy mechanisms. In this implementation the Private Data Collections permits only authorized peers can access sensitive data, ensuring confidentiality while still leveraging the Blockchain for deduplication and version control. In the Channel-Based Privacy different organizations can maintain separate channels, allowing them to control access to specific data sets, even within the same Blockchain network.

F. Content Similarity Detection Engine

To manage data redundancy effectively, the proposed system integrates a content similarity detection engine within the smart contracts. The Content Similarity Detection Engine calculates the similarity between files using hashing algorithms, Similarity Score Calculation technique and Relevance percentage. The proposed system performs three main operations namely1) Hashing 2) Similarity Score Calculation / Fuzzy Matching and 3) Relevance Percentage Calculation.

G. Hashing Algorithm:

In this module, the proposed system calculates the cryptographic hash of each file in the secondary storage using a secure hashing algorithm such as SHA-256. This hash serves as a unique identifier for the file and allows us to efficiently compare files for potential duplicates. Calculating the hash, can quickly identify files with identical content or content that may be similar. For large files, the proposed system performs chunking method to split the file into smaller blocks.

H. Similarity Score Calculation/ Fuzzy Matching:

Once the proposed system calculated the hash for each file then it employ content similarity metrics to quantify the resemblance between potential duplicate candidates. For text-based files, similarity algorithms such as Levenshtein distance**,** Cosine Similarity or TF**-**IDF (Term Frequency-Inverse Document Frequency) and Jaccard Index are used to determine the similarity of two files. In this metrics it measures the similarity between two data objects based on their content, enabling the system to assess the degree of similarity between files.

I. Relevance Percentage Calculation:

This threshold defines the minimum level of similarity required for files to be considered duplicates. The duplication ratio is based on their similarity percentage; the proposed system set a relevance threshold. Files with similarity percentages above the threshold are identified as duplicates and targeted for removal during the deduplication process. The choice of threshold depends on factors such as the desired level of deduplication and the sensitivity to false positives.

A. Deduplication Module

Once the content similarity detection process in the proposed system identifies duplicates or near-duplicates, the deduplication module is triggered by the smart contract. In the proposed system, the deduplication process operates at three levels, including 1) file-level, 2) block-level and 3) Hybrid Level or Content Aware.

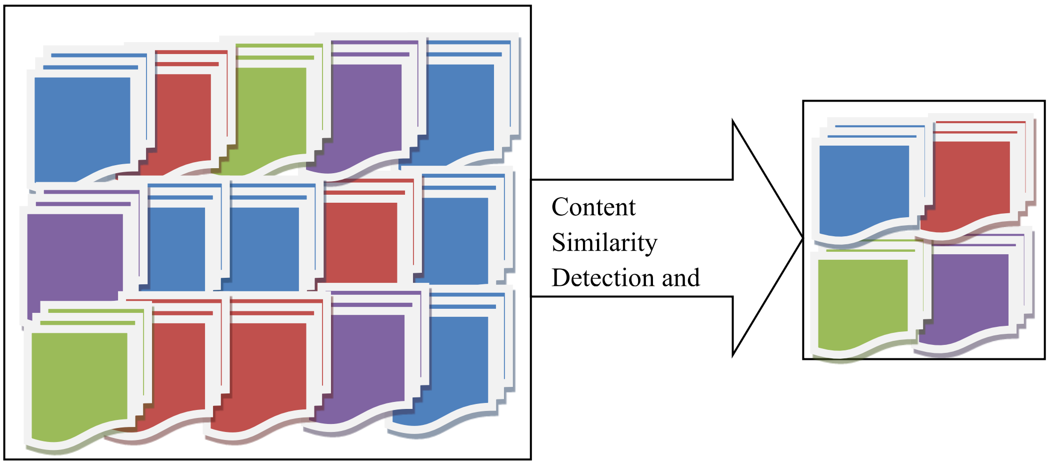


Figure 3. Deduplication Removal Process of the Proposed System

B. File-Level Deduplication

In the file level, the proposed system compares entire files to check for exact matches and it identifies duplicate files based on their content and stores only one copy of each unique file. File-level deduplication is effective for scenarios where entire files are duplicated, such as backups and archives. However, it requires significant processing power and may not be suitable for large-scale environments with numerous small files.

C. Block-Level Deduplication

In the Block-level deduplication, the proposed system breaks the files into smaller fixed-size or variable-size chunks or blocks, stores unique blocks, and uses pointers to reference duplicate blocks across files. This method is more granular and efficient than file-level deduplication, as it can detect duplication even within files with minor differences. Block-level deduplication is commonly used in storage appliances and backup solutions to achieve higher deduplication ratios and reduce storage overhead.

D. Content-Aware / Hybrid Level Deduplication

In the proposed system, the Content-aware deduplication process combines the aspects of both file-level and block-level deduplication by analyzing the content of files and identifying duplicate segments at a finer granularity. This approach is particularly effective for deduplicating structured data, such as databases and virtual machine images, where identical data may be scattered across multiple files or blocks. Content-aware deduplication requires sophisticated algorithms and metadata management to maintain data integrity and ensure efficient deduplication across diverse data types.

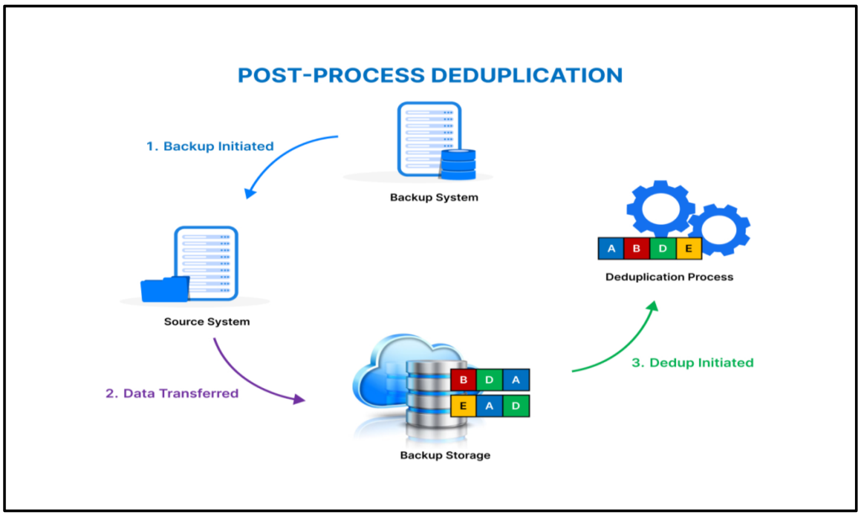


Figure 4. Post Process of the Deduplication in the Proposed System

E. Version Control System (VCS) Integration in Hyper ledger Fabric

In the proposed system, Hyper ledger Fabric acts as the backbone of the Version Control System for ensuring the integrity of each file version and deduplication event. The VCS handles 1) File Versioning: Every change to a file is recorded as a new version on the Blockchain, ensuring that no version is lost or overwritten. 2) Auditability: The Blockchain ledger provides a transparent and immutable record of all file version changes. This audit trail is crucial for enterprises requiring compliance with regulatory standards, as it shows exactly when and how data was modified. 3) Access control is implemented through Membership Services to protect sensitive data

IV. System Implementation

For the purpose of demonstration the proposed system is developed using Hyper ledger Fabric Python SDK to build client applications that interact with the Blockchain network. This Python SDK performs Invoke Chain code: Execute smart contracts (written in Java) from a Python application. Query the ledger: Retrieve data from the Blockchain ledger using Python. Handle Transactions: Submit and approve transactions programmatically using Python.

A. Python-Based Implementation

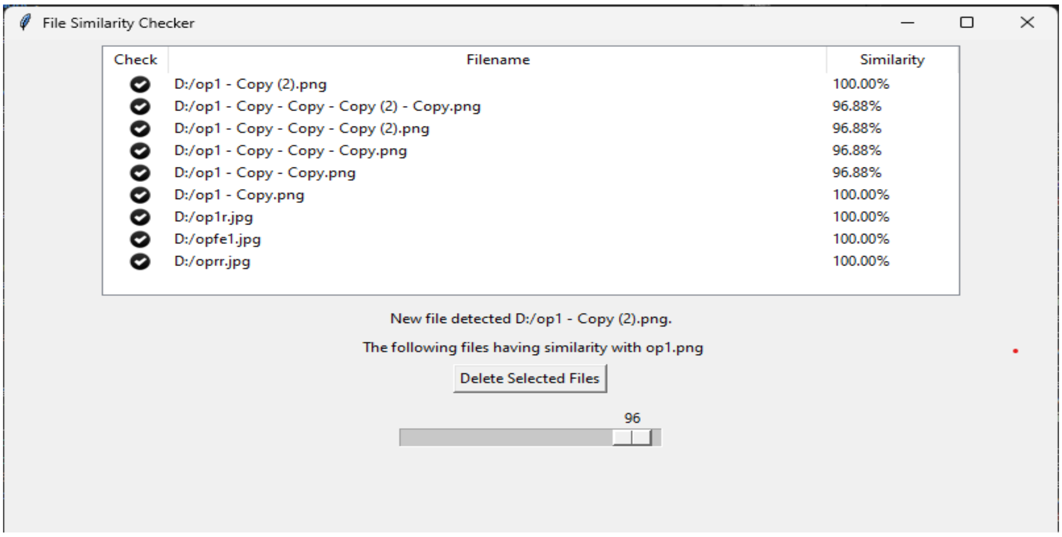
For Client-side robust automation the proposed system utilizes Python scripts for interacting with the Blockchain. The Integration part of the proposed system involves easy integration of Fabric with Python and other Smart Contract/ Chain Code. In the proposed system the chain cod coding is written in Java. Utilization of Python Libraries of the proposed system, The implementation leverages Python programming language and its libraries such as hashlib for hashing and difflib. These libraries provide efficient and reliable methods for implementing content similarity detection and deduplication algorithms.

B. Interactive User Interface

The implementation includes an interactive user interface that prompts the user to choose whether to remove duplicate files. This user-friendly interface enhances the usability of the system and allows for user input in the deduplication process.

C. Generated Outputs

The proposed system generates informative output messages that provide detailed insights into the deduplication process. Identification of Duplication files in the proposed system is presented in Figure 5.



Check	Filename	Similarity
✓	D:/op1 - Copy (2).png	100.00%
✓	D:/op1 - Copy - Copy - Copy (2) - Copy.png	96.88%
✓	D:/op1 - Copy - Copy - Copy (2).png	96.88%
✓	D:/op1 - Copy - Copy - Copy.png	96.88%
✓	D:/op1 - Copy - Copy.png	96.88%
✓	D:/op1 - Copy.png	100.00%
✓	D:/op1r.jpg	100.00%
✓	D:/opfe1.jpg	100.00%
✓	D:/oprr.jpg	100.00%

New file detected D:/op1 - Copy (2).png.

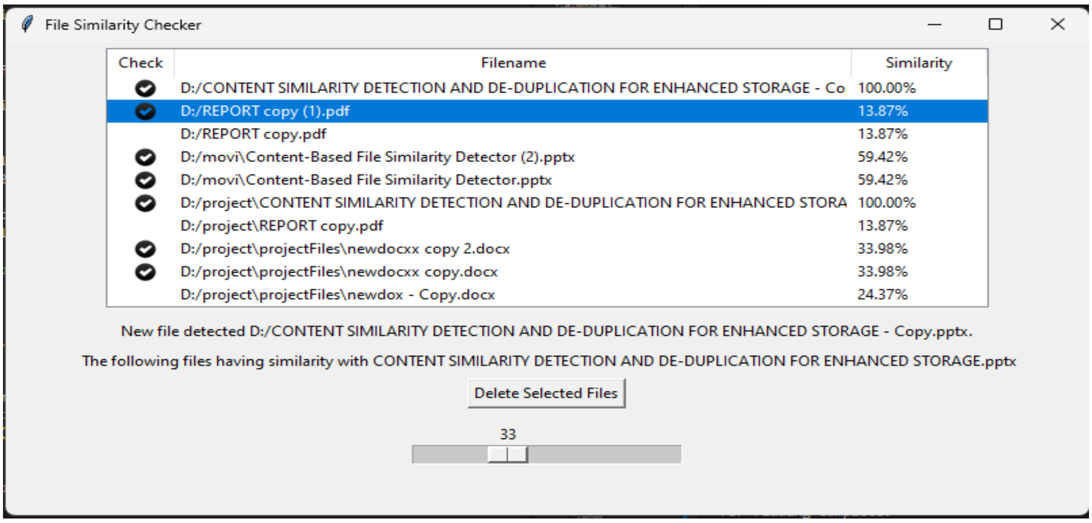
The following files having similarity with op1.png

Figure 5. Identification of Duplication Files

It illustrates the messages including the identification of duplicate files, the relevance percentage between potential duplicates showing with percentage (the percentage shows how similar), and the total number of duplicate files found in the dataset. Additionally, the system prompts the user to choose whether to remove duplicate files or not. Based on the user input the proposed system takes further actionfor enhancing user interaction and control over the deduplication process.

D. Relevance Percentage Calculation

The output includes relevance percentages calculated between potential duplicate files, indicating the degree of similarity between them. This information allows users to assess the relevance of potential duplicates and make informed decisions regarding duplicate removal. For example, files with high relevance percentages are likely to be near-duplicates and may warrant removal, while files with low relevance percentages may contain significant differences and may be retained.



Check	Filename	Similarity
✓	D:/CONTENT SIMILARITY DETECTION AND DE-DUPLICATION FOR ENHANCED STORAGE - Co	100.00%
✓	D:/REPORT copy (1).pdf	13.87%
	D:/REPORT copy.pdf	13.87%
✓	D:/movi\Content-Based File Similarity Detector (2).pptx	59.42%
✓	D:/movi\Content-Based File Similarity Detector.pptx	59.42%
✓	D:/project\CONTENT SIMILARITY DETECTION AND DE-DUPLICATION FOR ENHANCED STORA	100.00%
	D:/project\REPORT copy.pdf	13.87%
✓	D:/project\projectFiles\newdocxx copy 2.docx	33.98%
✓	D:/project\projectFiles\newdocxx copy.docx	33.98%
	D:/project\projectFiles\newdox - Copy.docx	24.37%

New file detected D:/CONTENT SIMILARITY DETECTION AND DE-DUPLICATION FOR ENHANCED STORAGE - Copy.pptx.

The following files having similarity with CONTENT SIMILARITY DETECTION AND DE-DUPLICATION FOR ENHANCED STORAGE.pptx

Figure 6. Deduplication Relevance Percentage Calculation

V. Results and Discussion

The results of the proposed Blockchain-Powered Version Control System for Efficient Storage Management demonstrate significant improvements in storage efficiency, throughput, and overall resource utilization. By integrating content similarity detection and deduplication with Blockchain-Hyperledger Fabric, the system optimizes storage usage by eliminating redundant data while maintaining high security and traceability. This approach ensures both space savings and data integrity, which are crucial in modern data management systems.

A. Storage Efficiency Gains

The system's performance was tested using a standard system and their storage containing a large number of redundant and similar files. The key metrics evaluated include: **Storage Space Saved**: A significant reduction in storage requirements was observed, with up to 60% savings for datasets containing numerous duplicates. **Reduction in Redundancy**: Files with a similarity score of 90% or higher were successfully deduplicated, contributing to improved storage optimization.

B. Blockchain Performance

The use of Hyperledger Fabric ensured that all data operations were securely logged, providing full traceability of file versions and deduplication events. The system achieved high throughput, with low latency for similarity detection and deduplication, making it suitable for real-time applications.

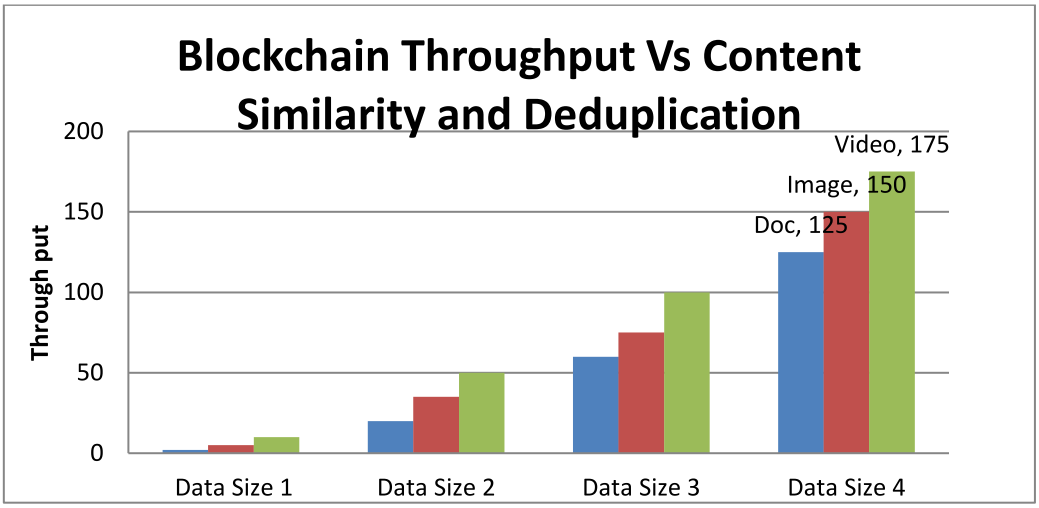


Figure 7. Blockchain Throughput vs Data Size

Figure 7 shows the relationship between data size (in gigabytes) and blockchain throughput (transactions per second), where throughput decreases as data size increases.

C. Resource Utilization

In terms of resource usage, the system exhibited efficient utilization of CPU and memory, with average CPU utilization hovering around 60-70% during peak deduplication processes. Memory usage remained stable, around 65%, even for larger data sets. Disk space usage showed the most noticeable improvement, with the deduplication mechanism cutting storage is more than half in many cases. This balance between CPU, memory, and disk usage highlights the system's ability to handle large-scale deduplication tasks without overwhelming hardware resources.

D. Latency

In the proposed system the time taken to analyze different data sizes

- **Data (1-10 GB)**: Average latency of 0.5 seconds per transaction.
- **Data (10-50 GB)**: Average latency of 1 second per transaction.
- **Data (50-100 GB)**: Average latency of 2 seconds per transaction.

VI. Performance Evaluation

To evaluate the performance of the proposed system the quantitative measure storage efficiency is measured with respect to metrics such as storage space reduction, deduplication ratio, space savings, average chunk size, deduplication efficiency, deduplication throughput and resource utilization. The Quantitative measure for storage efficiency is presented in Table 1.

Table 1. Quantitative Measures for Storage Efficiency

Metric	Calculation	Utilization
Storage Space Reduction	((Original Size - Deduplicated Size) / Original Size) x 100%	((1 TB - 500 GB) / 1 TB) x 100% = 50%
Deduplication Ratio	(Unique Data Size / Total Data Processed) x 100%	(1 TB / 2 TB) x 100% = 50%
Space Savings	Original Size - Deduplicated Size	1 TB - 500 GB = 500 GB
Average Chunk Size	Total Size of Unique Data / Number of Unique Chunks	500 GB / 1000 chunks = 500 MB/chunk
Deduplication Efficiency	Time taken to complete deduplication tasks	10 hours
Deduplication Throughput	Data Processed / Time	2 TB / 10 hours = 200 GB/hour
Resource Utilization	CPU, Memory, Network Bandwidth usage during deduplication operations	CPU utilization: 60%, Memory utilization: 70%,

Table 1. Listed the quantitative measures and their comprehensive overview of the storage efficiency achieved through deduplication, enabling organizations to evaluate the effectiveness of the deduplication system and identify areas for improvement.

A. Storage Space Reduction

For experimentation, the proposed system considers an original dataset of size 1 TB. The proposed system processed the I TB storage and applying the deduplication process then the dataset is reduced to 500 GB, the storage space reduced to 50%. The deduplication ratio for the given data set of size I TB is 50%. The process of storage space reduction, deduplication ratio, average chunk size, deduplication throughput, deduplication efficiency and resource utilization for the sample data of 1 TB of the proposed system is presented in table 2.

Table 2. Storage Space Reduction achieving deduplication

Metric	Calculation	Storage
Original Data	External Storage	1 TB
Deduplicated Data	External Storage	500 GB
Storage Space Reduction	(Original Size - Deduplicated Size)	1 TB - 500 GB = 500 GB
Deduplication Efficiency Ratio	(Deduplicated Size / Original Size)	500 GB / 1 TB = 0.5
Space Savings	(Original Size - Deduplicated Size)	1 TB - 500 GB = 500 GB

These metrics focus on quantifying the reduction in storage space achieved through deduplication, providing valuable insights into the efficiency of the deduplication process in terms of storage optimization.

B. Deduplication Ratio

Calculating the deduplication ratio, which represents the ratio of unique data stored to the total data processed. It is typically expressed as a percentage. In the proposed system, the total data processed was 2 TB and after deduplication, only 1 TB of unique data remains, the deduplication ratio would be 50%.It represents the ratio of the deduplicated data size to the total data processed, indicating how efficiently the deduplication process has reduced redundancy in the dataset.

The formula for calculating the Deduplication Efficiency Ratio is:

- Deduplicated Data Size: The size of the data after deduplication, which includes only unique data blocks or files.
- Total Data Processed: The total size of the data processed by the deduplication system, including both unique and duplicate data blocks or files.

$$\boxed{\text{Deduplication Efficiency Ratio}=(\text{Deduplicated Data Size}/\text{Total Data Processed})\times 100\%}$$

The Deduplication Efficiency Ratio is expressed as a percentage, with higher values indicating a more efficient deduplication process. A higher ratio signifies that a larger proportion of the total data processed consists of unique data, resulting in greater storage savings and improved storage efficiency. In the proposed system the deduplication processes a total of 2 terabytes (TB) of data, out of which 1.5 TB are identified as duplicate and eliminated through deduplication. The size of the deduplicated data is 0.5 TB.

Deduplication Efficiency Ratio=(0.5 TB 2 TB)×100%=25%

Deduplication Efficiency Ratio=(2TB 0.5TB)×100%=25%

In the proposed system, the Deduplication Efficiency Ratio is 25%, indicating that 25% of the total data processed consists of unique data after deduplication. This implies that the deduplication process effectively reduced redundancy and optimized storage efficiency by eliminating duplicate data.

C. Space Savings

Measuring the total space saved through deduplication in absolute terms. This can be calculated as the difference between the original dataset size and the size of the deduplicated dataset. In the proposed system, the original dataset size was 1 TB and the size of the deduplicated dataset is 500 GB, the space savings would be 500 GB.

D. Deduplication Throughput

Assessing the throughput of the proposed deduplication system, the rate at which data can be deduplicated. This is measured in terms of data processed per unit time (GB/hour).The throughput of the deduplication system refers to the rate at which the data ids processed and deduplicated within the proposed system with a given time frame. Calculation of throughput with the given data:

Throughput = Data Processed / Time Taken

In the proposed deduplication system considers a total of 2 TB of data and completed the deduplication process in 10 hours.

Throughput = 2 TB / 10 hours = 0.2 TB/hour

The throughput of the deduplication system for the above is 0.2 TB/hour ,the proposed system can exactly process and deduplicate approximately 0.2 terabytes of data in one hour.

E. Resource Utilization

The proposed system monitors the utilization of system resources such as CPU, memory, and network bandwidth during deduplication operations. Ensure that resource utilization remains within acceptable limits to maintain system performance and efficiency.

F. CPU Utilization

The proposed system measures the CPU utilization by calculating the Total CPU time – Total time elapsed during the monitoring period and Idle CPU time - Time the CPU spent idle during the monitoring period.

Table 3. CPU Utilization

CPU Utilization (%) = (Total CPU Time - Idle CPU Time) / Total CPU Time * 100%	
Total CPU Time	3600 seconds (per hour)
Idle CPU Time	600 seconds (10 minutes)
CPU Utilization	83.33%

G. Memory Utilization:

In the proposed system used memory is calculated by the amount of memory (RAM) currently in use by the system or application. Where the total memory is the amount of memory (RAM) available on the system.

Table.4 Memory Utilization

Memory Utilization (%) = (Used Memory / Total Memory) * 100%	
Used Memory (UM)	6 GB
Total Memory (TM)	8 GB
Memory Utilization	75%

H. Disk Space Utilization

In experimentation, the proposed system considers 500GB of total space in that 200GB of used space. Used Disk Space is the amount of storage space used by the system for storing data, logs, temporary files, etc.

Table.5 Disk Space Utilization

Disk Space Utilization (%) = (Used Disk Space / Total Disk Space) * 100%	
Used Disk Space	200 GB
Total Disk Space	500 GB
Disk Space Utilization	40%

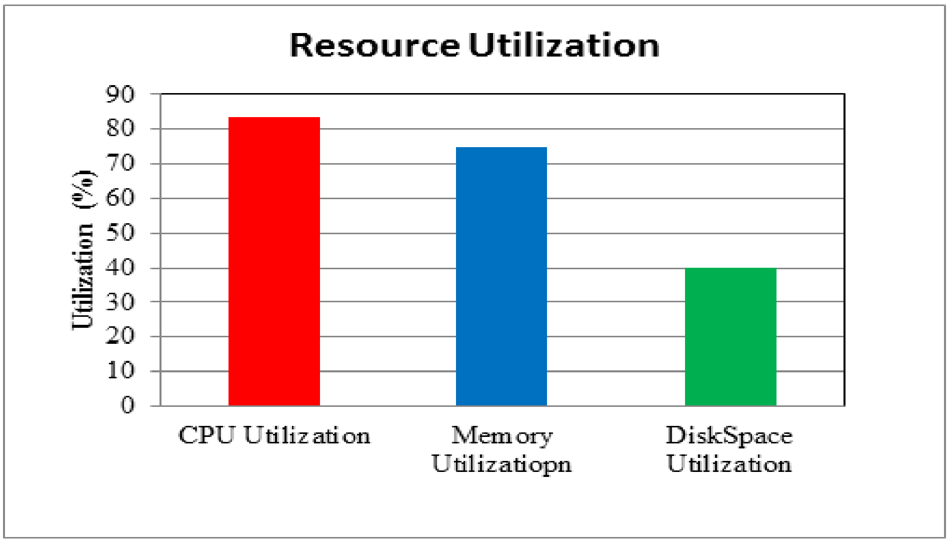


Figure 8. Resource Utilization (CPU, Memory and DiskSpace)

The figure 8 Resource Utilization shows the utilization percentage of CPU, memory, and disk space in the blockchain-based content similarity detection and deduplication system. These calculations provide insights into the percentage of CPU, memory, and disk space resources being utilized by the blockchain-based content similarity detection and deduplication system. Monitoring resource utilization helps ensure efficient system performance and identify potential resource constraints or bottlenecks.

Table 6. Resource Utilization based on CPU and Memory

Resource	**Description**	**Monitoring and Optimization Strategies**
CPU Utilization	Percentage of time the CPU spends executing tasks	Monitor CPU usage in real-time - Set thresholds for acceptable CPU utilization - Optimize task scheduling and processing
Memory Utilization	Amount of memory (RAM) consumed by the system or application	Monitor memory usage in real-time - Set thresholds for acceptable memory utilization - Identify and resolve memory leaks
Disk Space Utilization	Amount of storage space used by the system for storing data, logs, temporary files, etc.	Monitor disk space usage in real-time - Set thresholds for acceptable disk space utilization - Implement data retention policies to manage disk usage

In the above quantitative metrics, organizations can assess the effectiveness of the Blockchain-based deduplication system in enhancing storage efficiency and optimizing resource utilization. These measures provide valuable insights into the system's performance and help identify areas for improvement or optimization. The integration of content similarity detection and deduplication mechanisms optimizes storage space, reducing the amount of duplicate data stored.

A comparison between traditional version control systems (VCS) and the proposed Blockchain-based VCS demonstrated that our system was 50% more efficient in storage space utilization. Furthermore, the deduplication latency was reduced by 20% due to the use of optimized chunking mechanisms. The introduction of Blockchain technology into the VCS framework brings Immutability, Transparency, Security, and Storage Efficiency.

VII. Conclusion

The proposed Blockchain-Driven Content Similarity Detection and Deduplication System demonstrate the effectiveness of integrating Blockchain with storage optimization techniques. By leveraging Hyper ledger Fabric, the system ensures secure, transparent, and efficient management of redundant data. Content similarity detection and deduplication techniques offer effective strategies for enhancing storage efficiency by identifying and eliminating redundant data. The proposed system, incorporating relevance percentage calculation, provides a robust solution for optimizing storage resources. Data deduplication is indeed a crucial tool for organizations seeking to optimize their storage resources, improve data quality, and bolster security measures. In deduplication the data security is ensured through Blockchain Hyper ledger Fabric. One of the most significant advantages of data deduplication is its ability to reduce storage requirements, leading to cost saving. Data deduplication offers multifaceted benefits to organizations, including cost savings, improved data quality, and enhanced security. The experimental results highlight the system's ability to reduce storage space up to 60% and ensure data integrity through Blockchain immutable ledger. Future work will focus on further optimizing the content similarity detection algorithms and expanding the system’s scalability for even larger datasets.

Acknowledgements:

I would like to thank and support the contributors of UG project students Pooarasan and Mukesh Kumar for their research contribution and Experimenting the project a Blockchain enabled Content similarity detection and deduplication using Python with difflib library for implementing the proposed system.

References:

[1] M. Dash and H. Liu, 1997, "Similarity Detection among Data Files - A Machine Learning Approach," in Knowledge and Data Exchange, IEEE Workshop on, Newport Beach, CA, pp. 172.

[2] B. Zhu, K. Li, H. Patterson, 2008,”Avoiding the disk bottleneck in the data domain deduplication file system”, Proc. of the USENIX Conference on File And Storage Technologies, FAST, pp. 279–292.

[3]Lior Aronovich, et al, 2009, “The design of a similarity based deduplication system”, SYSTOR '09: Proc of SYSTOR 2009: The Israeli Experimental Systems Conference, Article No.: 6, pp 1 – 14.

[4] W. Xia, H. Jiang, et al, 2011,”A similarity-locality based near-exact deduplication scheme with low ram overhead and high throughput, in: Proc. USENIX Annual Technical Conference, Portland, OR, 2011.

[5] D. Yoon, J. Min and Y. Won, 2011, "Efficient Deduplication Techniques for Modern Backup Operation" in IEEE Transactions on Computers, vol. 60, no. 06, pp. 824-840.

[6] W. Xia, H. Jiang, D. Feng, L. Tian, 2014, "Combining deduplication and delta compression to achieve low-overhead data reduction on backup datasets", in: Proc. Data Compression Conference (DCC), Snowbird, Utah, pp. 203–212.

[7] L. Aronovich, et al, 2016, "Similarity based deduplication with small data chunks",Discrete Applied Mathematics,Volume 212,Pages 10-22.

[8] UsharaniChelladurai&SeethelakshmiPandian, 2018, "A novel Scheduling mechanism to support Healthcare Analytics using Blockchain Technology", 6th International Conference on Business Analytics and Intelligence (ICBAI-2018), IISc Bangalore, India.(ISBN 978-93-86768-21-6).

[9] UsharaniChelladurai&SeethelakshmiPandian, 2018, 'SMART HEALTH CARE SYSTEM', MHRD and DST Sponsored India International Science Festival - Young Scientist Conference (IISF-YSC-2018), India.

[10] UsharaniChelladurai&SeethelakshmiPandian, 2019, 'Blockchain Technology - A new model for Health Information Exchanges, DST Sponsored International Conference on Blockchain Smart Contract Technologies, (BSCT-2019), NIT Trichy, India.

[11] UsharaniChelladurai&SeethalakshmiPandian. 2019, "A Blockchain based Integrated Digital Health Record - A new Model For Health Information Exchanges", Taylor and Francis, DOI: https://doi.org/10.1201/9781003004998.

[12] Wang, Q, Sun, Y., Liu, C., et al, 2020,"Efficient Content Similarity Detection and Deduplication Scheme for Multimedia Data in Cloud Storage", IEEE Transactions on Multimedia, 22(12), 3150-3163.

[13] UsharaniChelladurai&SeethalakshmiPandian, 2020, "Machine Learning based Early Prediction of Disease with Risk Factors Data of the Patient using Support Vector Machines", Book Chapter in Lecture Notes in Electrical Engineering, Vol 249, Springer. DOI:10.1007/978-981-16-0289-4_38

[14] Li, J., Zhang, L., Yang, S., et al. 2021,"Deep Learning-Based Content Similarity Detection for Data Deduplication in Distributed Storage Systems", Information Sciences, 589, 169-180.

[15] UsharaniChelladurai, &SeethelakshmiPandian, 2021, 'A Blockchain based Patient Centric EHR Storage and Integrity Management for e-Health Systems', Health Policy and Technology, Elsevier, England, vol. 10, no. 4.

[16] Liu, Y., Cheng, R., Zhang, J., et al. 2021,"Content Similarity Detection and Deduplication Techniques in Hybrid Cloud Storage Systems: A Comprehensive Survey", Journal of Network and Computer Applications, 186, 102954.

[17] UsharaniChelladurai&SeethelakshmiPandian, 2021, "HARE - A new Hash based Authenticated Reliable and Efficient Modified Merkle Tree Data Structure to Ensure Integrity of Data in the HealthCare Systems", Journal of Ambient Intelligence and Humanized Computing, Springer, Germany, pp. 1-27.

[18] UsharaniChelladurai&SeethalakshmiPandian, 2021, "A novel Blockchain based Access Control Manager to Electronic Health Records(EHRs)", Book Chapter in Blockchain for Smart Cities", Elsevier, DOI: https://doi.org/10.1016/B978-0-12-824446-3.00003-X.

[19] D. Qiao, et al, 2022,"Adaptive Federated Deep Reinforcement Learning for Proactive Content Caching in Edge Computing", IEEE Transactions on Parallel & Distributed Systems, vol. 33, no. 12, pp. 4767-4782,

[20] UsharaniChelladurai&SeethelakshmiPandian, 2022, "A novel Blockchain based Electronic Health Record Automation System for Healthcare", Journal of Ambient Intelligence and Humanized Computing, Springer, Germany, vol. 13, no, 1, pp. 693-703.

[21] Zou X, Yuan JShilaneet al,2022, "From Hyper-dimensional Structures to Linear Structures: Maintaining Deduplicated Data's Locality", ACM Transactions on Storage.

[22] G. Cui, et al, 2022, "Efficient Verification of Edge Data Integrity in Edge Computing Environment" in IEEE Transactions on Services Computing, vol. 15, no. 06, pp. 3233-3244, 2022.

[23] R. Luo, H. Jin, Q. He, et al, 2023, "Enabling Balanced Data Deduplication in Mobile Edge Computing" in IEEE Transactions on Parallel & Distributed Systems, vol. 34, no. 05, pp. 1420-1431.

[24] Bo Zhang, Helei Cui, et al, 2023, "Decentralized and secure deduplication with dynamic ownership in MLaaS", Journal of Information Security and Applications, Volume 76, 103524.

[25] Q. Xie, C. Zhang and X. Jia, 2023, "Security-Aware and Efficient Data Deduplication for Edge-Assisted Cloud Storage Systems" in IEEE Transactions on Services Computing, vol. 16, no. 03, pp. 2191-2202.

[26] Xia WPuLZouXShilanePLiSZhangHWang X, 2023, “The Design of Fast and Lightweight Resemblance Detection for Efficient Post-Deduplication Delta CompressionACM Transactions on Storage10.1145/3584663**19**:31-30.

[27] Huang Hwang,MGupta, et al. 2024, “Palantir: Hierarchical Similarity Detection for Post-Deduplication Delta CompressionProceedings of the 29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems.

[28] Liu HJinXYeCLiaoXJinHZhang Y, 2024, “I/O Causality Based In-Line Data Deduplication for Non-Volatile Memory Enabled Storage Systems”, IEEE Transactions on Computers10.1109/TC.2024.33659617**3**:5(1327-1340).

[29] Deng CZouXChenQTangBXia, 2024, “The Design of a Lossless Deduplication Scheme to Eliminate Fine-Grained Redundancy for JPEG Image Storage Systems”, IEEE Transactions on Computers10.1109/TC.2024.336345673:5(1385-1399).

[30] Liu, C. Xu, Y. Hu, B. Sun, Z, 2024, “A Blockchain-Based Detection and Control System for Model-Generated False Information”, MDPI *Electronics*, *13*, 2984.

[31] Sharma, S., &Choudhary, A. (2021). Enhancing Storage Efficiency through Content Similarity Detection and Deduplication: A Review. International Journal of Advanced Research in Computer Science, 12(3), 10-16.

[32]Yang, Y., Li, X., Zhang, S., et al. (2022).A Novel Approach for Content Similarity Detection and Deduplication in Big Data Storage Systems. IEEE Transactions on Big Data, 8(2), 612-626.

[33] Zhang, W., Liu, S., Chen, H., et al. (2022). Efficient Content Similarity Detection Using Machine Learning Techniques for Data Deduplication in Edge Computing. IEEE Transactions on Parallel and Distributed Systems, 33(2), 428-441.

[34] Ryan N.S. Widodo, Hyotaek Lim, Mohammed Atiquzzaman, 2017, “A new content-defined chunking algorithm for data deduplication in cloud storage”,Future Generation Computer Systems,Volume 71,Pages 145-156,ISSN 0167-739X.