Secure And Efficient Cloud Datastorage Using Deduplication And Blockchain Technology

Arunadevi.S[1], Dharanya.S[2], Dharanya.V[3], Gurupriya.R[4], Ms.S. Gowthami[5]
[1, 2, 3, 4, 5] Dept of Computer Science and Engineering
[1, 2, 3, 4, 5] Vivekanandha College of Technology for Women Tiruchengode, Namakkal, Tamil Nadu, India

Abstract- *This paper presents a secure and efficient cloud data storage using deduplication and blockchain technology to reduce storage requirements and enhance data security.Cloud computing has become a fundamental backbone for modern organizations, but the rapid and continuous growth of digital data has led to significant challenges such as increased storage costs, redundant data accumulation, and heightened security risks. To overcome these limitations, this project proposes an integrated cloud storage framework that combines data deduplication, semantic encryption, and blockchain technology into a unified architecture. The deduplication mechanism focuses on identifying and eliminating duplicate data at the block level, ensuring that only unique data is stored within the system. Semantic encryption is employed to protect sensitive information, ensuring that encrypted data remains confidential even after deduplication and compression processes. The framework is designed such that unauthorized users cannot derive meaningful information from stored ciphertexts, thereby safeguarding user privacy. The blockchain technology is incorporated to manage authentication, access control, and metadata in a decentralized and tamper-proof manner. By maintaining an immutable ledger of transactions and access logs, the system ensures transparency, traceability, and resistance against unauthorized modifications.*

Keywords: Cloud computing, Data Deduplication, Semantic Encryption, Blockchain Technology, data security, Traceability, transparency, Temper-proof manner.

I. INTRODUCTION

The proposed project focuses on developing a secure and efficient cloud-based storage framework that integrates data deduplication, semantic encryption, and blockchain technology.The core functionality of the system lies in its deduplication and encryption processes. Deduplication is performed at the block level, where incoming data is divided into chunks and compared with existing stored data to identify duplicates. Only unique chunks are stored, while references are maintained for repeated data, significantly reducing storage space requirements. Alongside this, semantic encryption is applied to ensure data confidentiality before storage in the cloud. This encryption technique guarantees that even if data is accessed by unauthorized entities, it remains unintelligible and secure.

To further strengthen security and trust, the system incorporates blockchain technology as a decentralized mechanism for managing authentication, transactions, and metadata. The distributed nature of blockchain eliminates single points of failure and ensures that all transactions are validated through consensus mechanisms.

The implementation of this cloud-based deduplication and security framework has a significant impact on storage efficiency and resource optimization. By eliminating redundant data through deduplication techniques, the system drastically reduces the amount of physical storage required in cloud environments. Organizations and individuals are more likely to adopt cloud storage solutions when they are assured of both cost-effectiveness and robust protection of their data, the system contributes to building a more secure, efficient, and trustworthy cloud ecosystem.

II. LITERATURE REVIEW

Cloud computing has become an essential platform for storing and managing large volumes of data. As the adoption of cloud storage continues to increase, significant research has been carried out to improve storage efficiency, data privacy, and system security. Various techniques such as secure authentication models, advanced encryption methods, blockchain technology, and data deduplication mechanisms have been introduced to overcome the limitations of traditional cloud storage systems.Several existing studies have focused on secure data sharing and controlled user access in cloud environments. These approaches emphasize strong authentication methods, secure communication channels, and confidentiality of user data. At the same time, many researchers have highlighted the importance of balancing security requirements with system performance, scalability, and availability in large-scale cloud infrastructures.

Several existing studies have focused on secure data sharing and controlled user access in cloud environments. These approaches emphasize strong authentication methods, secure communication channels, and confidentiality of user data. At the same time, many researchers have highlighted the importance of balancing security requirements with system performance, scalability, and availability in large-scale cloud infrastructures.

Therefore,there is agrowing need for a comprehensive cloud storage solution that ensures efficiency, confidentiality, integrity, and trust within a single architecture.

III. PROBLEM STATEMENT

Cloud storage systems have become a popular solution for storing and managing large amounts of digital data due to their flexibility, scalability, and cost-effectiveness. However, the rapid growth of data in cloud environments has created several challenges related to storage efficiency, security, and trust. Large volumes of duplicate data consume unnecessary storage space, increase operational costs, and reduce overall system performance.

Although data deduplication techniques are used to eliminate redundant data, many existing approaches require high computational resources and may create security concerns when combined with traditional encryption methods. At the same time, centralized cloud storage architectures depend heavily on third-party service providers, making them vulnerable to unauthorized access, data tampering, and single points of failure. In addition, users often lack transparency and direct control over how their data is managed and accessed.

Therefore, there is a need to design a secure and efficient cloud storage framework that can reduce redundant data while ensuring confidentiality, integrity, and trust. The proposed solution aims to integrate intelligent deduplication for storage optimization, semantic encryption for privacy preservation, and blockchain technology for decentralized authentication, transparency, and tamper-proof data management.

IV. EXISTING SYSTEM

The existing cloud storage systems primarily rely on Content-Defined Chunking (CDC) techniques for data deduplication, with Rabin-based fingerprinting being one of the most widely used approaches. In this method, data streams are processed byte by byte to compute Rabin fingerprints, which are then compared against predefined conditions to determine chunk boundaries. While this approach is effective in identifying duplicate data, it is computationally intensive due to the continuous hashing and evaluation required for each byte in the data stream. As a result, the overall performance of the system is impacted, especially when handling large-scale datasets in cloud environments.

The Content-Defined Chunking (CDC) process is highly time-consuming because it performs hash computations and evaluations for the data stream on a byte-by-byte basis, leading to increased processing overhead. Traditional Rabin-based fingerprinting requires continuous hashing and comparison operations, which significantly consume CPU resources and reduce the overall efficiency of the deduplication system. The hash judgment mechanisms used in current systems can be complex and computationally expensive, adding additional latency during the chunking and deduplication stages.

Conventional cloud storage architectures are typically centralized and lack integrated advanced security mechanisms, making them vulnerable to unauthorized access, single points of failure, and limited transparency in data handling and transactions.

A. DRAWBACKS OF EXISTING SYSTEM

The Content-Defined Chunking (CDC) process is highly time-consuming because it performs hash computations and evaluations for the data stream on a byte-by-byte basis, leading to increased processing overhead. Traditional Rabin-based fingerprinting requires continuous hashing and comparison operations, which significantly consume CPU resources and reduce the overall efficiency of the deduplication system. Existing approaches such as FastCDC and Gear-based CDC still face limitations in achieving optimal deduplication ratios, resulting in less effective elimination of redundant data in some scenarios.

The hash judgment mechanisms used in current systems can be complex and computationally expensive, adding additional latency during the chunking and deduplication stages. Conventional cloud storage architectures are typically centralized and lack integrated advanced security mechanisms, making them vulnerable to unauthorized access, single points of failure, and limited transparency in data handling and transactions.

B. METHODOLOGY

The proposed system introduces a secure, scalable, and cost-efficient cloud storage framework that integrates data deduplication, semantic encryption, and blockchain

technology into a unified architecture. The primary objective of this system is to overcome the limitations of traditional cloud storage, including excessive data redundancy, high storage and operational costs, and vulnerabilities related to data security and integrity. In this framework, when users upload data to the cloud, the system first performs preprocessing by dividing files into smaller blocks or chunks, which enables fine-grained analysis and efficient storage handling. Each block is assigned a unique identifier, typically generated using a hashing technique.

The system then performs block-level deduplication by comparing the hash values of incoming data blocks with those already stored in the cloud. If a match is found, the system identifies the block as a duplicate and avoids storing it again, instead creating a reference pointer to the existing block. This process significantly reduces storage redundancy, optimizes resource utilization, and improves scalability, especially in environments where similar data is frequently uploaded by multiple users. To ensure data confidentiality, the system incorporates semantic encryption, where each data block is encrypted before being stored. This ensures that sensitive information remains protected from unauthorized access, and even identical data blocks produce secure encrypted outputs, enhancing resistance against attacks. In addition to encryption, the system integrates blockchain technology to provide a decentralized and tamper-proof mechanism for managing user authentication, file metadata, and transaction records. Every operation, such as file uploads, access requests, or modifications, is recorded on the blockchain as a transaction, ensuring transparency, traceability, and immutability of system activities. The blockchain also supports secure access control by verifying user identities and permissions before allowing data retrieval.

Metadata associated with each file, including ownership details, storage references, and access rights, is securely maintained and protected from unauthorized modifications. During data retrieval, the system authenticates the user through the blockchain, identifies the required data blocks using metadata, reconstructs the original file from deduplicated blocks, and decrypts it using the appropriate key before delivering it to the user. Overall, the proposed system enhances storage efficiency by eliminating redundant data, reduces operational costs, strengthens data security through encryption, and ensures data integrity and transparency through blockchain integration, making it a robust solution for modern cloud storage challenges.

V. PROPOSED ARCHITECTURE

The proposed architecture of the system consists of an integrated cloud storage framework that combines data deduplication, semantic encryption, and blockchain technology to ensure efficient, secure, and reliable data management. When a user uploads a file, it is first processed at the client or server side where it is divided into smaller data blocks for deduplication. These blocks are checked against existing stored data to identify and eliminate duplicates, ensuring only unique data is stored in the cloud. Before storage, the data is encrypted using semantic encryption to maintain confidentiality and protect against unauthorized access. Simultaneously, blockchain technology is used to record user authentication details, file metadata, and transaction logs in a decentralized and tamper-proof ledger. This architecture ensures optimized storage utilization, strong data privacy, and transparent, secure management of all operations within the cloud environment.

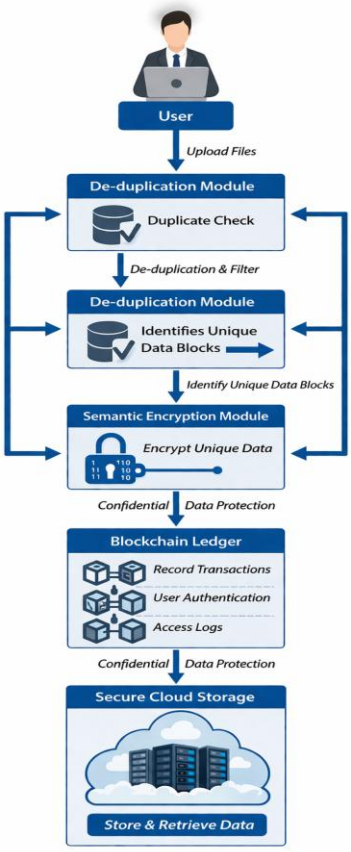

Fig:1 Proposed Architecture

VI. MODULES DESCRIPTION

The cloud storage framework serves as the foundational module of the proposed system, providing the environment in which all other operations are executed. It

enables users to upload, store, retrieve, and manage their data through a centralized or distributed cloud infrastructure. This module handles user requests and ensures seamless interaction between the client and the backend storage system. It is designed to support scalability, allowing the system to handle a large volume of data efficiently. The framework also manages data flow between different modules such as deduplication, encryption, and blockchain.

The data deduplication module is responsible for identifying and eliminating redundant data before it is stored in the cloud. It works by dividing incoming files into smaller chunks and generating unique identifiers or hashes for each chunk. These hashes are compared with existing entries in the storage system to detect duplicates. The semantic encryption module ensures the confidentiality and privacy of user data before it is stored in the cloud. It encrypts data in such a way that identical plaintexts may produce indistinguishable ciphertexts, preventing attackers from inferring information from stored data.

The access control module is responsible for managing user authentication and authorization within the system. It ensures that only legitimate users can access the cloud storage resources and perform permitted operations. Users are required to register and log in using secure credentials, which are verified by the system before granting access.

VII. RESULTS AND DISCUSSION

The results of the proposed system demonstrate its effectiveness in achieving secure and efficient cloud data storage through deduplication and proxy encryption techniques. When a user uploads a file, the system analyses the content rather than relying solely on the file name. If the same content already exists in the storage, the system identifies it as duplicate data, even if the file is renamed, and avoids redundant storage by indicating that the content is identical. This significantly reduces storage overhead and improves efficiency. Furthermore, the system ensures high-level data security through encryption mechanisms. In addition, controlled data sharing is enabled using proxy encryption, where a user can securely grant access to another authorized user for downloading specific files. This approach enhances both data privacy and access flexibility while maintaining system integrity and preventing unauthorized usage.

A. User Registration and Authentication

The system begins with the user is required to register in the system by providing necessary details. This registration process ensures that only authorized users can access the cloud storage environment and utilize its services securely.

Fig:2 User Registration and Authentication

B. User Login and Authentication

After successful registration, the user can log in to the system using the credentials created during the registration process. The system verifies the entered username and password to authenticate the user. Upon successful authentication, the user is granted secure access to the cloud storage environment, where they can upload, download, and manage files. This process ensures that only authorized users can access the system and perform operations.

Fig:3 User Login and Authentication

C. Storage Allocation and Subscription Selection

After successful login, the user is provided with options to select the required storage capacity. The system offers a default free storage space of 15 GB for basic usage. If additional storage is required, users can choose from higher

storage plans such as 20 GB or 25 GB by opting for a paid subscription. This flexible storage allocation mechanism allows users to select storage based on their requirements, ensuring efficient resource utilization and scalability.

Fig:4 Storage Allocation and Subscription Selection

D. File Upload

The Users can upload files to the cloud storage system. The system allows easy file selection and ensures that the uploaded data is stored successfully in the user's assigned storage space for future access and management.

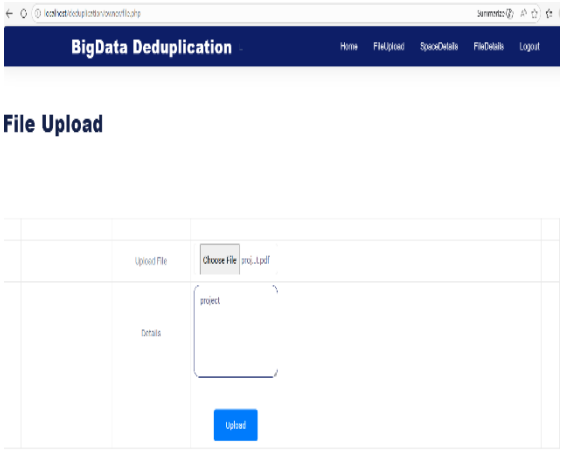

Fig:5 File Upload

E. File Upload Confirmation and Storage Status

Upon successful completion of the file upload process, the system confirms the status of the uploaded file to the user. A notification or message is displayed indicating that the file has been successfully stored in the cloud environment. The system updates the storage usage details and reflects the current page of the user's allocated storage space. This step ensures transparency by informing the user about successful data storage and system response.

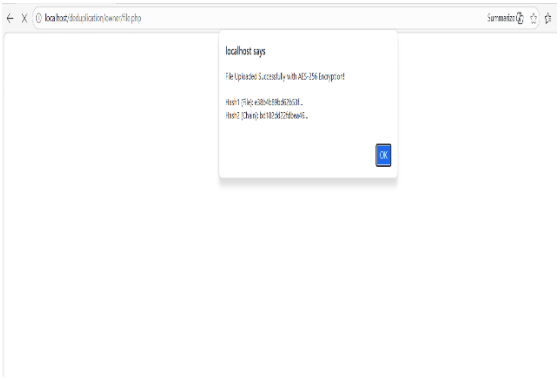

Fig:6 File Upload Confirmation and Storage Status

F. Data Deduplication Process

The user uploading data to the cloud through the client interface. Once the data is received, it undergoes preprocessing where it is divided into smaller chunks or blocks. These chunks are analysed using deduplication techniques to identify whether identical data already exists in the storage system. If duplicate chunks are detected, the system avoids storing them again and instead maintains references to the existing data, ensuring that only unique content is physically stored in the cloud. This process significantly reduces storage redundancy and improves overall storage efficiency.

After deduplication, the unique data chunks are passed through a semantic encryption module to ensure confidentiality and security. The encryption process converts the readable data into ciphertext using secure cryptographic techniques, making it inaccessible to unauthorized users. Even if duplicate data exists, the encrypted form remains secure and does not reveal any meaningful information to potential attackers.

Fig:7 Data deduplication process

G. Secure File Access and Download Request

At this stage, once the file upload and deduplication processes are completed, users may require access to files owned by other users. In this step, the system provides a secure mechanism where the requesting user must send a

download request to the respective file owner. The owner is notified of the request and has the authority to approve or reject it based on access requirements. Upon approval, the system grants permission to the requesting user to download the file securely. This approach ensures controlled data sharing, maintains user privacy, and prevents unauthorized access within the cloud storage environment.

Fig:8 Secure File Access and Download Request

H. Email Notification and Secure Key Generation

When a download request is initiated, the system sends a notification to the file owner through the registered email address. The email informs the user about the request to access the document. Upon approval, the system automatically generates a secure access key, which is shared with the authorized user. This key is required to complete the download process, ensuring that only permitted users can access the requested file. This mechanism enhances security by combining user authorization with key-based access control.

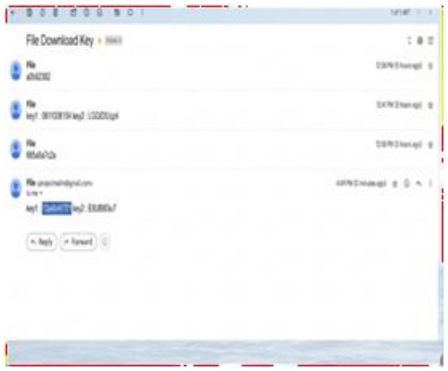

Fig:9 Email Notification and Secure Key Generation

I. Access Key Verification for Secure Download

To ensure secure file access, the system requires the user to enter a valid access key before initiating the download process. The access key is generated and shared with the authorized user upon approval of the request. The system verifies the entered key with the stored credentials to authenticate the user. Only when the key is validated successfully, the user is permitted to proceed further. This verification step enhances data security by preventing unauthorized access to sensitive files.

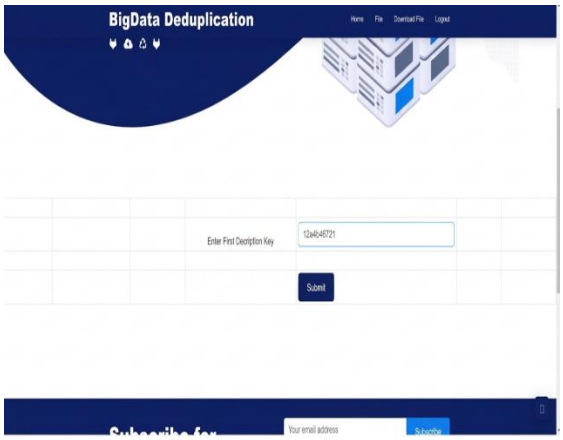

Fig:10 Access Key Verification For Secure Download

J. Secure File Download

Upon successful verification of the access key, the system allows the authorized user to download the requested file. The file is retrieved securely from the cloud storage and made available through the download interface. This process ensures that only authenticated users with valid permissions can access and retrieve the data, thereby maintaining confidentiality and integrity of the stored information.

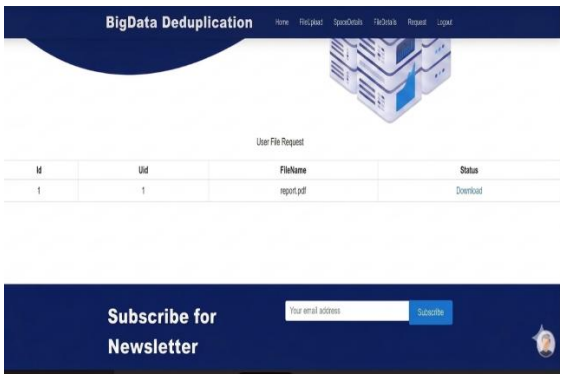

Fig:11 Secure File Download

VIII. CONCLUSION

The proposed cloud-based storage system successfully integrates data deduplication, semantic encryption, and blockchain technology to address the key

challenges of storage efficiency, data security, and system transparency. By eliminating redundant data through deduplication, the system optimizes storage utilization and reduces operational costs for cloud service providers. The incorporation of semantic encryption ensures that sensitive data remains confidential and protected from unauthorized access, even in the presence of duplicate data. Additionally, blockchain technology enhances the system by providing a decentralized and tamper-proof mechanism for managing user authentication, metadata, and transaction logs, thereby ensuring data integrity and trustworthiness across the platform.

Overall, the system presents a comprehensive and robust solution for modern cloud storage environments by combining efficiency with strong security mechanisms. It eliminates the limitations of traditional centralized storage systems by reducing redundancy, improving scalability, and enhancing transparency through distributed ledger technology.

The proposed architecture ensures reliable data management while maintaining privacy and integrity, making it suitable for real-world cloud applications. This integrated approach not only improves performance but also builds user confidence in cloud-based services by providing a secure, efficient, and transparent data storage framework.

IX. FUTURE SCOPE

The future enhancement of the proposed system can focus on improving the efficiency and scalability of the deduplication process by incorporating more advanced and intelligent algorithms. Machine learning techniques can be integrated to predict duplicate patterns and optimize chunking strategies, thereby reducing computational overhead and improving processing speed. Additionally, adaptive deduplication mechanisms can be developed to dynamically adjust chunk sizes based on data characteristics, which would further enhance storage optimization and system performance in large-scale cloud environments. Another important area of future work involves strengthening the security mechanisms of the system.

While semantic encryption and blockchain provide a solid foundation, additional layers of security such as homomorphic encryption or attribute-based encryption can be explored to enable secure computations on encrypted data. Enhancements in key management techniques and multi-factor authentication can also be implemented to further protect user credentials and sensitive information. Moreover, integrating intrusion detection systems and anomaly detection techniques can help in identifying and mitigating potential cyber threats in real time. Future developments can also aim at improving the interoperability and usability of the system by integrating it with other cloud platforms and services.

The implementing cross-cloud data sharing and hybrid cloud support would enable better flexibility and scalability for enterprises. Performance optimization techniques such as load balancing, edge computing integration, and caching mechanisms can be incorporated to reduce latency and improve response times. Additionally, enhancing the blockchain module with more efficient consensus algorithms like Proof of Stake or Delegated Proof of Stake can reduce energy consumption and improve transaction processing speed, making the system more sustainable and efficient.

REFERENCES

[1] Alabdulatif, Abdullah, NavodNeranjan Thilakarathne, and Kassim Kalinaki. "A novel cloud enabled access control model for preserving the security and privacy of medical big data." Electronics 12.12 (2023): 2646.

[2] Alsuqaih, Hanan Naser, et al. "An efficient privacy-preserving control mechanism based on blockchain for E-health applications." Alexandria Engineering Journal 73 (2023): 159-172.

[3] Athanere, Smita, and Ramesh Thakur. "Blockchain based hierarchical semi-decentralized approach using IPFS for secure and efficient data sharing." Journal of King Saud University-Computer and Information Sciences 34.4 (2022): 1523-1534.

[4] Dhinakaran, D., et al. "Privacy-preserving data in IoT-based cloud systems: A comprehensive survey with AI integration." arXiv preprint arXiv:2401.00794 (2024).

[5] Gupta, Ishu, et al. "Secure data storage and sharing techniques for data protection in cloud environments: A systematic review, analysis, and future directions." IEEe Access 10 (2022): 71247-71277.

[6] Kumar, Prabhat, et al. "A blockchain-orchestrated deep learning approach for secure data transmission in IoT-enabled healthcare system." Journal of Parallel and Distributed Computing 172 (2023): 69-83.

[7] Liu, Xiaoguang, Jun Yan, Shuqiang Shan, and Rongjun Wu. "A blockchain-assisted electronic medical records by using proxy reencryption and multisignature." Security and Communication Networks 2022 (2022).

[8] Narayanan, Uma, Varghese Paul, and Shelbi Joseph. "A novel system architecture for secure authentication and data sharing in cloud enabled Big Data Environment." Journal of King Saud University-Computer and Information Sciences 34.6 (2022): 3121-3135.

[9] Prajapati, Priteshkumar, and Parth Shah. "A review on secure data deduplication: Cloud storage security issue." Journal of King Saud University-Computer and Information Sciences 34.7 (2022): 3996-4007.

[10] Wang, Weizheng, et al. "Smart contract token-based privacy-preserving access control system for industrial Internet of Things." Digital Communications and Networks 9.2 (2023): 337-346.