strengthening network security in hadoop-based secure storage solutions for enhanced authentication

Ms.T.Kalaiselvi[1], Mr.R. Pavin Rohith[2],Mr.T. C. Stalin[3],Mr.S.V. Nivethan[4]

Associate Professor [1], Final Year[2,3,4], Department of Computer Science and Engineering,

Erode Sengunthar Engineering College (Autonomous)

Thudupathi, Erode, Tamilnadu, India

ABSTRACT

Cloud users can ensure the integrity of their data without having to retrieve the entire file, thanks to Provable Data Possession (PDP) schemes leveraging Public Key Infrastructure (PKI). These schemes offer efficiency, flexibility, and support for various verification methods including private, delegated, and public verification. However, one such scheme, ID-DPDP, exhibits a flaw that compromises its soundness. To address this issue, a novel approach is introduced, resulting in an improved ID-DPDP protocol that extends the basic ID-PDP to accommodate multiple clouds. With the proliferation of data storage and sharing services in the cloud, collaborative data modification and sharing among users have become commonplace. Maintaining the integrity of shared data for public verification necessitates users within the group to generate signatures for all data blocks being shared. Given that different users may modify different blocks within the shared data, each block is typically signed by a different user. When a user is revoked from the group, the blocks previously signed by that user must be reassigned to another existing user. However, the conventional method of downloading the corresponding segment of shared data for re-signing during user revocation proves inefficient due to the substantial size of data stored in the cloud.

Keywords: Ensuring data integrity in cloud storage via Provable Data Possession (PDP) schemes leveraging Public Key Infrastructure (PKI), addressing flaws in ID-DPDP, and accommodating multi-cloud environments amidst collaborative data sharing and user revocation challenges.

1. INTRODUCTION

Cloud computing is continuously evolving, presenting organizations with both challenges and opportunities, particularly with the advent of big data. To efficiently process and analyze large-scale data, secure storage solutions based on Hadoop have become indispensable. These solutions cater to the critical requirement for scalable, reliable, and secure storage of vast datasets. Hadoop, as an open-source distributed computing framework, lays the groundwork for handling massive volumes of data, yet its efficacy hinges on the presence of a robust storage system. In the domain of cloud computing, where data security and privacy are paramount, amalgamating Hadoop with advanced security measures presents a holistic approach to big data management while mitigating potential threats. A secure storage solution for big data in the cloud, rooted in Hadoop, encompasses a plethora of technologies and methodologies. It encompasses the utilization of distributed file systems such as the Hadoop Distributed File System (HDFS) for efficient data storage and management across a cluster of machines. Furthermore, these solutions integrate encryption methodologies to safeguard data both at

rest and in transit, implement access control mechanisms to thwart unauthorized access, and incorporate monitoring and auditing functionalities for compliance adherence and threat identification. With the escalating demand for big data storage in the cloud, organizations must prioritize the implementation of secure storage solutions to uphold the confidentiality, integrity, and availability of their data, all while leveraging the full potential of their big data analytics endeavors.

2. LITERATURE REVIEW

In their work [1], Gang Li et al. aim to establish a distributed, collaborative, and automated design and manufacturing workflow. They leverage Industry 4.0 principles along with associated technologies like Cyber-Physical Systems, Internet of Things (IoT), Cloud Computing, and Big Data Analytics to achieve this goal. The integration of cyber-physical systems and IoT facilitates the collection and transmission of industrial data through various peripherals such as sensors and software. Cloud computing techniques are utilized for centralized data storage and collaboration, streamlining resource allocation and enhancing industry-wide research efforts. Big data analytics plays a pivotal role in organizing digital assets and extracting valuable insights. This interdisciplinary approach has garnered attention from both industry and academia, with active research exploring the intricate relationship between Industry 4.0 and Big Data.

Chunhui Wen et al. [2] propose the effective integration of big data technology in the Internet of Things, particularly in the financial sector where it aids in managing internal and external data related to credit risks. By employing efficient machine learning algorithms, the prediction of credit risk is enhanced, leading to reduced losses and increased profits. Their approach involves utilizing distributed search engine technology to gather bank card and transaction data from diverse sources within the Internet of Things financial industry. Additionally, they design a Spark parallel algorithm for data preprocessing and establish an inverted table and two-level index file for big data analysis platforms. The authors employ the Mutually Exclusive Collectively Exhaustive (MECE) analysis method to derive indicators and quantification methods for evaluating financial credit risk in the Internet of Things, analyzing the correlation between these indicators and risk grading.

Xiangfan Zhang et al. [3] introduce a system aimed at enhancing the intelligence of the medical system, particularly focusing on the security of medical big data ecosystems. Their paper presents the design and implementation of a secure medical big data ecosystem on the Hadoop platform, ensuring data independence even when distributed across different medical institutions. They explore the potential of blockchain technology for multi-party maintenance and backup information security. The system includes a personalized health information system for patients to access their treatment and rehabilitation status remotely.

Xin Huang et al. [4] propose a system addressing challenges related to the storage and retrieval of massive electronic medical data. Their approach involves using electronic images for diagnosis within electronic medical data systems, improving access for patients through various electronic means. They suggest different merging strategies based on the characteristics of image files and introduce a two-level model combined with medical imaging information for enhanced storage performance. Their improved 2Q algorithm enhances file reading efficiency, as demonstrated through experimental comparisons.

In addressing the challenges posed by the exponential growth of data processed by computing systems, a shift towards knowledge-centric computing is observed. This survey paper [5] provides an overview of Cloud-centric Big Data placement and storage methodologies, focusing on non-functional properties. It analyzes various technologies related to Big Data management, guiding readers in selecting suitable solutions based on non-functional application requirements. Gaps and challenges in this field are also discussed.

W. Rajeh [6] explores security issues in Hadoop distributed file systems, investigating unauthorized access problems and proposing measures to enhance security and mitigate risks.

G. S. Bhathal and A. Singh [7] discuss weaknesses in the Hadoop framework, analyzing security challenges and potential attacks, while proposing strategies to address vulnerabilities.

Kapil, A. Agrawal, and R. A. Khan [8] address big data security challenges within Hadoop environments, identifying security risks and proposing mitigation strategies to safeguard data assets.

B. H. Husain, S. R. Zeebaree et al. [9] review enhanced distributions frameworks for Hadoop, examining developments to meet evolving requirements and challenges in big data management.

M. Naisuty, A. N. Hidayanto, N. C. Harahap, A. Rosyiq, and G. M. S. Hartono [10] conduct a systematic literature review on data protection in Hadoop distributed file systems, evaluating encryption algorithms' effectiveness in securing data stored within these environments. They analyze encryption techniques and their impact on data security within Hadoop ecosystems.

3. EXISTING SYSTEM

To address the challenges associated with using a single encryption algorithm in cloud computing environments, which can lead to low encryption efficiency and unreliable metadata for static data storage, we propose a secure storage scheme for big data based on the Hadoop framework. Our approach involves distributing the Name Node service across multiple servers using HDFS federation and HDFS high-availability mechanisms. Each node is coordinated using the Zookeeper distributed coordination mechanism to establish dual-channel storage. Additionally, we enhance the ECC encryption algorithm for standard data encryption and incorporate a homomorphic encryption algorithm for data requiring computation. To accelerate the encryption process, we implement a dual-thread encryption mode. Furthermore, we design the HDFS control module to seamlessly integrate the encryption algorithm with the storage model. Through extensive experimentation, our results validate that our proposed solution effectively mitigates the single point of failure issue with metadata, enhances metadata reliability, and ensures server fault tolerance.

3.1 DISADVANTAGES

•Introducing a secure storage scheme built upon Hadoop, which incorporates a distributed Name Node service, coordination facilitated by Zookeeper, and the integration of diverse encryption algorithms, introduces significant complexity to the system. This complexity increases maintenance demands and requires specialized expertise for effective deployment and management.

•Utilizing multiple encryption algorithms, especially homomorphic encryption for data requiring calculations, may lead to considerable consumption of computational resources. This could elevate processing overhead and potentially impact the performance of the storage system, particularly in environments with limited computational capabilities.

•The implementation of a dual-thread encryption mode and the integration of encryption algorithms with the storage model may introduce additional latency into the data storage and retrieval process. Consequently, this could result in delayed response times for data access operations, posing challenges for applications requiring real-time or low-latency access to data.

•Introducing custom encryption algorithms and modifications to the Hadoop storage model could potentially lead to compatibility and interoperability issues. This might limit the system's ability to seamlessly integrate with established tools, frameworks, or third-party services that rely on traditional Hadoop implementations.

4. PROPOSED SYSTEM

Addressing the evolving security landscape of large-scale network systems, CP-HABE (Hierarchical Attribute-Based Distributed Provable Data Possession) emerges as a pioneering authentication protocol. It introduces a novel dual scheme approach, wherein each subgroup is autonomously managed by a trusted group security intermediary. This approach treats each subgroup akin to a distinct multi-cloud group, thereby dispersing the workload across multiple entities and mitigating reliance on a single entity. The essence of CP-HABE lies in its capability to distribute authentication and data possession processes, thereby bolstering both security and scalability. Leveraging hierarchical attribute-based access control, CP-HABE not only ensures secure data access but also streamlines the management of permissions and access policies within intricate network systems. This innovative protocol sets the stage for secure, streamlined, and distributed authentication and data possession in the realm of multi-cloud and large-scale network environments.

4.1 ADVANTAGES

•CP-HABE introduces a novel strategy by distributing authentication and data possession tasks among multiple trusted entities, reducing reliance on a single point of vulnerability and bolstering system security.

•By treating each subgroup as its own multi-cloud group, CP-HABE efficiently spreads the workload, alleviating strain on any single entity. This adaptability is crucial in expansive network systems with dynamic user numbers and data volumes.

•In CP-HABE, authentication tasks are distributed across several entities, preventing congestion and enhancing system performance. This ensures smooth authentication even in widely distributed and ever-changing networks.

•The hierarchical attribute-based access control in CP-HABE simplifies permissions and access rule management for administrators in complex network configurations. This streamlined management enables more effective establishment and enforcement of access controls, resulting in improved overall security.

•CP-HABE's hierarchical attribute-based approach allows for tailored access policies based on user attributes, ensuring precise control over data access. This guarantees that users only access information they are authorized to view, enhancing data security.

4.2 HADOOP

Hadoop, an innovative open-source framework, has transformed the landscape of big data processing and storage. Leveraging its distributed computing architecture, Hadoop enables organizations to efficiently handle and analyze vast datasets across clusters of commodity hardware. By harnessing Hadoop's scalability and fault tolerance capabilities, businesses can glean valuable insights from a variety of data sources, empowering informed decision-making and driving innovation. The extensive adoption of Hadoop underscores its essential contribution in confronting the formidable obstacles posed by the exponential growth of data in today's digital environments.

4.3 CP-HABE

Cypher text Hierarchical Attribute-Based Encryption (CP-HABE) represents an innovative cryptographic technique poised to revolutionize data security in contemporary computing environments. This cutting-edge approach provides a robust solution for enforcing access control policies based on user attributes within hierarchical structures. Through CP-HABE, organizations can ensure secure and efficient data sharing while maintaining precise control over access permissions. This technology enables administrators to define intricate access policies encompassing multiple levels of hierarchical attributes, thus safeguarding sensitive information from unauthorized access. CP-HABE's seamless integration with existing systems makes it a versatile solution for protecting data across various applications, from cloud computing to IoT ecosystems. As the digital landscape evolves, CP-HABE emerges as a fundamental component in the pursuit of resilient and adaptable data security solutions.

4.4 BIG DATA

Big data serves as a transformative catalyst reshaping industries and fundamentally altering how organizations manage and analyze vast quantities of information. Defined by the 'three Vs'—volume, velocity, and variety—big data encompasses datasets of unparalleled size, generated at remarkable speeds, and spanning diverse sources and formats. This inundation of data presents both challenges and opportunities for enterprises striving to derive actionable insights and drive informed decision-making. By harnessing sophisticated analytics tools and technologies, organizations can unlock the potential of big data to uncover concealed patterns, trends, and correlations that traditional methods might overlook. From predictive analytics and machine learning to real-time data processing and sentiment analysis, big data empowers businesses to gain a competitive advantage, optimize operations, enrich customer experiences, and foster innovation. Nevertheless, realizing the full potential of big data necessitates robust data management strategies, scalable infrastructure, and adept data professionals capable of navigating the intricacies of large-scale data ecosystems. As big data continues to evolve and proliferate, organizations must adapt and embrace data-driven approaches to thrive in an increasingly interconnected and data-rich environment.

5.MODULE DESCRIPTION

Key Generation

Each member of the group generates their own public and private keys within the Key Generation module. Using a random key, users output both their public and private keys. Let's assume user u1 is the original creator of the shared data and is therefore the first user in the group. Additionally, the original user creates a public user list (UL) containing the IDs of all users in the group. This list is then signed by the original user and made public.

Encryption

The plaintext data is divided into multiple blocks, with each block encrypted using the public key. A signature is generated for authentication purposes. The ciphertext of each block, along with its signature, block ID, and signer ID, is uploaded by the user. Metadata and key details are stored in the Public Verifier for public auditing.

Authentication

The subsequent user provides the filename to retrieve a file and receives the confidential key. Upon inputting the confidential key, if valid, the user successfully deciphers the downloaded file. If incorrect, the Public Verifier blocks the user. When the confidential key is valid, each block is decrypted, and the signature is verified. If signatures match, all blocks are merged to obtain the original file.

Privacy-Preserving Public Auditing

The Public Verifier method involves blocking users entering incorrect secret keys, adding them to the collision user list. When attempting to download a file, the Data Cloud Server provides information about blocked users. To resolve collisions, users seek assistance from the Public Verifier. Once unrevoked, they can download files using their corresponding secret key. Proxy re-signatures allow the Data Cloud Server to re-sign blocks previously signed by collision users with a resigning key.

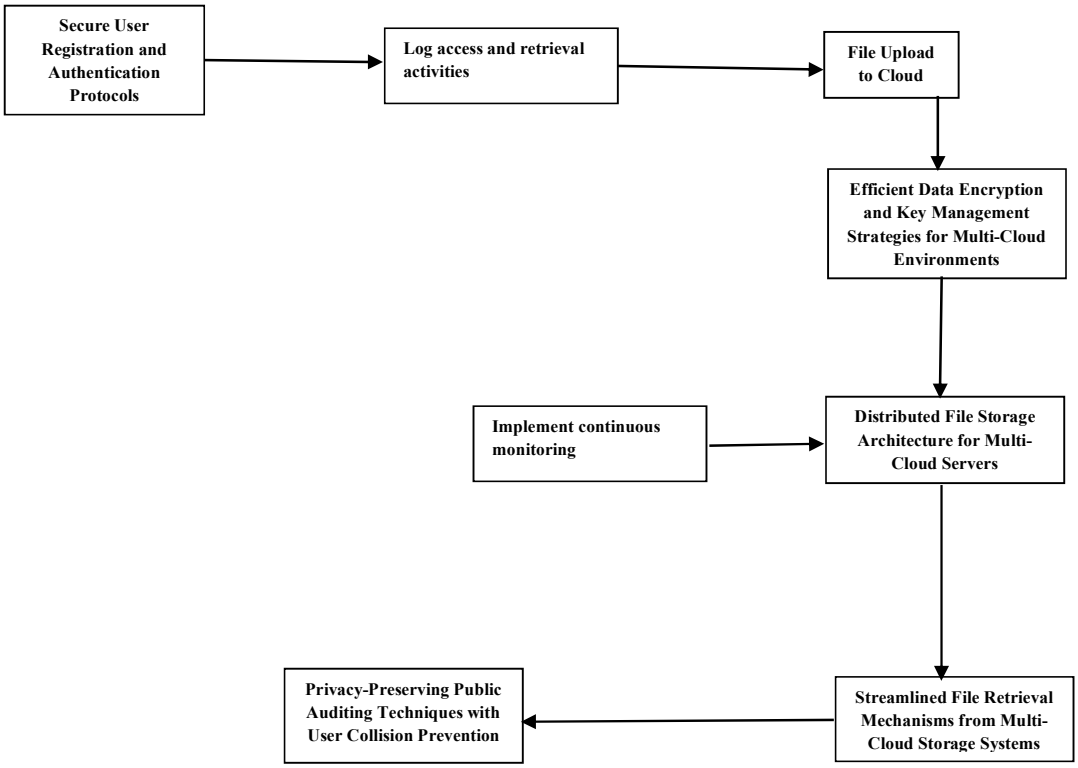


Figure 1. Block diagram

6. RESULT ANALYSIS

In large-scale network systems, the CP-HABE protocol introduces a dual scheme approach to authentication and data custody, distributing responsibility across multiple trusted organizations. This decentralized structure enhances security and scalability by optimizing resource utilization and reducing the risk of a single point of failure. CP-HABE enables secure data access and simplifies permission management in complex networks by leveraging hierarchical attribute-based access control.

algorithm	accuracy
Existing system	75
Proposed system	81

Table 1. Comparison table

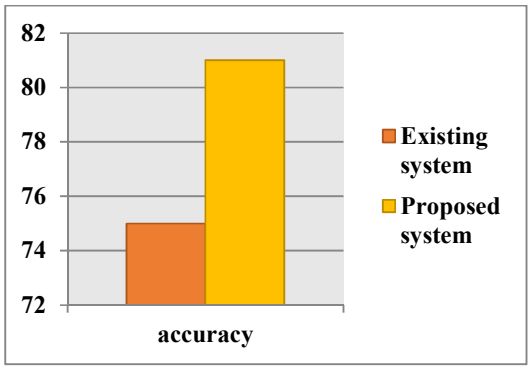
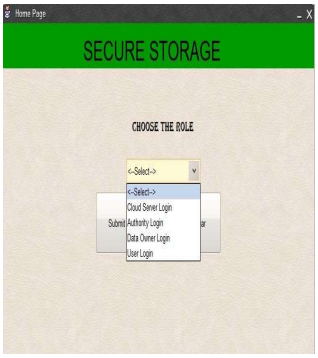
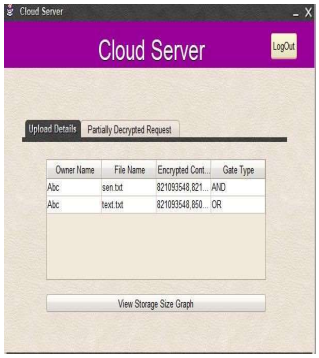
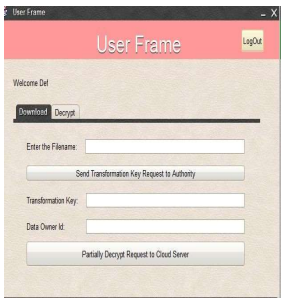


Figure 2. Comparison graph

6.1 SCREENSHOTS







7. CONCLUSION

Multi-cloud storage can benefit from a provable data possession (PDP) scheme, allowing users to ensure data integrity without needing to download the entire dataset. This is particularly valuable for safeguarding sensitive data and validating shared data across multiple clouds. The PDP scheme for multi-cloud environments typically involves three key components: a user client, a cloud server, and an identity management system (IMS). The user client is responsible for generating and verifying proofs of possession (POPs) and public proofs of possession (PPOPs). Meanwhile, the cloud server stores user data and generates POPs and PPOPs upon request. The IMS plays a crucial role in managing user identities and their associated public keys, as well as maintaining a revocation list (RL) to track revoked users.

8. FUTURE WORK

PDP schemes pose significant computational demands, especially when handling large data volumes. Future research should prioritize the development of more efficient PDP schemes and techniques to reduce the overhead associated with generating and verifying POPs and PPOPs. Presently, PDP schemes primarily target simple data structures like files and databases. However, there is an opportunity for future investigations to focus on creating PDP schemes capable of accommodating more complex data structures, such as graphs and networks. Furthermore, current PDP schemes assume static data storage in the cloud. To overcome this limitation, future research could explore the development of PDP schemes capable of supporting dynamic data, including datasets subject to frequent updates or deletions.

9. REFERENCES

[1] Enterprise Information Systems published an article on Industry 4.0 and big data innovations authored by G. Li, J. Tan, and S. S. Chaudhry in 2019.

[2] Future Generation Computer Systems featured a paper on big data driven internet of things for credit evaluation and early warning in finance by C. Wen, J. Yang, L. Gan, and Y. Pan in 2021.

[3] EURASIP Journal on Wireless Communications and Networking published a research paper on intelligent medical big data system based on Hadoop and blockchain by X. Zhang and Y. Wang in 2021.

[4] Mathematical Problems in Engineering presented a Hadoop-based medical image storage and access method for examination series.

[5] Journal of Big Data conducted a survey on data storage and placement methodologies for cloud big data ecosystem.

6. W. Rajeh, Journal of Information Security, 13 (2) (2022) 23–42, Hadoop distributed file system security problems and investigation of unauthorized access issue

7. Array. 1-2 (4) (2019) 1–8, G. S. Bhathal, A. Singh, Big data: Hadoop framework weaknesses, security challenges and attacks.

8. International Journal of Pure and Applied Mathematics, 120 (6) (2020), 11767–11784; Kapil, A. Agrawal, R. A. Khan, Big data security challenges: Hadoop perspective.

9. "Improvised distributions framework of hadoop: A review," by B. H. Husain, S. R. Zeebaree, et al., International Journal of Science and Business, 5 (2) (2021), 31–41.

10. Data protection on Hadoop distributed file system by applying encryption algorithms: a systematic literature review by M. Naisuty, A. N. Hidayanto, N. C. Harahap, A. Rosyiq, and G. M. S. Hartono, Journal of Physics Conference Series, 1444 (4) (2020) 1–8.