International Journal of Engineering Research and Science & Technology

VERIFIABLE AND SECURE DATA DEDUPLICATION WITH REAL TIME DATA INTEGRITY CHECKING WITH A CLOUD

BIRU SHIVA[1], RAVANAM AMRUTHA[2], R. RAGHAVENDAR REDDY[3],

TIRUPATI VENKATESH [4], KONDURU HARINATH [5]

[1]Assistant Professor, Dept. of CSE, Guru Nanak Institutions Technical Campus**,** Ibrahimpatnam, Telangana, India

[2, 3, 4, 5] B.Tech, Dept. of CSE, Guru Nanak Institutions Technical Campus**,** Ibrahimpatnam, Telangana, India

ABSTRACT— With the rapid growth of technologies, vast volumes of data are being generated and stored in the cloud. To reduce storage costs, data deduplication has emerged as a crucial technique. However, most existing blockchain-based deduplication schemes treat the blockchain as a passive storage medium and rely heavily on edge-level security protocols, making the system susceptible to data leakage and unauthorized access during the deduplication process. Furthermore, these schemes often lack integrated mechanisms for real-time data integrity verification and database authentication. This paper proposes a novel deduplication framework that ensures both verifiable and secure data deduplication with real-time data integrity checking in a cloud environment. The proposed system leverages blockchain technology not just for storage, but as an active trust layer by incorporating smart contracts and cryptographic proofs. These components enable real-time verification of data integrity and authentication of data before deduplication occurs. By addressing the limitations of existing methods, our framework significantly enhances data security, reduces redundancy, and ensures trustworthiness in cloud-based IoT data management.

Index Terms – Data deduplication, blockchain, cloud storage, Internet of Things (IoT), smart contracts, verifiable deduplication, data integrity verification, cryptographic proofs, cloud security, secure data storage, authentication, blockchain-based cloud computing.

I. INTRODUCTION

In recent years, IoT technology has seen significant development and advancement, enabling a multitude of applications supported by IoT devices. This rapid evolution is driving widespread development and innovation in IoT devices. Many industries are leveraging IoT devices to generate, collect, and process data for their business operations. This approach facilitates fast and efficient fine-grained data collection through advanced processing systems. In this context, data produced and collected by IoT devices is both vast and continuously evolving. The industrial IoT (IIoT) environment is one of the prominent areas where IoT data, generated from various sensors, is utilized for data reception, tracking, processing, and storage in numerous applications. In essence, IIoT represents a transformative leap in industrial operations, embedding advanced sensor technologies, connectivity, and data analytics into manufacturing and industrial environments. By interlinking machinery, processes, and personnel through the Internet, IIoT enhances operational efficiency, predictive maintenance, and overall productivity. This interconnected ecosystem generates massive volumes of data, necessitating sophisticated data management and security solutions to harness its full potential. Typically, cloud computing is employed to store this large volume of data for further querying or processing. However, the data generated by IoT devices in industrial settings often contains sensitive information related to product details. Therefore, it is crucial to preserve the privacy of this data before storing it in the cloud. A practical solution is to encrypt the data before outsourcing it to cloud storage. In practice, data from these devices may result in duplicate entries. When encrypted, these duplicates lead to redundant encrypted data, which incurs additional computational costs, communication costs, and storage overhead. To date, numerous solutions have been proposed for the deduplication of encrypted data. Convergent encryption, introduced by Douceur et al. and Storer et al, is a cryptographic technique that aims to combine data confidentiality with data deduplication capabilities.

II. RELATED WORK

A. Achieving efficient and privacy-preserving cross-domain big data deduplication in cloud

Secure data deduplication can significantly reduce communication and storage

overheads in cloud storage services and has potential applications in our big data-driven society. Existing data deduplication schemes are generally designed to either resist brute-force attacks or ensure efficiency and data availability, but not both conditions. We are also not aware of any existing scheme that achieves accountability, in the sense of reducing duplicate information disclosure (e.g., to determine whether plaintexts of two encrypted messages are identical). In this paper, we investigate a three-tier cross domain architecture and propose an efficient and privacy-preserving big data deduplication in cloud storage (hereafter referred to as EPCDD). EPCDD achieves both privacy-preserving and data availability, and resists brute-force attacks. In addition, we take accountability into consideration to offer better privacy assurances than existing schemes. We then demonstrate that EPCDD outperforms existing competing schemes, in terms of computation, communication and storage overheads. In addition, the time complexity of duplicate search in EPCDD is logarithmic.

B. *Efficient data aggregation and duplicate removal using grid-based hashing in cloud-assisted industrial IoT*

Industrial Internet of Things (IIoT) involves the incorporation of sensors, devices, and equipment with internet connectivity and data processing abilities. This connectivity allows sensors to collect and exchange data and communicate with each other. The proliferation of sensors and data-producing devices in the industrial sector has led to high volumes of data, causing data duplication and other issues in storage, processing, and resource consumption. To address data duplication becomes vital to guarantee effective operation and resource optimization. Existing data deduplication techniques in IIoT environments often struggle with efficiency and scalability. Current deduplication approaches may involve frequent and unnecessary hash generation for minor variations in sensor readings, resulting in excessive computational overhead, higher energy consumption, and reduced network lifetime. These limitations highlight the need to propose the Grid Hashing-based Efficient Data Aggregation (GH-EDA) scheme, a comprehensive solution that uses effective data aggregation, preprocessing, and region splitting, and employs an Extended Merkle Grid for efficient deduplication. The scheme begins with the aggregation of sensor data, followed by preprocessing steps to filter out

irrelevant or noisy data. Subsequently, the data is partitioned into regions and refined to improve resource utilization, thereby enabling fast duplicate detection while minimizing the number of comparisons. Key features of the proposed scheme include a threshold based approach to hash generation, guaranteeing that only substantial changes produce new hash values. Extensive simulations are conducted using Network Simulator-3. The performance of the proposed scheme is evaluated using metrics such as space reduction, search time, network lifetime, computation time, average latency, and energy utilization. Comparisons with existing techniques demonstrate the superior performance of the GH-EDA.

C. GreedyGD: Enhanced generalized deduplication for direct analytics in IoT

The exponential growth of data generated by the Internet of Things presents significant challenges for data communication, storage, and analytics. Consequently, organizations often face high costs when attempting to leverage their own data. Novel techniques that holistically optimize data storage and analytics in IoT systems are therefore required. One promising approach is generalized deduplication (GD), which is a lossless compression technique that delivers high compression while also enabling low-cost random access directly on compressed data. In this article, we introduce GreedyGD, a novel GD data compression algorithm that offers reliable, efficient data analytics, along with more compression and faster runtime compared to previous GD compressors. Evaluating GreedyGD on 18 real-world datasets revealed excellent performance: a 11.2× speed-up, 1.6× more compression, and more accurate and reliable analytics while using 4× less data compared to previous GD compressors.

D. Secure data deduplication with dynamic access control for mobile cloud storage

Data deduplication is of vital importance for mobile cloud computing to cope with the explosive growth of outsourced mobile data. In order to ensure the privacy of sensitive mobile data against an untrusted cloud, Message-Locked Encryption (MLE) has been proposed to enable deduplication over ciphertext. However, MLE prohibits data access control since it uses deterministic content-derived encryption keys. Recently, a lightweight rekeying-aware encrypted deduplication system (REED) has been proposed to achieve dynamic access control for secure data deduplication. However, REED is vulnerable to key-retaining attack

and stub-retaining attack, which leads to insecure access revocation, and thus cannot support secure dynamic access control. In response, we present AC-Dedup, an encrypted deduplication storage system that supports secure dynamic access control for mobile cloud storage. At the core of AC-Dedup are two novel encryption techniques named mixed message locked encryption and random stub re-encryption to resist the two types of attacks, respectively. To the best of our knowledge, AC-Dedup is the first practical system that achieves secure data deduplication and secure dynamic access control simultaneously. We conduct security analysis and experimental evaluation on mobile device and cloud platform with real world IoT datasets. The results show that AC-Dedup enables secure and efficient dynamic access control while preserving deduplication effectiveness.

III. PROPOSED METHODOLOGY

We propose an integrity verification algorithm that cross-references hash values stored in the cloud and blockchain, enabling fast and lightweight integrity checks. Finally, extensive experiments and comparative evaluations demonstrate that VERDUP significantly improves deduplication performance compared to existing methods. Instead of relying solely on edge-level security, the system uses cryptographic proofs and smart contracts to verify data authenticity and integrity during the deduplication process. This ensures that only legitimate data from authenticated IoT devices is stored and deduplicated, thereby reducing the risk of data leakage and improving overall trust and reliability in cloud storage.

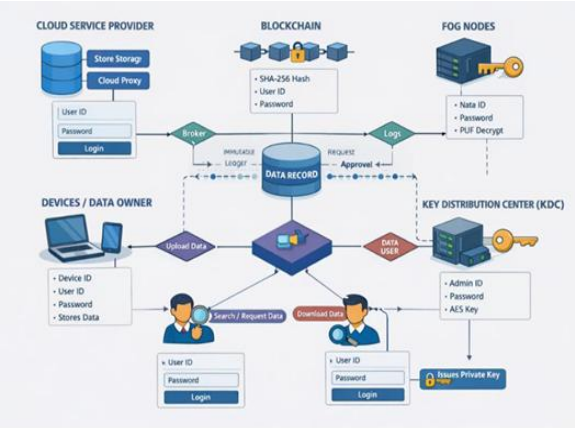


Fig. 1: System Overview

1. CSP

The Cloud service provider has a login with a user id and password. The CSP has a store in a cloud storage in a database. The CSP has a cloud proxy server in a data.

2. BLOCKCHAIN

The Blockchain has a login with a sha hash - 256 with a user id and password. The blockchain has a data integrity with a data.

The blockchain has a user authentication approval a data.

3. FOG NODES

The fog nodes has a register with a details and login with a user id and password. The fog nodes has a record a data. The fog nodes has a puf storage a decrypt a data.

4. DEVICES

Devices is a shows a system information. The devices have in a data owner module. The data owner register with a devices and login with a user id and password. The Data owner has a store a data. The data stores in a database.

5. DATA USER

The data user has a register with a user details and login with a user id and password. The Data user has directly cannot login it has to approve a data. The data user has a search a data. The data will be encrypted and get a request. After it will get a private key then has to download a data.

6. KEY DISTRIBUTING CENTER

The KDC has a login with a user id and password. The login it will has a AES to a private key to secure a data. The KDC has a key share to a user.

IV. EXPERIMENTAL RESULTS AND ANALYSIS

The experimental results show that the proposed blockchain-based deduplication framework effectively performs secure and verifiable data deduplication while ensuring real-time data integrity verification. The integration of smart contracts and cryptographic proofs enhances authentication and prevents unauthorized access during the deduplication process. The framework significantly reduces redundant data storage while maintaining data security and trustworthiness.

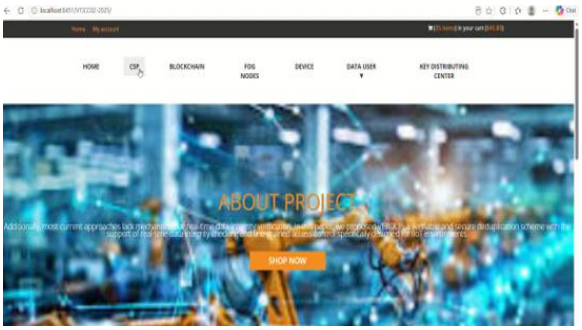


Fig. 2: Home Page

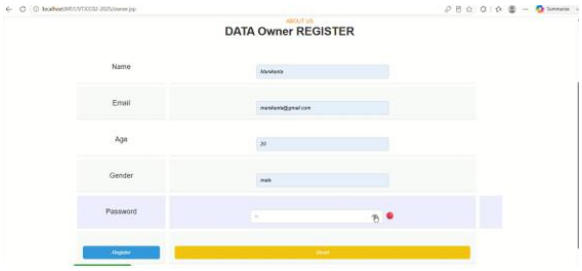


Fig. 3: Registration

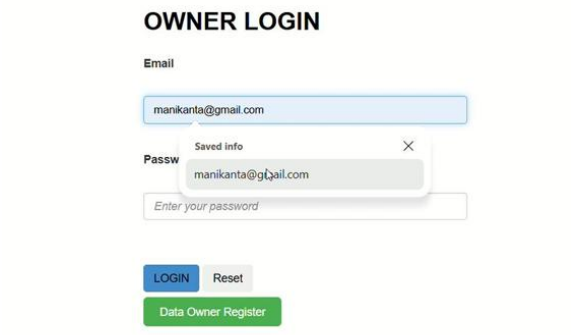


Fig. 4: Login

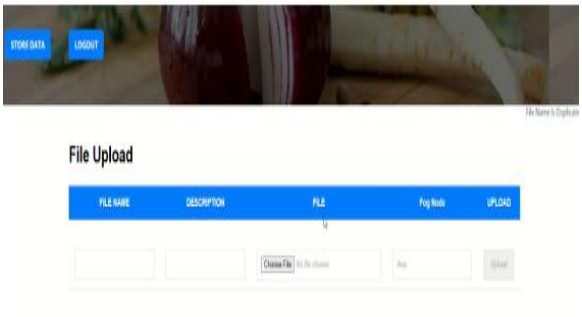


Fig. 5: Admin Dashboard

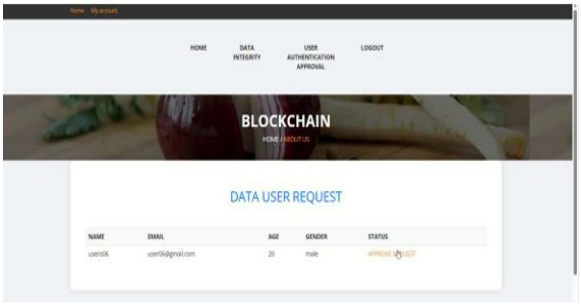


Fig. 6: Data User Request



Fig. 6: Download Data

V. CONCLUSION

In this work, we proposed VERDUP, a secure and efficient data deduplication scheme integrated with PUF-based authentication, which leverages an auto-incremental technique to enhance the verification process. The scheme incorporates fog and cloud deduplication using a tree classification structure, which significantly reduces the search space during deduplication operations. At the fog level, VERDUP introduces temporal removal techniques in the recording list, optimizing storage utilization by eliminating redundant data. Additionally, integrity checking is enforced through smart contracts, ensuring the authenticity and security of data transactions in a decentralized manner. This comprehensive framework improves both performance and security in IIoT environments by efficiently managing resources across fog and cloud infrastructures while maintaining strong data integrity and user authentication.

REFERENCES

[1] J. R. Douceur, A. Adya, W. J. Bolosky, P. Simon, and M. Theimer, ‘‘Reclaiming space from duplicate files in a serverless distributed file system,’’ in Proc. 22nd Int.

Conf. Distrib. Comput. Syst., 2002, pp. 617–624.

[2] M. W. Storer, K. M. Greenan, D. D. E. Long, and E. L. Miller, ‘‘Secure data deduplication,’’ in Proc. 4th ACM Int. Workshop Storage Secur. Survivability, Oct. 2008, pp. 1–10.

[3] S. Keelveedhi, M. Bellare, and T. Ristenpart, ‘‘DupLESS: Server-aided encryption for deduplicated storage,’’ in Proc. 22nd USENIX Secur. Symp., Aug. 2013, pp. 179 194.

[4] J. Liu, N. Asokan, and B. Pinkas, ‘‘Secure deduplication of encrypted data without additional independent servers,’’ in Proc. 22nd ACM SIGSAC Conf. Comput. Commun. Secur., Oct. 2015, pp. 874–885.

[5] Z. Pooranian, K.-C. Chen, C.-M. Yu, and M. Conti, ‘‘RARE: Defeating side channels based on data-deduplication in cloud storage,’’ in Proc. IEEE Conf. Comput. Commun. Workshops (INFOCOM WKSHPS), Apr. 2018, pp. 444–449.

[6] Z. Yan, W. Ding, X. Yu, H. Zhu, and R. H. Deng, ‘‘Deduplication on encrypted big data in cloud,’’ IEEE Trans. Big Data, vol. 2, no. 2, pp. 138–150, Jun. 2016.

[7] J. Xu, E.-C. Chang, and J. Zhou, ‘‘Weak leakage-resilient client-side deduplication of encrypted data in cloud storage,’’ in Proc. 8th ACM SIGSAC Symp. Inf., Comput. Commun. Secur., May 2013, pp. 195–206.

[8] N. Kaaniche and M. Laurent, ‘‘Asecure client side deduplication scheme in cloud storage environments,’’ in Proc. 6th Int. Conf. New Technol., Mobility Secur. (NTMS), Mar. 2014, pp. 1 7.

[9] T.-Y. Youn, K.-Y. Chang, K.-H. Rhee, and S. U. Shin, ‘‘Efficient clientside deduplication of encrypted data with public auditing in cloud storage,’’ IEEE Access, vol. 6, pp. 26578–26587, 2018.

[10] X. Yang, R. Lu, K. K. R. Choo, F. Yin, and X. Tang, ‘‘Achieving efficient and privacy-preserving cross-domain big data deduplication in cloud,’’ IEEE Trans. Big Data, vol. 8, no. 1, pp. 73–84, Feb. 2022.

AUTHORS Profile



Mr. BIRU SHIVA has received his B. Tech CSE from JNTUH, Hyderabad in 2013, M. Tech in Computer science and engineering from JNTUH, Hyderabad in

2016 and He has been dedicated to the teaching field for the last 6 years. present he is working as Assistant Professor in CSE Department at Guru Nanak Institutions Technical Campus, Ibrahimpatnam, Telangana, India.



Ms. RAVANAM AMRUTHA pursuing B.Tech degree in Computer Science and Engineering at Guru Nanak Institutions Technical Campus affiliated to JNTUH in 2022-2026.



Mr. R. RAGHAVENDAR REDDY pursuing B.Tech degree in Computer Science and Engineering at Guru Nanak Institutions Technical Campus affiliated to JNTUH in 2022-2026.



Mr. TIRUPATI VENKATESH pursuing B.Tech degree in Computer Science and Engineering at Guru Nanak Institutions Technical Campus affiliated to JNTUH in 2022-2026.



Mr. KONDURU HARINATH pursuing B.Tech degree in Computer Science and Engineering at Guru Nanak Institutions Technical Campus affiliated to JNTUH in 2022-2026.