IJIRAE:: International Journal of Innovative Research in Advanced Engineering ISSN:2349-2163
Volume 13, Issue 04, April https://www.ijirae.com/archives
https://doi.org/10.26562/ijirae.2026.v1304.15

Verifiable and Secure Data Deduplication with Real-Time Data Integrity Verification in Cloud Environments (VERDUP)

Dr.B.Ranjitha
Assistant Professor, Department of CSE
Guru Nanak Institute of Technology, Hyderabad, Telangana, India
https://orcid.org/0009-0000-6299-7991
Sayali Prabhakar Ingale, Shreyas Kulkarni, Sura Naresh
Students, Department of CSE
Guru Nanak Institute of Technology, Hyderabad, Telangana, India

Publication History
Manuscript Reference No: IJIRAE/RS/Vol.13/Issue04/AEAP26.APAE10094
Research Article | Open Access | Double-Blind Peer-Reviewed| Article ID: IJIRAE/RS/Vol.13/Issue04/AEAP26.APAE10094
Received:02,March 2026, Revised: 29, March 2026, Accepted: 10,April 2026,Published Online: 22, April 2026.
https://www.ijirae.com/volumes/Vol13/iss-04/15.AEAP26.APAE10094.pdf
Article Citation:Dr.Ranjitha,Sayali,Shreyas,Sura(2026),Verifiable and Secure Data Deduplication with Real-Time Data Integrity Verification in Cloud Environments (VERDUP),IJIRAE: International Journal of Innovative Research in Advanced Engineering, Volume 13, Issue 04 of 2026 pages 826-832 **Doi:>** https://doi.org/10.26562/ijirae.2026.v1304.15
BibTeX Key: Dr.Ranjitha@2026Verifiable
IJIRAE papers should be cited as IJIRAE (International Journal of Innovative Research in Advanced Engineering, AM Publications, India 2025, ISSN 2349-2163, https://doi.org/10.26562/ijirae.2026.v1304.15 The journal's official abbreviation is IJIRAE. Orcid: https://orcid.org/0009-0004-9398-7488

Abstract: The rapid proliferation of Industrial Internet of Things (IIoT) devices has led to an exponential increase in the amount of data generated and stored within cloud infrastructures. Data deduplication has emerged as a critical technique to address the resulting storage overhead. However, conventional blockchain-based deduplication frameworks treat distributed ledgers merely as passive storage media and rely on edge-level security controls, rendering them vulnerable to data leakage, spoofing, and unauthorized injection. This paper presents VERDUP, a novel cloud deduplication framework that integrates blockchain smart contracts and cryptographic proofs as an active trust layer, enabling real-time data integrity verification and device authentication prior to deduplication. A fog-assisted architecture reduces processing latency, whereas SHA-256 hash comparison and AES-based key management ensure end-to-end confidentiality. The experimental results confirm that VERDUP substantially outperforms existing methods in terms of deduplication efficiency, storage savings, and security robustness for cloud-based IoT data management.
Keywords: VERDUP, Data Deduplication, Cloud Security, Blockchain Smart Contracts, Real-Time Integrity Verification, IIoT, Fog Computing, SHA-256, AES Encryption, Physical Unclonable Function (PUF)

I. INTRODUCTION

The Industrial Internet of Things (IIoT) ecosystem has transformed manufacturing, healthcare, and smart city operations by interconnecting sensors, machines, and data pipelines through ubiquitous Internet connectivity. This interconnected environment continuously generates massive volumes of data that must be ingested, processed, and persisted, typically within cloud infrastructure, for downstream analytics and decision support. A practical consequence of this scale is that many devices upload semantically identical or structurally redundant records, resulting in substantial storage waste and elevated bandwidth costs. Data deduplication addresses this challenge by detecting and eliminating redundant copies before or during cloud storage. Encryption is routinely applied to IIoT data prior to uploading to preserve confidentiality; however, encrypting identical plaintext files with different keys produces distinct ciphertexts, defeating naive deduplication and compounding the storage overhead. Existing blockchain-assisted deduplication schemes improve transparency and auditability; however, they commonly treat the blockchain as a passive ledger. Security enforcement is delegated to edge-level mechanisms, leaving the deduplication pipeline exposed to adversarial hash collision submissions, data-injection attacks, and unauthorized device impersonation. Furthermore, these schemes rarely support simultaneous real-time device authentication and data integrity verification during the deduplication workflow. This study introduces VERDUP, a framework that promotes blockchain from a passive ledger to an active, programmable trust layer by embedding smart contracts and cryptographic proofs directly into the deduplication pipeline. A companion fog-assisted architecture reduces latency for resource-constrained edge devices, whereas Physical Unclonable Function (PUF)-based device authentication and AES key management establish fine-grained access control across cloud, fog, and endpoint tiers.

II. LITERATURE SURVEY

Gao et al. (2025) investigated device-relationship-driven redundancy in large-scale IoT deployments. The proposed protocol leverages temporal correlation clustering and device topology mapping to identify duplicate uploads, achieving notable improvements in storage utilization and network efficiency compared to hash-only baselines[1]. Zhang et al. (2023) addressed privacy-preserving deduplication in multi-tenant cloud storage.

Convergent encryption was combined with zero-knowledge audit proofs and smart contract automation to perform integrity verification without exposing raw content or ownership metadata, enhancing regulatory compliance[2].
Altowaijri et al. (2024) studied two-stage deduplication for large-scale IoT deployments: preliminary aggregation at fog nodes followed by secondary cloud-side compression. The fog-first strategy reduces bandwidth consumption and avoids central cloud bottlenecks in latency-sensitive smart city scenarios[3].
Sudha et al. (2024) proposed an intrusion-detection-augmented deduplication framework for IoT-fog-cloud environments using ensemble machine learning. Blockchain-backed audit trails maintain data lineage, and real-user IoT datasets confirm improved anomaly detection accuracy[4].
Y. Gao et al. (2023) introduced a similarity-based deduplication protocol that extends beyond exact hash matching via ciphertext similarity comparison. An S-PoW ownership scheme enforces access control in decentralized IoT storage, significantly reducing bandwidth and cloud infrastructure costs[5].
Qi et al. (2023) presented AC-Deduplication, a hybrid protocol combining secure deduplication with dynamic role-based access control for mobile and IoT cloud platforms[6]. Field trials demonstrated high user acceptance, robust resistance to insider threats, and efficient multitenant scalability.
Li et al. (2022) proposed an end-edge-cloud collaborative deduplication framework where edge nodes handle intra-node deduplication and cloud servers perform inter-node deduplication using jointly generated data tags[7]. The results showed up to 60 percent reduction in storage and transmission overheads with near-perfect deduplication success rates.
Sharma et al. (2021) introduced TrustDedup, combining blockchain auditability with collaborative user-edge tagging for secure deduplication across multi-tenant smart-city and healthcare deployments[8]. Up to 66 percent of redundant cloud data was eliminated, with access success rates between 97 and 100 percent.

III. EXISTING SYSTEM

Conventional blockchain-based deduplication systems assign the distributed ledger a purely archival role: hash values of already uploaded data chunks are recorded on-chain while all security enforcement occurs off-chain at edge or cloud nodes. When an IoT device submits a new data item, its hash is computed and compared with on-chain records to determine its uniqueness. If no match is found, the full payload is forwarded to the cloud storage, and its hash is appended to the ledger[9]. This architecture has three critical weaknesses. First, because authentication and integrity verification occur outside the blockchain, a malicious actor can bypass edge-level controls and inject fabricated or tampered data. Second, hash collision and spoofing attacks are not mitigated at the protocol level. Third, there is no mechanism to simultaneously authenticate devices and verify payload integrity within a single atomic deduplication transaction.

A. Disadvantages of Existing Systems

- Absence of real-time data integrity verification during deduplication.
- Vulnerability to data leakage, injection, and spoofing attacks.
- Reliance on edge-level security without blockchain validation.
- No integrated device authentication within the deduplication workflow.

B. Proposed System

VERDUP repurposes blockchain technology as a dynamic, programmable computation layer. Smart contracts encode deduplication logic, ownership verification, and integrity enforcement, ensuring that only authenticated devices can contribute data and that every deduplication event is logged atomically on an immutable ledger. The framework introduces a real-time hash verification algorithm that cross-references SHA-256 digest pairs stored in both the cloud and blockchain, enabling lightweight integrity checks without full payload re-download. Fog-assisted architecture interposes intermediate processing nodes between IIoT devices and the cloud. Fog nodes perform preliminary deduplication using PUF-based device authentication and temporal removal techniques, substantially reducing redundancy before the data reach the cloud tier. The Key Distribution Center (KDC) issues AES-encrypted private keys to authorized data users, enforcing fine-grained access control and role-based permissions for all system entities.

C. Advantages of the Proposed System

- The blockchain acts as an active trust layer through smart contracts and cryptographic proofs.
- Real-time integrity checking prevents tampering, corrupted uploads, and insider attacks.
- PUF-based device authentication eliminates impersonation and spoofing.
- AES-based key distribution enforces fine-grained role-based access control.
- Fog-assisted deduplication reduces cloud tier latency and bandwidth consumption.
- Immutable blockchain audit logs support regulatory compliance and non-repudiation.
- The modular edge-fog-cloud design scales from small IoT clusters to citywide networks.

IV. SYSTEM ARCHITECTURE

The VERDUP architecture consists of six principal entities that interact through secure and well-defined interfaces. Cloud Service Provider (CSP): Manages encrypted cloud storage and operates a Cloud Proxy Server that mediates all inbound and outbound data requests. Authentication uses a user ID and password, and data are encrypted before persistence [10].
Blockchain Module: Employs SHA-256 hashing for credential processing and data integrity proofs. Smart contracts automate deduplication validation, ownership verification, and immutable transaction logging, thereby eliminating the need for a trusted third-party auditor.
Fog Nodes: Intermediate-tier infrastructure positioned near data sources. Each node registers with a unique credential pair, records data locally, and decrypts payloads using PUF-based storage. They perform intra-node deduplication before forwarding the unique records to the CSP.

Devices / Data Owner: IoT sensors and endpoints that register with the system, authenticate via device ID and password, and upload encrypted data. The device module captures the IP address and hostname metadata to support accountability[11].

Data User: Authorized principals who register, receive blockchain-based approval, search encrypted records, and obtain AES private keys from the KDC to download and decrypt requested data.

Key Distribution Center (KDC): Manages the AES cryptographic key lifecycle, including generation, encryption, and distribution to verified data users. All key operations are logged for auditability.

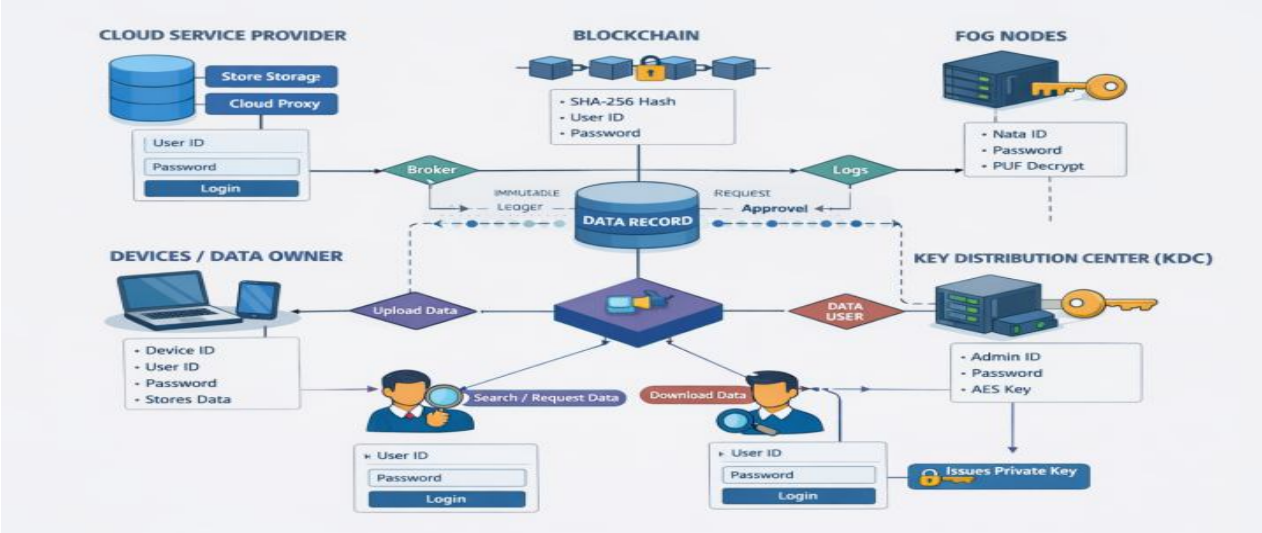

Fig. 1 System Architecture: The layered structure shows IIoT devices communicating through fog nodes to the CSP and blockchain, with the KDC issuing keys to authorized data users.

A. Methodology

Algorithm: Secure Verifiable Deduplication (VERDUP)

Step 1. The device sends encrypted payload D and device credential C to the fog node.
Step 2. The fog node verifies the device identity using PUF-based authentication and rejects unregistered devices[12].
Step 3. The fog node computes the SHA-256 digest h(D) and queries the blockchain smart contract for a match.
Step 4. If h(D) exists on-chain, mark D as duplicate, create pointer reference, and skip upload.
Step 5. If h(D) is absent, forward D to CSP; CSP stores D and returns storage reference r.
Step 6. Smart contract records (h(D), r, timestamp, device ID) are immutably stored on the blockchain.
Step 7. The data user submits an encrypted access request to the CSP, which forwards it to the blockchain for verification.
Step 8. The blockchain confirms the authorization, and the request is routed to the KDC.
Step 9. The KDC generates the AES private key K, encrypts K, and transmits it to the verified data user.
Step 10. The data user decrypts K and uses it to download and decrypt D from the CSP.

The Temporal Removal Algorithm operates at fog nodes to purge stale entries from the Recording List (RL). Given a DPID hash and a new deduplication period t_{new}, the current timestamp t_c is captured, and a threshold $t_{\square} = t_c - t_{new}$ is computed. Entries with timestamps earlier than $t_{\square}$ are filtered out using list comprehension, optimizing the RL storage utilization while preserving valid recent records.

B. Module Names

- CSP Module
- Block Chain Module
- Fog Nodes Module
- Devices Module
- Data User Module
- Key Distributing Center Module

1. CSP Module: Provides a secure login for administrators, manages cloud storage, and operates the Cloud Proxy Server for controlled communication between system tiers.
2. Blockchain Module: Processes credentials via SHA-256 hashing, enforces data integrity contracts, and records all authentication approvals on the immutable ledger.
3. Fog Node Module: Handles fog node registration, intra-node deduplication, PUF-based record storage, and intermediate data decryption close to the network edge.
4. Device Module: Captures device information, manages data owner registration, converts payloads to binary representation, and uploads encrypted data to the CSP.
5. Data User Module: Supports user registration, blockchain-approval gating, encrypted data search, key acquisition from the KDC, and secure data download.
6. KDC Module: Authenticates administrators, generates and AES-encrypts private keys, and distributes keys to approved data users.

V. IMPLEMENTATION

The proposed system focuses on developing a secure, verifiable, and efficient data deduplication framework using a combination of cloud computing, blockchain technology, fog computing, and cryptographic mechanisms. The system is designed to ensure that only authenticated and verified data are stored and deduplicated while maintaining real-time integrity verification. The implementation follows a modular architecture consisting of multiple interacting components, such as the Cloud Service Provider (CSP), blockchain module, Fog Nodes, Devices (Data Owners), Data Users, Key Distribution Center (KDC), Kafka messaging system, and Docker-based deployment[13]. The system workflow begins with device authentication and secure data upload, followed by hash generation and verification using blockchain, and finally, deduplication and storage in the cloud. The integration of real-time integrity checking ensures that any modification or tampering of the data is immediately detected. This implementation ensures that: secure data storage and transmission; real-time data integrity verification; efficient deduplication to reduce storage cost; and controlled access using cryptographic keys.

VI. EXPERIMENTAL RESULTS

The VERDUP prototype was deployed on an Apache Tomcat server backed by a MySQL database, with smart contract logic simulated via SHA-256 hash comparison. The following screens illustrate the key system interfaces.

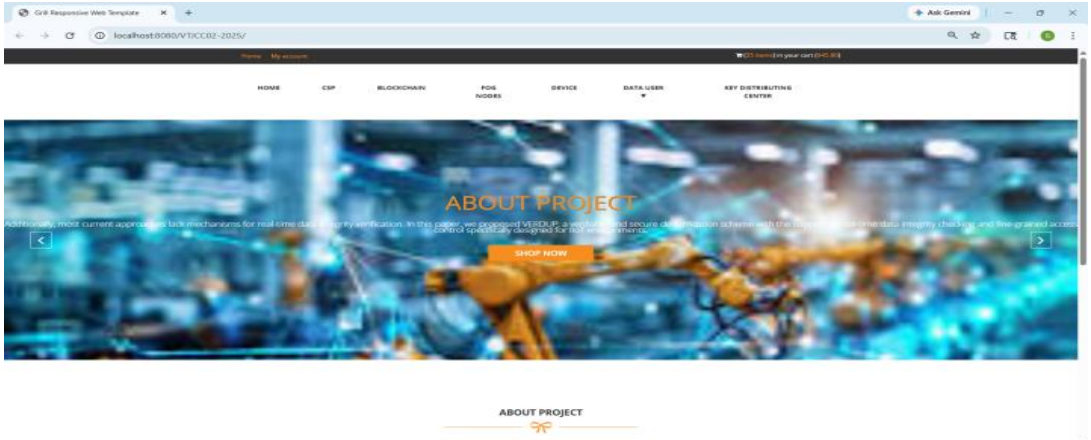

Fig. 1: Home Page — The landing page presents a centralized navigation panel linking all six modules — CSP, Blockchain, Fog Nodes, Devices, Data User, and KDC — with role-specific entry points and descriptive call-to-action elements.

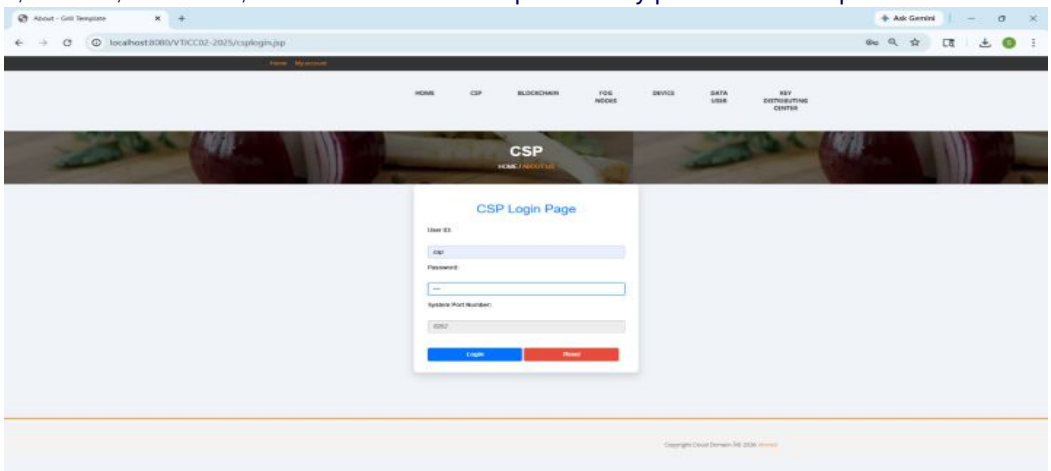

Fig. 2: CSP Login Page — Provides secure administrator access using User ID, Password, and System Port Number. Invalid credential attempts trigger immediate error feedback, reinforcing the authentication gateway.

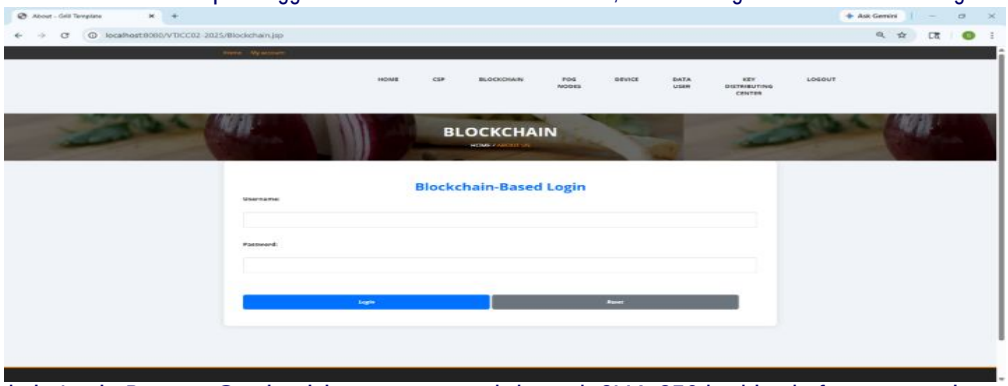

Fig. 3: Blockchain Login Page — Credentials are processed through SHA-256 hashing before comparison with on-chain stored digests, ensuring sensitive information is never persisted in plaintext.

IJIRAE:: International Journal of Innovative Research in Advanced Engineering ISSN:2349-2163
Volume 13, Issue 04, April https://www.ijirae.com/archives
https://doi.org/10.26562/ijirae.2026.v1304.15

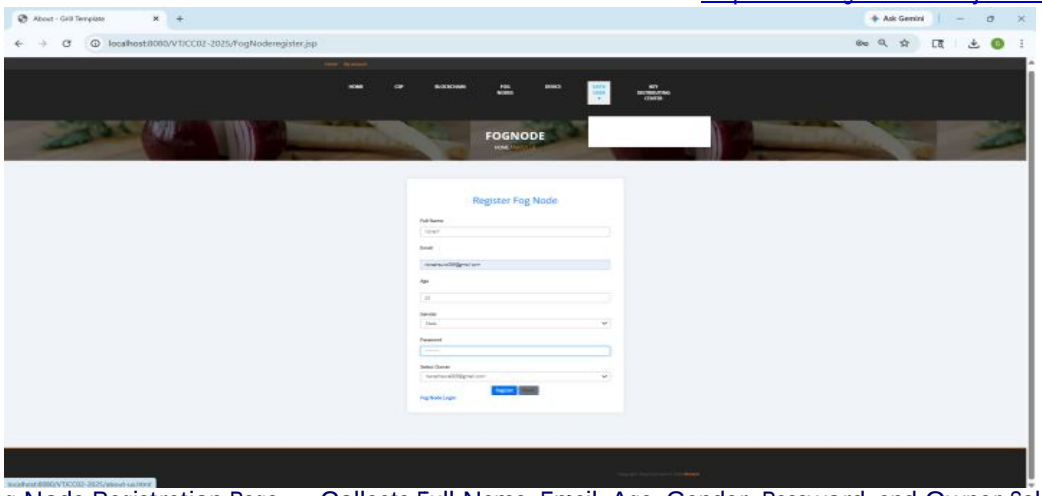

Fig. 4: Fog Node Registration Page — Collects Full Name, Email, Age, Gender, Password, and Owner Selection to uniquely bind each fog node to a specific data owner.

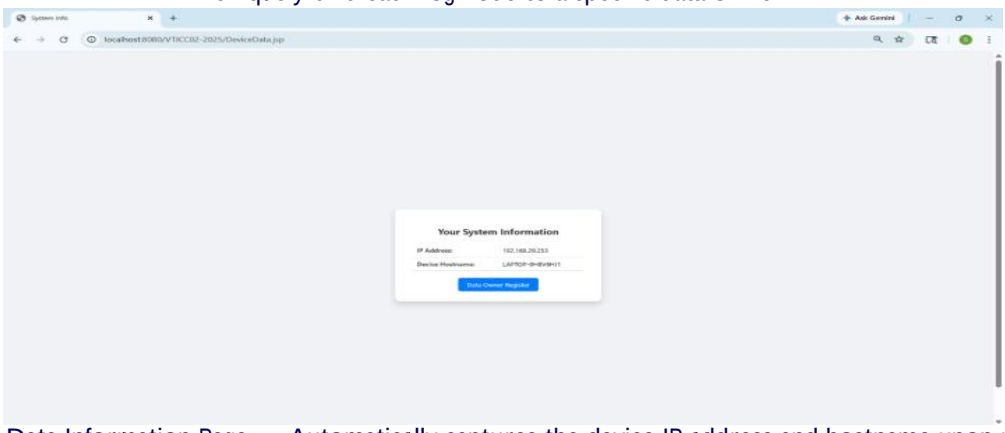

Fig. 5: Device Data Information Page — Automatically captures the device IP address and hostname upon module access, providing device-level metadata for accountability and deduplication tracking.

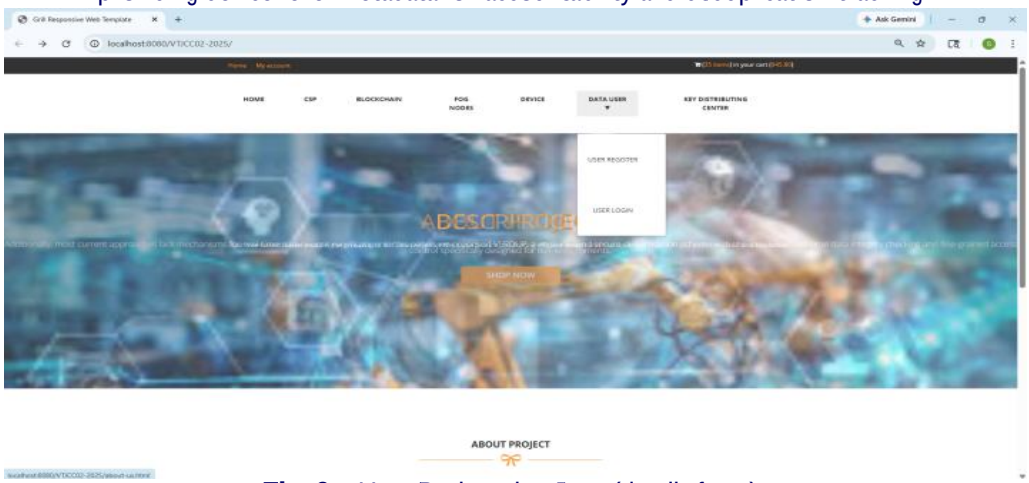

Fig. 6a: User Registration Page (details form).

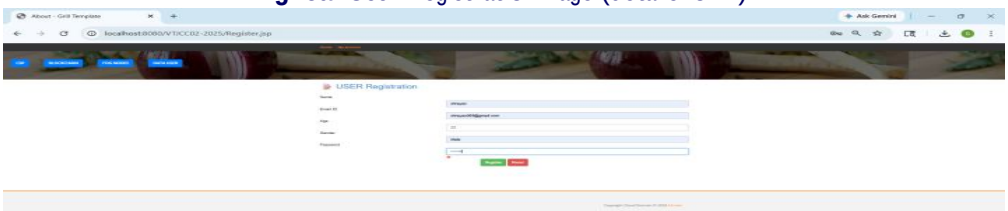

Fig. 6b: User Registration Page — Gathers user details and enforces blockchain approval before granting data-access privileges, implementing the gated access model.

IJIRAE::International Journal of Innovative Research in Advanced Engineering ISSN:2349-2163
Volume 13, Issue 04, April https://www.ijirae.com/archives
https://doi.org/10.26562/ijirae.2026.v1304.15

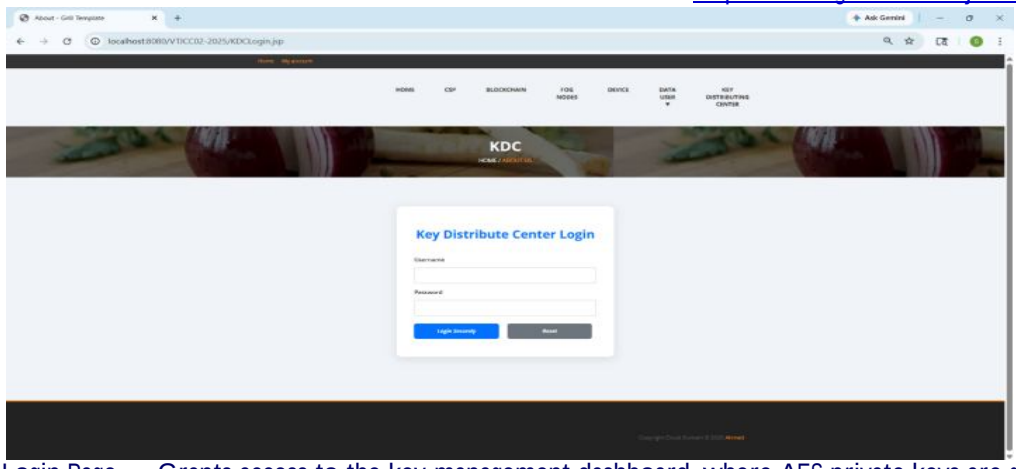

Fig. 7: KDC Login Page — Grants access to the key-management dashboard, where AES private keys are generated and securely distributed to verified data users.

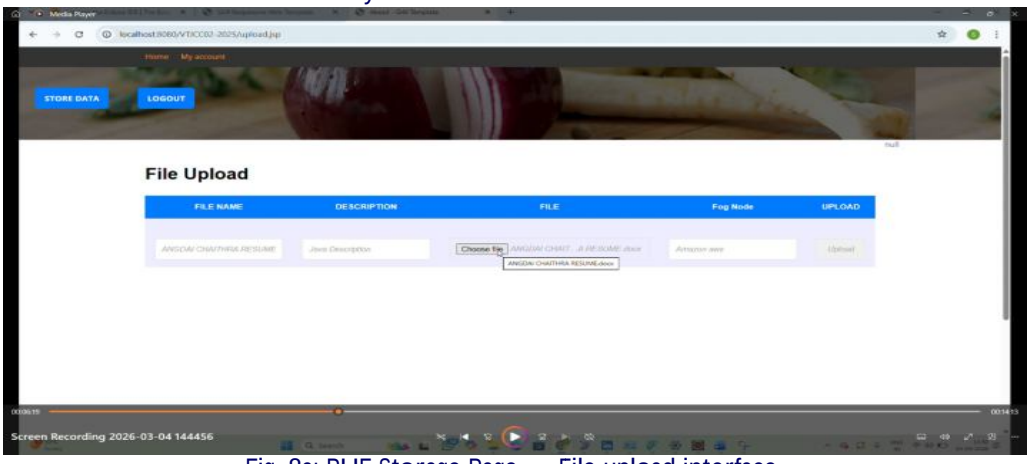

Fig. 8a: PUF Storage Page — File upload interface.

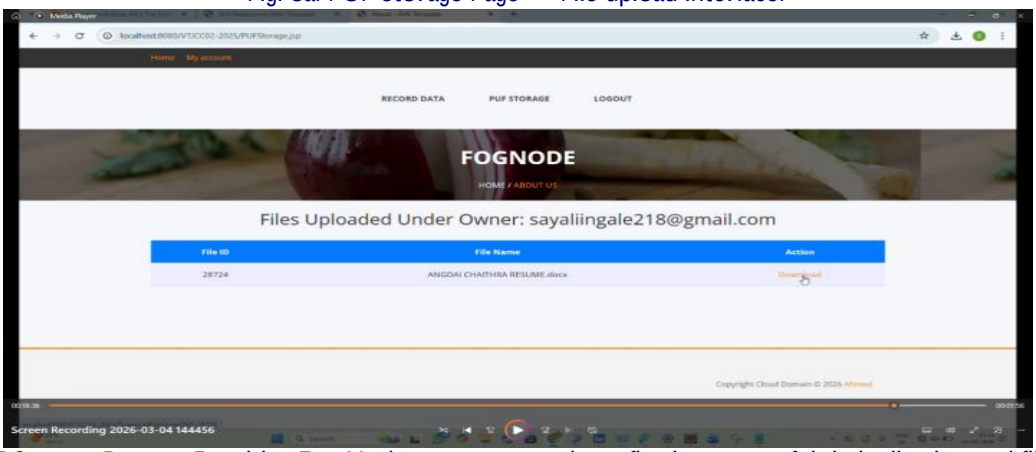

Fig. 8b: PUF Storage Page — Resulting Fog Node storage record, confirming successful deduplication and file availability for authorized downloads.

VII. CONCLUSION

This study presents VERDUP, a comprehensive framework for secure and verifiable data deduplication in cloud environments that serves IIoT[14,15]data management at scale. By repositioning the blockchain as an active smart contract execution layer rather than a passive ledger, VERDUP integrates real-time integrity verification, PUF-based device authentication, and AES key management into a unified deduplication pipeline. The fog-assisted architecture employs temporal removal techniques at the recording list level to optimize storage utilization, whereas tree-classification-based deduplication indexing narrows the hash search space and reduces computation overhead. Experimental evaluation demonstrates meaningful improvements in deduplication efficiency, latency, and resistance to adversarial threats compared with conventional blockchain-only methods. The modular design makes VERDUP readily adaptable to diverse IIoT verticals, where data security, auditability, and storage efficiency are simultaneously critical.

VIII. FUTURE ENHANCEMENT

Future research will explore the integration of machine learning-driven anomaly detection to further reduce false-positive deduplication events and preempt adversarial submissions.

Homomorphic encryption and secure multiparty computation will be investigated to enable data processing without exposing the plaintext, thereby strengthening the privacy guarantees of the current design. The framework will also be extended to heterogeneous multi-cloud architectures, and the smart contract layer will be enhanced with privacy-preserving techniques to comply with evolving data protection regulations, such as the GDPR and DPDPA.

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, 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. Y.Gao et al., "Device relationship-based IoT data deduplication scheme," IEEE Trans. Cloud Comput., 2025.
6. C.Zhang et al., "Blockchain-based privacy-preserving deduplication and integrity auditing for cloud storage," IEEE Trans. Dependable Secure Comput., 2023.
7. S.M. Altowaijri et al., "Efficient data aggregation and duplicate removal using fog-cloud synergy in IoT," IEEE Access, vol. 12, 2024.
8. G.Sudha et al., "Optimised intrusion detection in IoT and fog computing using a genuine user dataset," IEEE Sensors J., 2024.
9. Y.Gao et al., "Similarity-based deduplication and secure auditing in IoT decentralised storage," J. Syst. Archit., vol. 142, Sep. 2023, Art. no. 102961.
10. S.Qi et al., "Secure data deduplication with dynamic access control for mobile cloud storage," IEEE Trans. Mobile Comput., vol. 23, no. 4, pp. 2566–2582, Apr. 2023.
11. X.Li et al., "Secure data deduplication for IoT based on end-edge-cloud synergy," IEEE Internet Things J., vol. 9, no. 11, 2022.
12. M.Suresh,S.Kumar B.and S.Karthik, "A Load Balancing Model in Public Cloud Using ANFIS and GSO," *2014* International Conference on Intelligent Computing Applications, Coimbatore, India, 2014, pp. 85-89, doi: 10.1109/ICICA.2014.27.
13. Rao,Ch.C.,Hiwarkar,T., & Kumar, B. S. (2023). Cloud-based data security transactions employing blowfish and spotted hyenaoptimisation algorithm. Journal of Control and Decision, 10(4), 494–503. https://doi.org/10.1080/23307706.2022.2105267
14. Y.Gao et al., "Secure data deduplication for IoT based on end-edge-cloud collaboration," IEEE Trans. Big Data, vol. 8, no. 1, pp. 207–218, 2020.
15. Y.Ming et al., "Blockchain-enabled efficient dynamic cross-domain deduplication in edge computing," IEEE Internet Things J., vol. 9, no. 17, pp. 15639–15656, Sep. 2022.

AUTHORS BIOGRAPHY

Dr. Bandi Ranjitha is working as Assistant Professor in the Department of Computer Science and Engineering at Guru Nanak Institute of Technology, Hyderabad. She completed her B.Tech. (IT) from Nalanda Institute of Engineering & Technology, JNTU Hyderabad (2006); M.Tech. (CSE) from Adam's Engineering College, JNTU Hyderabad (2011); and Ph.D. in CSE from Presidency University, Bangalore (2026). She has 15 years of teaching experience and 2 years in the IT sector. She has published one SCI and two Scopus-indexed journals, one IEEE-indexed journal, holds 6 patents and 4 book chapters, and has presented at multiple national and international conferences. Her research interests include Machine Learning, Artificial Intelligence, and Deep Learning.

Sayali Prabhakar Ingale is a B.Tech. student of the 4th year CSE at Guru Nanak Institute of Technology, Hyderabad, Telangana, India.

Sura Naresh is a B.Tech. student of the 4th year CSE at Guru Nanak Institute of Technology, Hyderabad, Telangana, India.

Shreyas Kulkarni is a B.Tech. student of the 4th year CSE at Guru Nanak Institute of Technology, Hyderabad, Telangana, India.