IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 1, May 2025
Impact Factor: 7.67

Textsafe : Secure Deduplication in Cloud

Dr. Harika.B[1], Chitta Srinivas[2], Gangam Rithvik Reddy[3]
Associate Professor, Department of IT, Mahatma Gandhi Institute of Technology, Hyderabad, India[1]
B.Tech Student, Department of IT, Mahatma Gandhi Institute of Technology, Hyderabad, India[2,3]

Abstract: *In today's digital era, the exponential growth of textual data presents significant challenges in terms of both storage efficiency and security, especially in cloud environments. TextSafe offers a robust solution by integrating secure data deduplication with Advanced Encryption Standard (AES) encryption to ensure that sensitive information remains confidential and protected from unauthorized access. AES, a widely trusted encryption protocol, enables both encryption and decryption processes to occur seamlessly within the system, securing data throughout its lifecycle. TextSafe defines distinct roles—data owners, who manage and audit files; users, who request access under strict protocols; cloud administrators, who maintain infrastructure without access to decrypted data; and attackers, whose potential threats are mitigated through layered security defenses. A reliable SQL database supports the backend, ensuring data integrity, availability, and secure access logging. The deduplication mechanism effectively identifies and removes redundant data, significantly reducing storage needs—often by up to 70%—which results in both space savings and improved system performance. Overall, TextSafe combines encryption, access control, role-based management, and smart storage optimization to provide a comprehensive, secure, and scalable cloud data management solution.*

Keywords: TextSafe, data deduplication, AES encryption, cloud security, data owners, SQL database, access control, data integrity, storage optimization, secure file management

I. INTRODUCTION

In an age where digital information is expanding at an unprecedented pace, managing vast amounts of textual data—while keeping it secure—has become a major concern for individuals and organizations alike. Cloud environments, though highly convenient, introduce risks related to data breaches, unauthorized access, and overwhelming storage demands. **TEXT SAFE** rises to meet this challenge by introducing a thoughtfully designed, secure data deduplication system that leverages the power of **Advanced Encryption Standard (AES)**. This ensures that data is not only stored efficiently but also remains confidential and protected throughout its lifecycle.

At the heart of TEXT SAFE is a clear and practical division of roles that enhances both security and usability. **Data owners** can upload, manage, and audit files with full control, deciding who gets access and when. **Users** can request access, but only through rigorous security protocols, ensuring that sensitive information never falls into the wrong hands. **Cloud administrators** support the system infrastructure but are restricted from viewing encrypted data, preserving user privacy. The model also proactively considers the presence of **potential attackers**, embedding multiple layers of defense to prevent unauthorized intrusion.

Underpinning these operations is a strong **SQL database** that acts as the system's backbone—ensuring data remains intact, accessible, and traceable. One of the most compelling features of TEXT SAFE is its **deduplication capability**, which identifies and eliminates redundant copies of data. Research has shown that this process can reduce storage requirements by an impressive 90–95% (Kwon et al., 2020; Hur et al., 2016), making it not only a smart solution but also a cost-effective one. By streamlining storage and securing data with industry-standard encryption, TEXT SAFE represents a balanced and human-centric approach to modern cloud data management—keeping data safe, accessible, and responsibly handled.

DOI: 10.48175/IJARSCT-26140

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 1, May 2025
Impact Factor: 7.67

II. METHODOLOGY

1. Advanced Encryption Standard:

Advanced Encryption Standard (AES) is a specification for the encryption of electronic data established by the U.S National Institute of Standards and Technology (NIST) in 2001. AES is widely used today as it is a much stronger than DES and triple DES despite being harder to implement.

Advanced Encryption Standard (AES) is a specification for the encryption of electronic data established by the U.S National Institute of Standards and Technology (NIST) in 2001. AES is widely used today as it is a much stronger than DES and triple DES despite being harder to implement.

Working of the cipher :

AES performs operations on bytes of data rather than in bits. Since the block size is 128 bits, the cipher processes 128 bits (or 16 bytes) of the input data at a time.

The number of rounds depends on the key length as follows :

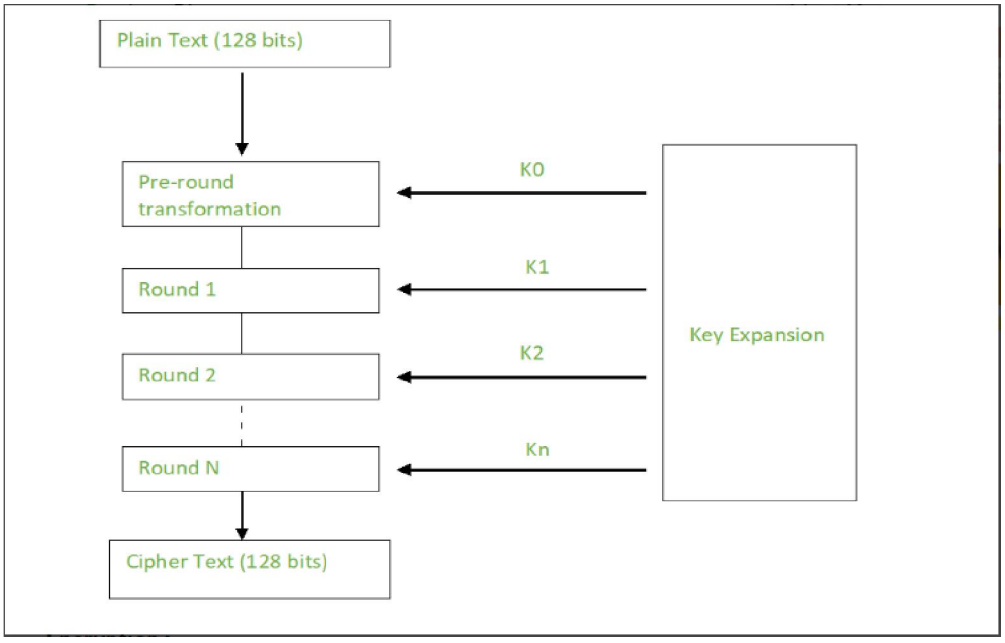
128 bit key – 10 rounds

192 bit key – 12 rounds

256 bit key – 14 rounds

Creation of Round keys :

A Key Schedule algorithm is used to calculate all the round keys from the key. So the initial key is used to create many different round keys which will be used in the corresponding round of the encryption.



Encryption :

AES considers each block as a 16 byte (4 byte x 4 byte = 128) grid in a column major arrangement.

[b0 | b4 | b8 | b12 |
| b1 | b5 | b9 | b13 |
| b2 | b6 | b10| b14 |
| b3 | b7 | b11| b15]

Each round comprises of 4 steps :

SubBytes

Shift Rows

MixColumns

DOI: 10.48175/IJARSCT-26140

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 5, Issue 1, May 2025 Impact Factor: 7.67

Add Round Key
The last round doesn't have the MixColumns round.
The SubBytes does the substitution and Shift Rows and MixColumns performs the permutation in the algorithm.

SubBytes:

This step implements the substitution.
In this step each byte is substituted by another byte. Its performed using a lookup table also called the S-box. This substitution is done in a way that a byte is never substituted by itself and also not substituted by another byte which is a compliment of the current byte. The result of this step is a 16 byte (4 x 4) matrix like before.
The next two steps implement the permutation.

Shift Rows :

This step is just as it sounds. Each row is shifted a particular number of times.
The first row is not shifted
The second row is shifted once to the left.
The third row is shifted twice to the left.
The fourth row is shifted thrice to the left.
(A left circular shift is performed.)ss

[b0	b1	b2	b3]		[b0	b1	b2	b3]
b4	b5	b6	b7	- >	b5	b6	b7	b4
b8	b9	b10	b11		b10	b11	b8	b9
[b12	b13	b14	b15]		[b15	b12	b13	b14]

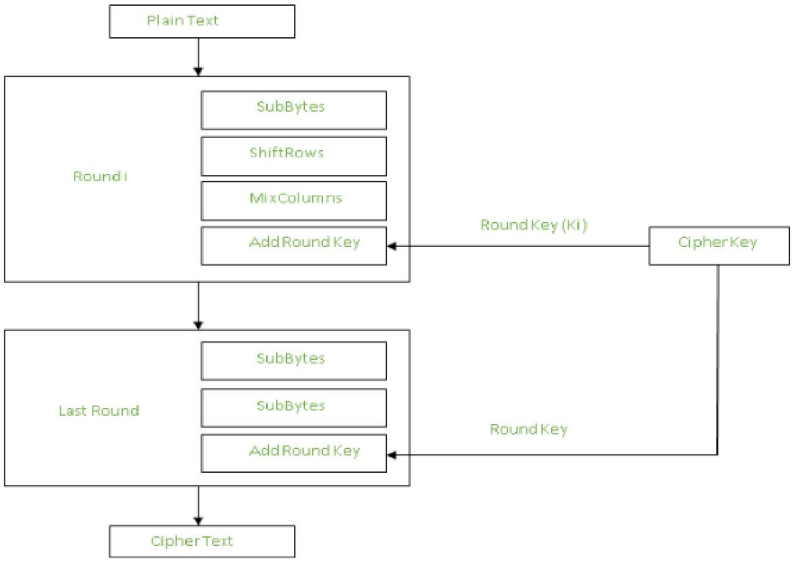
MixColumns :

This step is basically a matrix multiplication. Each column is multiplied with a specific matrix and thus the position of each byte in the column is changed as a result.
This step is skipped in the last round.

[c0]		[2	3	1	1]	b0
c1	=	1	2	3	1	b1
c2		1	1	2	3	b2
[c3]		[3	1	1	2]	b3

Add Round Keys :

Now the resultant output of the previous stage is XOR-ed with the corresponding round key. Here, the 16 bytes is not considered as a grid but just as 128 bits of data.



DOI: 10.48175/IJARSCT-26140

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 5, Issue 1, May 2025 Impact Factor: 7.67

After all these rounds 128 bits of encrypted data is given back as output. This process is repeated until all the data to be encrypted undergoes this process.

Decryption:

The stages in the rounds can be easily undone as these stages have an opposite to it which when performed reverts the changes. Each 128 blocks goes through the 10,12 or 14 rounds depending on the key size.

The stages of each round in decryption is as follows :

Add round key

Inverse MixColumns

ShiftRows

Inverse SubByte

The decryption process is the encryption process done in reverse so i will explain the steps with notable differences.

Inverse MixColumns:

This step is similar to the MixColumns step in encryption, but differs in the matrix used to carry out the operation.

| b0 | [14 11 13 9] [c0]
| b1 | | 9 14 11 13 | | c1 |
| b2 | | 13 9 14 11 | | c2 |
| b3 | [11 13 9 14] [c3]

Inverse SubBytes :

Inverse S-box is used as a lookup table and using which the bytes are substituted during decryption.

III. LITERATURE SURVEY

Secure data deduplication has become an essential aspect of efficient and secure cloud storage, especially as digital content continues to grow at an exponential rate. Researchers across the world have explored its development, challenges, and innovative approaches to ensure both **storage optimization** and **data security**. The following studies shed light on key advancements and issues in the field:

[1] Ghassabi et al. explored a novel angle by incorporating **Natural Language Processing (NLP)** into deduplication, particularly for unstructured textual data. Their approach uses semantic analysis to detect redundancy more accurately, especially in large textual datasets like news archives, legal documents, and academic repositories. By combining NLP with conventional deduplication techniques, they demonstrated improved accuracy in identifying meaningful duplicates beyond exact text matches.

[2] Yu et al. introduced **VeriDedup**, a verifiable deduplication scheme that addresses a growing concern—**trust in cloud service providers**. Their solution allows users to verify that deduplication is being performed correctly and honestly using **proof-of-duplication** techniques. This not only ensures **data integrity** but also increases transparency, which is vital in sectors like healthcare and finance where trust and auditability are non-negotiable.

[3] Khan and Raza provided a broad overview of **secure deduplication techniques**, mapping out the evolution from basic chunking and hashing to more advanced encryption-aware approaches. They emphasize the need for deduplication systems to support **cross-user scenarios**, allowing shared content across different users to be deduplicated without compromising privacy. They also highlight the importance of integrating deduplication with other cloud security components like **access control** and **authentication mechanisms**.

[4] Patel and Thakkar conducted a comprehensive survey of secure deduplication frameworks, especially in **multi-tenant cloud environments** where multiple users share the same infrastructure. They stressed the importance of balancing **efficiency and security**, pointing to encryption, hashing, and client-server deduplication models as key enablers. A standout point in their survey is the call for more effective cross-user deduplication that doesn't expose user data.

[5] Sharma and Kaul proposed a more **fine-grained deduplication method** using data segmentation combined with hashing. By breaking files into smaller chunks, their system increases the precision of duplicate detection while still supporting encryption. Their framework enhances performance by reducing network traffic and speeding up the deduplication process—making it especially suitable for **cloud backup and archival systems**.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 5, Issue 1, May 2025 Impact Factor: 7.67

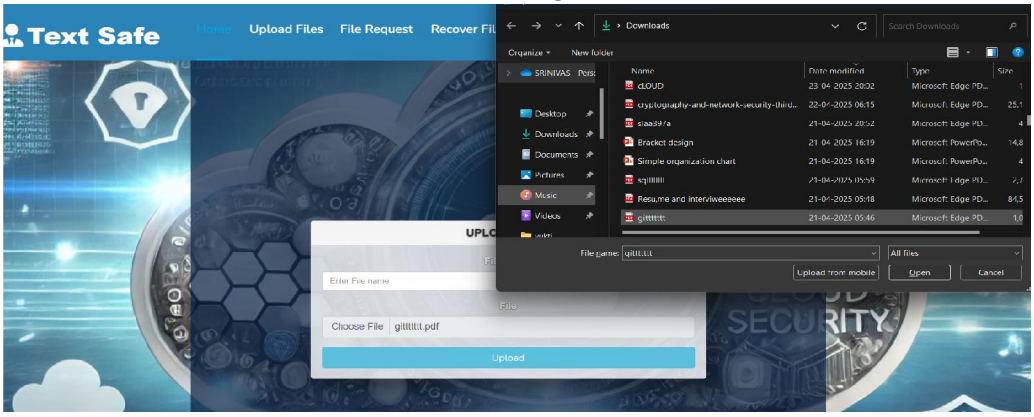
[6] Zhang et al. tackled the challenge of **dynamic data operations** in the cloud, such as updates and deletions, which often complicate deduplication. Their framework integrates secure deduplication with **dynamic data handling**, ensuring that deduplication indices remain consistent even when data changes. This makes their approach particularly relevant for collaborative platforms and enterprise environments with frequent file updates.

[7] Prajapati and Shah zoomed in on **security vulnerabilities** introduced by deduplication—such as side-channel attacks and potential data leakage. While deduplication is excellent for reducing storage and bandwidth usage, they argue that without proper encryption, it can expose sensitive data. They advocate for **encryption-aware deduplication schemes** that ensure privacy is maintained throughout the process and highlight the importance of **scalability** in managing large, growing datasets.

In addition to these research contributions, **blockchain** has emerged as a promising technology in this space. Recent studies suggest that blockchain can help manage **deduplication metadata** in a decentralized, tamper-proof manner. It brings transparency and auditability, particularly valuable in **multi-cloud environments** where multiple service providers are involved.

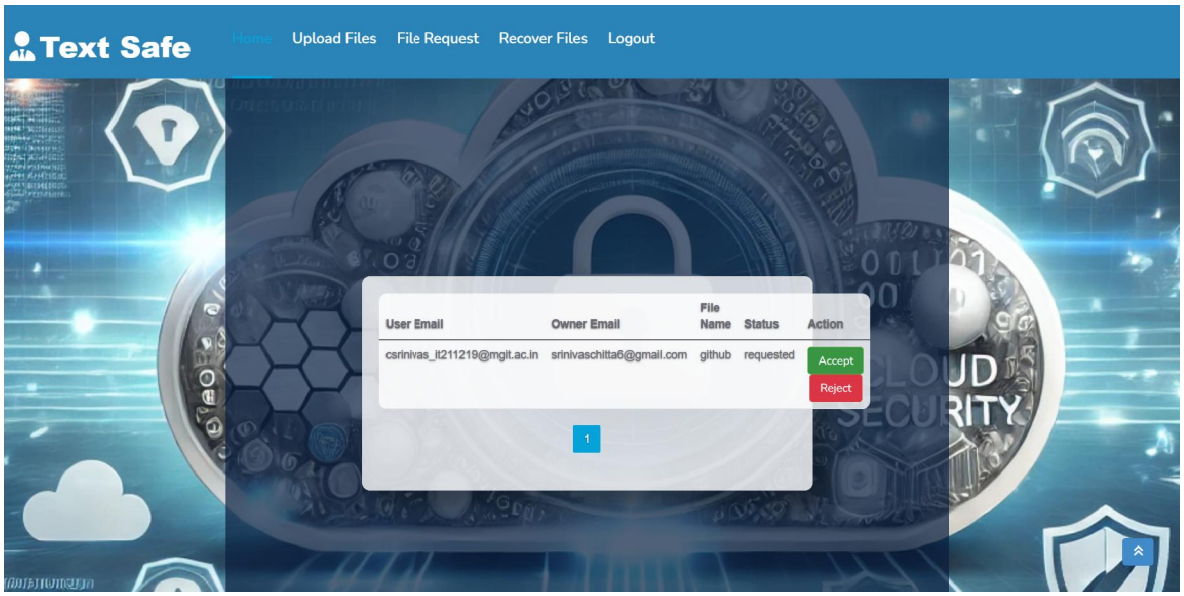
IV. RESULTS

Home Page

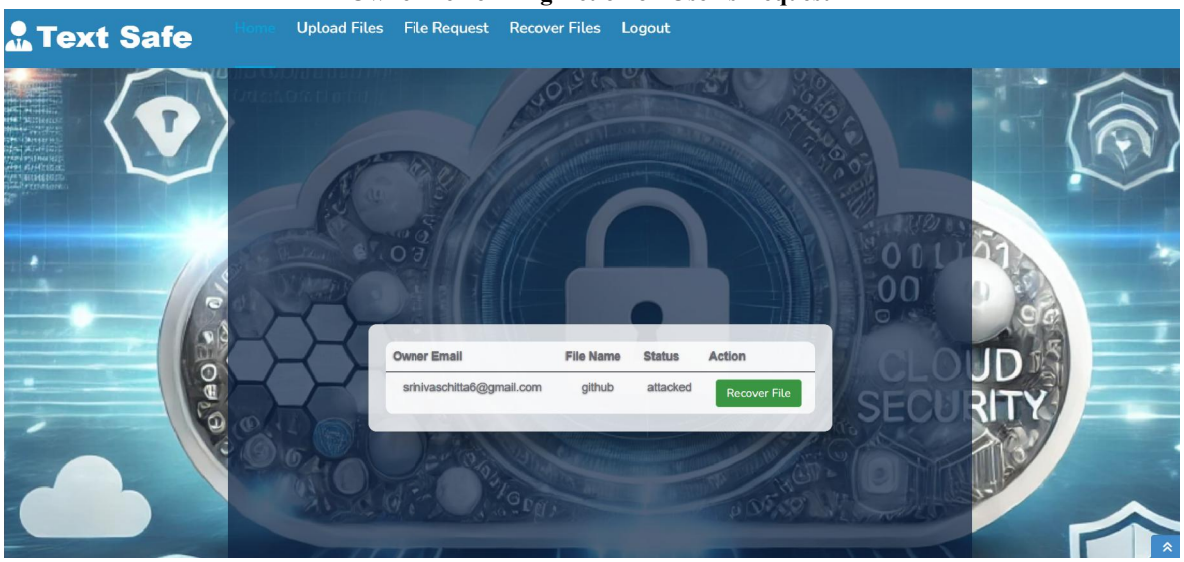


Owner Uploading Files

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 1, May 2025
Impact Factor: 7.67

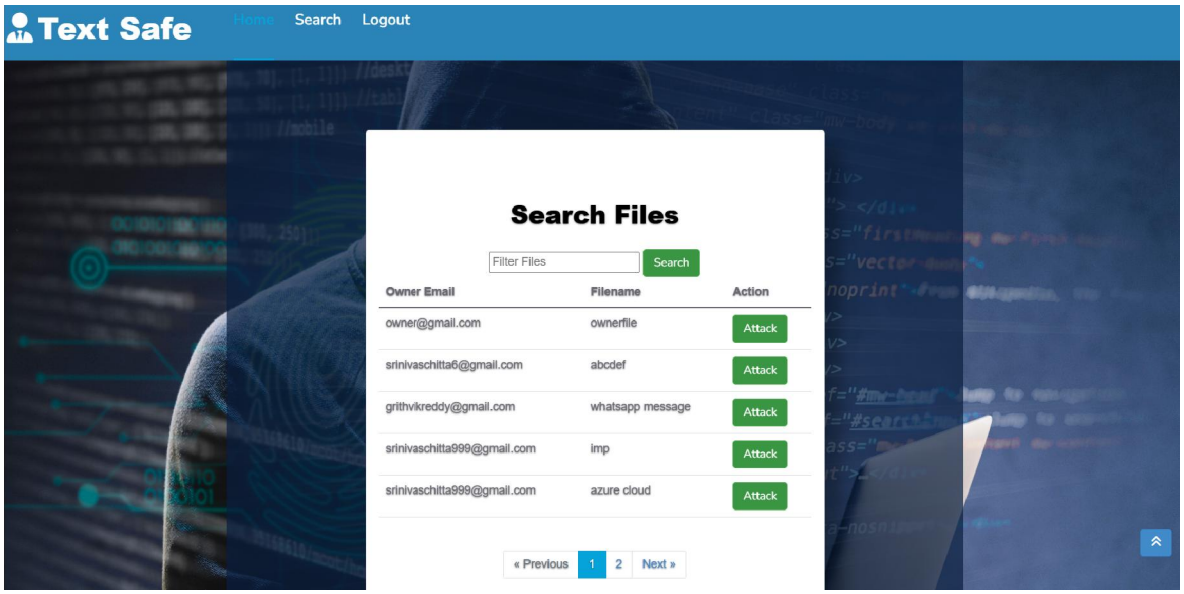


Owner Performing Action on User's Request



Owner Recovering Files Attacked By Attacker

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 1, May 2025
Impact Factor: 7.67



Attacker Searching Files To attack

V. CONCLUSION

TextSafe presents a robust solution leveraging AES encryption for secure data deduplication in cloud storage. By effectively managing textual data growth and enhancing storage efficiency, TextSafe ensures confidentiality and mitigates risks of unauthorized access. Future enhancements could explore adaptive encryption and blockchain integration to further strengthen security and transparency, making it a pivotal choice for secure cloud storage environments.

TextSafe is an innovative solution designed to tackle two critical challenges in cloud storage: keeping your data secure and making storage more efficient. With the rapid growth of digital information, especially text-based data, many organizations struggle to manage storage costs and ensure that their data is safe from unauthorized access. That's where TextSafe steps in.

At its core, TextSafe uses Advanced Encryption Standard (AES) technology to keep your data private and secure. This means that even if someone gains access to your cloud storage, the encrypted data remains unreadable without the proper decryption key. It's like having a digital lockbox for your most important information, giving you peace of mind knowing your data is protected.

What really sets TextSafe apart is its smart approach to data deduplication. Think about how often we save multiple versions of the same document or store repetitive information. Over time, this duplication clogs up storage, driving up costs and reducing efficiency. TextSafe solves this problem by identifying and removing duplicate data while keeping everything encrypted. The result? You get more storage space and save money without ever compromising security.

VI. ACKNOWLEDGMENTS

Indeed, the thanks are extended to all those involved in bringing this project on blockchain architecture for pharmaceutical supply chain traceability to completion. This project would not have been made possible without the strong support of dedicated team members, domain experts, and stakeholders intently guiding and collaborating with the project.

To our technical advisors: thank you very much for invaluable input in Ethereum smart contract development and secure system design; it is that which literally defined this tamper-proof, transparent, and decentralized solution. Lastly, we appreciate verbal givings from the pharmaceutical partners and regulatory consultants regarding their experiences from the field.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 1, May 2025
Impact Factor: 7.67

Thanks to the research and development team for their unending efforts towards formulation of a sturdy architecture that will make changing supply chain visibility, minimizing counterfeiting risks, and beefing public confidence with immutable transaction logs.

It is a collective event meant to raise the bar for integrity, safety, and scalability in the global pharmaceutical supply chains

.

REFERENCES

[1]. K. Ghassabi, P. Pahlevani, and D. E. Lucani, "Deduplication of textual data by NLP approaches," in *Proc. IEEE 97th Veh. Technol. Conf. (VTC-Spring)*, Florence, Italy, Jun. 2023, pp. 1–6, doi: 10.1109/vtc2023-spring57618.2023.10199538.

[2]. X. Yu, H. Bai, Z. Yan, and R. Zhang, "VeriDedup: A Verifiable Cloud Data Deduplication Scheme With Integrity and Duplication Proof," 2022.

[3]. M. A. Khan and M. A. Raza, "A Review on Secure Data Deduplication in Cloud Computing: Techniques and Applications," 2022.

[4]. H. Patel and P. Thakkar, "Secure Data Deduplication in Cloud Computing: A Survey," 2021.

[5]. S. Sharma and R. Kaul, "A Secure Deduplication Scheme for Cloud Storage Using Data Segmentation and Hashing," 2023