Journal of Science Engineering Technology and Management Science
Volume 02, Issue 08, August 2025 *ISSN: 3049-0952*
www.jsetms.com *DOI:10.63590/jsetms.2025.v02.i08.411-419*

IMPROVED CLOUD STORAGE AUDITING SCHEME USING DEDUPLICATION

[1]N.SAI SRUJANA,[2]M.ANUSHA,[3]M.MANOJ,[4]Mrs. O. NAGA KUMARI
[123] Students, [4] Assistant Professor
Department Of Information Technology
Teegala Krishna Reddy Engineering College, Meerpet, Balapur, Hyderabad-500097

To Cite this Article
N.Sai Srujana, M.Anusha, M.Manoj, Mrs. O. Naga Kumari, "Improved Cloud Storage Auditing Scheme Using Deduplication", Journal of Science Engineering Technology and Management Science, Vol. 02, Issue 08, August 2025,pp: 411-419, DOI: http://doi.org/10.63590/jsetms.2025.v02.i08.pp411-419
Submitted: 12-07-2025 *Accepted: 18-08-2025* *Published: 25-08-2025*

ABSTRACT
Cloud storage has become an essential service for most users handling large amounts of data. An enhanced cloud storage auditing scheme with deduplication guarantees data integrity and saves storage costs by eliminating redundant copies of the same data while ensuring secure and efficient verification techniques. This project is aimed at designing a system where users can safely store gigabytes of data on cloud servers without worrying that their data is compromised. The deduplication mechanism also ensures that a single copy of duplicate data is stored, greatly conserving storage space and minimizing costs. The system also facilitates dynamic operations like data updating, deletion, and modification while upholding high security. To avoid unauthorized access during deduplication, the scheme utilizes encryption methods that permit only authorized users to access and authenticate their data. This method improves data confidentiality, integrity, and availability in cloud systems. The project also maintains low computational overhead for both cloud servers and users, hence a viable solution for contemporary cloud storage requirements.

I. INTRODUCTION

With the rampant increase in digital data, cloud storage has emerged as a crucial solution for individuals and businesses looking for scalable and affordable data management. Yet, as more sensitive data is outsourced to third- party cloud servers, maintaining data integrity and privacy becomes a critical issue. One such essential way of solving this problem is through cloud storage auditing schemes, where users can ensure the accuracy of their stored data without downloading the entire dataset.

At the same time, data deduplication has become a major method to maximize storage efficiency by discarding duplicate copies of repeating data. Storing only single instances of data and pointing to them in different users or systems, deduplication greatly minimizes storage space and bandwidth consumption. While bringing deduplication and auditing together poses special security and privacy issues, including how to authenticate shared data without leaking confidentiality or ownership of data.

A cloud storage auditing design based on deduplication would seek to capture the advantages of both secure auditability and optimized storage. Data integrity is maintained, public or private auditing can be

supported, and multiple parties can securely share and verify information, all in a way that reduces redundancy to a minimum. This work describes the architecture, cryptographic underpinnings, and efficiency improvement of such a system, while considering both engineering implementation issues as well as security assurances.

Back ground of cloud storage

Cloud storage means on-demand data storage and retrieval over the Internet, as a service. Rather than having files hosted locally on hard drives or on-premises servers, organizations and end users outsource data storage, management, and maintenance to third-party vendors.

1. Evolution and Motivation

Early Days (Pre-2000s): Information was kept on local disks, tape libraries, and network-attached storage (NAS). Expansion meant huge capital outlays in hardware and real estate.

Grid and Utility Computing (Early 2000s):

As high-speed networks and virtualization became more prevalent, researchers began to explore the idea of aggregating computing and storage resources into shared "utility" services.

Commercial Cloud Storage (Mid-2000s onwards):

Led by companies such as Amazon (with S3 in 2006), Google, and Microsoft, cloud storage was established as a commercially acceptable service. Providers constructed large data centers, provided APIs for programmatic use, and implemented pay-as-you-go pricing schemes.

Drivers for Adoption:

Scalability: Essentially limitless storage capacity that can be expanded or contracted based on requirement.

Cost-Effectiveness: Transfers capital expenditures (CapEx) to operational expenditures (OpEx), only paying for consumed amounts.

Accessibility: Data that can be accessed from any location with Internet connectivity.

Reliability & Durability: Internal replication over several geographic locations for high availability and data integrity.

PROBLEM STATEMENT

In the age of information overload, precise categorization of news stories into particular classes is essential for improving targeted ads and personalized content delivery. Conventional methods of classification tend to fall short due to the intricacies of language, resulting in less-than- optimal outcomes. This project seeks to overcome the shortcomings of traditional machine learning algorithms by using state-of-the-art language models (LLMs) to develop a powerful classification framework. The core issue is to examine how well LLMs can classify news articles in comparison to conventional methods, in particular with respect to the use of prompt engineering. By measuring performance on various datasets, this work aims to show that LLMs can yield better accuracy and efficiency in news classification, ultimately leading to more relevant user experiences across content consumption and ad delivery.

Although conventional cloud storage auditing protocols (e.g., Provable Data Possession or Proofs of Retrievability) offer methods for clients or third-party auditors to ensure that a cloud server holds their data intact, they rely on each user having files stored separately. Data deduplication reduces storage, however, by sensing and retaining only a single instance of duplicate data blocks among many users. This underlying paradigm presents some interrelated challenges:

Integrity Verification in a Shared-Block Environment

Deduplication violates one-to-one mapping: One data block on the server can be shared by many users. Naively verifying "your" block could only guarantee that some user's copy still exists, not yours.

Malicious de-duplication:

The server might delete or corrupt a shared block, but still pass per-user audits if it generates proofs for another user's copy.

Leakage via deduplication queries:

Classic deduplication forces the server to verify identical data (usually through hashing). An attacker auditor or server can deduce if a target user possesses a certain file, if not carefully implemented, probing deduplication answers.

Cross-user inference:

Audit protocols should not leak information about the other users having a shared block or the availability of other files on the server.

Efficiency vs. Security Trade-offs Communication overhead:

Auditing of deduplicated data at the granularity of individual users' logical view can significantly grow challenge/response messages.

Computation burden:

Having proofs represent ownership of only the user's logical file, as opposed to raw deduplicated copy, requires additional cryptographic operations (e.g., keyed hashes or tags).

Dynamic Data Operations Updates and modifications:

Enabling file append, delete, or modify operations without re-uploading whole deduplicated chunks makes auditing and management of deduplication metadata harder.

Revocation and access control:

When a user revokes access to a shared block (e.g., due to subscription expiration), the scheme should prevent them from being able to successfully audit or derive proofs for that data anymore.

MOTIVATION

The constant expansion of data—fueled by multimedia, big-data analytics, and continuously growing backup needs—has put tremendous strain on storage systems. Organizations increasingly create petabytes of data, much of it with considerable redundancy (e.g., system images, log files, shared documents). Deduplication becomes a potent antidote, reducing duplicate data blocks within files or users to recover 70–80% of storage capacity and significantly decrease the bandwidth required for backups and synchronization.

However, while deduplication addresses the economic and performance aspects of large-scale storage, it also poses new security problems. In a deduplicated system, one physical chunk can represent many logical owners, so classical auditing schemes that check that "your" data is complete can be fooled by proofs created from another user's copy. Without a deduplication-aware protocol, clients forfeit the very guarantees that auditing is intended to offer, leaving them susceptible to silent deletion or corruption by a malfunctioning or malicious cloud provider.

Outside of strict integrity issues, regulatory systems like GDPR, HIPAA, and SOX now increasingly call for strict evidence of data integrity and access retention. Companies bound by these systems can no longer merely rely on service-level agreements; they require cryptographic validation that data stays unchanged and intact. A design that unifies deduplication with provable data possession consequently becomes a cornerstone for compliance without sacrificing the savings benefits of deduplication.

From the customer's point of view, a safe deduplication-aware auditing service is a differentiator in a highly competitive market. Customers receive quicker backups, leaner infrastructure expense, and transparent guarantees of data security—all serving to build more customer loyalty and fewer churns. Simultaneously, extending the frontiers of applied cryptography and metadata management fosters

innovation: creating lightweight tag-based proofs, secure indexing structures, and dynamic update mechanisms advances both academic research and practical system design.

In total, the incentive for a cloud storage auditing scheme that is completely deduplication-friendly extends across economic, security, regulatory, and technological realms. By providing both significant resource savings and unyielding data integrity assurances, such a scheme meets the urgent demands of contemporary businesses while forging new paths in secure, scalable storage systems.

OBJECTIVE

The primary goal of a deduplication-aware cloud storage auditing protocol is to offer every client **strong, verifiable guarantees** that their logical files are not damaged and accessible—even when the service provider retains only a single physical copy of any particular data block common across multiple users. To do this, the scheme needs to cryptographically associate each user's ownership rights with the underlying deduplicated blocks, for instance, by creating **lightweight, user-specific tags** (or keyed hashes) on each block prior to upload. At the time of an audit, the server generates proofs that establish ownership of these tagged blocks without disclosing the block contents or other users' tags so that a client cannot be deceived by proofs generated from another user's data.

No less crucial is **efficiency**: the auditing protocol must have minimal computation and communication overhead. Instead of downloading or re-uploading whole files, a client sends a small set of randomized challenges, and the server returns brief cryptographic proofs on the applicable block tags. This challenge–response exchange needs to scale well even as the number of users and common blocks increases into the millions without sacrificing the **storage and bandwidth savings** that make deduplication desirable in the beginning.

The scheme should also ensure **strong privacy assurances**. It should ensure that it is not possible for an attacker server—or even for an auditor—to find out if a specific user has a certain file (or which other users) by probing the deduplication system. Utilizing methods like

private set membership tests or **oblivious data structures**, the protocol is able to check for ownership of blocks without divulging metadata regarding sharing relationships or file contents.

Lastly, an effective system needs to facilitate **dynamic data operations**—file updates, deletions, and user revocations—without requiring bulk re-upload of deduplicated blocks. This necessitates strong metadata management that can effectively add or delete user references to shared blocks and update tags when content is modified. By achieving the proper balance between cryptographic rigor, operational efficiency, and user privacy, a deduplication-aware auditing framework provides unassailable integrity proofs, regulatory-compliant audit trails, and cost-effectiveness today's cloud customers expect.

SCOPE

The scope of this project includes the design, implementation, and evaluation of a cloud storage auditing framework that natively supports block-level deduplication. At its center, the framework will include three tightly coupled modules: a client-side tagging module that computes user-specific cryptographic tags for each data block before uploading; a storage-server module that securely deduplicates duplicate blocks with reference counts and tag metadata preserved; and an auditor interface that provides randomized, challenge–response proofs of possession. Combined, these elements make it possible for every client to check the integrity of their logical files even if the provider holds just a single physical copy of common blocks.

Functionally, the scheme will support both **static** and **dynamic** data operations. For static data, we shall concentrate on effective proof generation and verification protocols—keeping both computational overhead on the server and client communication bandwidth at a minimum. For dynamic

data, the scope is to support file appends, updates, and deletes, and user revocation cases (e.g., subscription termination or access-right modification). Metadata structures—such as secure reference counts, tag catalogs, and update logs—will be engineered to support these operations incrementally, without requiring full re-uploads of deduplicated content.

In addition to correctness and efficiency, the project will address

privacy and security goals with stringent rigor. The auditing protocol must leak no information about which other users share a given block or what other files are on the server. To this end, we will consider privacy- preserving primitives—e.g., oblivious lookups or zero-knowledge set membership proofs—to separate deduplication detection from the auditing task. Adversary models will take malicious storage provider and possibly inquisitive third-party auditors as their assumptions, and the scheme will be shown secure under normal cryptographic assumptions (such as collision resistance of hash functions, unforgeability of tags).

Lastly, the scope of the project will define neat limits: we **will not** address full-disk encryption, side-channel leakage on the server, or denial-of-service attacks that would aim to overwhelm the audit process. Neither will we consider cross-data-center synchronization or multi-tenant isolation more than trivial reference counting. We will test performance in a controlled testbed—quantifying proof sizes, challenge latency, and storage efficiency—instead of in a live, public data center. By concentrating on these rigorously defined functional, security, and operational facets, the project hopes to produce a solid, deployable design for deduplication-sensitive cloud storage auditing.

II. LITERATURE REVIEW

Cloud Storage Auditing

Cloud storage auditing is the mechanisms and procedures through which users or third-party auditors can validate the integrity and availability of cloud- stored data without first downloading the full dataset. Since users are outsourcing data to distant cloud servers, it becomes imperative that the data stored is not modified and remains accessible in the long term.

Some audit schemes have been suggested, such as provable data possession (PDP) and proofs of retrievability (PoR). These methods enable users to challenge the cloud server to verify that it has the entire, uncorrupted data set. PDP typically concentrates on the integrity and existence of data, whereas PoR guarantees data retrievability even in the event of corruption or loss.

Public auditing schemes, under which a third-party auditor confirms data on behalf of the user, have also been proposed in order to shift auditing work from users with limited resources. But these schemes should provide privacy- preserving properties so that the auditor will not be able to gather sensitive information during verification. Such operations are typically achieved using techniques such as homomorphic authenticators and random masking.

Deduplication in Cloud Storage

Deduplication is a space-saving method that removes duplicate copies of recurring data. In cloud storage systems, where several users might store duplicate or identical files (e.g., operating system images, backups), deduplication can drastically cut storage expense and bandwidth utilization.

Deduplication may be file-level or block-level. File-level deduplication deletes whole files if they are the same, whereas block-level deduplication breaks files into small blocks and deletes duplicates of the blocks.

Even though deduplication has its advantages, it poses several challenges that mainly concern the security and privacy of data. Standard encryption techniques are detrimental to deduplication because encrypting a file with different keys results in different ciphertexts. To overcome this, convergent encryption (or

content-defined encryption) has been employed, where a file is encrypted using the hash of the contents as a key. This is such that two identical files will always produce identical ciphertexts, allowing deduplication. Convergent encryption can, however, be susceptible to dictionary attacks, whereby a malicious attacker is able to guess file contents and verify their hashes.

Recent works, therefore, aim to integrate deduplication efficiency with data confidentiality through innovative cryptographic technologies, including proofs of ownership, oblivious key management, and secure multi-user deduplication.

Auditing Mechanisms and Security

Security is the foundation of any cloud auditing process. Auditing not only needs to check for the accuracy of stored data but also needs to ensure that the auditing process itself does not reveal sensitive information or provide unauthorized access.

Security objectives in cloud auditing are:

Data Integrity:

To ensure that data has not been altered or deleted without permission.

Privacy Preservation:

To ensure that the data content remains confidential, particularly during third- party audits.

Freshness Verification: Verifying that the most recent version of the data is stored and available.

III. PROPOSED SYSTEM AUDITING SCHEME USING DEDUPLICATION

These systems rely on hashing but do not involve third-party checks for data security. Proposed System The proposed system enhances cloud storage by integrating deduplication with third-party auditing to improve efficiency and security. When a user uploads a file, the system checks for duplicates using a hash-based deduplication method. If the file already exists, it is skipped to save storage space; otherwise, it is stored along with a unique hash value. Users can request integrity verification, prompting the Third-Party Auditor (TPA) to compare the stored hash value with a newly generated one. If they match, the file remains unchanged; otherwise, it indicates modified or duplicate. The metadata management system enables quick lookups, making verification faster and more efficient. This approach reduces redundancy, lowers storage costs, and ensures data integrity. By combining secure storage optimization with real-time verification, the system improves trust in cloud services. It is scalable, cost-effective, and ideal for large-scale cloud storage applications.

IV. EXISTING SYSTEM AUDITING SCHEME USING DEDUPLICATION

Traditional Cloud Storage Systems

Services like Google Drive, and OneDrive let users store and access files online. They ensure data availability but do not always include built-in security checks or auditing.

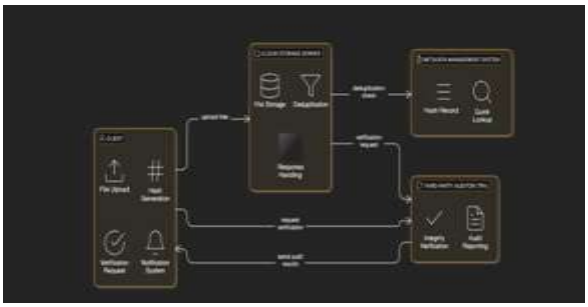
Basic Deduplication Systems

Some cloud services remove duplicate files by checking file names or metadata. However, they do not properly verify file integrity, making them less secure.

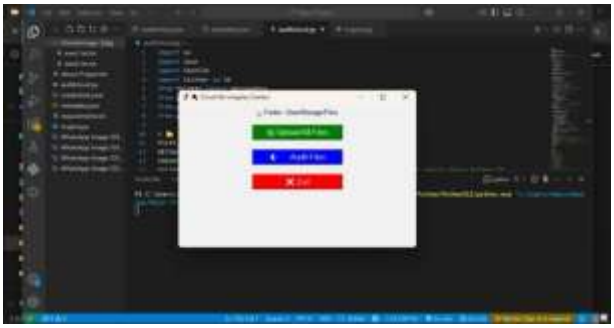
Server-Side Deduplication

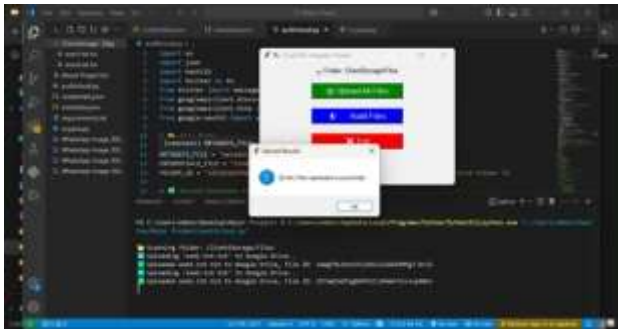
Cloud providers use deduplication to save storage space by identifying and removing duplicate files. These systems rely on hashing but do not involve third-party checks for data security

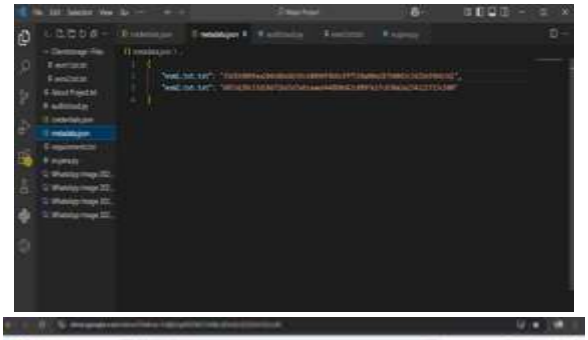
V. ARCHITECTURE DIAGRAM

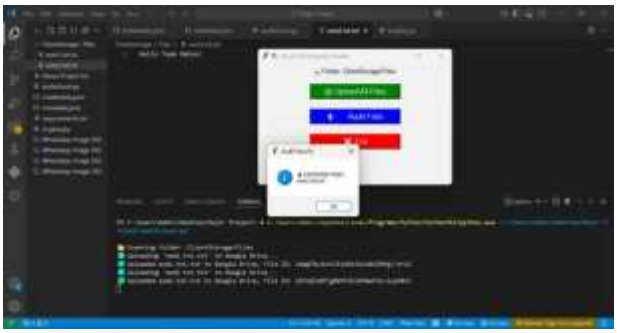


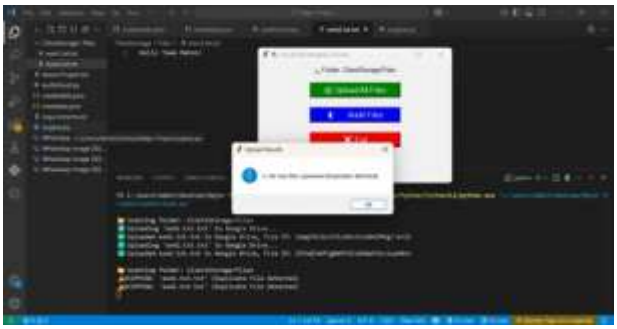
VI. OUTPUTS

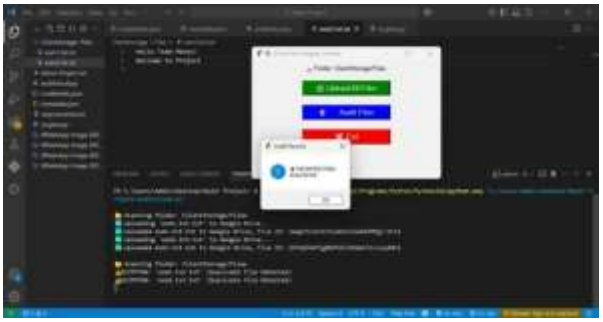


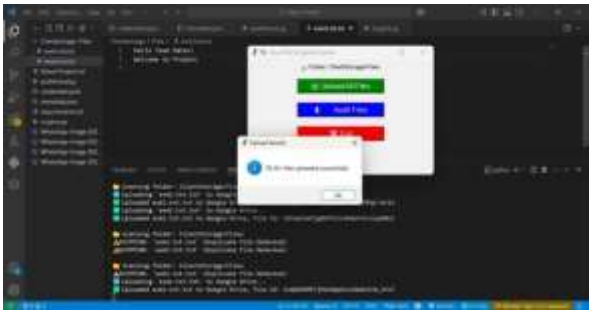












VII. CONCLUSION

By applying deduplication methods, the system eliminates duplicate storage of data, thus maximizing use of available resources and reducing operating expenses. Meanwhile, the provision of a highly advanced

auditing mechanism guarantees that stored data in the cloud are always accurate, unaltered, and can be traced at any given point in time. Through the application of cryptographic hashing (SHA-256) as well as metadata control, the system provides lightweight but secure verification of data without undermining user privacy.

The suggested architecture accommodates both client-side deduplication for bandwidth conservation and server-side integrity verification for preserving trust in third-party cloud providers. Both the efficiency and security considerations ensure that the solution is extremely relevant in real-world use cases, especially for those environments where data sizes are large and integrity is of critical concern, such as healthcare, enterprise IT, and government document repositories.

In summary, this project illustrates how deduplication and cloud auditing can be combined to create a scalable, secure, and storage-efficient system. It provides the foundation for future improvements in encrypted deduplication, multi-user ownership models, and privacy-preserving auditing protocols.

REFERENCES

J. Stanek, A. Sorniotti, E. Androulaki, L. Kencl, "A Secure Data Deduplication Scheme for Cloud Storage," IEEE Transactions on Cloud Computing, Vol. 5, No. 2,pp. 1-14, 2018.

D. Harnik, B. Pinkas, A. Shulman-Peleg, "Side Channels in Cloud Services: Deduplication in Cloud Storage," IEEE Security & Privacy, Vol. 8, No. 6, pp. 40-47, 2010.

M. Bellare, S. Keelveedhi, T. Ristenpart, "Message-Locked Encryption and Secure Deduplication," Advances in Cryptology - EUROCRYPT 2013, pp. 296-312, 2013.

P. Puzio, R. Molva, M. Onen, S. Loureiro, "ClouDedup: Secure Deduplication with Encrypted Data for Cloud Storage," IEEE Transactions on Computers, Vol. 65, No. 2,
pp. 356-365, 2016.

J. Li, X. Chen, M. Li, J. Li, P. Lee, W. Lou, "Secure Deduplication with Efficient and Reliable Convergent Key Management," IEEE Transactions on Parallel and Distributed Systems, Vol. 25, No. 6, pp. 1615-1625, 2014.