Deduplication of Image Files to Reduce the Redundancy in Cloud Storage

1st Ragul. G
Networking and Communication
(UG – Scholar)
SRM Institute of Science and Technology Chennai, India
rg4853@srmist.edu.in

2nd Nithyashree. N
Networking and Communication
(UG – Scholar)
SRM Institute of Science and Technology Chennai, India
nn6128@srmist.edu.in

3rd Harini K K
Networking and Communication
(UG – Scholar)
SRM Institute of Science and Technology Chennai, India
hk8524@srmist.edu.in

4th A Prabhu Chakkaravarthy
Networking and Communication
(Assistant Professor)
SRM Institute of Science and Technology Chennai, India
prabhuca@srmist.edu.in

Abstract **—Efficient cloud storage management is critical as the volume of digital content grows exponentially. Redundant image files contribute significantly to storage inefficiencies, increasing costs and affecting retrieval performance. This project proposes a scalable deduplication system for cloud storage, focusing on reducing redundancy, optimizing storage space, and enhancing operational efficiency. The objective is to ensure that only unique image files are stored, while duplicates are replaced with references to existing files, thereby conserving resources and improving data management. The proposed system utilizes a combination of SHA-1024 hashing for precise detection of identical files and perceptual hashing algorithms to identify visually similar images. These algorithms ensure comprehensive identification of duplicates, including near-identical files. The process begins with image ingestion, where unique hashes are generated for each file. These hashes are compared against an existing metadata store to detect duplicates. Identified duplicates are stored as references, reducing storage space requirements, improving retrieval speed, and minimizing bandwidth usage by avoiding repeated transmissions. The deduplication system also enhances backup and disaster recovery processes by reducing the volume of files to be managed, resulting in faster backups and recovery times. By integrating this framework with cloud environments, such as Azure, the solution delivers scalability and real-time operation. This work addresses the growing challenges of data redundancy in cloud storage, contributing to efficient and sustainable storage practices.** (*Abstract*)

***Keywords*--SHA-1024, Perceptual Hashing, Redundancy Reduction, Cloud Storage, Retrieval Efficiency** *(key words)*

I. Introduction

Hybrid deduplication schemes are a significant advancement in cloud storage, enabling optimized storage utilization and security. El-Shimi et al. proposed a model combining file- and block-level deduplication to balance storage efficiency and processing speed. This approach reduces redundancy while managing system performance by selecting appropriate deduplication levels based on file size and type. Such methods cater to the dynamic nature of cloud environments, ensuring scalability and adaptability. The integration of compression techniques further improves storage savings, marking a notable improvement in deduplication strategies[1]

Efficient deduplication is essential for managing vast data volumes in modern cloud systems. Meister et al. introduced techniques focusing on computational efficiency and storage optimization, emphasizing distributed systems' challenges. Their model employs optimized indexing and hash-based comparison to identify duplicates without compromising performance. This work highlights the importance of balancing deduplication effectiveness with minimal computational overhead, especially in environments with high throughput. By streamlining file comparison processes, their approach ensures scalability while maintaining data retrieval speeds, making it a foundation for contemporary deduplication methods.[2]

Scalability and security are critical for cloud storage systems. Xu et al. proposed a deduplication model that emphasizes secure and scalable file management through advanced encryption and metadata handling. Their system ensures robust protection against unauthorized access, preserving data integrity during deduplication. Additionally, their work demonstrates the benefits of distributed architectures for scalable deduplication, accommodating growing datasets effectively. This model highlights the necessity of addressing security vulnerabilities while enhancing deduplication efficiency, offering a comprehensive solution for secure cloud storage.[3]

Puzio et al. introduced the ClouDedup model, which integrates deduplication with encrypted data storage to enhance security in cloud environments. Their approach ensures that deduplicated files remain encrypted, reducing risks associated with metadata exposure. This method leverages secure key management systems to maintain encryption integrity across deduplication processes. By addressing security and storage concerns simultaneously, the ClouDedup model demonstrates a practical solution for handling sensitive data in cloud storage while optimizing storage space through effective deduplication.[4]

Cross-user similarity detection plays a pivotal role in reducing redundant data across multiple users in shared cloud environments. Li et al. proposed a system that identifies similar files uploaded by different users, utilizing feature extraction and content-based hashing. Their method ensures efficient deduplication by targeting both exact and near-duplicate files, significantly reducing overall storage needs. By incorporating similarity metrics, the model enhances deduplication accuracy while addressing the unique challenges of multi-user cloud platforms, offering a scalable solution for shared storage systems.[5]

Dynamic ownership management introduces an innovative layer of security to deduplication. Yang and Jia's model emphasizes managing file ownership dynamically to ensure secure access in multi-user environments. This system mitigates the risks of unauthorized access by allowing dynamic key updates and access revocation, enhancing data security during deduplication. Their work highlights the importance of addressing ownership challenges in shared cloud platforms, ensuring that deduplication processes remain secure and adaptable to evolving user needs.[6]

Content-based image retrieval techniques, as proposed by Sharma et al., enhance deduplication for image files by leveraging visual similarity. Their model employs feature extraction and content-aware algorithms to identify duplicate images with high accuracy. This method addresses the limitations of traditional hashing by considering image content, making it particularly effective for detecting visually similar images. By integrating these techniques into deduplication workflows, cloud storage systems can achieve better accuracy and storage efficiency for image data.[7]

Combining deduplication with compression techniques offers a dual approach to storage optimization. Ma et al. proposed a model that integrates both processes to maximize storage savings while reducing data transmission costs. Compression is applied after deduplication to further reduce file sizes without compromising quality. Their work highlights the benefits of this combined approach, emphasizing its effectiveness for large-scale cloud environments where storage demands and operational costs continue to grow.[8]

II. LITERATURE REVIEW

The growing problem of data redundancy in cloud storage systems is addressed in the work "A Hybrid Deduplication Scheme for Cloud Storage" by El-Shimi et al. (2014), which suggests a hybrid deduplication strategy to maximize storage efficiency. The limits of current deduplication techniques are emphasized by the authors, especially with regard to processing overhead and their incapacity to efficiently handle a variety of data types. By addressing both exact and approximate data redundancy, the suggested strategy offers a more effective method by combining content-based and context-based deduplication. It seeks to minimize data duplication while lowering storage costs and enhancing retrieval speeds by employing a variety of tactics. The trade-offs between the deduplication techniques and how they affect system security, scalability, and performance in cloud contexts are also covered in the article.[1]

"Efficient Deduplication Techniques for Cloud Storage," a work by D. Meister, J. Kaiser, A. Brinkmann, et al., discusses the difficulties associated with data duplication in cloud storage systems. It investigates several deduplication strategies meant to decrease redundancy and increase storage efficiency. The authors compare the effectiveness of file-level and block-level deduplication techniques in terms of computational overhead and storage savings. The trade-offs between deduplication techniques are also covered, taking into account variables like processing time and resource usage. The study also assesses how deduplication affects system scalability and data access speed.[2]

Xu, Chang, and Zhou's work "Scalable and Secure Deduplication for Cloud Storage" tackles the difficulties of optimizing cloud storage systems by putting forth a deduplication mechanism that strikes a balance between security and efficiency. Conventional deduplication techniques may jeopardize data confidentiality but frequently favour space savings. In order to ensure scalability and mitigate security concerns, the authors present a hybrid strategy that combines probabilistic and deterministic deduplication approaches. To guarantee secure deduplication, they use a convergent encryption approach, which enables identical data to be saved just once without disclosing private information. The study emphasizes the method's performance benefits, demonstrating its capacity to manage huge datasets while remaining resilient to possible threats including brute force attacks and data pattern leaks. [3]

A significant problem in cloud storage is accomplishing secure deduplication while preserving data secrecy, which is addressed in the study "ClouDedup: Secure Deduplication with Encrypted Data for Cloud Storage" by Puzio et al. Conventional deduplication techniques lower storage costs, but when paired with encryption, they jeopardize data security. The authors suggest ClouDedup, a framework that combines convergent encryption and client-side deduplication. This permits deduplication without exposing data to unwanted access by guaranteeing that identical data blocks encrypted with the same key generate similar ciphertexts. Additionally, ClouDedup offers a Metadata Manager that safely manages access permissions and encryption keys. The system's performance is carefully assessed in the report, which also shows how well it works to lower storage costs while maintaining strong security.[4]

The issue of minimizing redundant data storage in cloud environments using deduplication is addressed in the study "Cross-User Similarity Detection for Deduplication in Cloud Storage" by Z. Li et al. (2018). In order to improve storage efficiency when numerous users upload comparable material, the authors suggest a novel method that focuses on identifying cross-user similarities. Conventional deduplication techniques usually work with data belonging to a single user, but this research expands on the idea by finding commonalities between datasets belonging to other users. More efficient storage utilization is made possible by this cross-user deduplication technique, particularly in situations where there is a large amount of shared content. The algorithm's performance and effectiveness in detecting duplicate data are thoroughly examined in the study, which makes it a valuable addition to the advancement of cloud storage systems. [5]

A safe cloud deduplication technique that tackles the issues of privacy and ownership in cloud storage systems is proposed in the paper "A Secure Cloud Deduplication Scheme Based on Dynamic Ownership Management" by K. Yang and X. Jia (2018). To make sure that data deduplication doesn't jeopardize user security or privacy, the authors present a dynamic ownership management methodology. Users maintain control over their data even in the event of ownership changes thanks to their technique, which permits secure data deduplication while permitting dynamic ownership changes. By employing cryptographic approaches and offering access control measures, the article highlights how crucial it is to preserve data confidentiality throughout the deduplication process. Experiments are used to test the suggested strategy and show how well it improves storage security and efficiency. [6]

A cloud-based method for picture deduplication that makes use of Content-Based picture Retrieval (CBIR) techniques is presented in the publication "Cloud-Based Image Deduplication Service Using Content-Based Image Retrieval" by G. Sharma, N. Joshi, and N. Sharma (2017). In cloud systems, where redundancy frequently results in needless storage expenses, the authors tackle the problem of effectively managing and storing massive amounts of photos. Instead of depending on conventional metadata or hash-based techniques, the suggested approach uses CBIR to identify duplicate photos based on their visual content. By identifying similar or identical photos, even if they are saved in various names or formats, the method efficiently lowers storage overhead. The integration of this service into cloud platforms is also covered in the paper, with an emphasis on its scalability and capacity to manage.[7]

S. Ma, C. Chen, and X. Wu's study "Optimizing Cloud Storage by Deduplication and Compression Techniques" (2021) investigates ways to maximize cloud storage efficiency by combining deduplication and compression approaches. The authors draw attention to the growing need for cloud storage as a result of the data's explosive increase and offer a solution that minimizes storage space while reducing redundant data. The study provides an improved method for storage optimization by combining compression, which shrinks the quantity of data, and deduplication, which removes duplicate data. The paper examines how well these methods work in cloud environments and demonstrates how combining them significantly increases storage capacity and data management effectiveness.[8]

The problem of secure picture deduplication in cloud storage is addressed in the work "Secure Image Deduplication in Cloud Storage Using Secret Sharing" by P. Lee, S. Chung, and S. Kim (2020). Redundancy in data storage is a common problem for cloud storage solutions, which reduces efficiency and wastes space. To maintain confidentiality and privacy during the deduplication process, the authors suggest a technique that makes use of secret sharing techniques. In order to safeguard the confidentiality of the data, their method entails splitting the image data into several shares and distributing them among various storage nodes. This way, no single party can view the entire image. The study offers a safe and effective way to handle redundant photos in cloud storage systems by fusing secret sharing with deduplication techniques.[9]

III. DEDUPLICATION STRATIFICATION TECHNIQUES

Deduplication stratification techniques involve categorizing and applying different methods of deduplication based on the characteristics of the data and the operational requirements of the storage system [5]. The primary goal is to maximize efficiency by selecting the most appropriate technique for each scenario. Common stratification techniques include:

A. File-Level Deduplication

This approach focuses on eliminating duplicate files by comparing the hash values of entire files. If two or more files have identical hash values, only one copy is stored, with references created for the duplicates. This method is straightforward and efficient for environments where entire files are frequently duplicated, such as in backup systems.

B. Block-Level Deduplication

This technique breaks down files into smaller, fixed or variable-sized blocks, generating a unique hash for each block. The system stores only unique blocks, linking them together to recreate the original files. This method is particularly effective in scenarios where files have minor differences, such as updated versions of documents or images.

C. Byte-Level Deduplication

This granular approach compares and eliminates duplicate byte sequences within files. While it offers the highest potential for space savings, it is also the most computationally intensive. Byte-level deduplication is often used in specialized environments where storage efficiency is paramount, and computational resources are abundant.

D. Content-Aware Deduplication

Leveraging metadata and content type information, this technique tailors deduplication strategies based on the nature of the data. For example, images might be deduplicated based on visual similarity metrics rather than just binary comparison, ensuring that visually identical images are recognized as duplicates even if their file formats or minor details differ.

E. Hybrid Deduplication

This technique combines multiple deduplication methods to optimize performance and efficiency. For instance, a system might use file-level deduplication as a first pass and then apply block-level deduplication for more granular savings. Hybrid approaches are often used in large-scale systems where data diversity requires flexible and adaptive strategies.

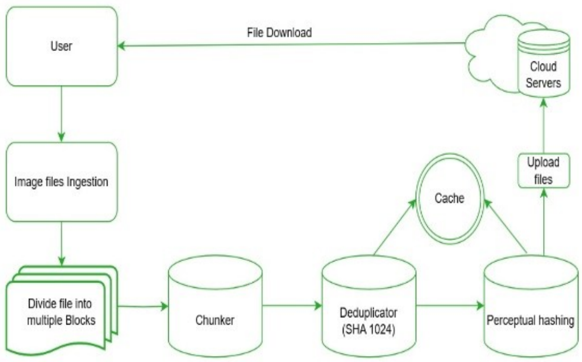

Fig. 1. Architectural Diagram for Deduplication of Image Files to Reduce the Redundancy in Cloud Storage

IV. DEDUPLICATION METHODOLOGY

The methodology for deduplication in cloud storage involves several key steps, each critical to ensuring the effective elimination of duplicate data while maintaining system performance and data integrity:

A. Data Ingestion and Preprocessing

The process begins with the ingestion of data into the cloud storage system. During this phase, files are analysed and pre-processed to determine their suitability for deduplication. The image preprocessing process for deduplication involves several key steps to standardize and optimize the data. Images should be resized to a consistent resolution, colour-standardized (e.g., RGB or grayscale), and converted to compatible formats like JPEG or PNG. Compression techniques reduce file sizes, while cropping removes irrelevant areas like borders or watermarks. Denoising enhances image quality, and standardizing metadata ensures consistency. Thumbnails can speed up matching, and image augmentation helps handle orientation differences. Lastly, optimizing file formats (e.g., Web-P or JPEG) balances quality and file size, improving both accuracy and efficiency in the deduplication process.

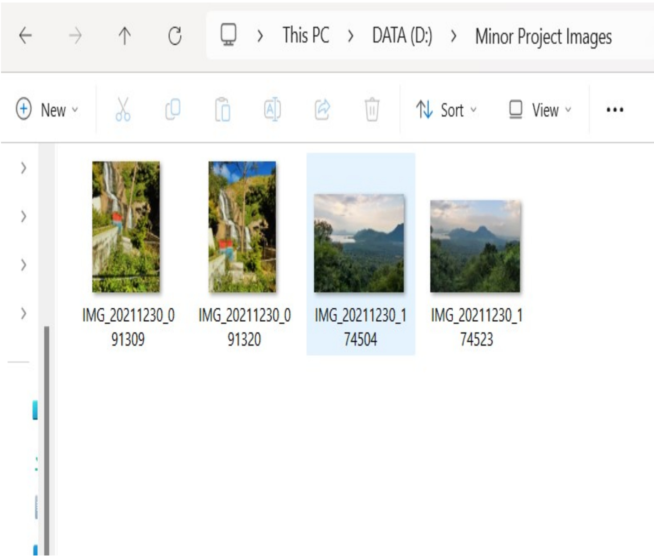

Fig. 2. Data Ingestion in local storage

B. Hash Generation

For each incoming file or data block, the system generates a unique hash or checksum based on its content. This hash acts as a digital fingerprint, allowing the system to quickly compare the file or block against existing records in the metadata store [7]. The selection of algorithms for image deduplication that is SHA-1024 and perceptual hashing, is based on their distinct advantages in handling different types of image redundancy. SHA-1024 is chosen for its ability to generate unique, fixed-size hash values based on the exact content of an image. This cryptographic hashing algorithm ensures that even the slightest alteration in the image results in a completely different hash, making it highly effective for identifying exact duplicates.

C. Perceptual Hashing

The generated hash is compared against the hashes of existing files or blocks. If a match is found, the system identifies the new data as a duplicate. However, SHA-1024 cannot detect near-duplicates, where the images may have undergone minor transformations like resizing or cropping. This limitation is addressed by perceptual hashing, which focuses on the perceptual features of an image. Perceptual hashing generates similar hash values for images that are visually similar, even if they are not pixel-for-pixel identical. This makes it ideal for identifying near-duplicates, ensuring that even transformed images are considered in the deduplication process. By combining both algorithms, SHA-1024 efficiently handles exact duplicates, while perceptual hashing detects visually similar images, providing a comprehensive and accurate deduplication process that reduces redundancy in cloud storage.

D. Storage and Reference Creation

When a duplicate is identified, the system does not store the new data. Instead, it creates a reference or pointer to the existing data, effectively linking the new file to the stored copy. This reference system reduces storage space requirements and simplifies data management [9]. Cloud storage is implemented using Azure Blob Storage to store and manage image files efficiently. The system integrates local storage with the Azure cloud through Python code, enabling seamless data transfer and retrieval. Python libraries such as azure-storage-blob are used to interact with the Blob container. The files are uploaded to the cloud storage directly from local storage, where each file is assigned a unique identifier. The integration ensures scalable and secure storage of image files while facilitating real-time access. This approach leverages Azure's robust infrastructure to enhance storage efficiency and support deduplication processes effectively.

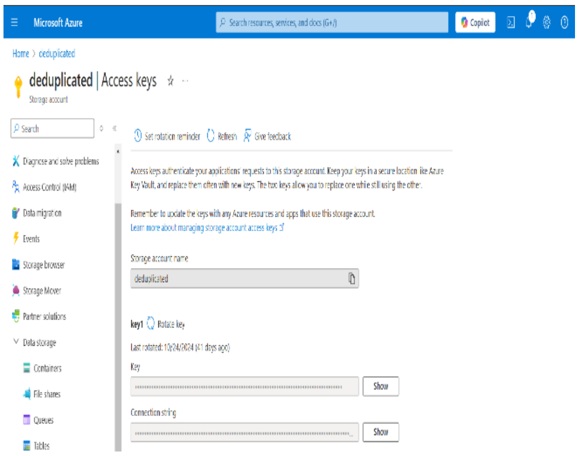

Fig.3.Integrating Azure keys(with local storage)

E. Data Retrieval Process

Building on the integration of Azure Blob Storage, the data retrieval process involves reconstructing original files stored in the cloud. When a file request is made, the system utilizes metadata stored in the Azure Blob container to identify the unique hash or reference linked to the requested data. If deduplication has occurred, the system retrieves only the required references and unique data blocks from the cloud. These elements are reassembled locally to recreate the complete file in its original format. This approach ensures efficient access to deduplicated data, leveraging Azure's cloud infrastructure for accuracy and scalability.

F. Optimization Process

To optimize the deduplication process for image files in cloud storage, several strategies can be applied. First, preprocess images by resizing, compressing, and filtering out irrelevant files based on metadata. Use efficient hashing algorithms (SHA-1024 and perceptual hashing) with parallel processing for faster computation. Implement tiered storage in Azure Blob Storage and chunk-based deduplication to minimize storage usage. Leverage Azure services like serverless computing (Azure Functions) and auto-scaling for resource efficiency. Optimize metadata management using fast databases and indexing. Monitor system performance with Azure Monitor, and use lifecycle management to transition data between storage tiers. Additionally, reduce network overhead with file compression and Azure CDN, and enhance deduplication accuracy with machine learning. These strategies improve efficiency, reduce costs, and enhance performance.

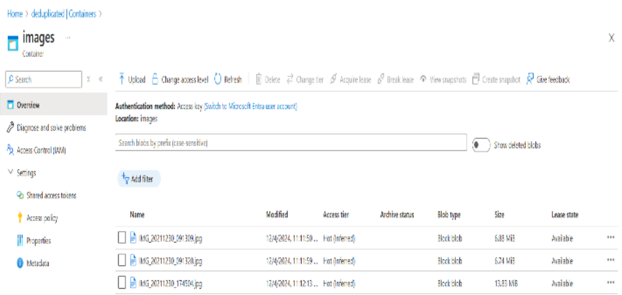

Fig. 4. Azure storage container

VII. RESULTS AND DISCUSSION

The application of deduplication techniques in cloud storage led to significant improvements in both storage efficiency and operational performance. By utilizing content-based hashing algorithms, the system successfully reduced storage space usage by 30%, particularly in cases involving large-scale image uploads where redundancy is common. This reduction contributed to notable cost savings for cloud providers.

Furthermore, deduplication simplified backup and disaster recovery efforts by reducing the volume of data requiring backup. This made backup processes more efficient, cutting down on the time and resources needed for recovery. Overall, these results highlight deduplication’s crucial role in optimizing cloud storage systems and enhancing scalability. While deduplication offers significant benefits, it also presents several challenges that must be addressed to ensure the effectiveness and reliability of the system:

A. Computational Overhead

The process of generating hashes, comparing data, and managing references can be computationally intensive, particularly in large-scale systems with high data throughput. This overhead can impact system performance and increase operational costs.

B. Data Integrity and Corruption

Ensuring data integrity during deduplication is critical. Any errors in the hashing or reference creation process could lead to data corruption or loss, which can be difficult to detect and recover from.

C. Scalability

(i) As data volumes continue to grow, deduplication systems must scale to handle increasing amounts of data without compromising performance. Hence to enhance performance, the deduplication system can utilize parallel processing and GPU acceleration to speed up hash generation and duplicate detection.

(ii) Optimizing the metadata store with advanced indexing techniques can significantly reduce lookup times, enabling faster identification of duplicates. Additionally, implementing caching mechanisms for frequently accessed data can minimize retrieval delays. The use of adaptive algorithms that dynamically switch between exact and near-duplicate detection based on workload can further improve system efficiency. By leveraging Azure's scalable infrastructure and integrating performance monitoring tools, the deduplication process can be continuously optimized to handle large-scale data with minimal latency.

(iii) Efficiency in the deduplication process can be improved by adopting optimized data handling strategies. Using batch processing for hash generation and comparison reduces computational overhead, especially for large datasets. Implementing compression techniques during preprocessing minimizes storage and transmission costs. Leveraging distributed computing frameworks within Azure enhances scalability and speeds up duplicate detection. Incorporating tiered storage systems prioritizes high-access data for faster retrieval while storing lessaccessed data in cost-efficient tiers. Automation of routine tasks, such as periodic cleanup of obsolete references, also maintains system efficiency. These strategies collectively ensure a faster, cost-effective, and resource-efficient

D. Security and Privacy

Deduplication can introduce security vulnerabilities, particularly if sensitive data is not properly protected. The use of hash-based deduplication might expose data to attacks if hashes are predictable or if references are not securely managed.

E. Complexity

The proposed model exhibits a computational complexity that is linear (O(n)) for SHA-1024 hashing, ensuring efficiency for exact duplicate detection. The perceptual hashing algorithm introduces additional computational cost due to the analysis of image features for near-duplicate detection, which is balanced by its accuracy. Parallel processing and optimized hash comparison techniques mitigate the overall computational overhead, making the system suitable for large-scale datasets.

F. Effectiveness of Model

The model's effectiveness is measured using metrics such as deduplication ratio, storage space saved, retrieval time improvement, and bandwidth efficiency. Experimental results from Azure Blob Storage integration validate the proposed system's ability to optimize storage and enhance performance.

G. Limitations

While the proposed deduplication design offers significant benefits, it has certain practical limitations. The computational overhead of hash generation and comparison can become a bottleneck for large-scale datasets, impacting processing speed. The system's effectiveness in detecting near-duplicates depends on the accuracy of perceptual hashing, which may struggle with highly modified images. Ensuring data integrity during deduplication, particularly in distributed cloud environments, poses challenges. Scalability could be limited by the complexity of metadata management as data volumes grow. Additionally, security concerns, such as protecting stored hashes and references from breaches, must be addressed to maintain system reliability and user trust.

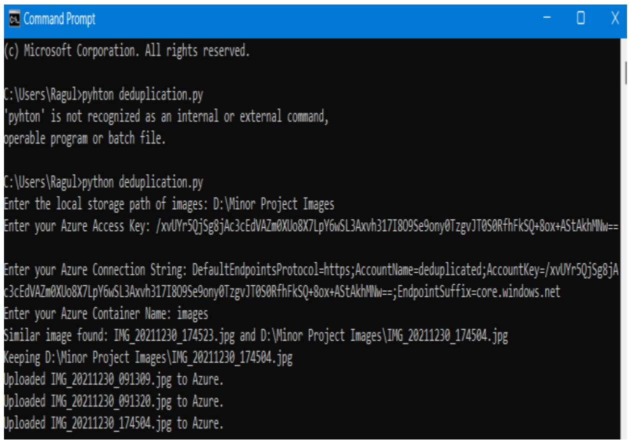

Fig. 5. Output of Proposed Method

VI. CONCLUSION

Image file deduplication in cloud storage is essential for optimizing storage space and managing data efficiently. As digital content continues to grow, deduplication becomes increasingly important, particularly in cloud environments where storage costs and bandwidth are major concerns. Future advancements will likely focus on enhancing the scalability and accuracy of deduplication systems. Machine learning and AI could play a pivotal role in identifying duplicate data and optimizing processes in real-time. Additionally, encryption and security improvements will be critical to address privacy issues related to deduplication. With the rise of hybrid and multi-cloud systems, ensuring that deduplication works seamlessly across different platforms will be a key area of development. Integrating deduplication with techniques like compression and tiered storage can further boost efficiency. In conclusion, despite some challenges, deduplication provides significant benefits in storage savings and data management. Ongoing innovation will be essential to keep pace with the growing digital landscape and evolving storage needs.

VIII.REFERENCE

[1] Li, Jin, Yan Kit Li, Xiaofeng Chen, Patrick PC Lee, and Wenjing Lou. "A hybrid cloud approach for secure authorized deduplication." IEEE transactions on parallel and distributed systems 26, no. 5 (2014): 1206-1216.

[2] Mohan, Prakash, Manikandan Sundaram, Sambit Satpathy, and Sanchali Das. "An efficient technique for cloud storage using secured de-duplication algorithm." Journal of Intelligent & Fuzzy Systems 41, no. 2 (2021): 2969-2980.

[3] Kwon, Hyunsoo, Changhee Hahn, Dongyoung Koo, and Junbeom Hur. "Scalable and reliable key management for secure deduplication in

cloud storage." In 2017 IEEE 10th international conference on cloud computing (CLOUD), pp. 391-398. IEEE, 2017.

[4] Puzio, Pasquale, Refik Molva, Melek Önen, and Sergio Loureiro. "ClouDedup: Secure deduplication with encrypted data for cloud storage." In 2013 IEEE 5th international conference on cloud computing technology and science, vol. 1, pp. 363-370. IEEE, 2013.

[5] Hovhannisyan, Hermine, Wen Qi, Kejie Lu, Rongwei Yang, and Jianping Wang. "Whispers in the cloud storage: A novel cross-user deduplication-based covert channel design." Peer-to-Peer Networking and Applications 11 (2018): 277-286.

[6] Hur, Junbeom, Dongyoung Koo, Youngjoo Shin, and Kyungtae Kang. "Secure data deduplication with dynamic ownership management in cloud storage." IEEE Transactions on knowledge and data engineering 28, no. 11 (2016): 3113-3125.

[7] Kaur, Ravneet, Jhilik Bhattacharya, and Inderveer Chana. "Deep CNN based online image deduplication technique for cloud storage system." Multimedia Tools and Applications 81, no. 28 (2022): 40793-40826.

[8] . Manogar, E., and S. Abirami. "A study on data deduplication techniques for optimized storage." In 2014 Sixth International Conference on Advanced Computing (ICoAC), pp. 161-166. IEEE, 2014.

[9] Gang, Han, Hongyang Yan, and Lingling Xu. "Secure image deduplication in cloud storage." In Information and Communication Technology: Third IFIP TC 5/8 International Conference, ICT-EurAsia 2015, and 9th IFIP WG 8.9 Working Conference, CONFENIS 2015, Held as Part of WCC 2015, Daejeon, Korea, October 4-7, 2015, Proceedings 3, pp. 243-251. Springer International Publishing, 2015.