WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Optimizing Storage Efficiency Observing De- Duplication Across Cloud Data

Saravanan C R[1]; Bhunaveshwari.M[2]

saravananr978@gmail.com : bhuvaneswari@drmgrdu.ac.in
PG Student; Professor
Department of Computer Applications,
Dr. M.G.R. Educational and Research Institute, Chennai – 6000 95

ABSTRACT:

Optimizing Storage Efficiency: Observing De-Duplication Across Cloud Data" explores the strategic application of data de-duplication techniques to enhance storage efficiency within cloud environments. This study investigates the efficacy of various de-duplication methods in reducing redundant data and optimizing resource utilization. By analysing patterns and trends in data de-duplication across diverse cloud datasets, this research aims to identify best practices and novel strategies for improving storage efficiency. The findings offer insights into how organizations can effectively manage their data footprint, minimize storage costs, and enhance overall performance within cloud infrastructures through targeted de-duplication approaches

Keywords: De-Duplication, Data Security, AE (Advanced Encryption Standard)

I.INTRODUCTION:

Services for cloud storage have grown in popularity recently. Client-side deduplication is used by several cloud storage services, such as Dropbox 1 and Waul 2, to lower resource usage in network bandwidth and storage [21, 39]. In other words, the server saves the uploading process if a file is already saved in cloud storage and checks if the file was already uploaded by another user when a user wants to upload it. Every single file will have a single copy in the cloud in this manner (also known as single-instance storage). Depending on the application, the deduplication technique can save up to 90% on storage, according to an SNIA white paper [34]. The following is an example of an existing client-side deduplication implementation, per Halevi et al. [20] and Dropship [17]: Alice, a cloud user, attempts to upload file F to cloud storage. Alice's computer will run the cloud storage service's client software, which will calculate and transmit the hash value hash(F) to the cloud server. The hash values of all incoming files are kept in a database by the cloud server, which consults this database to find the value hash (F). In the event that no match is discovered, file F is not yet stored in the cloud. The hash value hash (F) will be added to the look-up database, and Alice's client program will be needed to upload F to the cloud storage. If a match is discovered, it indicates that file F has previously been uploaded to the cloud storage by Alice or by other users. In this instance, Alice's computer uploads file F to the cloud storage, where it is saved and accessed by Alice via the cloud server. The client-side deduplication technique described above may be known as the "hash-as-a-proof" method. The hash value hash (F) has two uses in this method: (1) It serves as an index for file F, allowing the cloud server to locate information among a large number of files. (2) It is also considered "proof" that Alice owns file F.
Dropbox3 formerly used the "hash-as-a-proof" method for block-level cross-user deduplication [20][17]. If the cloud storage service's client software is trusted and the hash function is collision-resistant, the "hash-as-a-proof" method is secure enough. Malicious users can use the cloud service's public API4 to create their own client software and send modified communications, such as hash output, to the server. To avoid relying on client software, a more advanced solution is needed.
This research presents an effective RMA-ABE technique for cloud storage. Our major contributions are as follows:

- We present an efficient RMA-ABE technique for cloud storage using elliptic curve cryptography (ECC), eliminating the need for bilinear pairing processes. The proposed system utilizes linear secret sharing schemes (LSSS) to improve access policy expressiveness and incorporates a version key to enable quick attribute revocation.
- The proposed RMA-ABE method achieves indistinguishability against the choose plaintext attack (IND-CPA), as well as collision resistance and forward secrecy, under the decisional Diffie-Hellman (DDH) assumption.
- The scheme's performance evaluation shows reduced compute and storage costs compared to schemes.

II.LITERATURE SURVEY

According to Goyal, O. Pandey, A. Sahai, and B. Waters, et al., (2006), Encrypting data saved on third-party sites will become necessary as sensitive information is exchanged and stored online. Encrypting data allows for selective sharing, but only at a coarse level (e.g., sharing your private key). Our

new cryptosystem, Key-Policy Attribute-Based Encryption (KP-ABE), enables the fine-grained exchange of encrypted data. Our cryptosystem labels ciphertexts with qualities and associates private keys with access structures to determine which ciphertexts users can decipher. We show how our structure may be used for exchanging audit logs and broadcast encryption. Our structure enables delegation of private keys, which includes hierarchical identity-based encryption (HIBE).

According to S. Yu, K. Ren, and W. Lou, et al., (2011) In recent years, distributed sensor data storage and retrieval have become increasingly popular for numerous applications. Although distributed architecture provides a more robust and fault-tolerant wireless sensor network (WSN), it also presents security issues, particularly in mission-critical applications like war fields and e-healthcare. Individual sensors store and retain sensor data, and unattended sensors are vulnerable to physical attacks, making data security challenging. Fine-grained data access control is essential in mission-critical applications to prevent unauthorized access to sensitive data, which can have fatal consequences and may be unlawful. Sensors are often limited in resources, making immediate adoption challenging.

Acc0rding to Z. Wan, J. Liu, and R. H. Deng, et al., (2011) In recent years, cloud computing has become a major trend in the IT sector. Outsourcing valuable data to cloud providers raises security and privacy concerns. Many proposed techniques for access control of outsourced data in cloud computing use attribute-based encryption (ABE), but they are limited in their ability to execute complicated regulations. This paper proposes hierarchical attribute-set-based encryption (HASBE), which extends ciphertext-policy attribute-set-based encryption (ASBE) with a user hierarchy, for scalable, flexible, and fine-grained access control of outsourced data in cloud computing. The suggested approach enables scalability through a hierarchical framework. Additionally, it inherits the flexibility and fine-grained access control of ASBE to enable its compound attributes. HASBE's multiple value assignment for access expiration time improves user revocation efficiency compared to existing systems. We show the security of HASBE using the ciphertext-policy attribute-based encryption (CP-ABE) technique by Bethencourt et al. and evaluate its performance and computational cost. Our approach to access control for outsourced data in cloud computing is efficient and adaptable, as demonstrated by detailed trials.

III.PROPOSED METHOD

Performance-Based Data Deduplication for Cloud Primary Storage Existing primary storage data deduplication systems, such as died and Offline-Dedupe, are primarily concerned with saving storage capacity. These systems often prioritize recognizing and deduplicating large requests while ignoring smaller ones (e.g., 4KB, 8KB, or less) because they have little impact on total storage capacity. This capacity-oriented strategy is based on the idea that deduplicating short I/O requests is frequently unprofitable and can result in significant overheads that outweigh the benefits of storage efficiency.

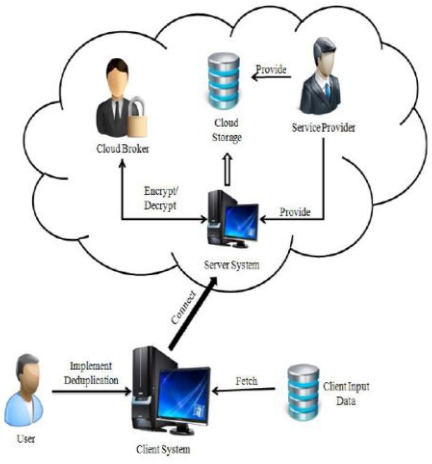
However, this capacity-focused strategy overlooks one important aspect of main storage performance, especially in cloud contexts. The performance of primary storage systems is critical, and any deduplication strategy employed must optimize storage capacity while also improving I/O speed.

POD has a two-pronged strategy for improving the performance of primary storage systems and lowering the overhead associated with deduplication operations. This method includes: Request-Based Selective Deduplication (Select-Dedupe): POD uses a selective deduplication approach called Select-Dedupe, which identifies and deduplicates tiny I/O requests based on workload properties. Unlike capacity-oriented techniques, which focus on large requests, Select-Dedupe recognizes the dominance of small-I/O-request workloads and improves deduplication strategies accordingly. Adaptive Memory Management (I Cache): To alleviate memory congestion caused by bursty read and write traffic, POD incorporates an adaptive memory management method known as cache. cache dynamically optimizes memory allocation to minimize conflicts between bursty read and write operations, hence enhancing overall system performance and responsiveness.

Select-Dedupe is a critical POD component that is specifically designed to improve deduplication performance by focusing on tiny I/O requests, which account for a substantial portion of primary storage workloads. The design concepts and mechanisms of Select-Dedupe are described as follows: Workload Characteristics Analysis: Select-Dedupe performs a thorough analysis of workload variables to find patterns related to modest I/O requests. Select-Dedupe enhances deduplication solutions by recognizing the type and frequency of these requests, focusing on tiny, often occurring data chunks. Request-based deduplication: Unlike typical techniques that just consider data segment size, Select-Dedupe employs a request-based deduplication mechanism. It assesses the significance and frequency of I/O requests, deduplicating data segments based on their impact on overall system performance and storage efficiency.

Dynamic Memory Allocation: In response to variations in workload, cache dynamically distributes memory resources, giving priority to those needed for quick read and write operations. This dynamic allocation technique reduces congestion and improves system responsiveness as a whole. Bursty Traffic Mitigation: I Cache prevents performance bottlenecks and ensures consistent I/O performance across a range of workloads by cleverly managing memory allocation to effectively alleviate the impact of bursty read and write traffic. Adaptive Policies: I Cache makes use of adaptive policies, which watch over system conditions continually and modify memory allocation techniques as necessary. Because of its adaptive nature, cache can react quickly to shifts in the dynamics of the workload, maximizing memory use and performance.

IV.EXPLANATION



The image you provided is a diagram of cloud computing architecture. It shows the different components of a cloud computing system and how they interact with each other.
The main components of a cloud computing system Hardware:
The servers, storage, network devices, and other hardware that power the cloud.
Virtualization:
An abstraction layer that creates a virtual representation of physical computing and storage resources. This allows multiple applications to use the same resources.
Front end:
The part of the cloud that users interact with. This includes the user interface, applications, and other tools that users need to access their data and resources.
Back end:
The part of the cloud that manages the infrastructure and resources. This includes the servers, storage, and network devices.
The diagram shows how the different components of a cloud computing system interact with each other. The client system sends a request to the server system. The server system then processes the request and sends the response back to the client system. The client system then displays the response to the user.
The diagram also shows how data is encrypted and decrypted in a cloud computing system. The data is encrypted before it is stored in the cloud. This protects the data from unauthorized access. The data is then decrypted when it is accessed by the authorized user.
Data deduplication is becoming more necessary for cloud storage providers because of the exponential increase in the number of users and the size of their data. Duplicate data can cause problems for businesses, resulting in diminishing profits due to fines and unnecessary cloud storage costs.

Data deduplication can occur at the source or target level:

Source-based deduplication Removes redundant blocks before transmitting data to a backup target, reducing bandwidth and storage use

Target-based deduplication:

Transmits backups across a network to disk-based hardware in a remote location, increasing costs but generally providing a performance advantage compared to source deduplication.
Data deduplication is becoming more necessary for cloud storage providers because of the exponential increase in the number of users and the size of their data. Duplicate data can cause problems for businesses, resulting in diminishing profits due to fines and unnecessary cloud storage costs.

V. MODULES

User:

In computing, a user refers to an individual or entity that interacts with a computer system, software application, or service. Users can perform various actions such as accessing files, running programs, or utilizing online services.

Data Owner:

The data owner is the individual or entity that has ultimate responsibility and control over a particular set of data. This includes determining who has access to the data, setting usage policies, ensuring data security, and making decisions regarding the data's lifecycle.

Public Cloud:

A public cloud refers to a cloud computing environment where computing services such as storage, networking, and applications are provided over the internet by third-party providers. These services are typically available to multiple users or organizations, and users can access them on a pay-per-use basis.

Private Cloud:

A private cloud, on the other hand, is a cloud computing environment that is dedicated to a single organization. It can be hosted on-premises or by a third-party provider, but its infrastructure and services are isolated and tailored to meet the specific needs and requirements of that organization. Private clouds offer greater control, security, and customization compared to public clouds

AA (Authentication and Authorization):

AA refers to the processes of authentication and authorization in the context of computer security. Authentication verifies the identity of a user or entity attempting to access a system or resource, typically through credentials like usernames and passwords, biometric data, or security tokens.

Authorization, on the other hand, determines what actions or resources a user is allowed to access or use based on their authenticated identity and assigned permissions.

VI. RESULT & DISCUSSION

- **Homepage**: The main landing page of a website, serving as an entry point for visitors to access content and navigate the site.
- **User**: An individual or entity interacting with a system, application, or service, often requiring authentication for access.
- **Data owner**: The entity responsible for supplying or managing data, offering access to information or resources.
- **Public Cloud**: A cloud computing environment providing services over the internet to multiple users or organizations.
- **Private Cloud**: A dedicated cloud computing environment tailored to a single organization's needs, offering enhanced control and security.

Fig 1: Home Page

User :

User: An individual or entity interacting with a system, application, or service, often requiring authentication for access.
Register: The process of creating a new account or profile on a platform by providing personal information and credentials, enabling access to personalized features.
Login: The action of accessing a previously registered account by entering valid credentials, typically a username or email and password, to gain authorized access.
Decryption: The process of converting encrypted data back into its original, readable format using an appropriate decryption key or algorithm, ensuring data accessibility.
Logout: The action of ending a user's session or connection with a system or platform, terminating access and enhancing security by preventing unauthorized use.
These terms are fundamental in user authentication, data security, and access control within digital platforms and systems.

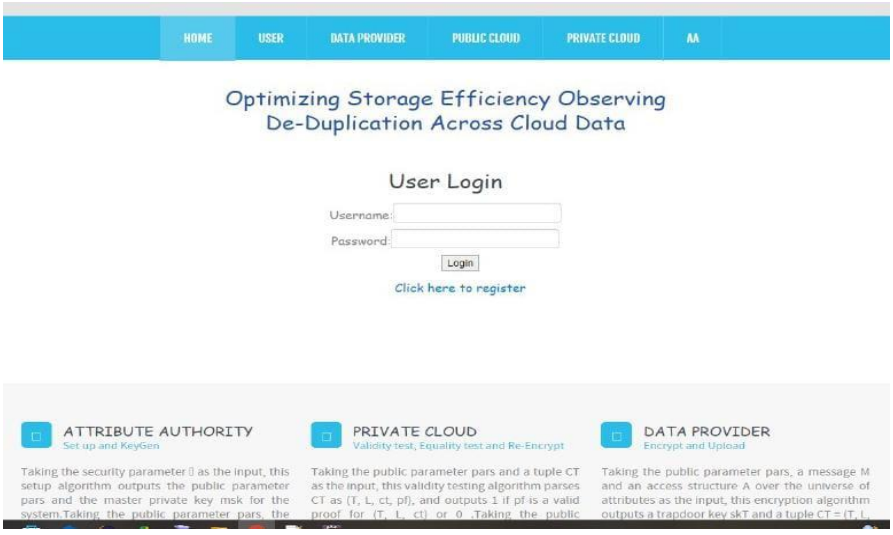


Fig 2: User Login Page

PRIVATE CLOUD:

Register: In a private cloud setting, registering refers to the process of creating a new user account within the private cloud infrastructure. This involves providing necessary information such as username, email, password, and possibly additional details depending on the cloud platform's requirements.
Login: After registering, users can log in to their accounts within the private cloud. Logging in requires entering the correct credentials (username and password) to gain access to the resources and services available within the private cloud environment.
Validity-Test: This term likely refers to a process within the private cloud that verifies the validity of user credentials during the login process. The validity test checks if the entered credentials match those stored in the private cloud's authentication system. If the credentials are valid, the user is granted access.
Equality-Test: In the context of a private cloud, an equality test could be a comparison check performed to ensure that user permissions and access rights are consistent and correctly assigned. For example, it may verify whether a user's role or group membership entitles them to access specific resources or perform certain actions.
Tags-Labels: Tags and labels are metadata attributes assigned to resources within the private cloud environment. They help categorize and organize resources for management, access control, and resource allocation purposes. Tags and labels can be used to group similar resources, apply policies, and streamline management tasks.
Logout: Similar to logging out in other contexts, logging out in a private cloud environment means ending the current user session and terminating access to the private cloud resources. It's an important security measure to ensure that unauthorized access is prevented once a user has finished their session
Certainly! Here's an explanation of each term in the context of a public cloud environment:

PUBLIC CLOUD:

Register: In a public cloud setting, registering typically involves creating a new user account or subscription with the cloud service provider. This process often requires providing information such as name, email address, payment details (if applicable), and creating login credentials (username and password).
Login: Once registered, users can log in to their accounts within the public cloud infrastructure. Logging in involves entering the correct credentials (username/email and password) to gain access to the cloud services, resources, and features associated with the user's account.
Storage Files: This term refers to the functionality of storing and managing files or data within the public cloud environment. Public cloud providers offer various storage solutions, such as object storage, block storage, and file storage, allowing users to upload, store, access, and manage their files securely over the internet.
Logout: Similar to logging out in other contexts, logging out in a public cloud environment means ending the current user session and terminating access to the cloud services and resources associated with the user's account. It's an essential security measure to prevent unauthorized access to the user's data and resources.
these terms relate to user management, authentication, data storage, and access control within a public cloud environment, highlighting key functionalities and security practices used in cloud computing.

DATA PROVIDER:

Register: This refers to the process of creating a new user account on the data provider platform. During registration, users typically provide information such as their name, email address, and create a password. Registering allows users to access the platform's features and service

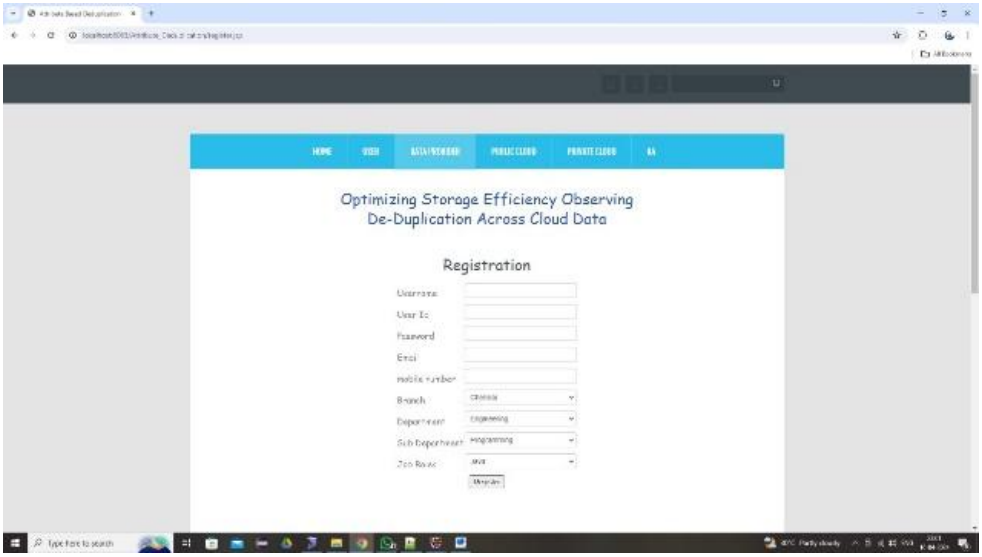


Fig 3: Data Provider Register Page.

Login: After registering, users can log in to their accounts by entering their login credentials, usually combination of a username/email and a password. Logging in verifies the user's identity and grants access to the data provider platform's functionalities, such as uploading, downloading, or managing data.

Upload File: This action involves transferring files or data from the user's device to the data provider platform. Users can upload various types of files, such as documents, images, videos, or datasets, depending on the platform's capabilities. Once uploaded, files are stored securely on the platform and can be accessed as needed.

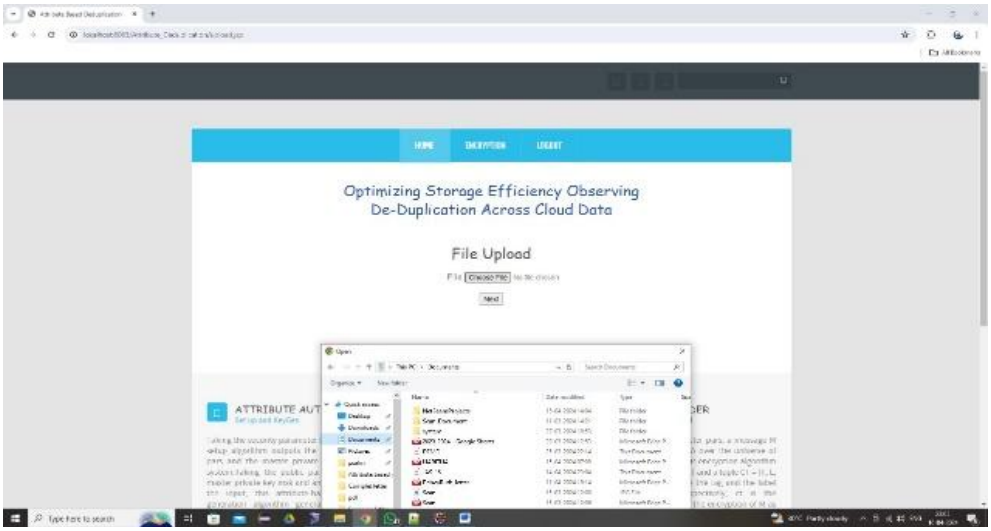


Fig 4: Data Provider Upload File.

It seems like you mentioned "View the file explantion." Could you please clarify what file or information you're referring to? If you need an explanation about viewing a specific file or type of information, please provide more details so I can assist you accurately

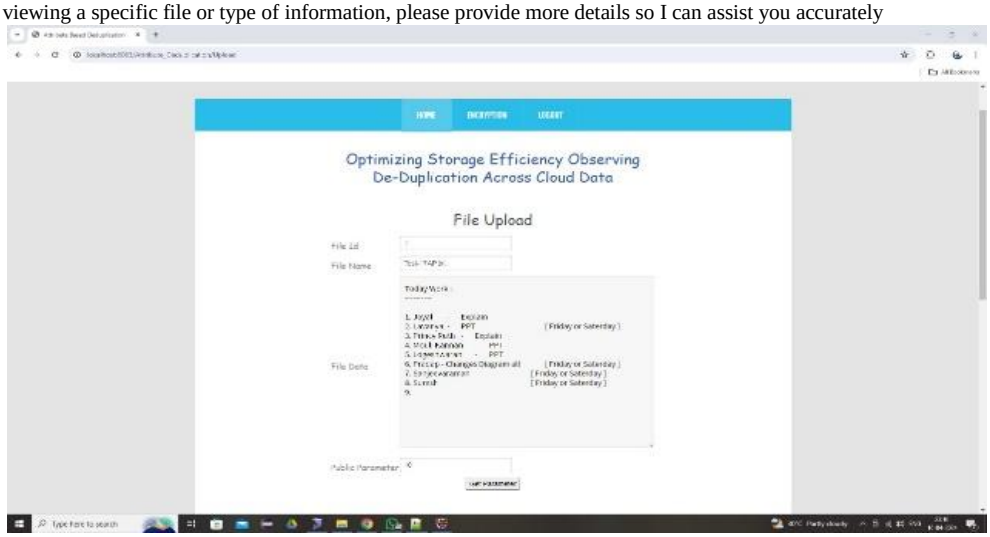


Fig 5: View the File

AA :

Login: This refers to the process of accessing a system, application, or platform by providing valid credentials, such as a username/email and password. Logging in verifies the user's identity and grants access to specific features, data, or functionalities based on their permissions and role within the system.

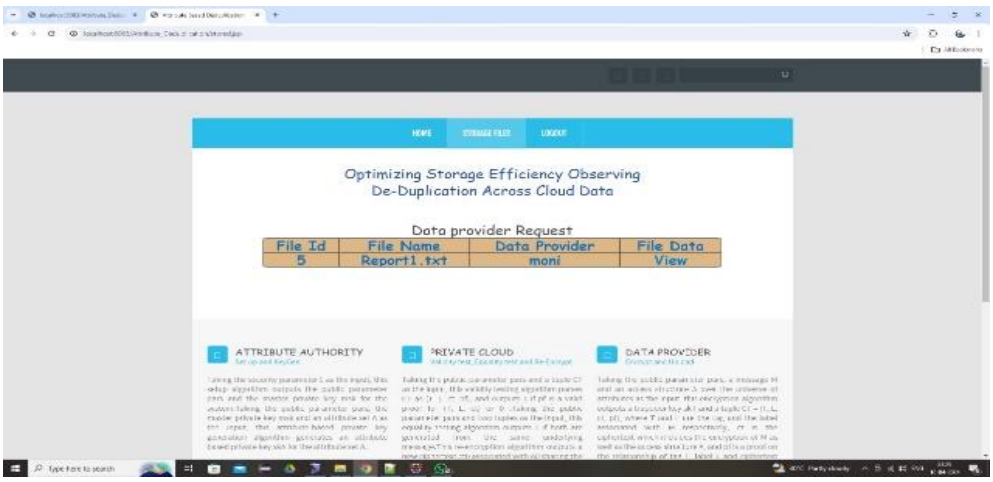


Fig 6: View the Storage Files

User: A "user" is an individual who has been granted access to a system, application, or platform. Users interact with the system by logging in and performing various tasks or activities based on their assigned permissions. Users can have different roles and levels of access, depending on the system's configuration.

Owner: The "owner" typically refers to a user who has special privileges or administrative rights within a system or platform. Owners often have the authority to manage user accounts, configure settings, and control access to resources. They may also have the ability to make critical decisions regarding the system's operation and maintenance.

Logout: Logging out is the action of ending the current session and disconnecting from the system or platform. When a user logs out, it closes their session and prevents unauthorized access to their account or data. Logging out is important for security purposes, especially when using shared or public devices.

VII.CONCLUSION :

We have presented an effective user-defined access control safe deduplication mechanism in this research. In particular, our technique achieves the permitted deduplication without requiring the usage of hybrid cloud architecture or the deployment of an extra authorized server. Only the CSP is able to oversee access privileges on behalf of data owners in our scheme without jeopardizing the privacy of the data. Additionally, our plan adds the Bloom filter to effectively finish the dual-category check. Thorough security evaluations show that our plan is capable of simultaneously achieving tag consistency, data secrecy, access control, and resistance against brute-force attacks. Our approach is efficient in terms of deduplication efficacy, computational cost, communication overhead, and storage cost, as demonstrated by comprehensive performance assessments on file-level and chunk-level deduplication.

VII. REFERENCES :

1. Amazon web services [Online]. Available: http://aws.amazon. com, 2014.
2. V. K. Adhikari, Y. Guo, F. Hao, M. Varvello, V. Hilt, M. Steiner, and Z.-L. Zhang, “Unreeling netflix: Understanding and improving multi-CDN movie delivery,” in Proc. IEEE Conf. Comput. Commun., 2012, pp. 1620–1628.
3. Cockcroft. (2011). Netflix in the cloud [Online]. Available: http://velocityconf.com/velocity 2011/public/schedule/detail/ 17785
4. B.Wong and E. G. Sirer, “Closestnode.com: An open access, scalable, shared geocast service for distributed systems,” Operating Syst. Rev., vol. 40, no. 1, pp. 62–64, 2006.
5. H. Ballani, P. Costa, T. Karagiannis, and A. Rowstron, “Towards predictable datacenter networks,” in Proc. ACM SIGCOMM Conf., Toronto, ON, Canada, 2011, pp. 242–253.
6. N. Laoutaris, M. Sirivianos, X. Yang, and P. Rodriguez, “Interdatacenter bulk transfers with netstitcher,” in Proc. ACM SIGCOMM Conf., Toronto, ON, Canada, 2011, pp. 74–85.
7. Greenberg, J. Hamilton, D. A. Maltz, and P. Patel, “The cost of a cloud: Research problems in data center networks,” ACM SIGCOMM Comput. Commun. Rev., vol. 39, no. 1, pp. 68–73, 2008.
8. Singh, M. Korupolu, and D. Mohapatra, “Server-storage virtualization: Integration and load balancing in data centers,” in Proc. ACM/IEEE Conf. Supercomput., 2008, p. 53.
9. R. Buyya, R. Ranjan, and R. N. Calheiros, “Intercloud: Utilityoriented federation of cloud computing environments for scaling of application services,” in Proc. 10th Int. Conf. Algorithms Archit. Parallel Process., 2010, pp. 13–31.
10. Qureshi, R. Weber, H. Balakrishnan, J. V. Guttag, and B. M. Maggs, “Cutting the electric bill for internet-scale systems,” in Proc. ACM SIGCOMM Conf., 2009, pp. 123–134.