IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

"DUPLICATE DATA IDENTIFICATION ON CLOUD"

Pradnya Pawar, Shalaka Deshpande, Gayatri Bagul and Prof. G.G.Raut
Department of MCA (Engg)
Gokhale Education Societies R.H. Sapat College of engineering
T. A. Kulkarni Vidyanagar, college road, Nashik,MH India-422005

Abstract:
Cloud data de-duplication means finding and removing duplicate copies of data in the cloud, which helps save space and makes the system work faster. However, some existing methods of de-duplication don't check if the data stored by the cloud service provider (CSP) is correct. So, this paper suggests a way to make sure that the data stored in the cloud is both de-duplicated and verified for accuracy. The suggested method includes checking the integrity (making sure the data hasn't been tampered with) of encrypted data stored in the cloud. This paper proposes a data integrity verification scheme of de-duplication for cloud cipher texts, including cloud cipher text de-duplication and cloud data integrity verification

Keywords: Data Duplication identification, hash algorithm, encryption, data integrity

1. Introduction

In recent years, more people and companies are choosing to store their data on cloud servers instead of their own computers. This saves money on storage and maintenance and avoids the hassle of managing complex networks and systems. However, when data is stored in the cloud, users lose control over it, and there's a risk of data loss or intentional deletion by the cloud service provider to save space.
To address these concerns, a method called the PDP mechanism was introduced to ensure that data stored on untrusted cloud servers remains intact. This allows third parties to challenge the cloud server to prove that it has the data. As cloud technology and big data grow, efficient management of resources becomes important, especially in dealing with duplicate data, which wastes storage space and complicates management.
Various methods have been proposed to handle duplicate data, but they often lack certain features like user revocation, data integrity verification, or support for multiple users. To tackle these issues, a new scheme is suggested. It stores data on the cloud server in encrypted form to protect privacy and uses block tags to verify ownership and avoid storing duplicate data. Additionally, a method called proxy re-signature is used for integrity verification, reducing the need for constant user involvement and decreasing the workload on the cloud server. Compared to similar schemes, this approach is more efficient.

2. Objectives

The main objective of the project is to deduplicate the data using various encryption standards, and to save the data in an encrypted format and give access to every user without leaking the main private key.

3. Literature review

Fine-grained data Duplication and proof of storage Scheme in Public Cloud Storage (2021) author: Hardik Gajera, Manik Lal Das **[1]**

This paper presents Verification of the integrity of the data stored on a public cloud server is a challenging research problem. While providing on-demand service to the service consumer, the cloud service provider also requires ensuring minimizing error in data stored in a cloud storage server. To address this issue, many schemes like proof of storage (POS), proof of data possession (PDP), and proof of irretrievability (POR) exist in the literature. However, for practical purpose, a scheme which verifies the data stored in the server should be compatible with the standard practice of data deduplication for an efficient storage system.

Secure Cloud Data Duplication with Efficient Re-Encryption (2021) Author-MohdAman, Prashant Verma D Rajeswari **[2]**

After the emergence of the cloud architecture, many companies migrate their data from conventional storage i.e., on bare metal to the cloud storage. Since then huge amount of data was stored on cloud servers, which later resulted in redundancy of huge amount of data. Hence in this cloud world, many data de-duplication techniques have been widely used. Not only the redundancy but also made data more secure and privacy of the existing data were also increased. Some techniques got limitations and some have their own advantages based on the requirements. Some of the attributes like data privacy, tag regularity and interruption to brute-force attacks. To make data deduplication technique more efficient based on the requirements. This paper will discuss schemes that brace user-defined access control, by allowing the service provider to get information of the information owners.

zDeduplication in cloud storage on the basis of proof of ownership (2020) author: Rupali Bhimrao Sirsat, Nitin R Talhar **[3]**

Nowadays, more and more corporate and private users outsource their data to cloud service provider. It is provide lots of services to the users. Due to the large amount of data stored by multiple users on cloud one critical problem must be faced by users. In this large amount of data there are so many copies of repeated data. Data duplication is compress repeated data. We can use this technique To make data management scalable and mitigate the increased amount of data. In the proposed system, we are going to used two cloud server, public cloud server and private cloud server. Public cloud server store data (hash, encrypted data) and compare hash which is generated by private cloud. After comparison hash is stored in the data base by private cloud. This proposed system used to mitigate the problem of data duplication. Based on this we can improve storage utilization and efficiency of cloud storage

CPDA: A Confidentiality-Preserving Deduplication Cloud Storage with Public Cloud Auditing (2019**)**Author: Jiaojiao Wu; Yanping Li; Tianyin Wang **[4]**

In this paper, we propose confidentiality-preserving deduplication cloud storage with public cloud auditing (CPDA). Firstly, our CPDA scheme achieves secure file deduplication on encrypted file, which supports public integrity auditing for the unique copy in the deduplication cloud storage system. Particularly, our CPDA scheme also realizes secure authentication tag deduplication. Secondly, our CPDA scheme utilizes the convergent encryption and random masking techniques to ensure data confidentiality during the file deduplication and integrity auditing process. Thirdly, our scheme not only supports each data owner to independently launch the integrity auditing of their own files, but also supports cloud server to periodically delegate the third party auditor to concurrently handle multiple auditing tasks to ensure the integrity of the outsourced files. Finally, the security of our scheme is formally proved and its performance is confirmed by numerical analyses and simulation experiments.

A secure data deduplication System For integrated cloud-edge networks(2020)Author: Nadesh R. K Varun G. Menon Mahadi Abbasi **[5]**

Data redundancy is a significant issue that wastes plenty of storage space in the cloud-fog storage integrated environments. Most of the current techniques, which mainly center around the static scenes, for example, the backup and archive systems, are not appropriate because of the dynamic nature of data in the cloud or integrated cloud environments. This problem can be effectively reduced and successfully managed by data deduplication techniques, eliminating duplicate data in cloud storage systems. Implementation of data deduplication (DD) over encrypted data is always a significant challenge in an integrated cloud-fog storage and computing environment to optimize the storage efficiently in a highly secured manner.

Hybrid cloud storage System With enhanced multilayer cryptosystem for secure duplication in cloud(2023)Author-Nagappan Mageshkumar, J. Swapna, R Rajkumar **[6]**
Data deduplication is a crucial technique in the field of data compression that aims to eliminate redundant copies of recurring data. This technique has gained significant popularity in the realm of cloud storage due to its ability to effectively reduce storage requirements and optimize bandwidth utilization. To ensure the safeguarding of sensitive data while simultaneously facilitating deduplication, researchers have put forth the concept of convergent encryption as a potential solution. This technique involves encrypting the data prior to its outsourcing, thereby enhancing the confidentiality of the information. In this work, an earnest endeavor is undertaken to formally tackle the issue of authorized data deduplication, with the aim of enhancing data security

Hash- indexing block -based deduplication Algorithm for reducing storage in the cloud (2023) Author- D. Viji and S. Revathy **[7]**
Cloud storage is essential for managing user data to store and retrieve from the distributed data centre. The storage service is distributed as pay a service for accessing the size to collect the data. Due to the massive amount of data store in the data centre containing similar information and file structures remaining in multi-copy, duplication leads to increase storage space. The potential deduplication system doesn't make efficient data reduction because of inaccuracy in finding similar data analysis. It creates a complex nature to increase the storage consumption under cost. To resolve this problem, this paper proposes an efficient storage reduction called Hash-Indexing Block-based Deduplication (HIBD) based on Segmented Bind Linkage (SBL) Methods for reducing storage in a cloud environment under cost. To resolve this problem, this paper proposes an efficient storage

A survey on DE – Duplication schemes in cloud servers for secured data analysis in various applications(2022) Author- K. Pragash , J. Jaya bharathya **[8]**
The process of eliminating repeated or redundant copies of data in order to reduce the storage requirements which in turn could improve the efficiency of the system is termed as data duplication. There are two major strategies through which deduplication can be implemented. The initial process is to run the deduplication scheme as an inline process which could avoid the redundant data occurrence before writing the data into the storage system. The secondary process to implement deduplication is use the scheme as a background process which can clear the redundant data from the storage unit after the data is identified from the disk

Data Deduplication for Efficient Cloud Storage and Retrieval (2019) Author- Rishikesh Misal and Boominathan Perumal **[9]**
Cloud services provide flawless service to the client by increasing the geographic availability of the data. Increasing availability of data induces high amount of redundancy and large amount of space required to store that data. Data compression techniques can reduce the amount of space required for that data to be store at various sites. Data compression will ensure that there is no loss of availability and consistency at any site. As there is huge demand for cloud services and storage due to this the amount of investment also increases. By using data Compression we can reduce the amount of investment required and this will also decrease the amount of physical space and data Centres required to store data. Various security protocols can be incorporated to secure these compressed files at various sites. We provide a reliable technique to store deduplicates and its management in a secure manner to accomplish high consistency as well as availability

Deduplication of Data in the Cloud-based Server for Chat/File Transfer Application (2021) Author- Ankit Rai1, Arpit Diwan, Harsh Tripathi, K. Rajkumar **[10]**
In this paper, our focus is on saving maximum possible storage and providing a secure deduplication mechanism in order to keep the client's trust on the system. For Block-level deduplication when the size of a specific file is greater than a predefined threshold, we made chunks of that specific size and applied crypto-hash function which significantly improved our deduplication percentage. After the comparison of our experimental results, we found our system showed 5% improvement to the previously discussed framework, in addition to improving results we also had made comparison of the File-level data deduplication and Block level data deduplication, the security model of our approach provides the same privacy and security measures to the user data

Data Deduplication On Encrypted Big data in cloud (2019) Author-V. Khanna , A .Kumaravel and A. Rama[**11**]
In this data duplication which involves identifying and eliminating duplicate copies of the data to optimize storage space and reduce transmission bandwidth requirements this extends to support data deduplication in hybrid cloud environments where data is stored across both private and public cloud infrastructure. data compression techniques are applied to further reduce storage space by minimizing the data. this paper includes security analysis to evaluate the effectiveness of the proposed approach in preventing unauthorized access to sensitive data.

Data deduplication for cloud storage (2021) Author-C.S.N. Koushik, Shruti choubey, Abhishek choubey and G.R. Sinha [**12**]
Cloud storage devices to efficiently stored and access large amount of data cloud storage offers easy access ability to the data and its duplicate data. Data duplication enables efficient backup and restoration process in the cloud duplicate data is back up once, and logical pointers are used to reference duplicate instances ,reducing storage requirement and ensuring consistency in backup and restoration .the deduplication manager is employed to handle the deduplication process efficiently.

Data Duplication removal in cloud computing based on the file checksum (2022) Author-James Adegboye and Folahan jiboku **[13]**
It focuses on the duplicate data at both the end file and sub-file levels, identifying and removing redundant content using file checksum algorithm. This algorithm compute enabling efficient identification of duplicate content. This works acknowledges the computational infrastructure required for effective data processing and analysis in addressing rising storage demands. this approach efficiently eliminates duplicate data from storage system. With the exponential growth of digital content and increasing numbers of cloud customers , deduplication becomes a necessity for cloud storage providers to optimize storage space and reduce costs.
Secure cloud storage with deduplication technique (2023) Author- komal kshirsagar , Pallavi patekar , Sarav Kolhe and prof. P.N. Pathak[**14**]
Data duplication technique proposes using data deduplication as a method to identify and eliminating redundant data before storing it in the cloud. This technique involves computing incoming data with existing data in the cloud to identify duplicate and only storing unique data. the paper suggest employing third party auditors to verify the integrity of the data and provide assurance to users regarding data accuracy. the correctness and accuracy of the data stored in the cloud.

HASH-INDEXING BLOCK-BASED DEDUPLICATION ALGORITHM FOR REDUCING STORAGE IN THE CLOUD (2023) AUTHOR-D. VIJI, S. REVATHY [15]

This proposed system intends a new deduplication technique to reduce the storage space using hash indexing forums. It reduces the content search capability between the documents to make an ordered list of hash value indexes for finding the file content. By reducing the file comparison to find the similarity, content analyses based on the segmentation concept are used to split the files into a block that makes the block-based comparison. The compared files contain similarity count terms based on the content presented in the document, which is estimated by distance vector weightage correlation. Then the files are grouped into clusters to make similarity indexing values in the cluster. Depending on the cluster weight, the record linkage concepts are intended in the segmented bind linkage to compare the cluster weight to find the duplicates.

A middle-ware approach to leverage the distributed data deduplication capability on HPC and Cloud storage systems (2020) Author-Hsing-bung Sihai tang Texas Denton **[16]**
A new middle-ware design and implementation approach, named D3M, to support distributed data de-duplication feature on existing file and object storage systems. We also incorporate this proposed D3M middle-ware with the RedHat Linux device layer de-duplication and compression driver, called VDO (Virtual Data Optimizer). With these two layers of data de-duplication support, we accommodate both client side and server-side data de-duplication features. Finally, we conduct various testing cases on HPC data sets and Enterprise data sets to illustrate the benefits and advantages of applying our bilayer data de-duplication middle-ware solution

4. Existing System

In existing, a random client-side data de-duplication scheme, which realized multi-user ownership management and data sharing by means of a dynamic encryption key tree. proposed Client-side secure de-duplication scheme for cipher text data in cloud storage, but did not achieve data integrity verification. A cloud data audit scheme that supports encrypted data de-duplication, and does not support multi-user operations. An integrity audit scheme that supports key update and cipher text data de-duplication, but requires users to participate in key update online, and does not support multiple users operating.

Disadvantages:

- Proposed an integrity audit scheme that supports key update and cipher text data de-duplication, but requires users to participate in key update online.
- A cloud data audit scheme that supports encrypted data de-duplication, and does not support multi-user operation.

5. System Architecture

Imagine you have a system where you can store your files on the internet, like Google Drive or Dropbox. In this system, we want to make sure that your files are kept safe and private, so we're using specialtechniquescalled cryptography to protect them.Here's how it works:

1. **Uploading Files**: When you upload a file to the cloud (like putting it in a digital storage locker), an admin (let's call them AP) keeps an eye on things to make sure everything is secure.
2. **Cloud Storage**: The files are actually stored on a service provided by a company (we'll call them CSP, for Cloud Storage Provider), like Google or Amazon. They have big computers called servers where they keep all the files safe.
3. **Keeping Data Safe**: Before your files are sent to the cloud, they're scrambled up using encryption. It's like putting your file in a digital safe with a special code that only you and the system know.
4. **Saving Space**: Now, here's where it gets interesting. We don't want to waste space on the cloud by storing the same file multiple times if different people upload it. So, we use a process called deduplication. This means the system checks if someone else has already uploaded the same file. If they have, it doesn't make a new copy—it just points to the original one.
5. **Getting Your File Back**: When you want to see your file again, the system takes the encrypted version from the cloud, and you use a special key to decrypt it. It's like unlocking that digital safe to get your file back in its original form.

So, in simple terms, this system keeps your files safe and private on the internet, makes sure we don't waste space by storing the same file multiple times, and lets you easily access your files whenever you need them.

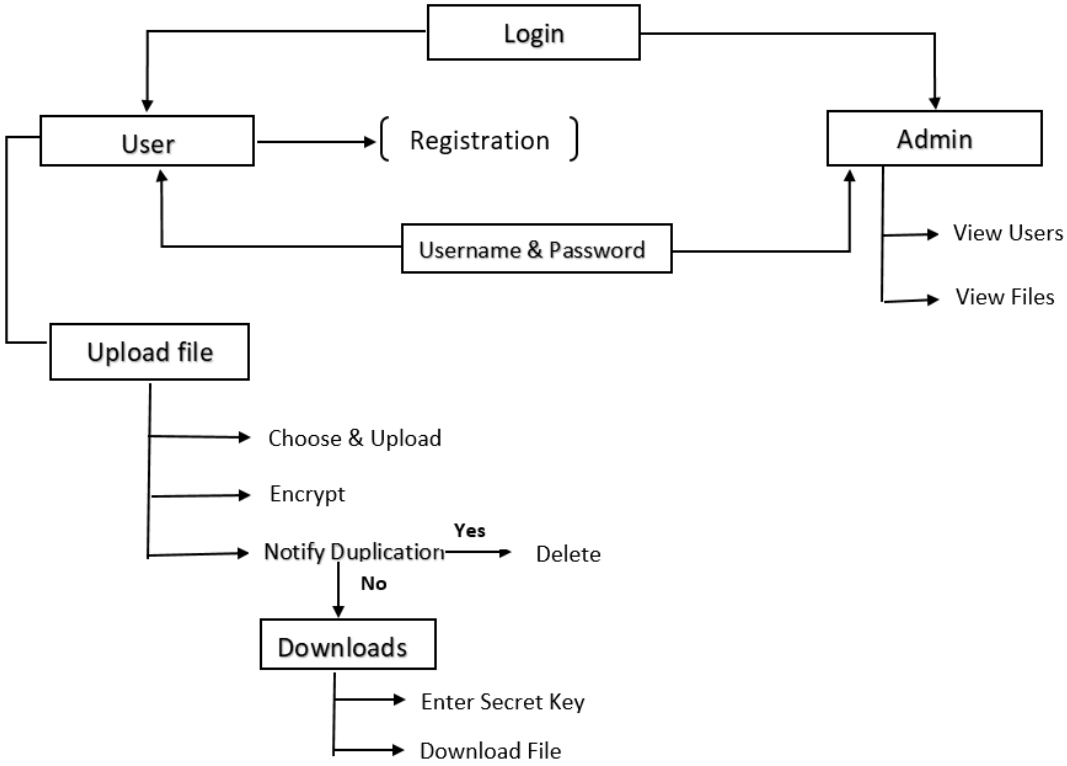

Fig 1: System Overview

Working of system

User:

- Registration:
  New User Registration is a framework that enables a user to sign in to the Campus Solutions system in order to complete a specific file sharing. Going through New User Registration, the user can either create a user ID and password, or use an existing user ID to Login to your system.

- Login:

Logging in is usually used to enter a specific page, which trespassers cannot see. Once the user is logged in, the login token may be used to track what actions the user has taken while connected to the site.

- Upload Files:

The user can upload their file into the cloud storage. In this process the cloud service provider monitors the files and detect stop the data duplication.

- Share A File:

The user can share their file to other user who givens the request for file. The user generates the private to key decrypt the file then only the other user can view the file

- Download File:
  After received a file user can download the files.

Cloud Server

- Login:
  Logging in is usually used to enter a specific page, which trespassers cannot see.
- View User Details:
  In this module, cloud server can be view the registered user details.
- View File Details:
  In this module, cloud server can view the file details which user downloads a file.

Cloud

- File Name Deduplicate:
  In this module, when user try to upload a same file then occur a filename deduplication message will popup.
- Content Level Deduplicate:

In this module, when user try to upload a different file but content as same then popup a content-level deduplication message.

6. TOOL USED

HARDWARE REQUIREMENTS:

- Processor : Intel Core i3 Processor
- Speed : 2.5 GHz
- RAM : 2GB(min)
- Hard Disk : 500MB
- Key Board : Standard Windows Keyboard
- Mouse : Two or Three Button Mouse
- Monitor : LCD

SOFTWARE REQUIREMENTS:

- Operating System : Windows7/10.
- Application Server : Tomcat6.0/7/8.X.
- Front End : Java , HTML,CSS
- Scripts : JavaScript.
- Server side Script : Java Server Pages.
- IDE : Net beans 8.2
- Back End : MYSQL 5.0/ Heidi SQL 8.1
- Database Connectivity : JDBC

7 .Algorithm

Secure Hash Algorithm (SHA)

SHA algorithm generates a hash code on the basis of file content. Cryptographic hash functions are like math tricks done on computer stuff. They produce a unique code called a "hash" for any data you give them. By checking this hash against a known one, you can see if the data has been changed. For instance, if you download a file and compare its hash with a verified one, you can tell if the file has been messed with. a key aspect of cryptographic hash functions is their collision resistance: nobody should be able to and two different input values that result in the same hash output.

Input:

1. Log in credentials.
2. File that is to be uploaded.
3. Secret pin on download.

Expected output:

1. Hash code of the file.
2. On download encrypted file will be downloaded.

8. Advantage

1. Enhanced cloud storage efficiency
2. Ensures data integrity and privacy in the cloud
3. Optimizes storage resource utilization
4. Reduces network communication bandwidth
5. Implements user revocation for added security
6. Data Consistency
7. Compliance and Security

9. Application

1. Cloud Storage Services
2. Enterprise Data Management
3. Government Agencies
4. Financial Institutions
5. Improve Search and Retrieval

10. Results

1. Home page:

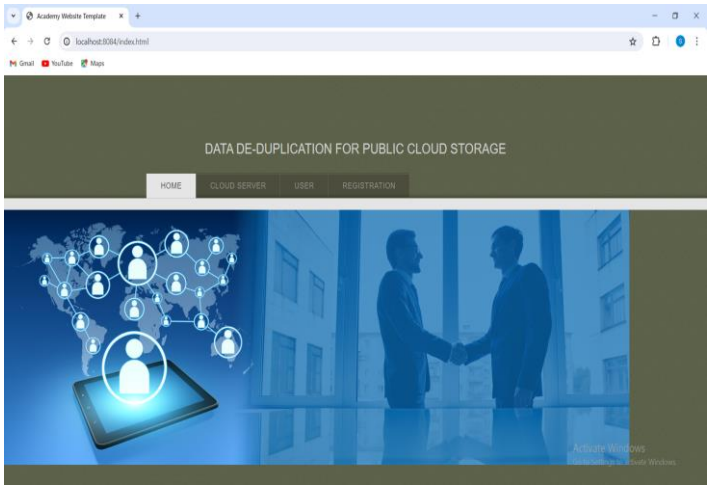

Fig 10.1:This is home page where user can view the details of the System.

2. User Registration page:

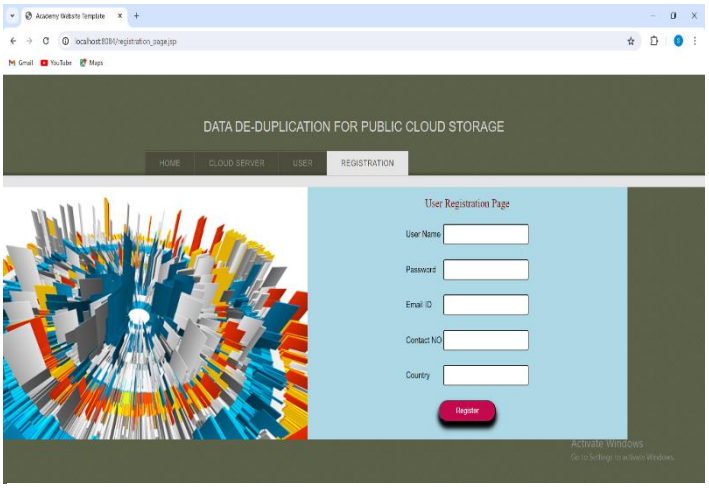

Fig 10.2: New User Registration, the user can either create a user ID and password, or use an exist user ID to Login to the system. And also fills the other

3. Registration Successful:

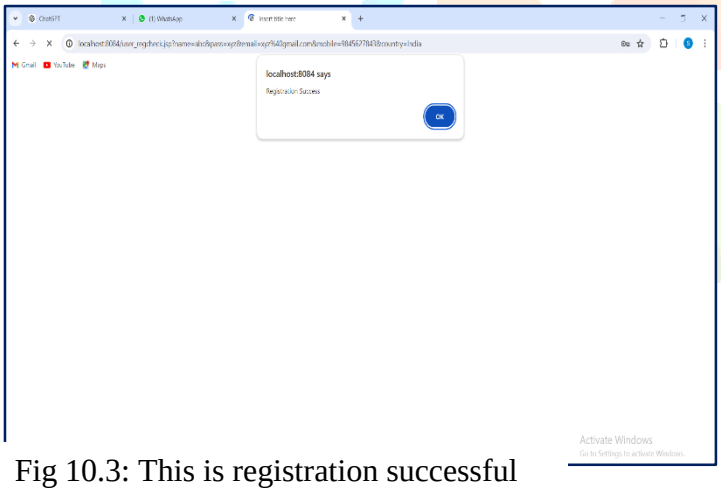

Fig 10.3: This is registration successful

4. User login page:

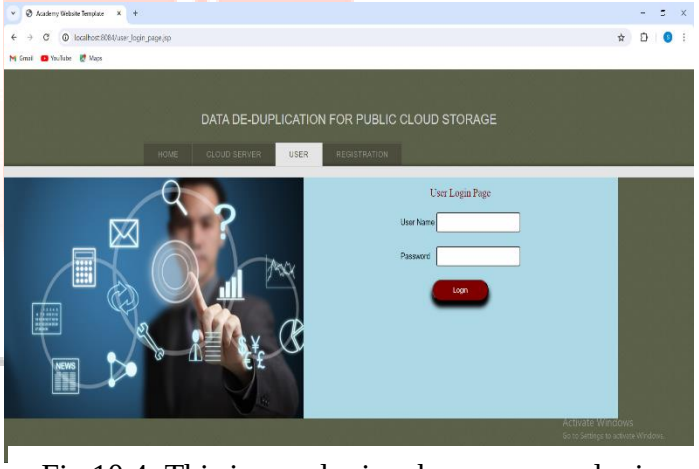

Fig 10.4: This is user login where user can login with their username and password.

5. User Login Successful:

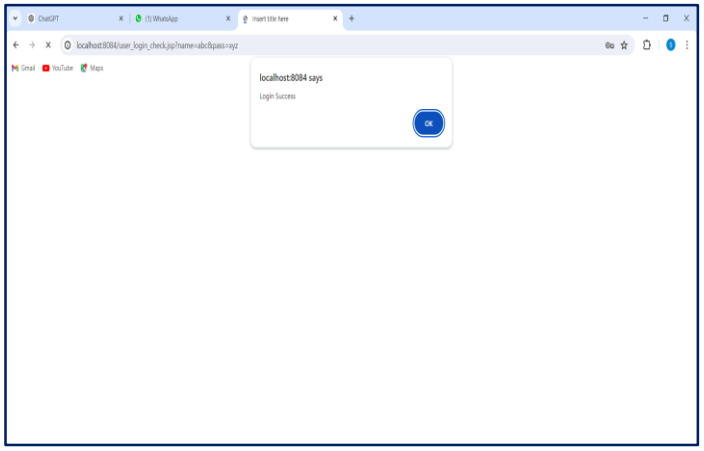

Fig 10.5: This is users successfully login screen.

6.Cloud Home Page:

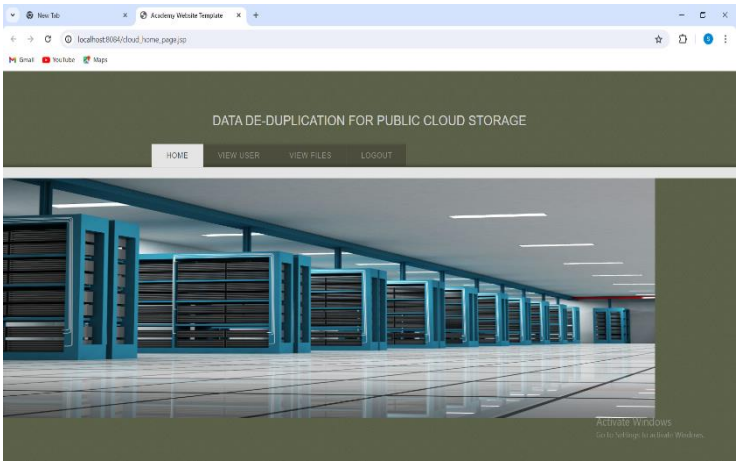

Fig 10.6: This is Admin home page where they manage the details of the System and admin response to the

7. Cloud Server Login Page:

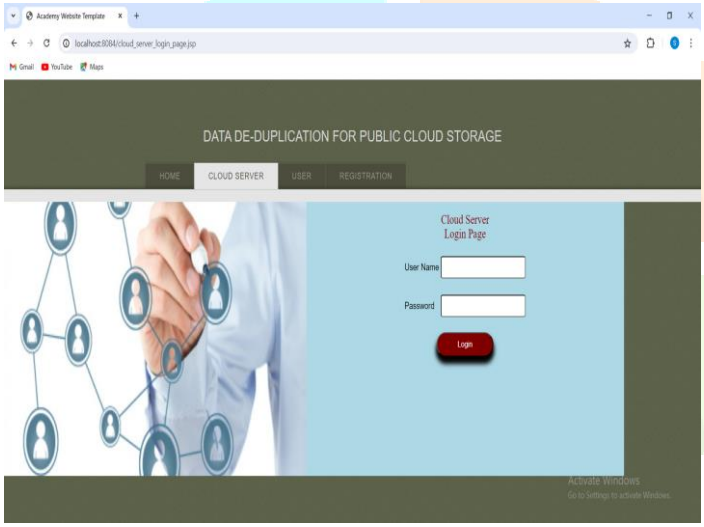

Fig 10.7: This is Admin login where admin can login with their username and password in this screen, cloud server can be view the user's registered details.

8. User Upload Page:

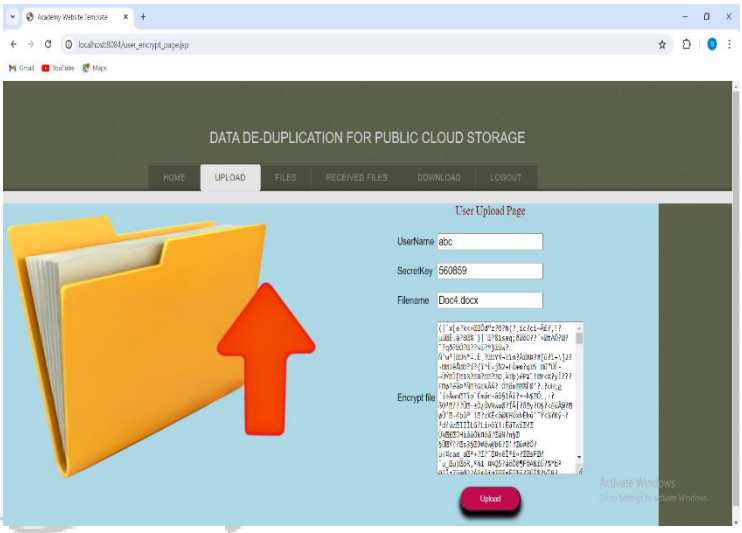

Fig 10.8: In this when the user login into the system successfully then they upload a file on cloud server with secrete key. they can also encrypt the file.

9. Upload success and no duplication screen:

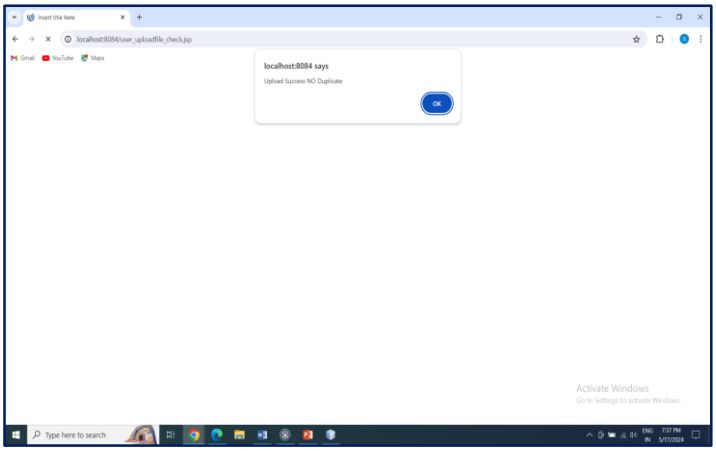

Fig 10.9: This screen shows the upload success no duplicate this message .after uploading the file .if file is duplicate or their content are exist then it shows the duplicate file or file already exist message.

11. Future Scope

The future system would offer a more flexible and robust solution for handling real-time data streams and multimedia content while ensuring security, integrity, and interoperability. By incorporating advanced technologies and security measures, it would address the evolving needs of users and organizations in an increasingly digital and interconnected world.

12. Conclusion

This paper designs a data integrity verification scheme of deduplication for cloud cipher texts, in order to achieve cloud cipher text data deduplication and data integrity verification. In the scheme, the user and TPA jointly generate the encryption key of the data block, which realizes the secondary encryption of the convergence key, making the data encryption more secure. And also data should be stored in the form of encrypted manner. It reduce the cloud storage using the algorithm. By comparing similar scheme, our scheme has higher efficiency in data deduplication and integrity verification.

13. Reference

[1] Hardik Gajera, Manik Lal Das. "Fine-grained data Duplication and proof of storage Scheme in Public Cloud Storage" (2021).
[2] MohdAman, Prashant Verma D Rajeswari . "Secure Cloud Data Duplication with Efficient Re-Encryption" (2021)
[3] Rupali Bhimrao Sirsat, Nitin R Talhar . "Deduplication in cloud storage on the basis of proof of ownership" (2020)
[4] Jiaojiao Wu, Yanping Li, Tianyin Wang. "CPDA: A Confidentiality-Preserving Deduplication Cloud Storage with Public Cloud Auditing" (2019**)**
[5] Nadesh R. K Varun G. Menon , Mahadi Abbasi . "A secure data deduplication System For integrated cloud-edge networks" (2020)
[6] Nagappan Mageshkumar, J. Swapna, R rajkumar. "Hybrid cloud storage System With enhanced multilayer cryptosystem for secure duplication in cloud" (2023)
[7] D. Viji and S. Revathy . "Hash- indexing block -based deduplication Algorithm for reducing storage in the cloud" (2023)

[8] K. Pragash , J. Jayabharathy . "A survey on DE – Duplication schemes in cloud servers for secured data analysis in various applications" (2022)
[9] Rishikesh Misal and Boominathan Perumal . "Data Deduplication for Efficient Cloud Storage and Retrieval" (2019)
[10] Ankit Rai1, Arpit Diwan, Harsh Tripathi, K. Rajkumar. "Deduplication of Data in the Cloud-based Server for Chat/File Transfer Application" (2021)
[11] V. Khanna, A. Kumaravel and A. Rama "Data deduplication On Encrypted Big data in cloud" (2019)
[12] C.S.N. Koushik, Shruti choubey . Abhishek choubey and G.R Sinha "Data deduplication for cloud storage" (2021)
[13]]James Adegboye and folahan jiboku "Data duplication Removal in Cloud Computing based on the file checksum" (2022)
[14]Komal Kshirsagar , Pallavi Patekar , Saurav kolhe , and P.N.Pathak "Secure Cloud storage with deduplication technique" (2023)
[15] D. Viji, S. Revathy "Hash-Indexing Block-Based Deduplication Algorithm for Reducing Storage in the Cloud" (2023)
[16] Hsing-bung, Sihai tang, texas Denton "A middle-ware approach to leverage the distributed data deduplication capability on HPC and Cloud storage systems" (2020)