International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 06 Issue: 04 | April - 2022 Impact Factor: 7.185 ISSN: 2582-3930

# FILE SHARING AND DATA DUPLICATION REMOVAL IN CLOUD USING FILE CHECKSUM

Gopi B[1] and Murugan R[2]

[1]Research Scholar, School of Computer Science and Information Technology, JAIN(Deemed to be University), Bangalore, India
[2]Associate Professor, School of Computer Science and Information Technology, JAIN(Deemed to be University), Bangalore, India

## Abstract:

Data duplication uses file checksum technique to identify the duplicate or redundant data rapidly and accurately. There may be the chance of inaccurate result which can be avoided by comparing the checksum of already exiting filewith newly uploaded file. The file can be stored using multiple attributes such as file name, date and time, checksum, user id, and so on. When the user uploads the new files the system will generates the checksum of the file and compare it with the check of file that has already been stored. If the match is found then it will update the old entry otherwise new entry will be created into the database.

Keywords: Database, Duplication, Entity, Data, Checksum, Redundant, User id.

## 1. INTRODUCTION

The collection of information is known as data. The data is increasing constantly in the digital universe. A study suggests that at end of 2020 each person will create 1.7 megabyte of data. It is also clear that the rate of data production per day is about 2.5 quintillion bytes of data.

The reasons behind thegrowth of multiple data are:

- Multiple backup of data or file by single person.
- Misuses of social media.

The hacking of the organisation system in 9/11 and loss of data caused by illegal activity proved that loss of data is major problem for the organization. This event forces the organization to implement data back of system in order to preserve their important data. The organizations started keeping regular backup of their data such as email, video audio etc. which increase their storage unit. While backing the data regularly, they end up with storing the duplicate data multiple times which is the misuse of storage.

So, the solution to above problem is proper implementation of data duplication removal system. The data duplication removal method stores the data or file to the system if they are not storedpreviously. If the match is found then it will update the old entry. So this system will remove the duplicate data quickly and saves the precious storage units.

## 2. RELATED WORKS

"Di Pietro, Roberto, and Alessandro Sorniotti" discussed the security concern raised by de- duplication and to address this security concern the author utilizes the idea of Proof of Ownership (POW). POW are intended to permit server to verifywhether a client possesses a file or not.

According To "Atishkathpal Matthew John Anf Gauravmakkar", data duplication removal is the method of eliminating the duplicate data from the storage devices in order to minimize the consumption of memory in storage devices. Since, the concepts were good but theirsystem cannot work as they intended due to poor management of hardware devices and not easy to use which result inthe under performance of the system.

## 3. PROBLEM FORMULATION

Many work has been done in past in order to save the storage problem that is caused by data duplication. Data duplication has been the major problem and thetechnology developed in past was not able to solve the problem due to improper management oftechnology.

Data de-duplication has been of great benefit. However, security and protection of client's data have not been guaranteed (Li et al., 2013). Furthermore, the current encryption though providing a wide scope of security is incompatible with information duplication. This type of encryption needs the diverse customers to scramble their information with their keys (Ng& Lee, 2013). Therefore, the purpose of this study is to make indistinguishable data duplicates of various customers. Furthermore, it will promptly distinguish code text, hence, making deduplicationimpossible.

## 4. SYSTEM ANALYSIS

The proposed system is Data deduplication increases the amount of unwanted data in the storage unit by storing the multiple copy of same file. Data duplication removal technique uses file checksum technique to find duplicate or redundant data quickly. The technique calculates the checksum of the file

when the file is uploaded and checks the newly calculated checksum with the checksum of file that are already store in database. If the file is already present it will modify the file else it willmake new entry of file. In this system we are going to use MD-5 hash algorithm, to detect the duplicate file. MD-5 refers to Message Digest algorithm which is 128 bit hash algorithm.

## 5. PROPOSED SYSTEM

Usually if any user wants to send any data to another user, the data which the user needs to send is first uploaded in the public cloud server and then it is delivered to the actual receiver. If another user wants to send that same data to some other user, he/she again uploads the same data once again in the cloud server. Here the same data is being uploaded twice in the server, which results in data duplication and thereby the storage space of the server is being wasted. In-order to overcome this data duplication problem occurring at the server end, we propose a new data de-duplication technique in which we introduce the concept of Convergent Encryption and proof of rights for the data with secure, more scalable and very efficient solution.

### Advantages:

- Faster file searching.
- Reduce storage space by eliminating data redundancy.
- Ease to download and upload file.

## 6. RESULTS AND DISCUSSION

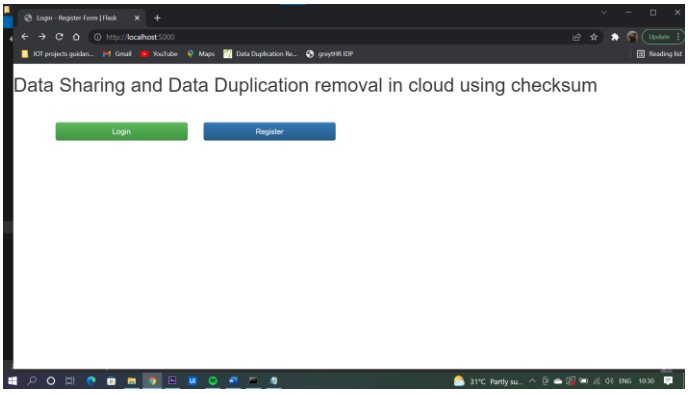
This technique focus in developing web based application that can find the redundant data quickly and easily using file checksum technique. For calculating the checksum of already existing files and new file Message Digest (MD-5) algorithm is used. MD-5 algorithm is used to calculate the checksum as well as to provide the better security and encryption to the valuable files of users. Hence, this system removes duplicate file easily and quickly by providing better security.

In this study, the concept of sanctioned data de-duplication was suggested to guarantee data security. Other that assuring on data security, the framework includes variance profits of customers in the copy check. It also showed a few new de-duplication advances aiding approved copy check in a half-breed cloud design. In this design, the copy check tokens of information are established using the private cloud server with private keys. The security evaluation displays that plans are protected as far as inside and outside case assaults designated in the projected security model. As evidence of an idea objectified a trial product of a suggested and permitted copy check plan and lead tested experiments on the prototype.It also demonstrated the accepted copy check plan brings about insignificant overhead juxtaposed with the United encryption and system exchange.

## 7. SCREENSHOTS

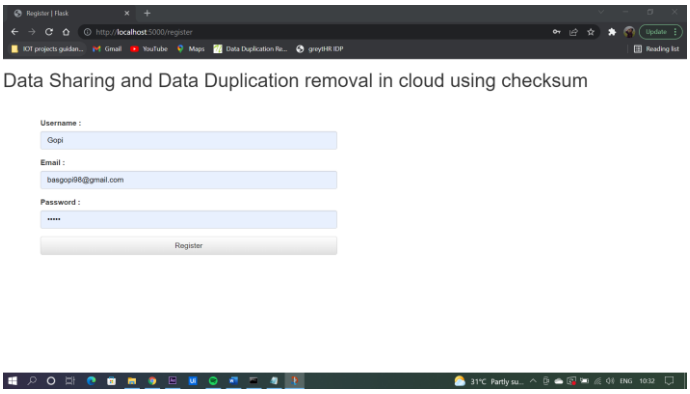
### Login form/ Register form

This is the home page where user can login or register.



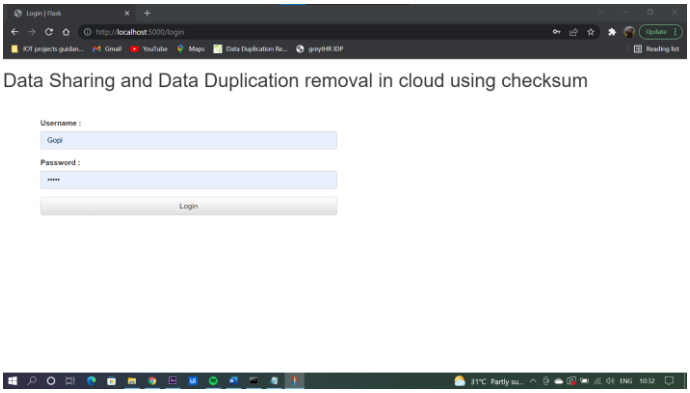

### Register form

Register form is created for the new user to register themselves to use the data duplication removal site.



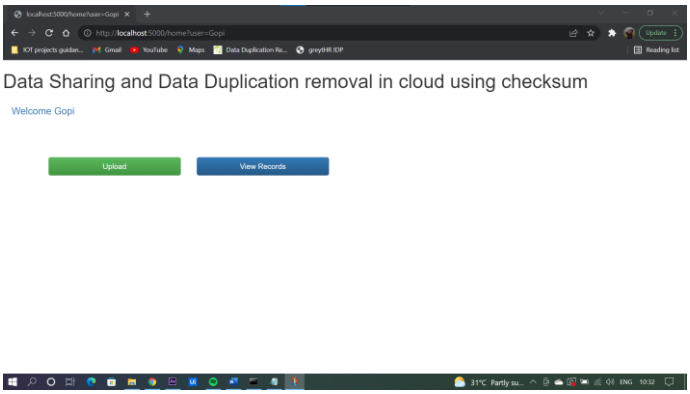

### Login Form

Login Form is used to login to the data duplication removal site.



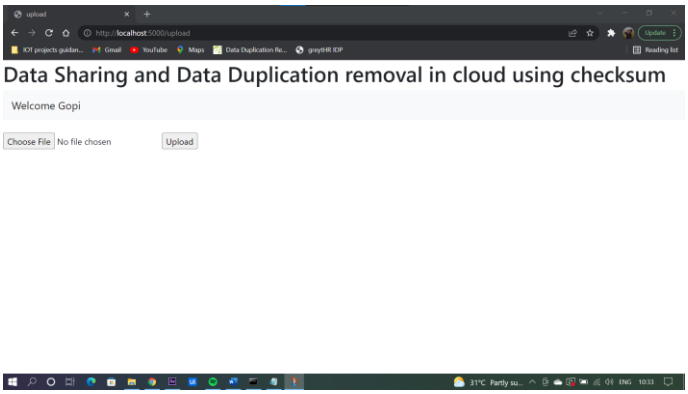

### Welcome page

This page is the main page where use can see upload the new file and the view the records that has been already stored in the database.



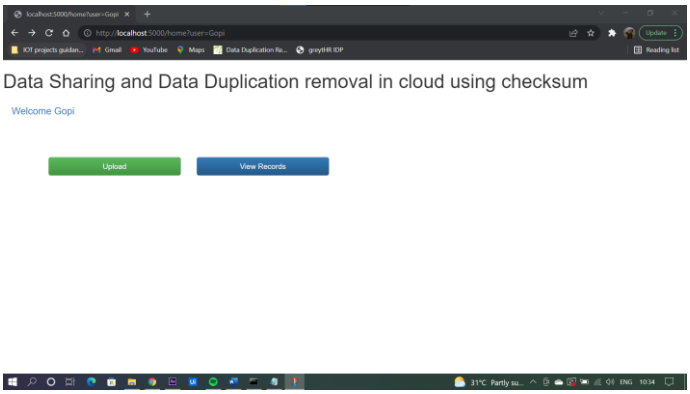
**Upload page**

Upload page is used to upload a new user to check whether the file is already stored.



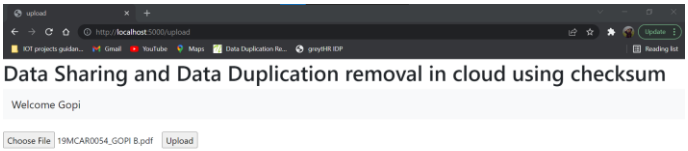

**Welcome Page (while clicking back to home)**

After successfull the page will redirect to the Welcome page.



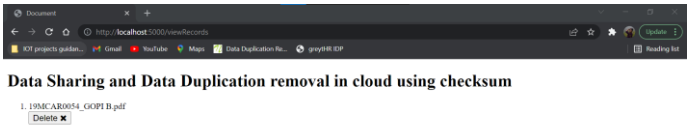

**Upload page-1**

Here you can the choose the file to upload & check.



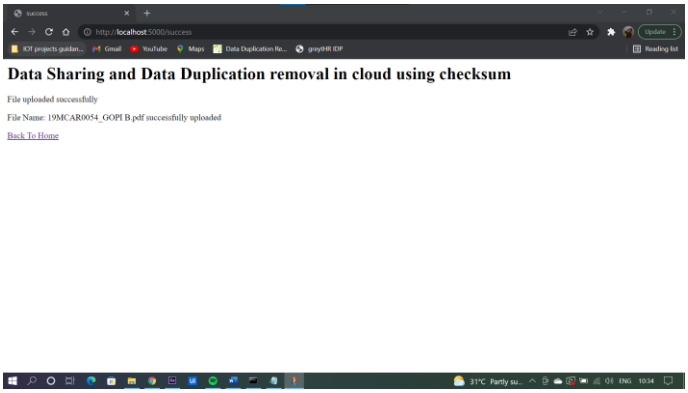

**View Records page**

This page is used to view the user stored files in the database.



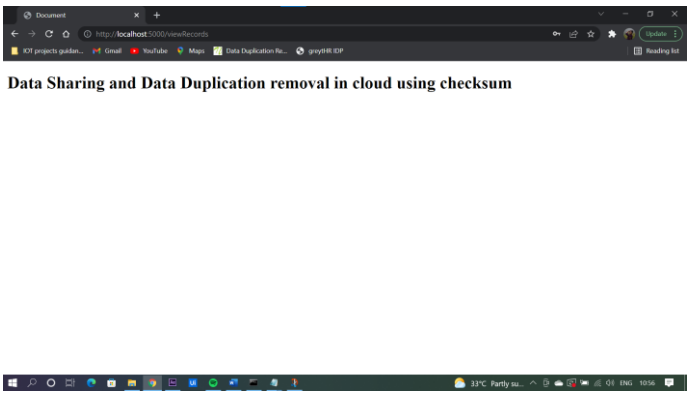

**Success Page**

This page will show the result of the uploaded file.



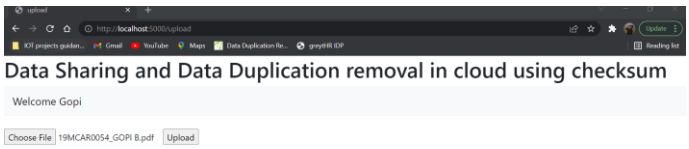

**View Records Page (After deletion)**

This page is used to see the file deleted from the database.



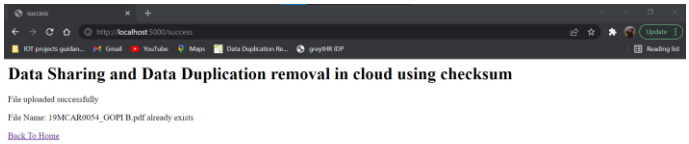
**Upload page (Trying the same file to upload again)**

We are trying to upload the same file into the database again.




**Failure page(File already exist)**

Failure page shows that the file is already exist.




## 8. CONCLUSION

In this study, the concept of sanctioned data de-duplication was suggested to guarantee data security. Other that assuring on data security, the framework includes variance profits of customers in the copy check. It also showed a few new de-duplication advances aiding approved copy check in a half-breed cloud design. In this design, the copy check tokens of information are established using the private cloud server with private keys. The security evaluation displays that plans are protected as far as inside and outside case assaults designated in the projected security model. As evidence of an idea objectified a trial product of a suggested and permitted copy check plan and lead tested experiments on the prototype. It also demonstrated the accepted copy check plan brings about insignificant overhead juxtaposed with the United encryption and system exchange.

## 9. ACKNOWLEDGEMENT

I should convey my real tendency and obligation to Dr M N Nachappa and Dr. Murugan R for undertaking facilitators for their effective steerage and consistent inspirations all through my assessment work. Their ideal bearing, absolute co-action and second discernment have made my work gainful.

## 10.SCREENSHOTS

[1]. Di Pietro, Roberto and Alessandro Sorniotti, "Proof of ownership for de-duplication systems: A secure, scalable, and efficient solution", Computer Communications, 15 May 2016.
[2]. M. Bellare,S. Keelveedhi, and T. Ristenpart, "Dupless**:** Server aided encryption for deduplicated storage", USENIX Security Symposium, 2013.
[3]. Harnik, Danny, Alexandra Shulman-Peleg and Benny Pinkas, "Side channels in cloud services, the case of deduplication in cloud storage ", IEEE Security & Privacy 8, 2014.
[4]. Atishkathpal, Matthew John and Gauravmakkar, "Distributed Duplicate Detection in Post-Process Data De-duplication", Conference: HiPC , 2011

[5]. X. Zhao, Y. Zhang, Y. Wu, K. Chen, J. Jiang, K. Li, "Liquid: A Scalable Deduplication File System for Virtual Machine Images",IEEE Transactions on Parallel and Distributed Systems, January 2013.
[6]. Stephen J. Bigelow, "Data Deduplication Explained: http://searchgate.org", February, 2018
[7]. http://www.computerweekly.com/report/ Data-duplication-technology-review
[8]. https://nevonprojects.com