Data Duplication Removal Using File Checksum

Raghavendra B[1], Sudip Dhakal[2], Simon Karki [3]

[1]Assistant Professor, Department of Computer Science and Engineering, Nagarjuna College of Engineering and Technology, Bangalore, India

[2,3]B.E. Student, Department of Information Science and Engineering, Nagarjuna College of Engineering and Technology, 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 file with 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 the growth 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.

As the data is increasing constantly storing them and managing them becomes more difficult. More data requires more storage and more storage require more cost as we have to increase the hardware or storage unit. Only increasing the storage unit is not the solution because we are not sure that how much storage unit we have to add. Adding more number of storage units makes system bulk and more costly.

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 stored previously. 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. LITERATURE SERVEY

"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 verify whether a client possesses a file or not.

According To "Atishkathpal Matthew John Anf Gauravmakkar" as referenced by [4], 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 their system cannot work as they intended due to poor management of hardware devices and not easy to use which result in the under performance of the system.

3. EXISTING SYSTEM

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 the technology developed in past was not able to solve the problem due to improper management of technology.

Drawbacks of the existing system

- More processing time.
- Chance of false result.
- Not user friendly.
- System maintenance is difficult.

4. PROPOSED SYSTEM

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 will make 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.

Advantages:

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

5. SYSTEM ARCHITECTURE

The system consists of 2 modules as follows:

I. Admin
II. User

I. Admin:

Admin is the person who has full access to the system. Admin can perform different task such as managing the security issue, maintaining system server, giving different access to users and soon.

a) Login: To get the control over the system the admin should first login to the system using his valid credentials.
b) View User: Once the admin get logged in he can view the entire user.
c) Block / Unblock User: Based on the activities of user the admin can block and unblock the users.
d) View Files: Files uploaded by the users can be viewed by the admin but he cannot modify it.

II. User:

a) Registration / Login: In any system registration is the primary concerned. The user needs to register to the system in order

to get access to that system. In registration phase user need to provide some information such as name, email, password, etc. Once the registration is complete the user can login into the system using the email and password which he/she has provided during registration time.

b) Upload a File: This section helps user to upload their files to system which then can share with their friends.

c) Download a File: Once the file is uploaded and if user wants to download he/she can download file by clicking on download file option.

d) Change Password: User can change his/her password any time they want.

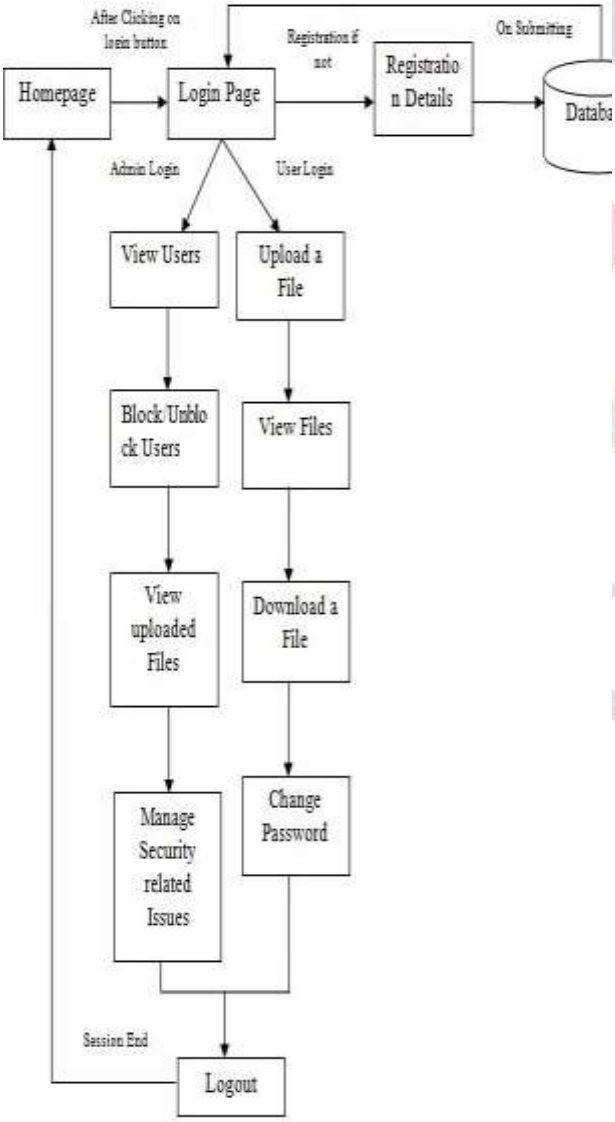

Fig: System Architecture

6. System Design

a) Use Case Diagram

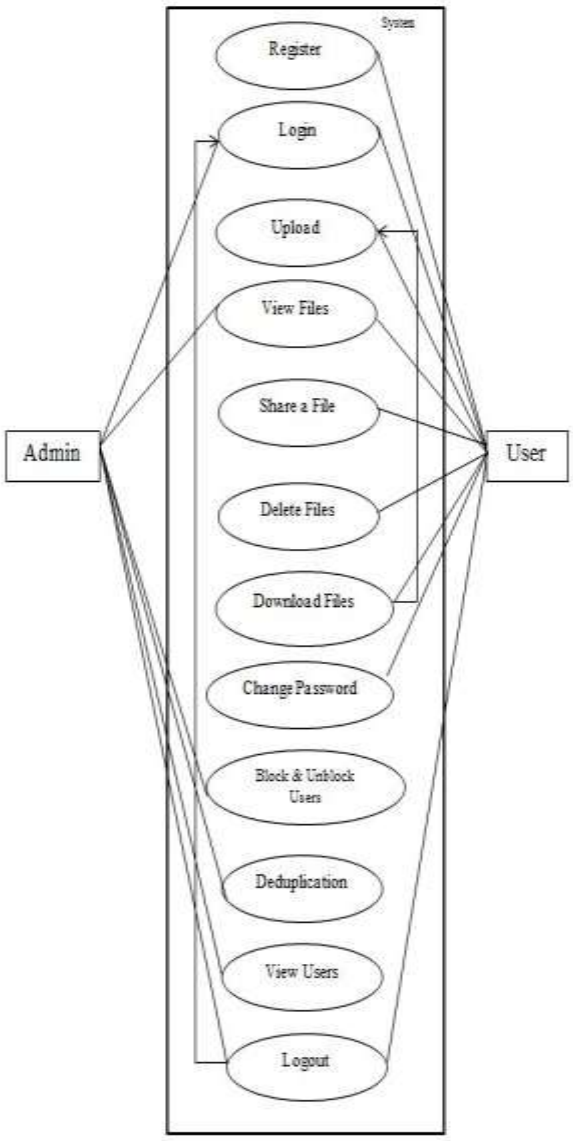

Fig: Use case for user and admin

b) Process Chart

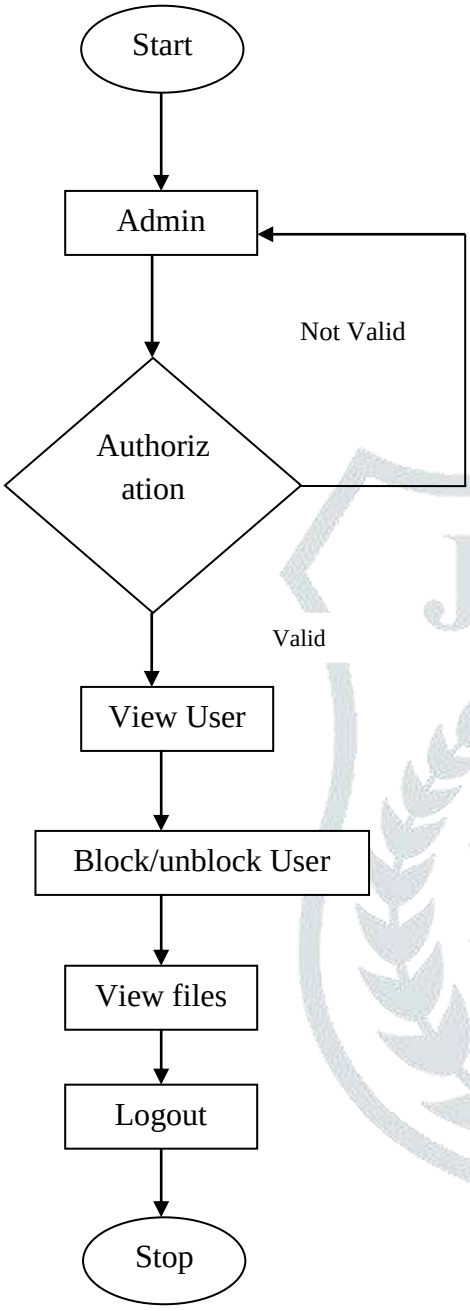

Fig: Process Chart for Admin

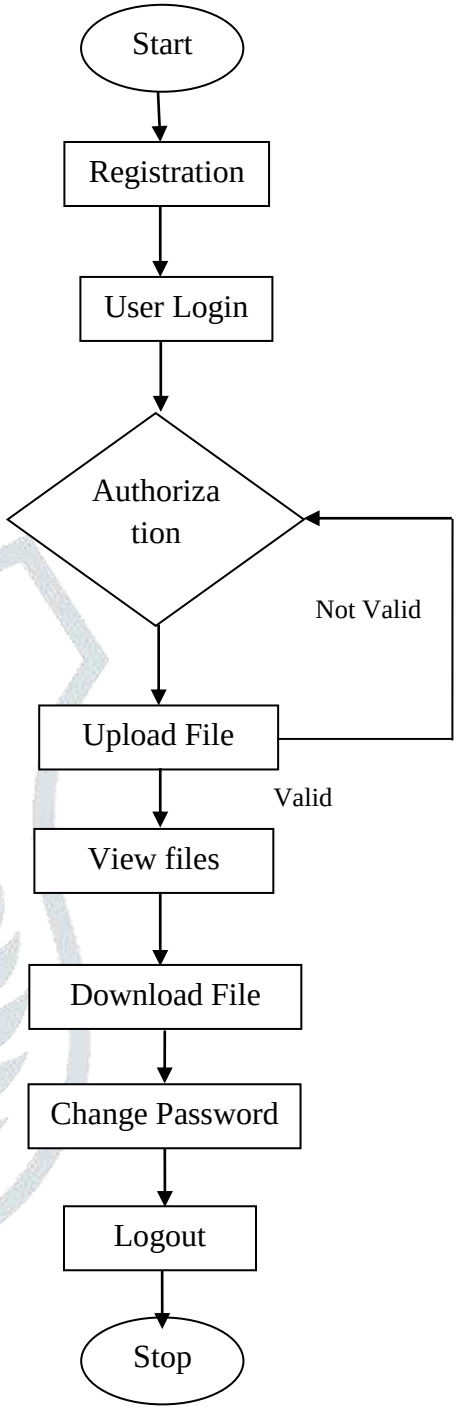

Fig: Process Chart for User

c) Sequence Diagram

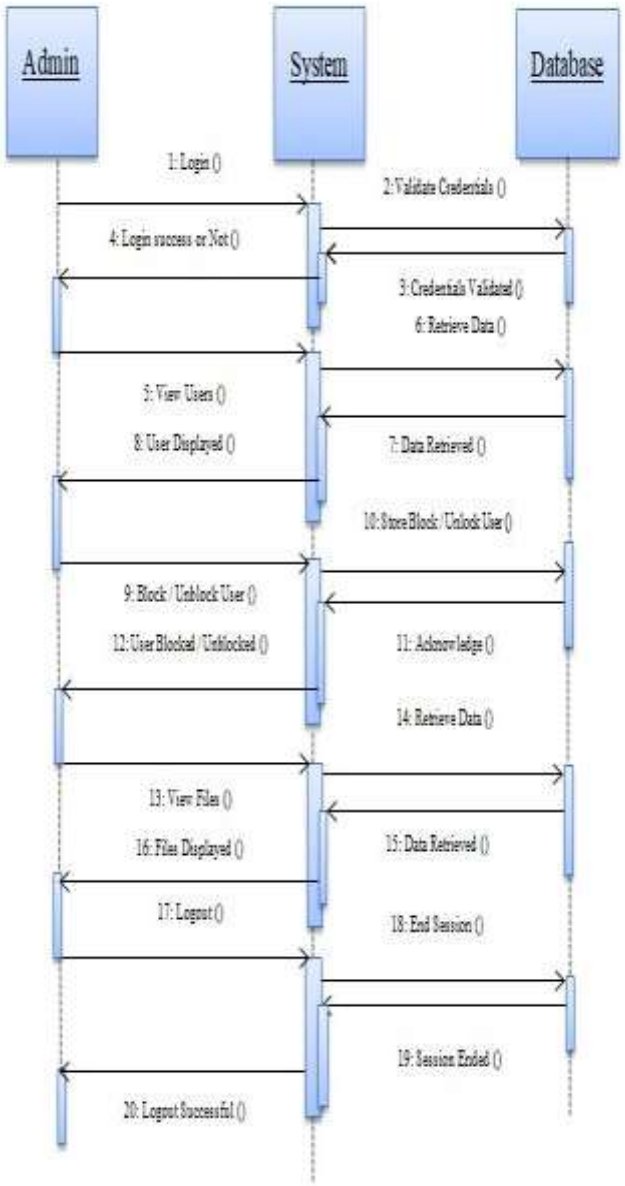

Fig: Sequence Diagram for Admin

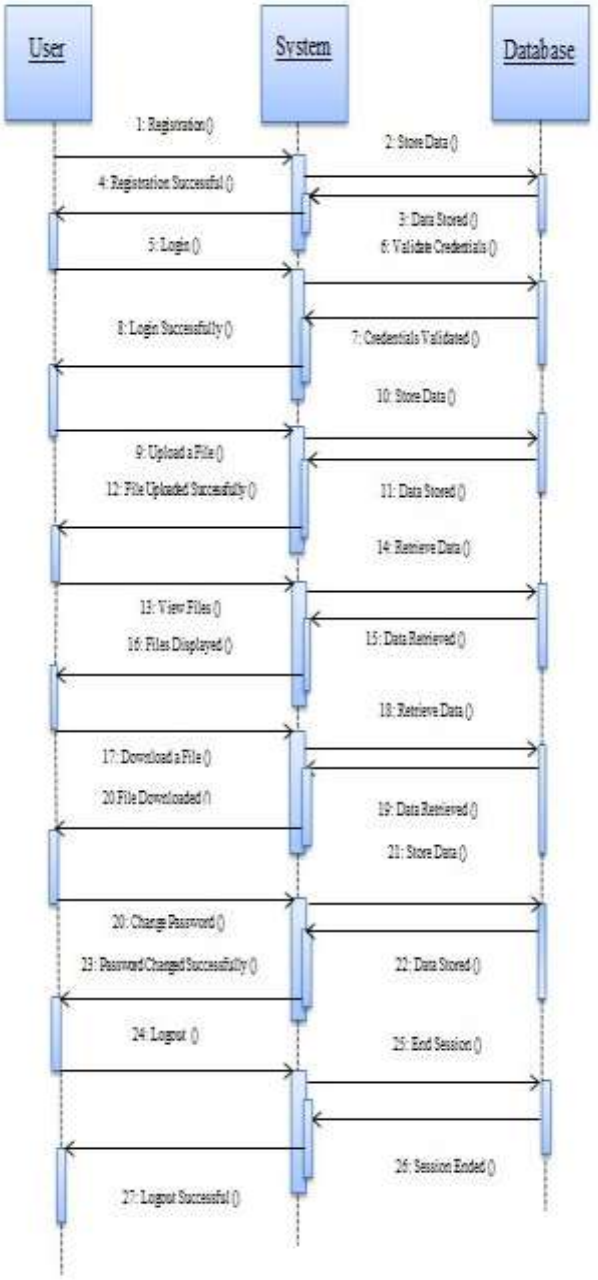

Fig: Sequence Diagram for User

7. SYSTEM REQUIREMENT

A. Hardware Requirements:

- Minimum 350MB Hard Disk space
- Central Processing Unit (CPU)
- Graphic Display
- Keyboard

- Mouse
- Internet Connection

B. Software Requirements:

- Windows 7 or higher
- PHP
- Google Chrome Browser
- MYSQL
- XAMPP Server
- Sublime Text

8. Result

From the graph below, we can conclude that the size of the file remain the same before and after the upload but the memory space is saved by removing the duplicate files from database.

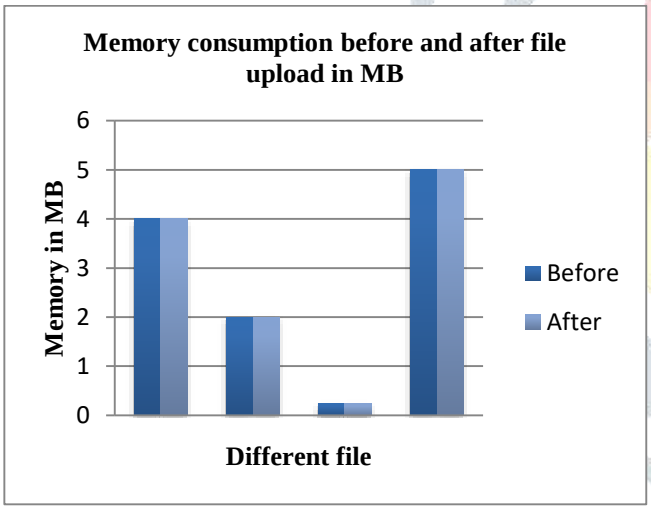

Fig: Memory consumption before and after file upload in MB

For any system the accuracy is the most important aspect. Accuracy refers to the state without any error. From the graph it is cleared that our system is more accurate and takes less time to detect duplicate file.

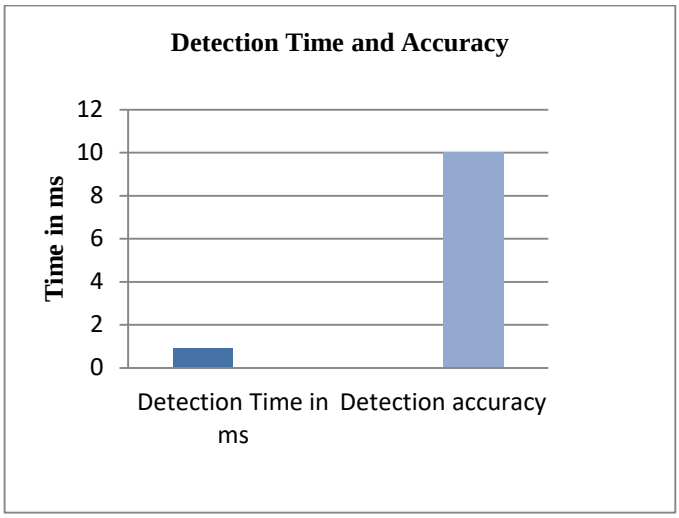

Fig: Detection Time and Accuracy

The third graph indicates the performance of different hash function. When comparing between MD2, MD4 and MD5, MD2 has the lowest performance and MD4 has the highest performance but MD5 has highest security level. So, the MD5 is considered as best hash algorithm because of its security reason.

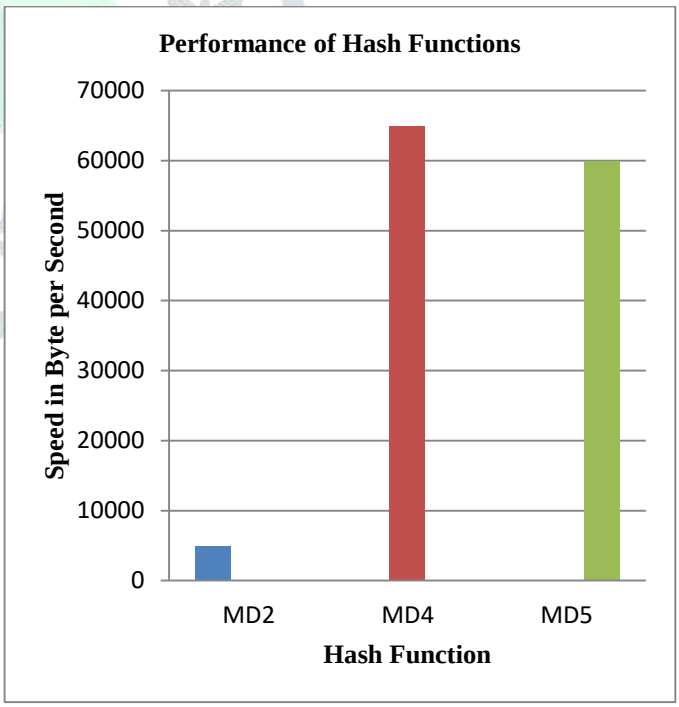

Fig: Performance of Hash Functions

9. CONCLUSION

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.

REFERENCES

[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