

# Cloud Data Duplication Identification using Hash Code

Sakshi Tadge<sup>1</sup>, Rutika Khainar<sup>2</sup>, Neha Jadhav<sup>3</sup>, Shraddha Banne<sup>4</sup>

Student, Dept. of Computer Engg, G.G.S.F, Nashik<sup>1</sup>

Student, Dept. of Computer Engg, G.G.S.F, Nashik<sup>2</sup>

Student, Dept. of Computer Engg, G.G.S.F, Nashik<sup>3</sup>

Assistant Professor, Dept. of Computer Engg., G.G.S.F, Nashik<sup>4</sup>

Department of computer

Guru Gobind Singh College of engineering and research center Nashik

## Abstract

Storage and security for all data is very important for cloud computing. Securing and privacy preserving of data is of high priority when it comes to cloud storage. De-duplication is a process that eliminates redundant copies of data and reduces storage overhead. Storage and data transfer costs will be reduced by their cloud providers by storing a unique copy of

duplicate data. The most often important and popular cloud service is data storage. Intrusion detection and prevention are performed manually by network operators in the existing system. When the utilization of such private cloud storage increases, there will be an increase in the connection problem, performance storage demand, privacy security, data integrity. The key idea behind the develop a dynamic load balancing algorithm based

on de-duplication to balance the load across the storage expansion of private cloud storage. For maintaining privacy, security data integrity we will use SHA algorithm. SHA algorithm also used to avoid duplication. The result also show the comparison of various hashing algorithms and show SHA is more secure and take less time than other hashing algorithms. The base scheme without de-duplication checking scheme take more storage space and time also increases.

**Keywords:** *Security, Cryptography, Hash, De-duplication, SHA, Cloud computing, cloud solution, load balancing, secure de-duplication.*

## I. INTRODUCTION

Cloud computing is very important in the Information Technology. The storage services provided to users are through internet. Cloud computing is one of the emerging technology, which helped several organizations to save money and time adding convenience to the end users. To solve connection problem we can implement offline storage sync mechanism. To improve performance, load balancing is important task for doing operations in cloud and de-duplication also. As cloud computing has been growing and many clients all over the world are demanding more services and better results and load balancing is necessary in the cloud computing. customers on their demand and build up the overall performance of cloud. The key idea behind this project is to develop a offline store sync mechanism, dynamic load balancing algorithm based on de-duplication to balance the load across the storage nodes

during the expansion of private cloud storage. The de-duplication process requires comparison of data 'chunks' (also known as 'byte patterns') which are unique, contiguous blocks of data. These chunks are identified and stored during a process of analysis, and compared to other chunks within existing data. Whenever a match occurs, the redundant chunk is replaced with a small reference that points to the stored chunk. Given that the same byte pattern may occur dozens, hundreds, or even thousands of times (the match

frequency is dependent on the chunk size), the amount of data that must be stored or transferred can be greatly reduced. high security is the only solution to retain strong trust relationship between the cloud users and cloud service providers. Thus to overcome the security threats and this paper proposes multiple cloud storage. Thus the common forms of data storage such as files and databases of a specific user is split and stored in the various cloud storages.

## II. OBJECTIVES

- Reduces the spending for extra disk or tape .
- Reduces storage requirements to an extent.
- Reduces the required bandwidth for backup process in a network .
- Speeds up the reserve process and recovery process .
- Saves time, storage and money .
- Reduces the volume of data that is sent over LAN.

## III. PURPOSE

- This application helps in easy maintenance of data on the cloud platform so that no duplicate files are Saved in the cloud.
- De-duplication aids in saving the Storage space.
- Save time and increase productivity.
- Data can help remove redundant data from your data stores, and as a result, it can reduce the amount of disk space required to store all that data.
- Unlimited storage
- Security

## IV. THEORETICAL BACKGROUND

As huge amounts of data are available and stored for various applications, then the problem arises in data storage as well as data duplication. The stored data is accessed in the least amount of time, despite the fact that there exists duplicates of data in the storage. This becomes easier by cloud storage devices due to easy access of data and their duplicates. Memory reutilization makes the storage efficient. So the data can be backed up in the cloud and restored easily when there is crash preventing inconsistency in data. Data De-duplication helps storage administrators reduce costs that are associated with duplicated data. Large datasets often have a lot of duplication, which increases the costs of storing the data. For example:

User file shares may have many copies of the same or similar files.

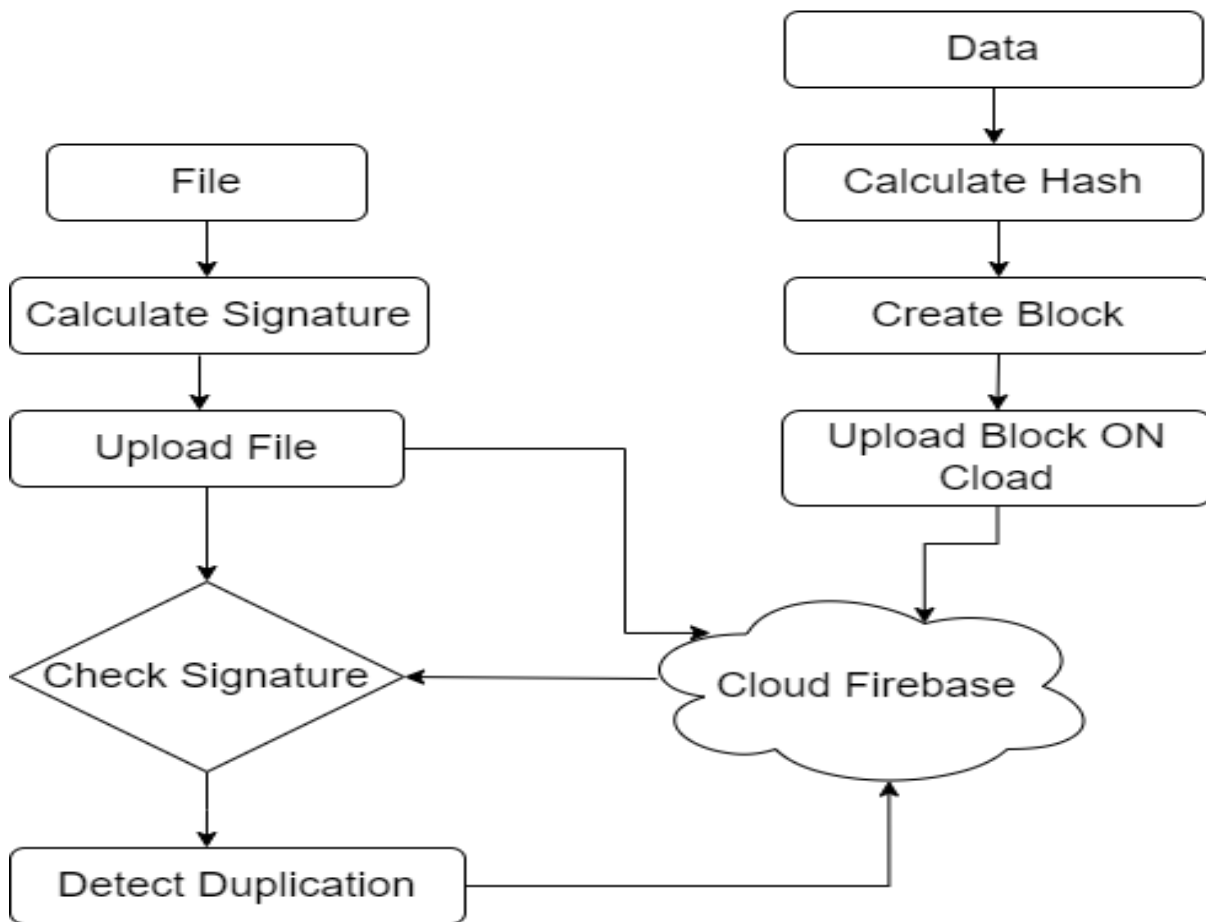
Virtualization guests might be almost identical from VM-to-VM.

Backup snapshots might have minor differences from day to day.

The space savings that you can gain from Data De-duplication depend on the dataset or workload on the volume. Datasets that have high duplication could see optimization rates of up to 95%, or a 20x reduction in storage utilization.

## V. DESIGN OF SYSTEM

These system propose the architecture to remove a load on cloud base servers and which will avoid data duplications using the some methodologies and algorithms. This system is basically performing Hash Code detection techniques which is used for avoiding multiple storage of the files on the Cloud Server.



## VI. IMPLEMENTAION

- Algorithm of Proposed System:-

Step 1: Start

Step 2: Login to the User

Step 3: Category & Subcategory are added

Step 4: Upload file

Step 5: Calculate Hash

Step 6: Hash code is generated

Step 7: Comparison Of Hash Value to Uploaded files.

If :

Hash value is same then go to step 8

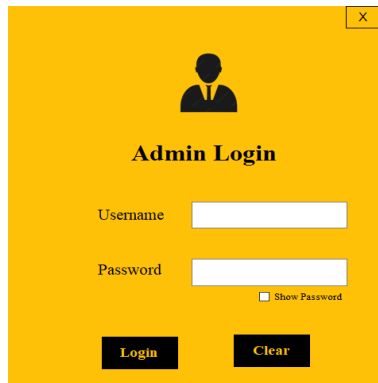
Else :

Hash value is unique then go to step 9

Step 8: Duplicate file message are arrive(detect duplication)

Step 9: File Upload.

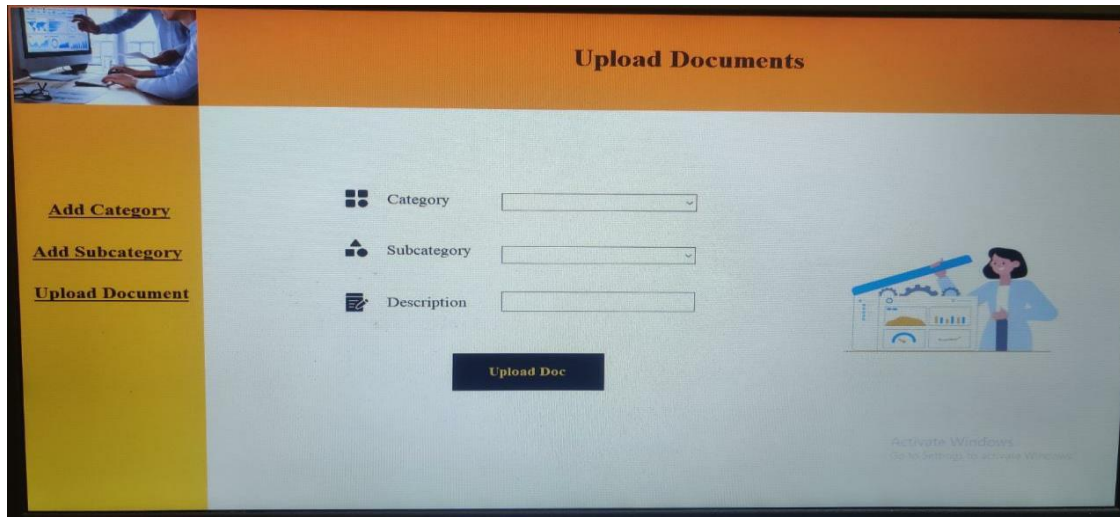
- Working Module Snapshot



A screenshot of a web application window titled "Admin Login". The window has a yellow background. At the top center is a black silhouette of a person. Below it, the text "Admin Login" is displayed. There are two input fields: "Username" and "Password". The "Password" field has a small checkbox labeled "Show Password" to its right. At the bottom, there are two buttons: "Login" and "Clear".

Activate Windows  
Go to Settings to activate Windc

### Admin login



### Upload Documents

## VII. CONCLUSION

By our Study we have come to a Conclusion that for our project we are expected to develop an application for the to avoid cloud disaster such as problems in connection, performance, privacy security, data management .To overcome this we are using mechanism like offline store & sync. load balancing, de-duplication, . De-duplication aids in saving the Storage space.

This application helps in easy maintenance of data on the cloud platform so that no duplicate files are .Saved in the Cloud. With the evolution of Cloud computing, storage resources of commodity machines can be efficiently utilized. This allows every organization to build its own private cloud for a variety of purposes. In order to .Better utilize the limited storage available in a private cloud, a suitable approach for optimization has to be used.

### VIII. REFERENCES

- [1]. M. K. Yoon, "A constant-time chunking algorithm for packet-level deduplication," *ICT Express*, vol. 5, no. 2, pp. 131–135, 2019, doi: 10.1016/j.ict.2018.05.005.
- [2]. M. W. Storer, K. Greenan, D. D. E. Long, and E. L. Miller, "Secure data deduplication," *Proc. ACM Conf. Comput. Commun. Secur.*, no. October, pp. 1–10, 2008, doi: 10.1145/1456469.1456471.
- [3]. T. R. Burramukku, "Available Online through A Comparative Study n Data Deduplication Techniques In Cloud Coden : IJPTFI Research Article," no. October, 2018.
- [4]. C. Vinothini, P. Balasubramanie, M. Jayanthi, J. Priya, and P. Anitha, "Swarm Intelligence Algorithms in cloud Computing : A Survey," vol. 29, no. 7, pp. 105698–105706, 2020.
- [5]. C. Vinothini, P. Balasubramanie, and K. S. Arvind, "Hybrid Fuzzy C Means Clustering ( Fcm ) And Improved Bat Optimization Algorithm For Multi-Servers Load Balancing In The Cloud Environment Department of Computer Science & Engineering , MVJ College of Engineering College ," vol. 29, no. 12, pp. 841–851, 2020.
- [6]. X. L. Liu, R. K. Sheu, S. M. Yuan, and Y. N. Wang, "A file-deduplicated private cloud storage service with CDMI standard," *Comput. Stand. Interfaces*, vol. 44, pp. 18–27, 2016, doi: 10.1016/j.csi.2015.09.010.
- [7]. X. Xu and Q. Tu, "Data Deduplication Mechanism for Cloud Storage Systems," *Proc. - 2015 Int. Conf. Cyber-Enabled Distrib. Comput. Knowl. Discov. CyberC 2015*, pp. 286–294, 2015, doi: 10.1109/CyberC.2015.71.
- [8]. J. Stanek, A. Sorniotti, E. Androulaki, and L. Kencl, "A secure data deduplication scheme for cloud storage," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 8437, no. ii, pp. 99–118, 2014, doi: 10.1007/978-3-662-45472-5\_8.
- [9]. S. Halevi, D. Harnik, B. Pinkas, and A. Shulman-Peleg, "Proofs of ownership in remote storage systems," *Proc. ACM Conf. Comput. Commun. Secur.*, pp. 491–500, 2011, doi: 10.1145/2046707.2046765.
- [10]. J. Xu, E. C. Chang, and J. Zhou, "Weak leakage-resilient client-side deduplication of encrypted data in cloud storage," *ASIA CCS 2013 - Proc. 8th ACM SIGSAC Symp. Information, Comput. Commun. Secur.*, pp. 195–206, 2013, doi: 10.1145/2484313.2484340.
- [11]. T. Y. Youn, N. S. Jho, K. H. Rhee, and S. U. Shin, "Authorized Client-Side Deduplication Using CP-ABE in Cloud Storage," *Wirel. Commun. Mob. Comput.*, vol. 2019, 2019, doi: 10.1155/2019/7840917.