iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 9 **Issue:** VI **Month of publication:** June 2021

DOI: https://doi.org/10.22214/ijraset.2021.35861

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 9 Issue VI Jun 2021- Available at www.ijraset.com

A Novel Technique to Remove Duplicacy from Cloud

Neha Verma[1], K Nawazia[2], Harendra Pratap Singh[3], Kumar Divyam[4], Jyoti Arya[5]
[1, 2, 3, 4, 5]*Department of Computer Science & Engineering, Inderprastha Engineering College, Ghaziabad, Uttar Pradesh, India*

Abstract: Cloud Computing is one of the most emerging technology, which helped several organizations to save time, resources and money by adding convenience to the end users. This project deals with removing duplicate data on cloud to save storage. The continuous development of the information technology and the requirement to make the data available for anytime and anywhere purpose, makes it necessary to remove duplicate files form the storage area to maximise storage. After uploading a particular file, the system crosschecks the data, and it will keep it inside the user's bucket on cloud if the data is new. This proposed method is more secure and utilizes less resources of cloud. Otherwise, the system will not keep the data if it's already there. The main concept of this technology is to reduce duplicate data as much as possible using Hashing technique.
Keywords: De-Duplication, Cloud Computing, Hashing.

I. INTRODUCTION

Current period is cloud computing era. These days cloud computing has wider range of scope in data sharing. Day-to-day use of cloud is increasing. Users use the resources for sharing information. But users have to pay as per the use of resources of cloud. User can upload huge amount of data on cloud and share data to millions of users. But the problem in cloud computing is, everyday data is uploaded on the cloud, so it increases similar data in cloud. Therefore, it becomes necessary to reduce the size of similar data in cloud using the data Deduplication method. This method's main aim is to remove duplicate data from cloud. It can also help to save storage space.

Information deduplication stores just one-of-akind occurrence of the information sort on the plate or tape. For illustration, an administration may contain a few instances of same information document, putting away every one of these occasions would require a lot of storage room. This issue can be penetrated by utilizing Data Deduplication procedure.

The proposed method helps to store data with a minimum number of duplications as possible so that the user gets maximum storage space, to reduce the complexity that happens with redundant files, as a result, the user doesn't need to delete copied files manually. Also helps to easily upload files with a connection of active internet, compares the newly uploaded chunk of data with the existing chunk of data and determines the decision and also increases the process performance so that the window gets affected and helps to run on a cloud platform in an easier manner.

II. LITERATURE SURVEY

So many researches have been done to secure duplication check of data on cloud.

One of the proposed methods was, an effective user authentication using fingerprint feature extraction, image-based authentication during file upload/ download, removing repetition of data in cloud server and implemented through multiple cloud storage. In this proposed system the intrusion detection and prevention are carried out automatically by defining rules for the major attacks and alerts the system automatically. To ensure data secrecy the data is stored in an encrypted type using Advanced Encryption Standard (AES) algorithm.

Another method was, Data Deduplication Using Hybrid Cloud. This proposed method removed the duplicate data but in which user have assigned some rights according to that duplication check & each user have their unique token. Cloud Deduplication is achieved using the hybrid cloud architecture. Content Level Deduplication as well as File Level Deduplication of file data is checked over the cloud.

One of the other techniques includes, removing Duplicate Data in Cloud Environment using Secure Inverted Index Method. In this method, they were removing duplicate data to save storage space and increase storage speed of network. Hear they have applied inverted index technique and tf-idf to identify the duplicate data in cloud environment. Encryption algorithm plays a main part in information security system. Security is achieved throw encryption and decryption on data.

Another method includes, Data Finding, Sharing and Duplication Removal in the Cloud Using File Checksum Algorithm. This project proposes a method for removing and detecting duplicates using file checksum algorithms by calculating the digest of files.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 9 Issue VI Jun 2021- Available at www.ijraset.com

III. SYSTEM ANALYSIS

System analysis is the term used to describe the activity of gathering and examining facts in respect of existing operation of the solution of the situation prevailing so that an effective computerized system may be designed and implemented to provide feasibility. It also identifies the problems and using that information suggests improvement to the system.

System analysis is the pruning of the entire system by studying the numerous operations performed and the relationship with the system and requirement of its successor. A system can be defined as an in-order grouping of unconstrained components associated together according to a plan to achieve a specific task.

System analysis may be contemplated as an linkage between the actual problem and computer. Before a computer can perform, it is important for investigations, are called system analyst. System analysis also enhances system design which is an activity concerned with the design of a computerized application based on the facts revealed during the analysis step. The same individual who knows as the system analyst convey out both activities. In feasibility study in most of the cases, project is being operated by a problem in the business.

IV. PROPOSED SYSTEM

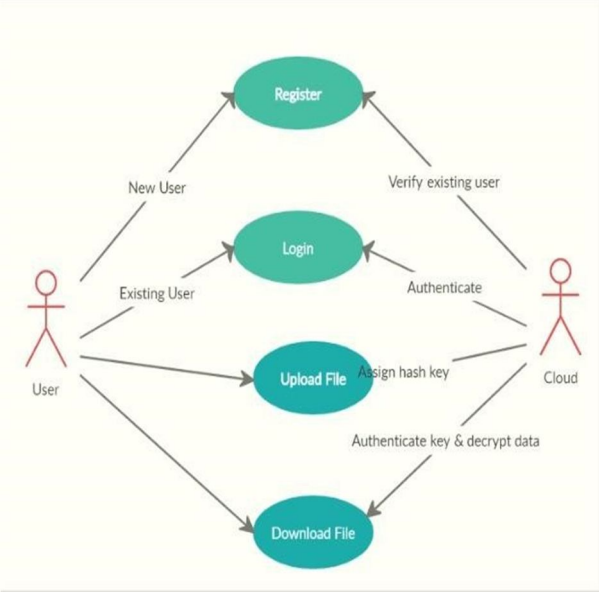
In the proposed system we are doing duplication check in substantiated way. If the user is a new user, then he has to register himself and all the login credentials will be stored in the database. After this, the user can perform sign in and if the user is an authorised user, then he gets the access to upload his file on the cloud. After sign in, the user can upload any file from his system and after uploading a unique hash value will be assigned to the file using MD5 hashing algorithm and the file will be stored inside the bucket. If the user uploads the same file again, then that file will not be uploaded because of the matching hash value. In this way, we are trying to save cloud storage space.

V. DETAILED DESCRIPTION

1) Step 1: Firstly, build a website which will act as an interface between the user and the cloud environment. A website is a collection of HTML documents that can be called up as individual webpages via one URL on the web with a client such as a browser. The website consists of manageable home section, about section and contact section with sign in and sign-up options.
2) Step 2: If the end user is a new user, then he has to register himself first to upload any file on the cloud. On doing sign up, he moves on to next page where he can register himself. All the registration credentials will be stored on to the phpMyAdmin database.
3) Step 3: After sign up, the user can sign in using the same login credentials. After successful sign in, the user can access the next page where he can find the option of uploading files over cloud.
4) Step 4: The authorized user can select a file from his machine to upload it over cloud.
5) Step 5: Create an instance on Google Cloud Platform and host all the web pages over cloud. Also, install php for running the php files, phpMyAdmin for the database, composer for bucket creation and Nginx server.
6) Step 6: For that instance, a bucket will be created where all the files will be uploaded or stored.
7) Step 7: When the user uploads any file, the file gets into the bucket and a unique hash value will be generated of 32 digits for that file.
8) Step 8: If the user again uploads the same file of the same name, then the unique value for that file will not be changed, it will remain the same depicting that the file is already there on the cloud's bucket and the file will not get uploaded.
9) Step 9: To perform the above de-duplication, MD5 hashing algorithm is used.
10) Step 10: In this way, de-duplication is being performed.

A. MD5 Algorithm

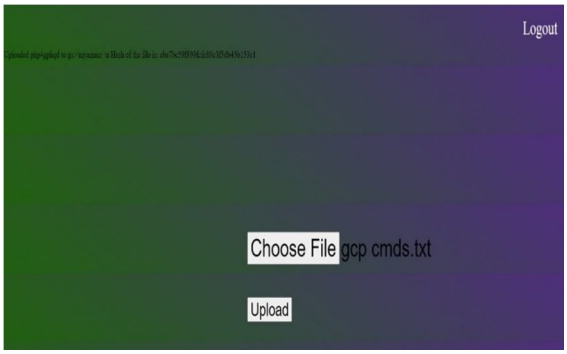
MD5 message-digest algorithm is the 5th version of the Message-Digest Algorithm developed by Ron Rivest to produce a 128-bit message digest. MD5 is quite fast than other versions of the message digest, which takes the plain text of 512-bit blocks, which is further divided into 16 blocks, each of 32 bit and produces the 128-bit message digest, which is a set of four blocks, each of 32 bits. MD5 produces the message digest through five steps, i.e. padding, append length, dividing the input into 512-bit blocks, initialising chaining variables a process blocks and 4 rounds, and using different constant it in each iteration. It was developed with the main motive of security as it takes an input of any size and produces an output if a 128-bit hash value. The 128-bit MD5 hashes typically are represented as 32digit hexadecimal numbers for example: ec55d3e698d289f2afd663725127bace.



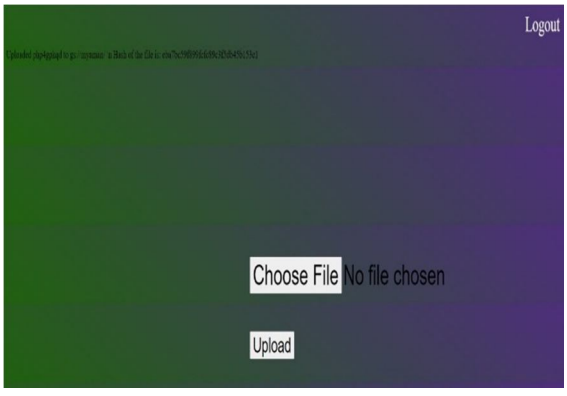
VI. RESULT

In this project, a method is implemented for removing the duplicate files using the MD5 algorithm, ensuring to reduce the deduplicate files being uploaded by the clients while using the cloud. The MD5 algorithm takes lesser time for removing duplicates of files. As the results show that the proposed approach is efficient. In future, further enhancements in this area for a greater number of deduplicated files would require by using various other encryption techniques with different block sizes which can be combined to obtain more efficient results.

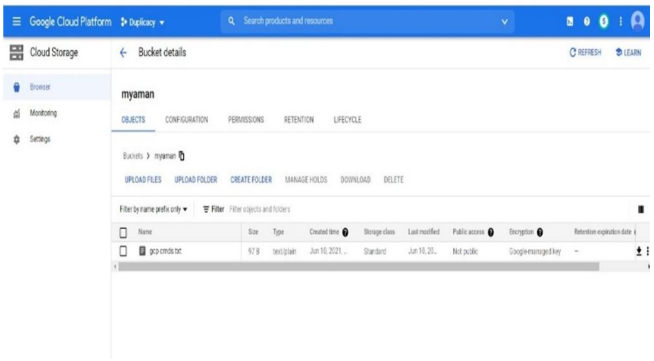
If we upload a file then a unique hash value will be generated for that file.



If we try to upload the same file again, then it will not generate any hash value for that same file again.



The file will be stored inside the bucket only once.

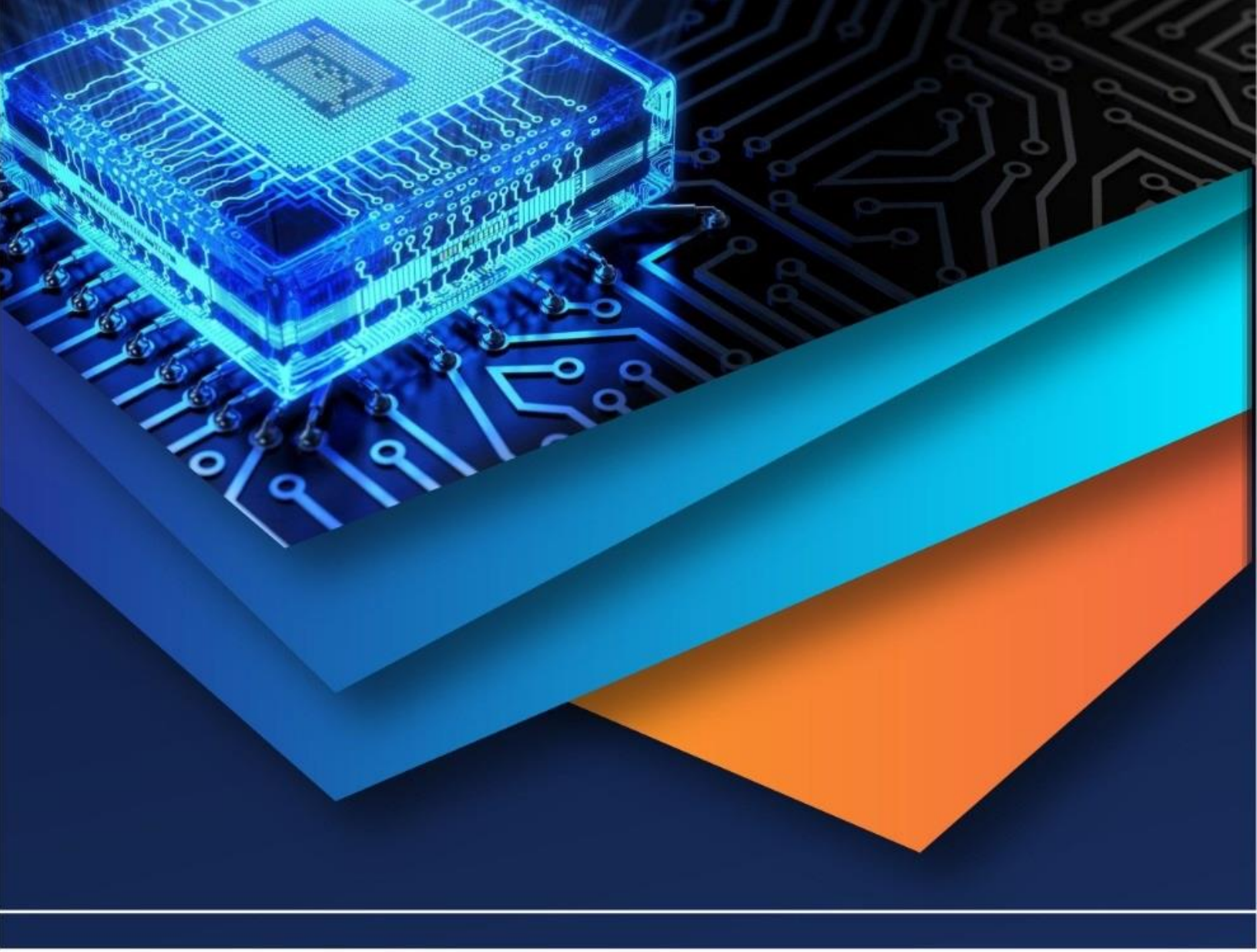


VII. CONCLUSION

Here, it is concluded that this proposed system of file de-duplication is done with authorized way and securely perform all operations. This paper compresses the data by removing the duplicate copies of identical data and it is extensively used in cloud storage to minimize the storage space. It solved more critical part of the cloud data storage which is only tolerated by different methods. Proposed method ensures the data duplication securely.

REFERENCES

[1] IEEE TRANSACTIONS ON PARALLEL AND DISTRIBUTED SYSTEM VOL:PP NO:99 YEAR 2014,A Hybrid cloud approach for secure authorized deduplication, Jin Li, Yan Kit Li, Xiaofeng Chen, Patrick P. C. Lee, Wenjing Lou
[2] M. Bellare, S. Keelveedhi, and T. Ristenpart. Dupless: Serveraided encryption for deduplicated storage. In USENIX Security Symposium, 2013.
[3] J. Li, X. Chen, M. Li, J. Li, P. Lee, and W. Lou. Secure deduplication with efficient and reliable convergent key management. In IEEE Transactions on Parallel and Distributed Systems, 2013.
[4] S. Halevi, D. Harnik, B. Pinkas, and A. Shulman-Peleg. Proofs of ownership in remote storage systems. In Y. Chen, G. Danezis, and V. Shmatikov, editors, ACM Conference on Computer and Communications Security, pages 491–500. ACM, 2013.
[5] C. Ng and P. Lee. Revdedup: A reverse deduplication storage system optimized for reads to latest backups.Proc.of APSYS, Apr 2013.
[6] S. Bugiel, S. Nurnberger, A. Sadeghi, and T. Schneider. Twin clouds: An architecture for secure cloud computing. In Workshop on Cryptography and Security in Clouds (WCSC 2011), 2012.
[7] W. K. Ng, Y. Wen, and H. Zhu. Private data deduplication protocols in cloud storage. In S. Ossowski and P. Lecca, editors, Proceedings of the 27th Annual ACM Symposium on AppliedComputing, pages 441–446. ACM, 2012.
[8] R. D. Pietro and A. Sorniotti. Boosting efficiency and security in proof of ownership for deduplication. In H. Y. Youm and Y. Won, editors, ACM Symposium on Information, Computer and communications Security, pages 81–82. ACM.
[9] P. Anderson and L. Zhang. Fast and secure laptop backups with encrypted de-duplication. In Proc. of USENIX LISA, 2010.
[10] C.-K Huang, L.-F Chien, and Y.-J Oyang, “Relevant Term Suggestion in Interactive Web Search Based on Contextual Information in Query Session Logs,” J. Am. Soc. for Information science and Technology, vol. 54, no. 7, pp. 638-649, 2003.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY