DATA FINDING, SHARING AND DUPLICATION REMOVAL IN THE CLOUD

Janhavi Rahul Patil, Tanishka Manoj Suryavanshi, Unnati Karbhari Bhamare, Arya Narendra Joshi, Prof.P.A.Agrawal

[1] *janhavipatil821@gmail.com, Student of Poly. Dept. of Computer Technology K.K.Wagh , Nashik, India*
[2] *tanusuryvanshi14@gmail.com, Student of Poly. Dept. of Computer Technology K.K.Wagh , Nashik, India*
[3] *unnatibhamare590@gmail.com, Student of Poly. Dept. of Computer Technology K.K.Wagh, Nashik, India*
[4] *aaryajoshi1807@gmail.com, Student of Poly. Dept. of Computer Technology K.K.Wagh, India*
[5]*Internal Guide Dept. of Computer Technology K.K.Wagh, Nashik, India*

---***---

Abstract - Deduplication involves eliminating duplicate or redundant data to reduce stored data volume, commonly used in data backup, network optimization, and storage management. However, traditional deduplication methods have limitations with encrypted data and security. The primary objective of this project is to develop new distributed deduplication systems that offer increased reliability. In these systems, data chunks are distributed across the Hadoop Distributed File System (HDFS), and a robust key management system is utilized to ensure secure deduplication with slave nodes. Instead of having multiple copies of the same content, deduplication removes redundant data by retaining only one physical copy and referring other instances to that copy. The granularity of deduplication can vary, ranging from an entire file to a data block. The MD5 and 3DES algorithms are used to enhance the deduplication process. The proposed approach in this project is the Proof of Ownership (POF) of the file. With this method, deduplication can effectively address the issues of reliability and label consistency in HDFS storage systems. The proposed system has successfully reduced the cost and time associated with uploading and downloading data, while also optimizing storage space.

Key Words: *Cloud computing, data storage, file checksum algorithms, computational infrastructure, duplication.*

1. INTRODUCTION

Cloud computing is an efficient technology for storing large amounts of data that can be easily accessed from anywhere at any time, eliminating the need for expensive hardware, dedicated space, and software maintenance. Meeting the growing demand for storage is a complex and time-consuming task that requires extensive computational infrastructure to ensure successful data processing and analysis. As the number of users and the size of their data continue to grow exponentially, data deduplication becomes increasingly essential for cloud storage providers. By storing a single copy of duplicate data, cloud providers can significantly reduce their storage and data transfer costs. This project provides an overview of cloud computing, cloud file services, accessibility, and storage, while also examining storage optimization through deduplication. It explores existing data deduplication strategies, processes, and implementations to benefit both cloud service providers and users. Additionally, the project proposes an efficient method for detecting and removing duplicates using file checksum algorithms, which requires less time than other pre-implemented methods.

2. LITERATURE SURVEY

2.1 Multi-level comparison of data deduplication in a backup scenario

Cloud computing is an emerging trend in new generation Information and Communication technology. Every user has some amount of data to store in an easily available secure storage space. The advantage of storing data to the cloud for giving out information among friends, to simplify moving data between different mobile devices, and for small commerce to back up and provide disaster recovery capabilities cannot be over-emphasized as the cloud computing concept is basically to make live easy by the solutions emerging with it.

2.2 Data Management in the Cloud: Limitations and Opportunities.

Cloud Computing offers services to users by revising different resources such as computational and storage services and giving them to clients in light of their resources. Cloud computing gives a very big resource pool by connecting network assets together. Data storage is the most vital and prevalent service provided by cloud. Users transfer individual or confidential information to the server form of a Cloud Service Provider and permit it to keep up these information. application.

.

2.3 The Relational Model for Database Management Version

The project enables the user to check for any duplicates in the database by checking the hash value of the file uploaded. If the file already exists in the database, it won't be stored otherwise the file will be saved in the database. The goal of the project is to develop software that uses file checksums to prevent data duplication. The project's main goal is to reduce the number of duplicates in the database, particularly the key-value store, to improve process performance so that the backup window is not impacted, and to design for horizontal scaling so that it can compete on a Cloud Platform.

2.4 Data Duplication Removal using File Checksum

The amount of digital data created throughout the world is massive, and it is continually increasing. According to research, the amount of data produced each year will more than six fold in the next decade, expanding at a rate of 57% every year. The storage infrastructure is being strained by the tremendous rise in data. Enterprise data includes images, audio, video, emails, and other sorts of data. As data grows at a rapid rate, traditional storage methods confront a slew of problems. A high amount of data necessitates the utilization of extra storage space. In truth, a significant size of the data in storage archives is redundant or has been slightly altered to another copy of data. There are a variety of approaches for removing redundancy from stored data.

3. MODULE DESCRIPTION

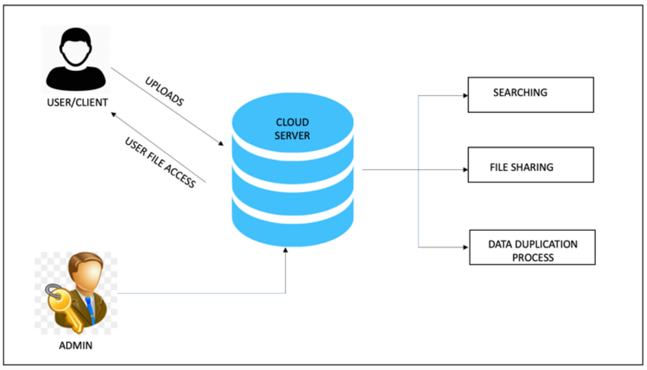


Fig 3.1 System Architecture

If the file already exists, it will update the entry; otherwise, it will make a new entry in the database. Data de-duplication plays a vital role in reducing storage consumption to make it affordable to manage given today's explosive data growth. To avoid false positives, we need to compare new chunks of data with chunks already stored. To reduce the time spent excluding false positives, current research uses extraction of file checksums. However, the target file stores multiple attributes including user ID, filename, size, extension, checksum, and date-time. Whenever a user uploads a file, the system first calculates the checksum and cross-verifies it against the checksums stored in the database.

Here are the key components:

1. Administrators:

Individuals with high-level access and control over the system. They are responsible for system configuration, maintenance, and monitoring. They establish deduplication policies and rules, manage user access and permissions, and monitor system performance while generating reports.

2. Data Managers:

Individuals responsible for overseeing data management and deduplication processes. They set up and schedule deduplication jobs, review and approve flagged duplicate data for removal, and ensure data consistency and integrity. They collaborate with administrators to fine-tune deduplication settings.

3. End Users:

Regular employees or team members who utilize the system to access and manage data. They collaborate on shared documents and files and may have limited permissions, mainly read and edit access. They receive communication and training about the system's impact on their work.

4. Legal and Compliance Officers:

Individuals focusing on the legal and compliance aspects of data management. They monitor and ensure adherence to data protection regulations, manage e-discovery and data retention processes, and work with administrators and data managers to implement governance policies. They also utilize machine learning and data analytics to identify patterns and predict potential issues.

5. IT Support and Help Desk:

Individuals who provide technical support and assistance to end users. They troubleshoot issues related to the duplication removal system, ensure system availability, and respond to user inquiries. They collaborate with administrators to resolve technical challenges.

3. PROJECT METHODOLOGY

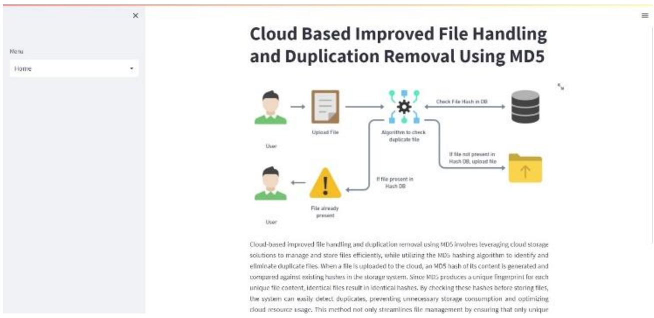
3.1 Checksum Algorithms Parity byte or parity word:

The simplest checksum algorithm is the so-called longitudinal parity check, which breaks the data into "words" with a fixed number n of bits, and then computes the exclusive or (XOR) of all those words. The result is appended to the message as an extra word. To check the integrity of a message, the receiver computes the exclusive or of all its words, including the checksum; if the result is not a word consisting of n zeros, the receiver knows a transmission error occurred. With this checksum, any transmission error which flips a single bit of the message, or an odd number of bits, will be detected as an incorrect checksum. However, an error which affects two bits will not be detected if those bits lie at the same position in two distinct words. Also swapping of two or more words will not be detected. If the affected bits are independently chosen at random, the probability of a two-bit error being undetected is 1/2 n .

3.2 Md5 Checksum Algorithm:

MD5 checksum algorithm which is known as MD5 message-digest is an algorithm that takes as input a message of random length and produces as output a 128-bit fingerprint or message digest of the input. It is estimated that it is computationally impossible to produce two messages having the same message digest, or to produce any message having a given pre-specified target message digest. The MD5 algorithm is proposed for digital signature applications, where a large file must be "compressed" in a secure manner before being encrypted with a private (secret) key under a public-key cryptosystem.

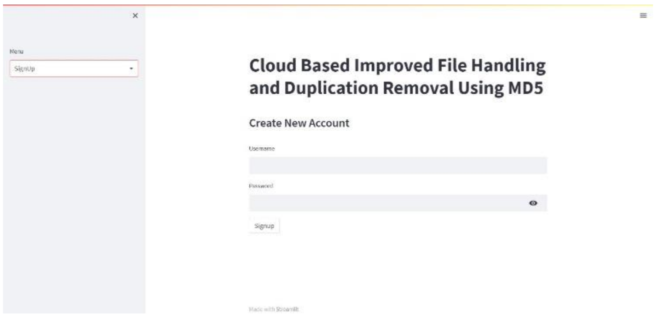
4. GUI/WORKING MODULES


×
≡
Menu
Home
User
Upload File
Algorithm to check duplicate file
Check File Hash in DB
If file not present in Hash DB, upload file
If file present in Hash DB
File already present
User

Cloud Based Improved File Handling and Duplication Removal Using MD5

Cloud-based improved file handling and duplication removal using MD5 involves leveraging cloud storage solutions to manage and store files efficiently, while utilizing the MD5 hashing algorithm to identify and eliminate duplicate files. When a file is uploaded to the cloud, an MD5 hash of its content is generated and compared against existing hashes in the storage system. Since MD5 produces a unique fingerprint for each unique file content, identical files result in identical hashes. By checking these hashes before storing files, the system can easily detect duplicates, preventing unnecessary storage consumption and optimizing cloud resource usage. This method not only streamlines file management by ensuring that only unique

Fig4.1 : Home Page



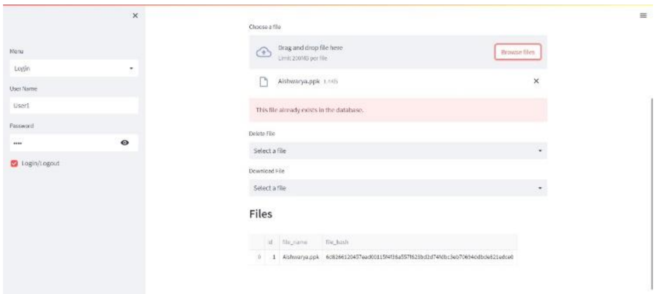
Cloud Based Improved File Handling and Duplication Removal Using MD5

Create New Account

Username

Password

Fig4.2 : Login Page



User Name: User1

Password: ••••

☑ Login/Logout

Choose a file

Aishwarya.ppk 1.1kB

This file already exists in the database.

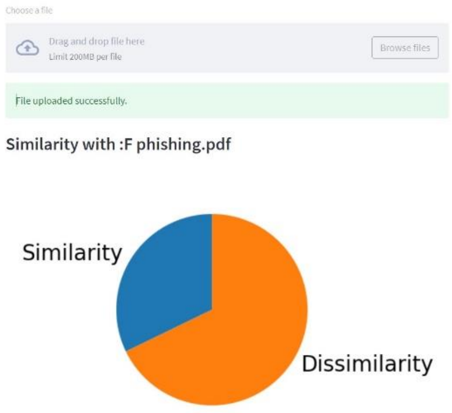
Delete File

Download File

Files

	id	file_name	file_hash
0	1	Aishwarya.ppk	6d8266120457ead00115f4f38a5571623bd2d74fdbc3eb70694ddbde821edce0

Fig4.3 : User Dashboard



File uploaded successfully.

Similarity with :F phishing.pdf

Fig4.3.4: Pie Chart

5. CONCLUSIONS

The web application has been designed to efficiently identify, share, and eliminate duplicate data in the cloud using file checksum. It offers valuable support for organizations engaged in highly repetitive operations that involve frequent data copying and storage for future reference or recovery purposes. This approach forms a critical part of the backup and disaster recovery solution, enabling enterprises to store data repeatedly and facilitate swift, reliable, and cost-effective data recovery. For example, when a file is backed up on a weekly basis, it generates a significant amount of duplicate data, consuming a considerable amount of disk space.

Employing file checksum for data duplicate removal involves conducting an analysis to eliminate these sets of duplicate data, retaining only unique and essential information, thereby freeing up storage space. The primary challenge was to ensure that all files stored in the database did not contain duplicates of themselves, which was addressed by utilizing PHP (Hypertext Pre-processor) software. The educational impact of this system lies in the development of a new and efficient method for identifying, sharing, and removing data duplicates using file checksum in the cloud.

6. REFERENCES

[1] D. Meister, A. Brinkmann, ”Multi-level comparison of data deduplication in a backup scenario”, Proceedings of SYSTOR 2009: The Israeli Experimental Systems Conference, ACM, pp. 8:1-8:12, 2009.

[2] Daniel J. Abadi, Data Management in the Cloud: Limitations and Opportunities, IEEE Data Engineering Bulletin, Volume 32, March 2009, 3-12.

[3] Edgar J. Codd, 1990; The Relational Model for Database Management Version 2; Addison Wesley Longman Publishing Co., incBooston, MA, USA ISBN:0-201-141-14192-2.

[4] Edwin Schouten; 2012; Cloud Computing Business Benefits; http://wired.com/insights/ 2012/10/5-cloud-business-benefits/

[5] Ames B., Rajkman B., Zahir T., 2009 MetaCAN: HanessingStorgae Clouds for High Performance Content Delivery; Journal of Network and Computer Application, 1012-1022, 2009.