File Sharing and Data Duplication Removal in Cloud Using File Checksum

Gopi B, Murugan R

School of Computer Science and Information Technology,
Jain (deemed to be University), Bangalore, Karnataka, 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*

How to cite this paper: Gopi B | Murugan R "File Sharing and Data Duplication Removal in Cloud Using File Checksum" Published in International Journal of Trend in Scientific Research and Development (ijtsrd), ISSN: 2456-6470, Volume-6 | Issue-2, February 2022, pp.1331-1333, URL: www.ijtsrd.com/papers/ijtsrd49416.pdf

IJTSRD49416

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. SURVEY MOTIVATION

"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", 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.

2.1. GOAL

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.

2.2. LIMITATION

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

2.3. KEYWORDS

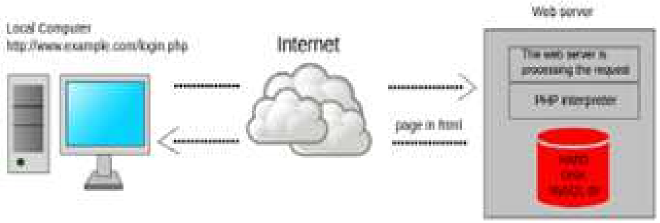
Cloud computing, data storage, file checksum algorithms, computational infrastructure, duplication.

3. SURVEY OUTCOMES

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.


Local Computer
http://www.example.com/login.php
Internet
Web server
The web server is processing the request
PHP interpreter
page in html

4. 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.

5. 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

[9] Morris Dworkin, 2015; NIST Policy on Hash Functions; Cryptographic Technology group,https://csrc.nist.gov/projects/hash-functions/nist-policy-on-hash-functions August 5, 2015; "National Institute of Standard and Technology NIST Special Publication 800-145

[10] NimalaBhadrappa, Mamatha G. S. 2017, Implementation of De-Duplication Algorithm, International Research Journal of Engineering and Technology (IRJET), Volume 04, Issue 09.

[11] O'Brien, J. A. &Marakas, G. M. (2011). Computer Software. Management Information Systems 10th ed. 145. McGraw-Hill/Irwin

[12] Peter Mel; The NIST definition of Cloud Computing, "National Institute of Standard and Technology NIST Special Publication 800-145

[13] PHP 5 tutorials; W3Schools, https://www.w3schools.com/pHP/default.asp

[14] Accessed June, 2018.

[15] Rivest R., 1992 The MD5 Message Digest Algorithm. RFC 1321 http://www.ietf.org/rfc/rfc321.txt

[16] Sandeep Sharma, 2015; 15 Best PHP Libraries Every Developer Should Know; published on; https://www.programmableweb.com/news/15-best-php-libraries-every-developer-should-know/analysis/2015/11/18 ; accessed June 12, 2018.

[17] Single Instance Storage in Microsoft Windows Storage Server 2003 R2Archived 2007-01-04 at the Way back Machine: https://archive.org/webTechnical White Paper: Published May 2006 access September, 2018.

[18] Stephen J. Bigelow, 2007 Data Deduplication Explained: http://searchgate.org; Accessed February, 2018

[19] Wenying Zeng, Yuelong K. O, Wei S., (2009) Research on Cloud Storage Architecture and Key Technologies, ICIS 2009 Proceedings of the 2nd International Conference on Interaction Sciences: Information Technology, Culture and Human Pages 1044-1048.

[20] What is PHP? PHP User contributory notes; http://php.net/manual/en/intro-whatis.php. Accessed June 6, 2018

[21] X. Zhao, Y. Zhang, Y. Wu, K. Chen, J. Jiang, K. Li, "Liquid: A scalable deduplication file system for virtual machine

[22] Images", *Parallel and Distributed Systems IEEE Transactions on*, vol. 25, no. 5, pp. 1257-1266, May 2014.