International Journal of Advanced Trends in Computer Applications
www.ijatca.com

A Review: De-Duplication Approach for Reducing Memory Consumption in Cloud Computing

Archana[1], Gurjot Singh Sodhi [2]

[1] Archana
Shaheed Udham Singh College of Engineering and Technology
Tangori (Mohali)
Research Scholar, CSE department
archanabeni@gmail.com

[2] Gurjot Singh Sodhi
Shaheed Udham Singh College of Engineering and Technology
Tangori (Mohali)
Assistant Professor, CSE department
er.gurjotsinghsodhi@gmail.com

Abstract: *Cloud computing is technology defined as a computing system that deals with various services as IAAS, PAAS, SAAS and handles all resources which takes part in sharing. In cloud computing cloud storage is that storage where the data is retained in logical pools, the physical storage compasses multiple servers and the physical environment is typically maintained and supervised by a hosting company. Huge amount of data over cloud storage leads to insufficient use of cloud storage and cost of using cloud storage service is increasing side by side. When huge amount of is transferred then network bandwidth is more utilized. To elucidate this issue of data storage data de-duplication is one of the modern technologies in storage. Here different de-duplication approaches are discussed which reduces the storage and bandwidth required for data transfer.*

Keywords: *Cloud computing, cloud storage, Data de-duplication, Metadata, Caching, De-duplicator*

1. Introduction

1.1 Cloud computing the word cloud is used for "*the Internet*," so *cloud computing* means "a type of Internet-based computing," where different services such as servers, storage and applications are used by an organization’s through the Internet. Cloud Computing is a technology which refers to manipulating, configuring, and accessing the applications online. It provides online data storage, infrastructure and application. Example of cloud computing is Gmail, drop box or Hotmail etc.

1.2 Cloud storage is that model where data can be placed, managed, backed up, stored and modified. Cloud storage makes available data to clients in any time, with high storage space and also makes it user friendly so that availability of data increases. Cloud storage is of three types: Public, Private and hybrid. As long as population of cloud service and data volume increasing people want to economize the capacity of cloud storage **[1].** Therefore how to make use of the cloud storage capacity well becomes far reaching issue now a day.

1.3 Data de-duplication is a data compression technique which reduces the storage capacity by eliminating duplicate copies of data or reduces the amount of data that has to be transferred over a network **[1].** Data de-duplication not only reduces the storage space requirements by eliminating redundant data but also minimizes the network transmission of duplicate data in the network storage systems. Data de-duplication done at client side and server side **[2].** In client side de-duplication is done before sending the data to a storage device. Only unique data is transferred to the device with the minimum available band width and it needs less space. At server side de-duplication is done after sending the data to storage device. De-duplication is also used in back up services to reduce network bandwidth.

2. Methods for De-duplication

There are some methods of de-duplication are discussed. To find the data de-duplication hashing algorithm and application aware local global de-duplication is used which will perform better results.

2.1 Methods of de-duplication Based on data

2.1.1 File Level De- duplication: - In file level de-duplication whole file is checked for de-duplication files **[3].** When clients want to upload the file then by using hashing algorithm like Rabin fingerprint, MD5, SHA-1, etc. a fingerprint of that file is generated. This fingerprint is unique for all different files. This fingerprint is used to store in place of whole file which reduces the storage space in data storage. A Pointer is used to point the original file for the subsequent copies. The advantage of this method is that it is simple and fast. If one byte of file is changed then its hash value is changed. In file level de-duplication is performed at the source level where de-duplication is done at the source without storing it into database. This method saves more space in the data storage.

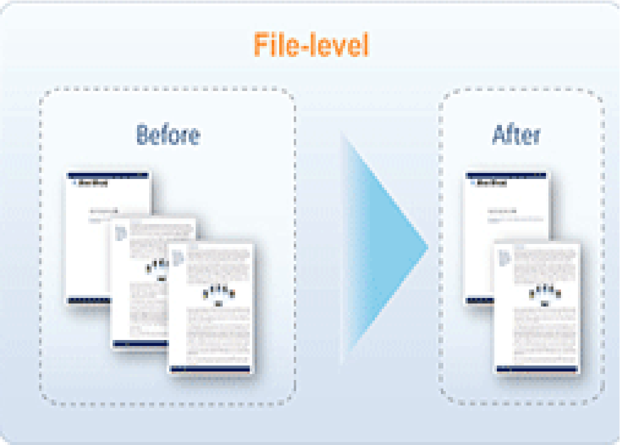

Figure 1: Shows the file level de-duplication

2.1.2 Block level De-duplication:- Block or Sub file De-duplication: - In block level whole file is divided into number of blocks or sub blocks **[3].** Then hash value of these blocks is calculated and then compared to hash value of each block. If the hash value of the block is unique then it is consider as a unique and stored in the database, otherwise a pointer is stored. Only pointers are stored in the storage not the block of file which saves storage. According to the size of block there are two processes in block De-duplication.

a) Fixed-Length block: - In Fixed length blocks are divided into fixed length which is defined by the user. After that its duplication is checked. This method is fast, simple and minimum CPU overhead.

b) Variable-Length block: - In which of blocks are divided into variable sizes. Blocked are stored according to the size and then its de-duplication is checked by using de-duplication techniques.

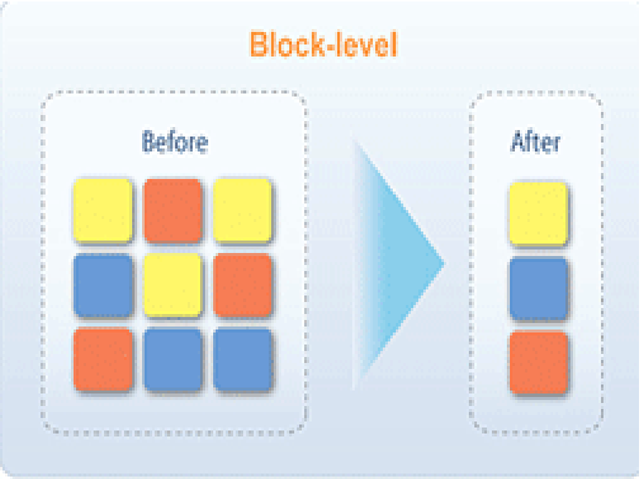

Figure 2: Shows the Block level de-duplication

2.2 Methods based on when de-duplication is performed

2.2.1 Inline De-duplication:- In this de-duplication method de-duplication is done as data is flowing. De-duplication is done before storing data into data storage. This method results in less storage space used and only unique data is transferred through the network. If duplicate data is found then its pointer is used for storage.

2.2.2 Post process De-duplication:- This method performs de-duplication after storing the data into data storage. In which all the data is transferred through the network and stored in the disk. Duplicate data is also stored into disk which takes more space in data storage.

2.3 Techniques of de-duplication

2.3.1 Hashing Technique: - In hashing technique hashing the data means creating a hash value or number of the file, block and byte which guarantee to be unique for all the above types **[2].** In hashing technique some hashing algorithms are used. These hashing algorithms have their own properties like their output size, block size, rounds and performance. Hashing technique is used after uploading the file. When fingerprint of the file is generated then it is stored in the metadata and used for the comparison purpose. If two files have same hash value then it is said to be a similar file otherwise it is considered as a different file. By using this technique storage space is reduced and time is also saved to find the duplicate files. Searching of files is also easy when we have customized data storage.

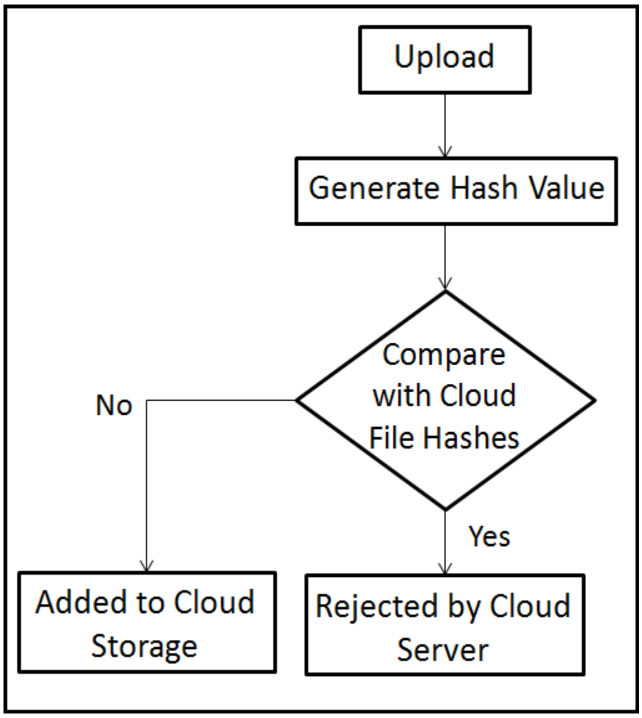

Figure 3: Basic Block Diagram

From the above block diagram the file to be uploaded is fed to the hashing algorithm which generates the hash value. The hash value is compared with already existing hash values. If a matching hash value is found the particular file will not be added to the cloud storage, else server will store the file.

Table 1: Comparison of hashing algorithms

Algorithms	O/P(bits)	Interna l state size	Rounds
WHIRLPOOL	512	512	10
PANAMA	256	8736	64
Havel	160	256	160
MD2	128	384	864
MD5	128	128	64
SHA-0	160	160	80
SHA-1	160	160	80
SHA-2	256	256	64

In the above table hashing algorithms are defined with their sizes. From these algorithms different sizes have been calculated. The cycle per byte and the rounds show the efficiency of above algorithms.

2.3.2 Application Aware de-duplication:- This technique of de-duplication is known as Byte Level De-duplication, because in which deepest level of de-duplication is performed. It is also known as Content Aware De-duplication where all the content of the objects, files, applications. Data is divided into blocks and then check the bytes and stored only those bytes which are not unique. This method is so time consuming method and some loss of data is possible.

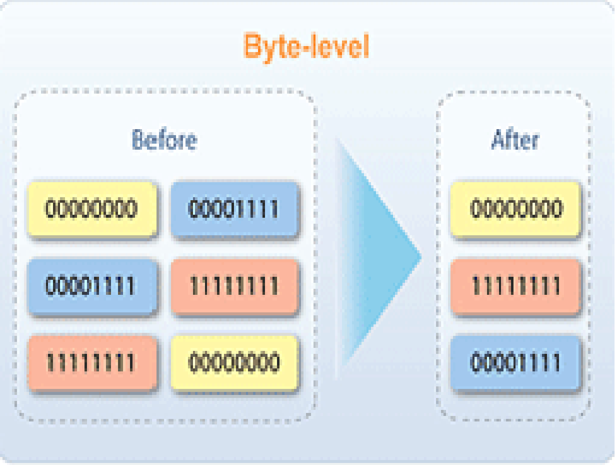

Figure 4: Byte level de-duplication

2.4 Meta data server: - Metadata server maintains information of fingerprints and storage server in the system. MM maintains a linked list and a log file in order to keep track of file and avoid the storage of multiple copies of the same files. In metadata fingerprint, file id, and format of the file is stored **[5].** Fingerprint of files are compared from the existing fingerprints which are stored in metadata server.

2.5 Caching: -To makes the speed high to access the data in addition to the storage place a CACHE memory will be added **[6].** So that speed will increase and also time will reduce to access the data. So in proposed work CACHE memory will be used to access the files. Doing this we need virtual memory that is also called cache memory.

2.6 De-duplicator instances: - De-duplicator instances are those which are used to detect that the file is de-duplicated or not by comparing from the metadata server **[5].** When large number of clients wants to upload the file then one de-duplicator is not enough. Load balance factor will be more cooperated in large number of de-duplicators. Time will be decreased when we have large number of de-duplicators.

3. Conclusion

Cloud is the costly storage provider, so the motivation is to use its storage area efficiently. De-duplication has been proved to reduce memory consumption by removing the useless duplicate files. So far from the previous studies file level de-duplication is the better approach to be used, the focus of the proposed work will be on file level de-duplication. A lot of research has been carried out over this by means on hashing algorithm. From the previous hashing algorithms sha2 will perform better than md5 and sha1.Our aim is to choose a well built algorithm which will generate a good hash value in turn reducing cloud storage.

REFERENCE

[1]. Neelaveni, P., and M. Vijayalakshmi. "A Survey on De-duplication in cloud storage." Asian Journal of Information Technology 13.6 (2014): 320-330.

[2]. NagaMalleswari, T. Y. J., D. Malathi, and G. Vadivu. "De-duplication Techniques: A Technical Survey."

[3]. Deepu, S. R. "Performance Comparison of De-duplication techniques for storage in Cloud computing Environment." Asian Journal of Computer Science & Information Technology 4.5 (2014).

[4]. Xing, Yu-xuan, et al. "AR-dedupe: An efficient de-duplication approach for cluster de-duplication system." Journal of Shanghai Jiaotong University (Science) 20 (2015): 76-81.

[5]. Leesakul, Waraporn, Paul Townend, and Jie Xu. "Dynamic Data De-duplication in Cloud Storage." Service Oriented System Engineering(SOSE), 2014 IEEE 8th International Symposium on. IEEE, 2014.

[6]. Jyoti Malhotra1,Priya Ghyare2, A Novel Way of De-duplication Approach for Cloud Backup Services Using Block Index Caching Technique, Vol. 3, Issue 7, July 2014 DOI: 10.15662/ijareeie.2014.0307040

[7]. Upadhyay, Amrita, et al. "Deduplication and compression techniques in cloud design." Systems Conference (SysCon), 2012 IEEE International. IEEE, 2012.

[8]. Li, Jin, et al. "Secure de-duplication with efficient and reliable convergent key management." Parallel and Distributed Systems, IEEE Transactions on 25.6 (2014): 1615-1625.

[9]. Paulo, João, and José Pereira. "A survey and classification of storage de-duplication systems." ACM Computing Surveys (CSUR) 47.1(2014)

[10]. Madhubala, G., et al. "Nature-Inspired enhanced data de-duplication for efficient cloud storage." Recent Trends in Information Technology (ICRTIT), 2014 International Conference on. IEEE, 2014.

[11]. Paulo, João, and José Pereira. "A survey and classification of storage de-duplication systems." ACM Computing Surveys (CSUR) 47.1(2014)