IJCSE International Journal of Computer Sciences and Engineering Open Access
Survey Paper Volume-6, Special Issue-2, March 2018 E-ISSN: 2347-2693

Survey on Enhancing Cloud Storage Using Deduplication Technique

A. Vijayakumar[1*], A. Nisha Jebaseeli[2]

[1*]Dept. of Computer Science, Arul Anandar College, Madurai, Tamilnadu
[2]Dept. of Computer Science, Bharathidasan University Constituent College, Lalgudi, Tamilnadu

Available online at: www.ijcseonline.org

***Abstract*—** Today we are in a highly informative and digital era, comparatively large amount of data are generated simultaneously day to day. All these data must be stored for processing the result and for future references. As the availability of data increases, we have to invest more amounts in maintaining the storage devices. To meet the cost on storage devices and to enhance storage efficiency. It is mandatory use a creative technique to store maximum data within the available storage devices. This can be achieved by data deduplication method. It eliminates redundant copies of data in order to minimize the storage requirement and achieves cost efficiency on storage devices. This paper will analyze various deduplication methods and give an ample survey.

***Keywords*—** Deduplication, genetic algorithm, Hash algorithm, bandwidth, chunking, compression, Jenkins hash function, file similarity.

I. INTRODUCTION

Cloud computing is a newly emerged technology, and the rapidly growing field of IT. The process of improving the storage efficiency and power consumption of data storage equipment is a complicated one, with a variety of factors to consider such as upgrading data storage equipment, combination of storage efficiency and data depuplication techniques. Data deduplication often called intelligent compression or single-instance storage. Data deduplication techniques ensure that only one unique instance of data is retained on storage media, such as disk, flash or tape. Redundant data blocks are replaced with a pointer to the unique data copy. In that way, data deduplication closely aligns with incremental backup, which copies only the data that has changed since the previous backup. Data depuplication can occur at the source or target level. Deduplication removes redundant blocks before transmitting data to a backup target at the client or server level. There is no additional hardware required. Deduplication at the source reduces the bandwidth and storage use. In deduplication, backups are transmitted across a network to disk-based hardware in a remote location. Using deduplication targets increases costs, although it generally provides a performance advantage compared to source dedupe, particularly for petabyte-scale data sets [1-10].

II. WORKING OF DEDUPLICATION

Data deduplication works by comparing objects (usually files or blocks) and removes objects (copies) that already exist in the data set. All the processes which are not unique are removed in this method. In Data deduplication method we divide the input data into blocks and a hash value is calculated for each of these blocks. Then using these hash values we can determine whether another block of same data has already been stored. If a similar data file is found then replace the duplicate data with a reference to the object already present in the database.

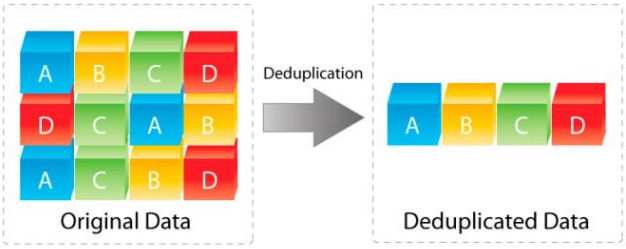


Fig. 1: data deduplication Diagram

IV. FUNCTIONS OF DATA DEDUPLICATION

It compares objects (usually files or blocks) and removes objects (copies) that already exist in the data set. The deduplication process removes blocks that are not unique.

1. Divide the input data into blocks or "chunks."

2. Calculate a hash value for each block of data.

3. Use these values to determine if another block of the same data has already been stored.

4. Replace the duplicate data with a reference to the object already in the database.

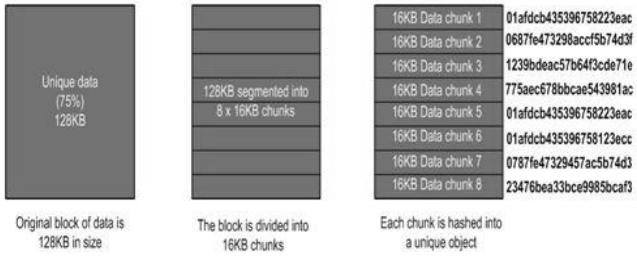


Fig. 2: working mode of data deduplication

V. DEDUPLICATION METHODS

A. *Dynamic Chunking Algorithm*

Efficient chunking is one of the key elements that decide the overall deduplication performance chunking refers to divide something into pieces. Generally, the chunking algorithms are divided into two; fixed-length chunking and variable-length chunking. The fixed-length chunking approach achieves very fast data deduplication result but the performance is not good, because boundary shift problem degrades the deduplication performance. On the contrary, the variable length chunking achieves high degree of performance while causing high computation. When we modify a file or append new data blocks to an old version file, the new version of a file usually contains lots of duplicated region of data blocks compared with previous version of a file. Therefore, if we find one duplicated data block then we can find lots of duplicated data blocks around this position. Overhead and longer processing time. Here deduplication is done by fixed length and file similarity techniques.[1]

B. *Genetic Algorithm*

After remove the duplicate records some of the useless documents are available in the results. Using genetic algorithm with particle swarm optimization accurately detects the all duplicates. It can remove useless data and finally produces the meaningful and high accurate quality data. Quality data is more helpful for organizations. [2]

C. *Jenkins Hash Function*

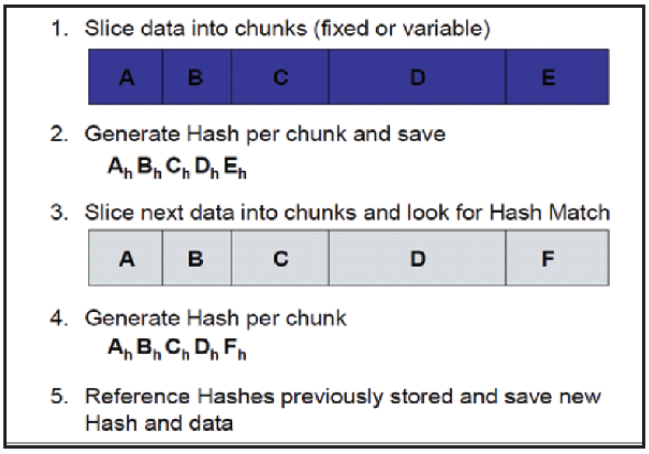
Data deduplication is a method of tumbling stowage needs by eliminating redundant data. Only one unique occasion of the data is actually retained on storage broadcasting. Redundant data is replaced with a pointer to the unique data copy and it has been widely used in cloud packing to reduce the amount of storage space and save bandwidth. Jenkins hash function (JHF) techniques have been proposed to encrypt the data before outsourcing. It is used for text file, pdf file, book, and image and video, and also it decrease the authorization steps parameters from above 60%.[3]

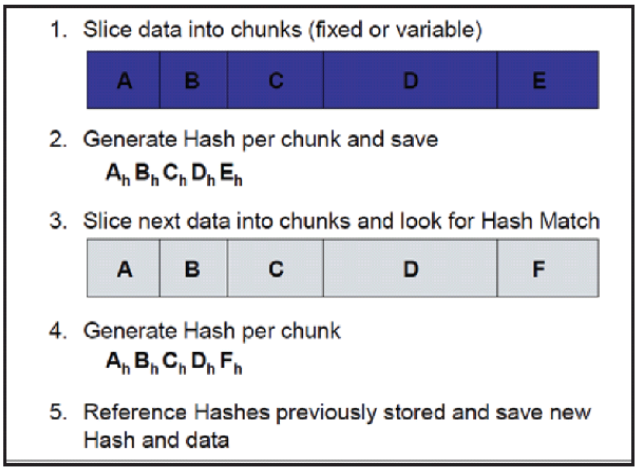
D. *Active Learning*

The main challenge in this task is designing a function that can resolve when a pair of records refers to the same entity in spite of various data inconsistencies. Most existing systems use hand-coded functions. One way to overcome the tedium of hand-coding is to train a classier to distinguish between duplicates and non-duplicates. The success of this method critically hinges on being able to provide a covering and challenging set of training pairs that bring out the subtlety of the deduplication function. [4]

E. *Hash-based Algorithms*

Hash based deduplication methods use algorithms to identify chunks of data. If the hash is already created, the data is identified as a duplicate and is not stored. Commonly used algorithms are Secure Hash Algorithm 1(SHA1) and Message-Digest Algorithm 5(MD5).SHA-1: This was devised to create cryptographic signatures for security application. The 160-bit value created by SHA-1 is unique for each piece of data. It breaks data into "chunks" which are either fixed or variable in length. This processes the "chunk" with hashing algorithm to create a hash, If the hash already exists, the data is deemed to a duplicate and is not stored. If the hash does not exist, then the data is stored and the hash index is updated with the hash. MD5: This 128-bit has also designed for cryptographic uses. In this method the 128-bit state is divided into four 32-bit words, denoted A, B, C and D. These are initialized to certain fixed constants[5]





F. Divide and Conquer Based Deduplication

In this approach, it first converts the attributes of data into numeric form. Then, this numeric form is used to create clusters by using K-Means clustering algorithm. The use of clustering reduces the number of comparisons. After that the divide and conquer technique is used in parallel with these clusters for identification and removal of duplicated records. Here, this technique identifies all type of duplicated records like fully duplicated records, erroneous duplicated records and partially duplicated records. This technique is only applicable for single table instead of multiple sorted tables. The performance is measured by using the terms like true positives, false positives, false negatives, precision, recall and F-Score. [5]

VI. Conclusion

An analysis of the existing deduplication techniques is done here. Deduplication is a crucial step in data integration. From this survey, it is possible to conclude that the existing algorithms require more memory for deduplication. It is also time consuming process. In future a deduplication algorithm can be designed for reducing time consumption and utilizes less memory space.

References

[1] Young Chan Moon1, Ho Min Jung1, Chuck Yoo2, and Young Woong Ko1 "*Data Deduplication Using Dynamic Chunking Algorithm*", Springer ICCI 2012.

[2] RavikanthM*1,Dr.D.Vasumathi*2, B.Mallikarjuna Reddy*3, "*Enhanced Duplicate Detection Using Genetic Algorithm With Particle Swarm Optimization"*, IJCSIET— International Journal of Computer Science information and Engg., Technologies ISSN 2277- 4408

[3] R.Supriya1, K. Sathyaseelan2," *Secure Deduplication Using Jenkins Hash Function*" Ncraccess -2015.

[4] Sunita Sarawagi, Anuradha Bhamidipaty," *Interactive Deduplication using Active Learning*"

[5] Amanpreet Kaur, 2Sonia Sharma "*An Efficient Framework and Techniques of Data Deduplication in Cloud Computing* ", IJCST Vol. 8, Iss ue 2, April - June 2017.

[6] Lalitha. L1, Maheswari.B2,Dr.Karthik.S3, "*A Detailed Survey on Various Record Deduplication Methods*", International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)Volume 1, Issue 8, October 2012

[7] R.Supriya1, K. Sathyaseelan2," *Secure Deduplication Using Jenkins Hash Function*" Ncraccess -2015.

[8] Amanpreet Kaur, 2Sonia Sharma "*An Efficient Framework and Techniques of Data Deduplication in Cloud Computing* ", IJCST Vol. 8, Iss ue 2, April - June 2017.

[9] Lalitha. L1, Maheswari.B2,Dr.Karthik.S3, "*A Detailed Survey on Various Record Deduplication Methods*", International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)Volume 1, Issue 8, October 2012.

[10] Poonam R. Wagh, Amol D. Potgantwar, "Providing Security to Data Stored on HDFS Using Security Protocol", International Journal of Scientific Research in Network Security and Communication, Vol.5, Issue.4, pp.20-25, 2017.

