ISSN No: 2309-4893
International Journal of Advanced Engineering and Global Technology
Vol-03, Issue-08, August 2015

Data Deduplication in Cloud Storage

Golthi Tharunn, Gowtham Kommineni, Sarpella Sasank Varma, Akash Singh Verma

Electronics and Communication Department, GITAM University
Rudraram, Hyderabad, Telangana, India
golthitarun@gmail.com, gowthamk63@gmail.com, sasank.vrm@gmail.com, verma.akash9506@gmail.com

Abstract: **With an increase in the usage of cloud storage, effective methods need to be employed to reduce hardware costs, meet the bandwidth requirements and to increase storage efficiency. This can be achieved using Data Deduplication. Data Deduplication is a method to reduce the storage need by eliminating redundant data. Thus by storing less data you would need less hardware and would be able to better utilize the existing storage space.**

I.INTRODUCTION

The use of cloud for storing data by companies for backup and common people for sharing information among friends has increased drastically over the past few years. This has created a challenge to the cloud service providers to maintain all this massive data and to offer these services at lower price to the customers. In reality most of the data stored in the servers is often repeated. For example, a service may contain several instances of same data file, storing all these instances would require a large amount of storage space. This problem can be solved by using Data Deduplication technique.

Data deduplication stores only one unique instance of the data type on the disk or tape. In this method redundant data is replaced with a pointer to the unique data copy. This reduces the hardware used to store data and the bandwidth costs required for transmitting and receiving purposes. Block and bit level deduplication methods are able to achieve compression ratios of 20x to 60x, or even higher, under the suitable conditions.

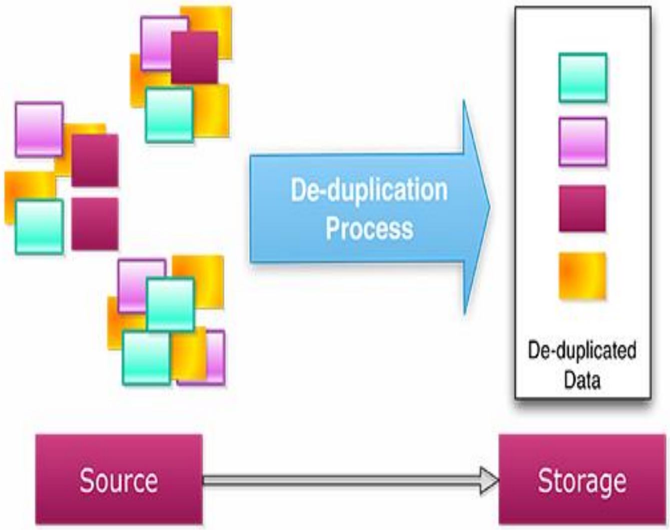
II. DEDUPLICATION vs. COMPRESSION

Deduplication is sometimes confused with compression, another method for reducing storage requirements. While deduplication eliminates redundant data, compression uses algorithms to save data more concisely. Compression may be lossless compression or lossy compression but when you consider the case of deduplication, no data is lost as it only eliminates extra copies of data. Deduplication often has a larger impact on bacakup file size than compression. In a typical enterprise backup situation, compression may reduce backup size by a ratio of 2:1 or 3:1, while deduplication can reduce backup size by up to 25:1, depending on how much duplicate data is in the systems

III. HOW DEDUPLICATION WORKS?

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.

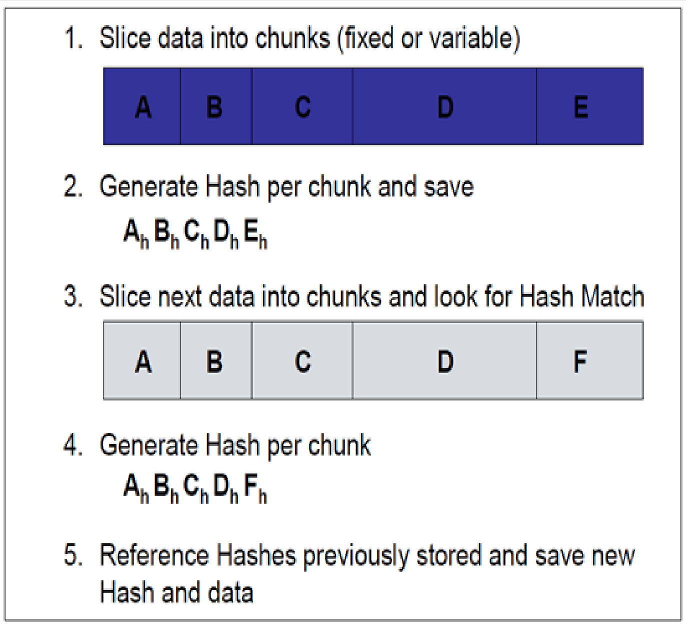


Hash-based algorithms:

Hash based deduplicatoin 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.


1. Slice data into chunks (fixed or variable)

2. Generate Hash per chunk and save

$A_h B_h C_h D_h E_h$

3. Slice next data into chunks and look for Hash Match

4. Generate Hash per chunk

$A_h B_h C_h D_h F_h$

5. Reference Hashes previously stored and save new Hash and data

MD5:

This 128-bit has was 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. The main algorithm then uses each of these messages in turn to modify the state. The processing of a message block consists of four similar stages, termed rounds. Each round consists of 16 similar operations based on a non- linear operation F. There are four possible functions of this function and a different one is used for each round.

$$F(B,C,D) = (B \wedge C) \vee (\neg B \wedge D)$$

$$G(B,C,D) = (B \wedge D) \vee (C \wedge \neg D)$$

$$H(B,C,D) = B \oplus C \oplus D$$

$$I(B,C,D) = C \oplus (B \vee \neg D)$$

$\oplus, \wedge, \vee, \neg$ Denote XOR, AND, OR and NOT operations respectively.

IV.Technological Classification

Post-process deduplication (PPD):

It is also known as asynchronous deduplication or offline deduplication. It involves the removal of redundant data after a backup is completed and data has already been written to storage. The benefit of this method is that backup data is straightforward and takes very less time because the calculations of hash values and lookup takes place only after all of the data is stored.

In-line deduplication:

This process involves the calculation of hash values as the data enters the system. The benefit of this process over post-process is that it will take very less space because the calculation of hash values and the lookup process is completed before the data enters the database. So only one instance of a particular data is stores and the duplicate data is reference to the data present in the server.

Source deduplication:

This type of deduplication is the best suited to use at remote offices for backup to the cloud. The deduplication takes place typically within the system by regularly scanning new files creating hashes and compares them to the hashes of existing files. It offers a number of benefits, including the reduction of bandwidth and the amount of data that has to be sent to the cloud server.

Target deduplication:

This is best suited for the use in the data center for the reduction of massive data sets. In this case, the client is unmodified and is not aware of any deduplication. Target deduplication requires that the target backup server or dedicated Hardware target appliance handles all of the deduplication. This process requires more network resources compared to source deduplication because the original data, with all its redundancy, must go over the network.

File Level and Sub-file Level Deduplication:

The full file level duplicates easily can be eliminated by calculating single checksum of complee file data and comparing it against existing checksums of the already backed up files. This method of deduplication is simple and fast, but the extent of deduplication is less, as this process does not address the problem of duplicate files or data-sets.

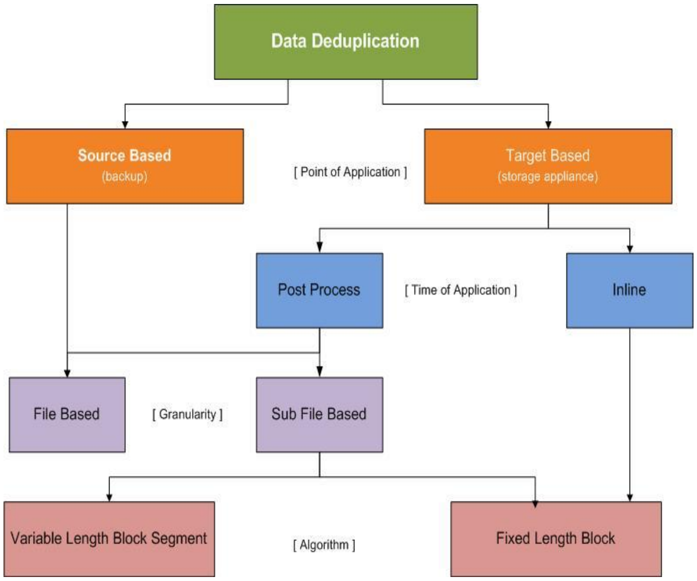
The sub-file level deduplication breaks the file into smaller fixed or variable size blocks, and uses hash based algorithm to compare these blocks and find similar blocks.

Fixed-Length Blocks:

A fixed-length block approach divides the files into fixed size length blocks and uses a simple checksum-based approach (MD5/SHA etc.) to find the duplicates. This process has a limited effectiveness. The reason for this is that the primary opportunity for data reduction is in finding duplicate blocks in two transmitted datasets that are mostly- but not completely of same data segment.

Variable Length Data Segment technology:

This technique divides the data stream into variable-length data segments using a methodology that can find the same block boundaries in different locations and contexts. This allows the boundaries to float within the data stream so that changes in one part of the dataset have little or no impact on the boundaries in other location of the dataset.



V.Practical Applications

Data deduplication helps to achieve data optimization and capacity scaling goals. It offers practical ways for the cloud service providers to achieve these goals. These ways include the following.

Capacity optimization: Data deduplication reduces the physical space used for storing data. This achieves greater storage efficiency than was possible by using features such as Single Instance Storage (SIS) or NTFS compression. This method uses subtle variable-size chunking and compression, which deliver optimization ratios of 2:1 for general file servers and up to 20:1 for virtualization data.

Scale and Performance: Data deduplication can process up to 50 MB per second in typical windows derver 2012 R2, about 20 MB of data per second in Windows Server 2012. It can work on multiple volumes simultaneously without effecting other workloads on the server.

Reliability and data integrity: Data deduplication process maintains the integrity of data. This method uses checksum, consistency and identity validation to ensure data integrity. For all metadata and most frequently referenced data, data deduplication maintains redundancy to ensure that the data is recoverable in the event of data corruption.

Bandwidth efficiency with BranchCache: Through integration with BrachCache, the same optimizator techniques are applied to data transferred over the WAN to a branch office. The result is faster file download times and reduced bandwidth consumption.

Optimization management with familiar tools: Data deduplication functionality built into Server Manager and Windows PowerShell. Default settings can provide savings immediately, or administrators can fine-tune the settings to see more gains. One can easily use Windows PowerShell cmdlets to start an optimization job or schedule one to run in the future. Installing the Data Deduplication feature and enabling deduplication on selected volumes can also be accomplished by using an Unattend.xml file that calls a Windows PowerShell script and can be used with Sysprep to deploy deduplication when a system first boots.

VI.SECURITY

The only major drawback with it is a security hole in one of its basic properties. Consider a file being uploaded, then the question arises "Has anyone stored a prior copy?” That means is this particular file already stored or not? This question is answered by the attacker, requesting to upload a copy of the file and checking whether de-duplication occurs. This being a restricted query, the answer is either true or false which does not provide any information about who performed the task. Also, in the basic form of attack the attacker can only request this query once. Once the query is requested by uploading the file, it is saved at the upload service and hence the answer to the query is always positive. Further, the information can be erased by the following method; the attacker starts uploading a file and checks if de-duplication occurs. If de-duplication does not occur, a full upload starts and the attacker shuts down the communication channel and the upload terminates. As a result, the copy of the file kept by the attacker is not saved at the server, this in turn enables the attacker to repeat the same test at a later time and check if the file was uploaded. Therefore, by using this strategy at regular intervals, the time window of the uploaded file can be obtained**.**

VII. CONCLUSION

This paper discusses the information about data deduplication for the cloud based systems. It includes the methods that are used to achieve cost effective storage and effective bandwidth usage by deduplication. The core concept involves eliminating the duplicate copies of the repeated data by using hashing algorithms. However, reliability and speed are at stake. The future challenge therefore lies in identifying more effective hashing algorithms for improving the speed of storing data and security. However, data deduplication is the most crucial element for improving efficiency of the cloud system. This technique will play a major role in the cloud based services for storing backup data by both medium and large enterprises.

VIII.REFERENCES

1. https://technet.microsoft.com/en-us/library/hh831602.aspx
2. http://www.computerworld.com/article/2474479/data-center/data-deduplication-in-the-cloud-explained--part-one.html
3. http://searchdatabackup.techtarget.com/definition/post-processing-deduplication
4. http://searchstorage.techtarget.com/definition/data-deduplication
5. https://en.wikipedia.org/wiki/Data_deduplication
6. http://www.webopedia.com/TERM/D/data_deduplication.html
7. http://www.druva.com/blog/understanding-data-deduplication/
8. https://pibytes.wordpress.com/2013/02/09/deduplication-internals-hash-based-part-2/
9. http://programming4.us/enterprise/12380.aspx