IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN : 2278-0661, p-ISSN : 2278-8727
PP 85-90
www.iosrjournals.org

A Survey on Deduplication Strategies and Storage Systems

Guljar Shaikh
((Information Technology,B.V.C.O.E.P/ B.V.C.O.E.P, INDIA)

Abstract : *Now a day there is raising demands for systems which provide the data storage by keeping in mind security and cost factor. A duplicate file not only occupies a high volume space but also increases access time so that removing duplicates records become very necessary. But to achieve this is not that simple since neither duplicate files don't have a common key nor they contain error which makes duplicate matching a tedious task. There are different ways to eliminate duplicate data first at file level & then at chunk levels that reduces duplicate lookup overhead. In this survey paper we have discussed and describes about some of the deduplication strategies and some of storage systems like MAD2, Venti, HYDRAstor, Extreme Binning, Duplicate Data Elimination (DDE).*

Keywords- *Deduplication, storage system, MAD2, Venti, HYDRAstor, Extreme Binning, Duplicate Data Elimination (DDE).*

I. INTRODUCTION

Cloud computing is model of the distribution of the information services in which the resources are the retrieved from the web through some of the interfaces and applications, instead forming direct connections to the server. The fast expansion in information sources has mandatory for the users to make use of some of the storage systems for storing their secret data. Cloud storage systems provide the management of the ever-increasing quantity of data by keeping in mind factors like reduce occupation storage space and the network bandwidth.To make the scalable and consistent management of the data in the cloud computing, deduplication technique plays an important role.

Data deduplication is one of compression method which is mostly used for deleting the repeated copies of files or data by keeping unique copy in storage system to reduce the space occupation. Data deduplication also helps to improve the results in efficiency term and searches are quicker. Deduplication identifies the files whichever replicated from the data repository or from the storage systems and explicitly it uses the "reference pointer" to find out inessential chunks; this is also known as the storage capacity optimization.Data deduplication may happen as file level de-deduplication or as block level data de-duplication. Instead of maintaining numerous duplicate copies of file or the data with alike content, de-duplication senses and remove the redundant data by keeping original physical copy.

There are many cloud storage services are in existence such as, mozy, dropbox and the memopal, they have been applying the deduplication strategies for the user's data. Outsourcing raises the issues like security and the privacy. And deduplication is designed in consideration of these factors which improves space of storing the data as well as improves the network bandwidth ability and is attuned with the new functionality of the convergent key management. From securing data from the unauthorized user Proof of protocols are used.

In this survey, we first introduce to the deduplication concept then the deduplication three important layers in section 2 ,proceeding further we have discussed about the implementation of the data deduplication i.e. strategies of the data deduplication such as data unit based (Fie level deduplication and block level deduplication), location based (source based deduplication and target based deduplication) and then the disk placement deduplication (forward reference deduplication and the backward reference deduplication) in section 3 and in part 4 we describe the current status of the storage system, in section 5 we have concluded i.e. summary of this survey paper.

II. THREE WAYS OF DEDUPLICAITON

Deduplication mainly performed into 3 ways i.e. it is also called as main layers of deduplication and those has mentioned below:

2.1. Chunking

During the data deduplication implementation, technology or the knowledge varies primarily into this method i.e. chunking method or in its architecture. In few of systems, chunks are defined by the physical layer constraints and in some of systems only entire files are compared, which is called as

single instance storage or SIS. One of the smartest methods to chunk is usually sliding block, where the window size is passed along with file to divide files stream to look for more obvious happening internal file limitations.

2.2. Client Side Backup Deduplication

In this kind of deduplication, the deduplication hash calculations are first generated on the source i.e. client machine as its name suggest, and files that have indistinguishable hashes to the files already in the target device are not yet sent. Target devices generate inside links references to do away with the repeated copies of files or the data. Benefit of this type of deduplication is that it avoids unnecessary transmitting of the files and data across the network thus dropping the traffic stress.

2.3. Primary storage and secondary storage:

Primary storage systems are developed for the finest performance except than the lowest possible cost. Furthermore primary storage systems are less tolerant of any of the application or operation and higher in cost that can negatively impact performance. So while developing secondary storage two things are to be considered: best performance and the lowest possible cost. Also, the secondary storage system contains primarily replicates and the secondary copies of the data. These copies of the data are not naturally used for the actual construction operations so they have more tolerant of some performance degradation, in switching for the efficiency increasing. Whenever data is sent over the network, it mainly concerns about the potential loss of the data. And by data deduplication systems, they store data differently from how it was written. As a result, vendors are concerned with the purity of their ambiguous data. Though, the integrity of the data will eventually based on the structure of the deduplication system, and the superiority used to develop those algorithms.

III. IMPLEMENTING DATA DEDUPLICAITON

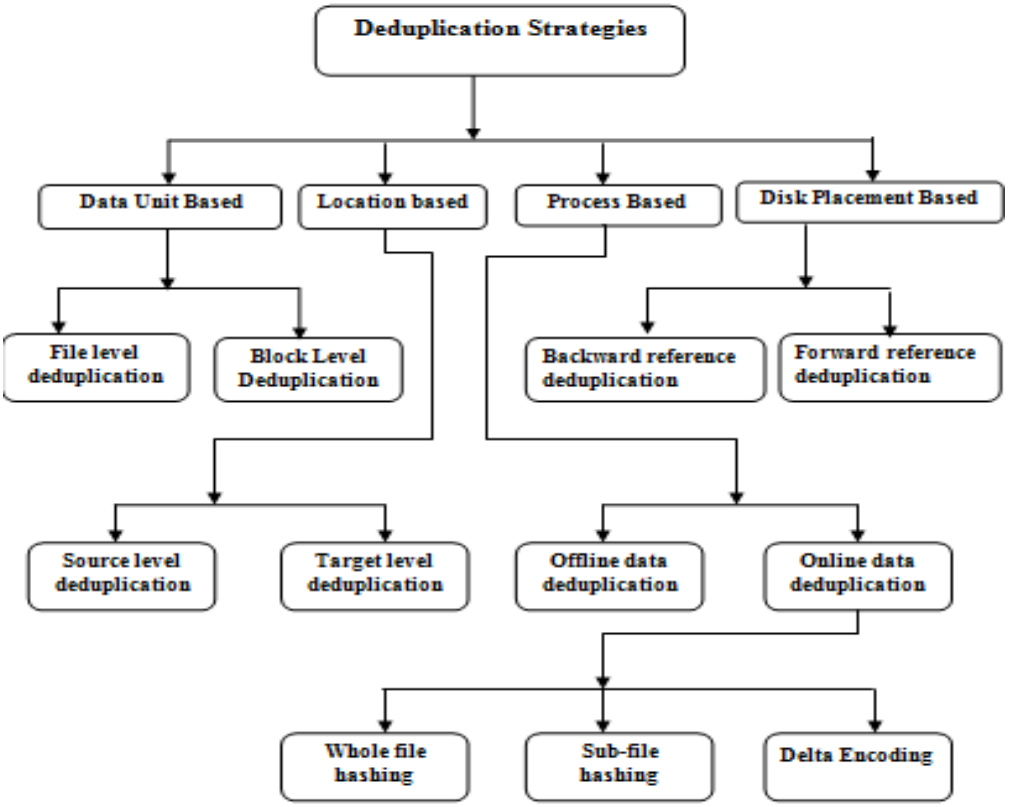

FIG.1.DEDUPLICATION STRATEGIES

The first and the foremost requirements are to identify the basic entity of data upon which the deduplication can be performed.

Strategies are explained in detail:

3.1. Data unit based deduplication (Levels of Deduplication):

There are fundamentally two levels of deduplication:

3.1.1. File level deduplication

In this level, deduplication performed over single file and it eliminates the duplicate copies of the same file. The file checking function is based on their hash values. The hash numbers are comparatively easier to generate so it does not require more processing power. Based on those hash numbers duplicated files are identified. i.e. if two or more files having the same hash values, they assumed to have the similar contents and only one copy of file is to be stored. They Searches for any files that are exactly alike and stores only one copy, where placing 'pointers' in place of the other copies. It is said to be more efficient than no deduplication whatsoever. Even a single slight alter to the file will effect in an supplementary replica being stored.

3.1.2. Block level deduplication

In this level of deduplication, it firstly divides the files into the blocks and stores only one copy of each block. It may also use fixed-sized blocks otherwise variable-sized chunks.

They compute hash values for each block for examining duplication blocks and then it eliminates duplicate blocks of data that occur in non-identical files. As the name indicates it performed over blocks and Analyzes entire blocks of data and then Allows for granularity without being overly time consuming and resource-intensive. While compared it to with whole file, the block level deduplication eliminates the tiny unnecessary chunk of data. The equivalent deduplication algorithm is used by each one and all file system in block level deduplication.

3.2. Location based Deduplication(Target area based Deduplication):-

Data deduplication can be further separated into target based and source based types:-

3.2.1. Target based de-duplication

Here Deduplication is performed on the target i.e. server side data storage center. In this case of deduplication the client is not modified and is totally not knowing or aware of any deduplication mechanism happening at the target side. Target-based deduplication accomplished by the backup servers and hardware appliances at target which having ability to perform and handle those deduplication activities. That means there is no overhead on the client or server being backed up. Thus the result of this kind of technology i.e. target based deduplication drastically raise the storage utilization, but at the same time it fails to meet bandwidth saving perspectives.

3.2.2. Source based de-duplication

In this scenario deduplication performed on client side i.e. before it transmitted By processing the data before transmitting we can reduce the transmitted amount of data and therefore it reduces the network bandwidth and this less bandwidth is required for the backup software.

Deduplication on source side uses the engine at client side which checks for the duplication against the deduplication index which is located on the backup server. This is done with the help of the backup agent who is aware of the deduplication which is located at the client side and who is responsible for backs up only unique data or blocks. And those unique blocks of data will be transmitted to the disk. The result of this kind of technology i.e. source based deduplication improves bandwidth as well as the storage utilization.

3.3. Process Based deduplication

The point at which the deduplication algorithms are performed is called as the deduplication timing. the timing of that algorithm always means a enormous constraints on how greatly time it has to achieve data deduplication and the stage of information the algorithm know about the original file information. With

reference to the timing, the deduplication circumstances can be divided into two types: offline deduplication and online deduplication.

Data deduplication process consisting two types of process:-

3.3.1. Offline data deduplication

This type of deduplication comes into picture when the data deduplication is performed offline that means all the data first written into the storage disk or data center and then the deduplication is carried out later. The major advantage of this orientation is that the system has a static view of the whole file system when the deduplication style is carried on, and it has a full knowledge about all the data it has access to and can significantly improve the deduplication efficiency and effectiveness. But it slows down the performance. Another time, the data written to the data center or storage disk must be framed until then scheduled deduplication time. This discovers that it cannot be a dispersible lag between when the data is written to disk and when space is recreated by cut out the duplicated data.

3.3.2. Online data deduplication

In this type of deduplication, the deduplication process is performed before data is being stored to the storage disk or to the data center. The key reward of this type of deduplication process is that this allows for the instant space recovery. But this leading to boost in the write intermission, as the write is blocked until all unnecessary file data is eliminated. There are numbers of approaches that have previously been organized which helps in deduplication technique once the timing of deduplication has been set. The most commonly used are WFH-whole file hashing, DE-delta encoding, and SFH-sub file hashing.

3.3.2.1. Whole File Hashing

As its name specify i.e. whole file hashing is applied to the whole file. The methodology used here is the hashing function, named MD-5, RC5 or SHA-1. The outcome of this functions is a cryptographic hash which forms the basis for the classification of the entire replicate files

Advantages:

1. Fast execution
2. Little metadata overhead
3. Low computation

Disadvantage: Rising granularity of replicate files.

3.3.2.2. Sub File Hashing

In this kind of deduplication, initially whole file is divided into numbers of small piece before actual deduplication takes place. The file separation mainly depends on the type of SFH i.e. sub file Hashing that fundamentally have the two type named :

1. Fixed-size chunking
2. Variable-length chunking.

In fixed-size chunking, the file is divided into a number of sections that may either stable or the fixed sized block i.e. chunks.
In variable-length chunking, the file is divided into the blocks i.e. chunks of varying length unlike another type of chunking named fixed size chunking.

3.3.2.3. Delta Encoding

The term Delta encoding mainly derived from the scientific and the mathematical symbol named "Delta symbol" .In this area delta is essentially used to compute the change or the rate of change in an object. It is also used to calculate the difference between the source object and the target object. Generally it is used when Sub File Hashing doesn't generate outcome except there is a tough enough match between two chunks that storing the differentiation would take fewer space than storing the no duplicate block

3.4. Disk placement based deduplication

In this kind of deduplication, deduplication is categorized into the 2 types:backward reference deduplication and forward reference deduplication planed on the how data is placed in the disks.

3.4.1. Backward reference deduplication

The current needless data blocks pointers are pointed backward to the earlier original data blocks.

3.4.2. Forward reference deduplication

All the earlier identical data blacks are pointed to the current or recent redundant data blocks to maintain it wholly. It also provides the fastest read performance on recent redundant data and also gives the more division on older data blocks which causes more index and metadata update operation that leads to the system performance. Because of this reason most of the traditional systems were based on the approach i.e. backward reference approach.

IV. DE-DUPLICATION STORAGE SYSTEM

` There are few deduplication storage systems that are being used with respect to the different storage purposes. Some of them have mentioned below:

4.1. Venti:

Venti is type of storage system that is basically referred by the network. The deduplication method in the network is uses the identical hash values of the data chunks for identification those hash values, so that it decreases the overall usage of the storing space .Venti follows the "write once" policy to avoid the collision of the data and this storage system is frequently related with the generation of chunks for vast storage applications

Disadvantages:

1. Not suitable to deal with a huge amount of data
2. Not offer scalability.

4.2. HYDRAstor:

This is secondary storage solution, which offers the decentralized hash index for the grid of storage nodes, which proceed as the back end and and a conventional file interface as the front end.

The back end HYDRAstor is brilliant to manage huge scale, changeable size and static addressed ,unchallengeable and extremely flexible data chunks with the help of the Directed Acyclic Graph.

The deduplication is detected based on the hash table. The eventual target of this type of storage system is to form a backup system

Disadvantage:

It ignored multiple users sharing files request.

4.3. Extreme Binning:

This kind of storage systems focus on non conventional workload. This storage system provides scalability in support of corresponding deduplication approach.

It's includes composition of low-locality individual files. The similarity of files gets priority over locality & permits one disk access for blocks lookup for file. This technique includes sorting of similar data into bins & removing duplicate chunks inside each bins. In this way deduplication achieved with respect to the different bins. In addition to this extreme binning keeps only primary index in memory that reduces RAM consumption.

4.4. MAD2:

MAD2 provides key feature i.e. its accuracy. It is an exact deduplication network backup services and not for the pure storage system, which often works on the both deduplication levels namely file level and the chunk (block) level. In order to achieve the desired performance they follows some of techniques - a hash bucket matrix, a distributed Hash Table based load Balancing.

4.5. Duplicate Data Elimination (DDE):

Duplicate Data Elimination supports both the combination of the content hashing then copy on write and lazy updates to identify and merge the indistinguishable data blocks in the SAN (storage area network system).The core difference between DDE storage system and other storage systems is that it accurately deduplicate and analyzes the analogous hash values of the blocks of the data right at the source side itself, before the actual transmission of the data. This kind of system always works in background.

V. CONCLUSION

This paper is mainly focuses on the data deduplication strategies and terminologies with respect to the some of the storage systems. Data deduplication technique is new trend in market for data compression. It is also said to single instance storage and one of the intelligent compression techniques. The main aim of data deduplication is removing repeated data and keeping single unique copy off data files. Data deduplication leads to reducing the storage utilization and improves the network bandwidth which is required for transferring the data.

References

[1] Qian Wang, Cong Wang, Jin Li, KuiRen, Wenjing Lou: Enabling Public Verifiability and Data Dynamics for Storage Security in Cloud Computing. ESORICS 2009:355-370.
[2] A Study of Practical Deduplication Dutch T. Meyer *† and William J. Bolosky * * Microsoft Research and † The University of British Columbia {dmeyer@cs.ubc. edu, bolosky @microsoft .com.
[3] K. Jin and E. Miller. The effectiveness of deduplication on virtual machine disk images. In Proc. SYSTOR 2009: The Israeli Experimental Systems Conference.
[4] Harnik, D., B. Pinkas and A. Shulman-Pelge, 2010 side chaneels in cloud services : Deduplication in cloud storage. IEEE Security Privacy, 8: 40-47 Zhu, B., K. Li and H. Patterson, "Avoiding Disk bottleneck in the data domain deduplication file system. Proceedings of the 6th USENIX Conference on File and Storage Technologies, February 26-29, 2008, San Joes, CA. USA., pp: 269-282.
[5] D. T. Meyer and W. J. Bolosky, "A Study of Practical Deduplication," in Proceedings of the 9th USENIX Conference on File and Storage Technologies (FAST'11), 2011, pp. 1-14.
[6] Deepak Mishra ,Dr. Sanjeev Sharma-Comprehensive study of data de-duplication International Conference on Cloud, Big Data and Trust 2013, Nov 13-15, RGPV