ISSN(Online): 2320 9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

*(An ISO 3297: 2007 Certified Organization)*

**Vol. 3, Issue 10, October 2015**

# De-Duplication on Cloud – Survey

S.Charumathy[1], A.Lalitha[2]

[1] II year M.E Student, Department of CSE, Valliammai Engineering College, Kattankulathur, India

[2]Assistant Professor, Department of CSE, Valliammai Engineering College, Kattankulathur, India

**ABSTRACT:** Now a days cloud has a rapid increase in demand for storage of data. So here is need for efficient storage mechanism. There are different de-duplication techniques to remove repeated data in cloud. This survey paper briefly explains the available de-duplication techniques and there drawbacks.

**KEYWORD**: storage, de-duplication

## I. INTRODUCTION

The user of cloud needs quick asses to the information stored in it. De-duplication is one among such technique to produce immediate result. De-duplication is a data compression technique where replication of data is prohibited. This technique is also called storage optimization. When there is a growth in the amount of information within the organisation, there is chance of repeated data storage. In future the volume of human generated digital information will decrease and the volume of automatically generated information will be high.

De-duplication has three advantages when compared to other techniques.

- **Storage** –storage can be reduced by eliminating repeated files.
- **Backup** - It can also make the service provider to reduce the number of copies of same data to be back up.
- **Network Traffic** – De-duplication will reduce the traffic flow in the network while transferring of same data many times to storage.

There are two ways of doing de-duplication

1) **Chunking:** *C*hunks compare only some files completely, this is called single instance storage [3]. Sliding block method is generally used in chunks. There are different types of chucking based on size of chuck namely variable and fixed size chucking. Sometimes both can be used as a mixture.

2) **Client backup de-duplication:** The hash function is used in this technique. The file with identical hash value same as the target is not sent to the server. This has a benefit of reduced traffic load.

Providing data security is the goal of a secure de-duplication without lack in the space efficiency [4].The encryption of files is done through keys. Then by matching the encrypted data the identical contents are eliminated. The same content encrypted by different key results in different cipher text, this is a drawback of using traditional encryption technique. This paper will narrate some of the de-duplication concepts.

## II. DE-DUPLICATION STUDIES

### A. De-duplication Techniques:

File objects are created using the convergent encryption and stored in the backup.Directory objects are stored in directories. These objects contain metadata and encryption for each child. Backups are stored in local disk and not stored in server. Backups are vulnerable to theft because they are not encrypted [1]. Personal data are not encrypted, so

ISSN(Online): 2320 9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 10, October 2015

it can be viewed by others [2]. Only user files are taken backups, because the system files can be recovered easily. Unique key for a block of data can be generated using hash function. Same key is generated for identical data. Convergent encryption [5] uses function similar to hash function which will identical data block that can be de-duplicated as normal. Now there will be separate encryption key and some mechanism is needed to access the data. Apart from convergent encryption, per user encryption can be used. The user needs to record key for each file. These keys can be separately stored in backup system and also file is encrypted with this key. For distributed environment the backups are at remote sites. In such cases the redundancy can be avoided by using Prompt Redundancy Elimination technique (PRUNE) [6]. The PRUNE will generate summary information for each chuck to speed up the process of comparing the chucks. The fingerprint is generated using Fingerprint Generator for each chunk. The hash function that can be used are SHA-1[7], MD5 [8], or SHA-256 [7] can be used to generate the fingerprint. Managing the fingerprint table efficiently is the success factor for de-duplication. For the purpose of managing the fingerprint table LRU-based index partitioning scheme can be used. The responsibility of Backup Generator is to arrange a sequence of chunk information with the metadata.

The other method of de-duplication is Post Process (PP); in this the new data is stored on the device and then being processed in the later stages. It will also analyze the duplicate data. The advantage of using this is that there is no need to wait for hash function calculation and lookup before storing. Thus the performance of the store is not degraded. Policy based operation offers the user the ability to optimize on running file or it can be based on the type and location. But the drawback is that there is a need to store duplicate data for a short duration, which can be a issue if the storage is nearing to be filled.

The often used method for de-duplication is the In Line (IL) de-duplication. In this method the hash value is calculated in the target device, which is discussed later. In this way the data enters the device in real time. When the device detects the duplicate file it does not store it, instead it refers to the block that already exists. The benefit of IL is that it requires less storage as it does not allow duplicate data. The disadvantage in IL is that it reduces the throughput of the device, because of hash value calculation and lookup takes more time to compute. But some vendors have proven the similar performance in IL as that of PP technique.

Another way of classifying data de-duplication is from where data occurs. When the de-duplication is done in the client side it is called “source de-duplication”. When the de-duplication is done in the client side it is called “target de-duplication”.

Generally Source de-duplication is executed directly within the file system. Then the system will scan periodically the new file’s hash value with that of existing value. When the same hash value is found then the new one is removed and it is made to point to the old file. The process of de-duplication is a transparent technique to both users and backup application. Backup file system will the bigger than the source data. This is because backup is a duplicate file.

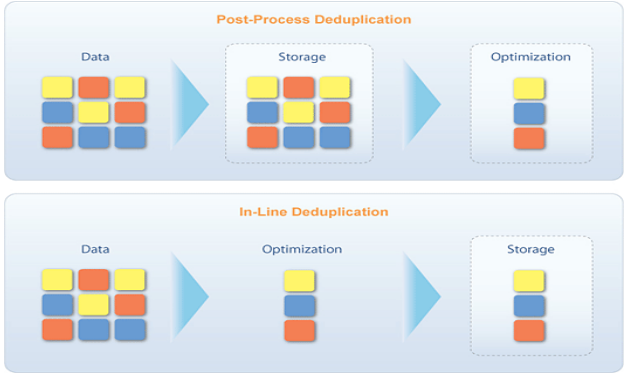


Fig 1 PP & IL de-duplication

ISSN(Online): 2320 9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

*(An ISO 3297: 2007 Certified Organization)*

**Vol. 3, Issue 10, October 2015**

The technique of removing duplicate data when it is not generated at that same location is called target de-duplication. The server will not know the de-duplication process. For example backup system with de-duplication. The backup can be like data repository.

### B. Draw backs in de-duplication techniques:

Data loss occurs during data transformation. Data de-duplication by definition writes data differently from how it is stored. So the user concern is more towards data integrity. All de-duplication technique discussed so for varies slightly from each other. The integrity of the data depends on the technique used and the quality with which it is implemented. Due to the technological growth, the integrity of the products has been proven.

One of the techniques depends on cryptographic hash function to detect the duplicate data. The term collision means if two different pieces of information generate the same hash value. The hash function used will decide the probability of collision, and even thought the probability is smaller, they are non zero values. Therefore there is a chance of data corruption if collision occurs. Addition verification needed to check the difference in the data. Both IL and PP offer bit-by-bit validation of original data, which guarantees the data integrity. The use of algorithm like SHA-1, SHA-256 and others, will reduce the probability of data loss and the risk of both uncorrected and undetected errors.

The intensity computational resource of the process can be issue in data de-duplication. But, this is a rare issue in a standalone appliance where the computation is done away from the original system. When the technique of de-duplication is incorporated within the device, this can be an issue. Both weak and strong hash value must be utilized for improved performance of the system. Weak hash has a greater risk of collision but it is faster to compute. The system that uses weak hash will also use the strong hash. The strong hash will determine the factor whether the data is actually same or not. There are overhead associated with calculation and lookup of hash value. The reassembly of data chuck is likely to impact the performance of the application.

Another area of concern is against primary storage, where de-duplication has a related effect on snapshots, backup, and archival. When reading the file from storage device there is a need to reconstruct the file. So this makes the secondary copy of any data to be larger than that of primary copy. The post de-duplication will preserve the entire original file, while the snapshot is taken prior to de-duplication. This makes the snapshot to consume more space than the primary copy.

Another major concern is the effect of encryption and compression. Even though de-duplication is a version of compression, it works in strain with traditional compression. The efficiency against the smaller chunk is achieved by de-duplication, but the efficiency against the larger chunk is achieved by compression. The encrypted data cannot be de-duplicated even though the content is redundant, because encryption is used to eliminate apparent pattern of data.

The biggest challenge for de-duplication is scaling. The de-duplication is affected only when multiple disk are being used. A de-duplication technique that is shared across devices will preserve space efficiently.

### C. Thinks to be remembered before data de-duplication:

1. **De-duplication ratio must be low.**

The effectiveness is measured in terms of ratio for de-duplication. Even though higher degree of ratio conveys more de-duplication they will mislead the technique. It is not possible to compress a file 100%.

2. **CPU intensive De-duplication.**

The hash function used find hash value, which when compared detects duplicates. All these hashes are CPU intensive. This will not affect if the de-duplication is done outside the source. It gets affected if source de-duplication is done on the production server. Therefore it affects the server's performance.

ISSN(Online): 2320 9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

*(An ISO 3297: 2007 Certified Organization)*

**Vol. 3, Issue 10, October 2015**

**3. Initial de-duplication does not save more space.**

Only PP de-duplication which occurs on secondary storage like disk is used in disk-by-disk-backups. In such architecture the data is initially stored in a normal format; later the de-duplication process is performed. Due to this process of work there is no initial space saving. This spacing depends on the type of software used.

**4. Rarely possible Hash collisions.**

The hash collision as discussed earlier depends on the strength of the algorithm used; sometimes dissimilar file of data can result in identical hashes. This is because weak hashing algorithm will not potentially identify the duplicates. Hence it has to be rehashed using strong hashing algorithm.

**5. Avoid multiplexing.**

There are many chances of using multiplexed data. It is recommended to use virtual tapes to store data instead of physical tapes. This is because it can occupy more space, so it can be prevented from wastage.

**6. Compare with other's de-duplication methods.**

Examine other's de-duplication ratio also in order to find the efficiency of yours de-duplication. The initial de-duplication ratio will be low and it grows rapidly over time. There is need to periodically check the method significantly as negative de-duplication ratio leads to something which is wrong.

**7. Do not encrypt data before de-duplication.**

As discussed earlier encryption is a technique to remove apparent pattern from the data. Hence when data are checked for de-duplication the encrypted data cannot be check because two dissimilar data can have same encryption value, or it may lead an attacker to track out the data.

**8. Do not compress before de-duplication.**

There are two reasons to do de-duplication before compression of data. The first reason is the de-duplication will automatically compress the data file. The other reason is that compression may scramble the data, while looking for pattern in the data.

**9. Learn what data de-duplicates well.**

The data that is generated by human will support de-duplication. Whereas data that are generated by system or other automatic generation will not support for de-duplication. Consider such type of data to be stored in non de-duplication area.

## III. CONCLUSION

This survey paper has compared all the de-duplication techniques and it has also analysed the advantages and disadvantages of the technique discussed. In addition it has also included the thing to be remembered before de-duplication of a data is processed.

## REFERENCES

1. P. Anderson and L. Zhang, ''Fast and Secure Laptop Backups with Encrypted De-Duplication,'' in Proc. USENIX LISA, 2010, pp. 1-8.
2. Backblaze: online backup.http://www.backblaze.com/.
3. Jiansheng Wei,1Ke Zhou, 2Lei Tian, 1Hua Wang, Dan Feng, ''A Fast Dual-level Fingerprinting Scheme for Data Deduplication''
4. Mark W. Storer Kevin Greenan Darrell D. E. Long Ethan L. Miller," Secure Data Deduplication"
5. J. R. Douceur, A. Adya, W. J. Bolosky, D. Simon, and M. Theimer, "Reclaiming space from duplicate files in a serverless distributed file system," in Proc. Int. Conf. Distrib. Comput. Syst., 2002, pp. 617–624.
6. IEEE TRANSACTIONS ON COMPUTERS, VOL. 60, NO. 6, JUNE 2011 Efficient Deduplication Techniques for Modern Backup Operation Jaehong Min, Daeyoung Yoon, and Youjip Won
7. J. Burrows and D.O.C.W. DC, "Secure Hash Standard," Federal Information Processing Standards Publication, Apr. 1995.
8. R. Rivest, "The MD5 Message Digest Algorithm, RFC 1321," Internet Activities Board, 1992.

ISSN(Online): 2320 9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

*(An ISO 3297: 2007 Certified Organization)*

**Vol. 3, Issue 10, October 2015**

9. Peter Christen."Probabilistic Data Generation for Deduplication and Data Linkage", http://datamining.anu.edu.au/linkage.html.
10. Mark W. Storer Kevin Greenan Darrell D. E. Long Ethan L. Miller," Secure Data Deduplication".

## BIOGRAPHY

**S. Charumathy** received Bachelor degree B.Tech Information Technology from Adhiparasakthi Engineering College, Anna University, Chennai. She is now pursuing Masters Degree M.E Computer Science and Engineering department at Valliammai Engineering College, Anna University, Chennai.

**A. Lalitha** received Bachelor degree B.E Computer Science and Engineering from Mookambigai Engineering College, Pudukotttai, and Masters Degree M.Tech Embedded System from Anna University, Guindy. She is now working as Assistant Professor in Computer Science and Engineering department at Valliammai Engineering College, Anna University, Chennai.