A Survey: Deduplication Ontologies

Sulakshana S.Patange
PG Scholar
JSPM Narhe Pune

Kanchan Varpe
Lecturer
JSPM Narhe,Pune

ABSTRACT
Cloud computing drive out the need of IT based companies to invest in high computing infrastructure and services used by them. In cloud, the data is dwell into set of networked resources that enable data to be accessed via virtual machines. These data centers are located in various parts of the world beyond the control and reach of the user, so there are multiple challenges and security issues that need to be addressed and understood. This review paper aims to analyze and elaborate deduplication issues in cloud computing which is the base for our future roadmap.

Keywords
Deduplication, Ontologies,

1. INTRODUCTION
The Cloud computing is a model to provide access to computing resources and applications available on the internet Cloud computing platform provide network resources and storage space to the remote users. The user can access information at anytime from anywhere via internet. So data and the user need not to be on same physical location and user do not need to manage the actual resources. Cloud computing enables users to access shared resources which provide services as per user requirement over the network to perform operations. Users can develop their own application and deploy on the cloud and also manage it with the help of cloud computing services. IT industry has widespread use of cloud computing, companies like Microsoft, IBM and Google deliver their services using cloud to its users. Cloud computing is highly scalable as its environment provides services to users on a large scale. Because of its availability it increases response time, which results in high performance.

Cloud computing era have lots of research issues. One of them is Deduplication. Data deduplication is compression technique which identifies duplicate data, eliminate duplicate copies of repeating data and reduce the data that needs to be physically store. Reduction in Storage Allocation and Efficient Volume of Replication are two big advantages of deduplication over a normal file system.

During this review, we first present the basic three layers of deduplication in section II . In section III, we briefly review the different strategies of data deduplication such as process based deduplication, operation based deduplication, target area based deduplication, disk placment based deduplication. and in section IV , we define Existing methodologies of deduplication. Section V descibes current status of deduplication by adding examples and product review. Section VI describes the summary of this review paper.

2. BASIC THREE LAYERS OF DEDUPLICATION
Deduplication commonly performed in 3 ways is described below:

2.1 Chunking
During the implementation of deduplication, primarily it divides into the chunking method or in architecture. Chunks are defined in physical layer constrained or in few systems whole file is compared which is called as single instance storage .The most famous technique for dividing files into chunk is sliding block, In which window size is passed along with file to divide files into chunks.

2.2 Client Side Backup Deduplication
In this, deduplication hash calculation are on source machine. Target machine creates internal links references to remove duplicate copies of data. The advantage of this is , it avoids unnecessary sending of data over the network.

2.3 Secondary Storage and Primary Storage
High cost and Less tolerant of any operations are drawbacks of primary storage. So while developing secondary storage two things are considered : Optimal performance and lowest possible cost. Secondary storage systems contains secondary copies of data. these copies are not actually used for operation. so the problem of tolerance is avoided here.

3. STRATEGIES OF DATA DE-DUPLICATION
3.1 Process based deduplication
Process based deduplications are classified into Online Data de-duplication and Offline Data de-duplication.

- **Offline Data Deduplication:** In this offline data deduplication technique, data is initially written to the storage disk and then moves forward for the deduplication process. The process of deduplication is always at the rear of the data writing process[1].
- **Online data Deduplication: In** this, repeated copies of data is deleted then it will written to the storage disk.

The data deduplication timing are decided, and then the prior existing techniques can be applied on it. Some of the often used data deduplication approaches are: Whole File Hashing, Sub File Hashing and Delta Encoding.

- **WFH (Whole File Hashing)**

Whole File Hashing function is applied on whole file. Some of the cryptographic hash functions are MD5,SHA-1 or RC5. These functions are used to find out whole replicate files. Advantages of this hashing are listed below:

- ✓ Minimum metadata overhead
- ✓ Speedy with low computation

- **SFH (Sub File Hashing)**

In this deduplication technique, File is divided into number of smaller sections before data deduplication check. SFH is categorized into fixed size chunking and variable size chunking[2].

✓ Fixed size chunking - File is divided into fixed size chunks.

✓ Variable size chunking - File is divided into variable size chunks.

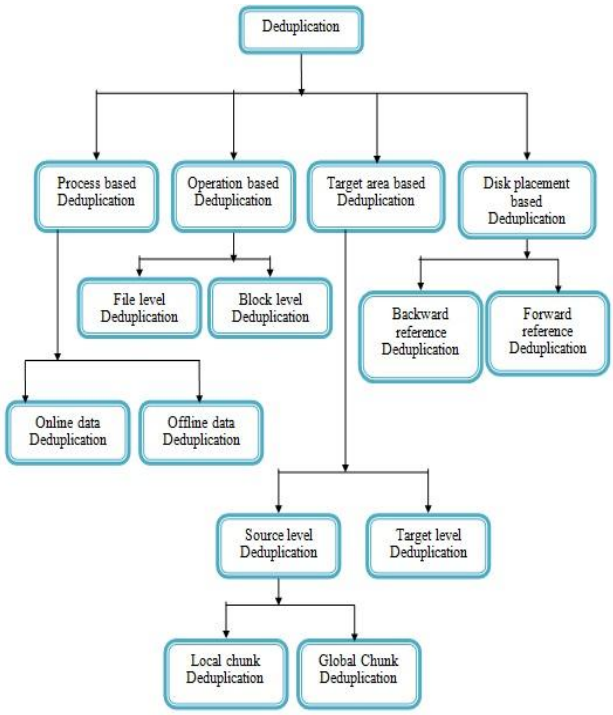
- **DE (Delta Encoding)**

This approach is designed from the mathematical term "delta". This stands for "change" or "rate of change" in an object. DE is used to show the difference between the target and source object. Normally it is used when SFH does not produce results but there is a strong enough similarity between two items/ locks / chunks that storing the difference would take less space than storing the nonduplicate block..

3.2 Operation based deduplication

Data deduplication can be categorized into File level Deduplication and Block level Deduplication[3].

- **File level Deduplication :** In File level Deduplication, the whole file is compared with other identical files. If match found, appropriate action performed to remove duplicate copies.
- **Block level Deduplication :** It performs over blocks. Primarily files are divided into blocks and store a single copy of block. These blocks are further divided into variable size blocks or fixed size blocks.



3.3 Target area deduplication

Data de-duplication can be categorized based on their location: Target based deduplication and Source based Deduplication [4].

- Target based deduplication
- Source based deduplication

These two deduplication are explained in following Table 1.

Type of Deduplication	Deduplication Performed on	Description	Advantages	Disadvantages
Target based deduplication	Target data storage center	In this case the client is unmodified and not aware of any deduplication.	Improves storage utilization	Does not save bandwidth
Source based deduplication	At the source data storage before it's transmitted	A deduplication aware backup agent is installed on the client which backs up only unique data.	Storage efficient, Increased bandwidth	Enforces extra computational load on the backup client. Replicates are changed by pointers and the actual replicate data is never sent over the network.

Further Source Deduplication approach [5] are categorized into Local and Global chunk level deduplication. In Local chunk level deduplication [4] , repeated data are eliminated before sending them to the remote destination within the same client. In Global chunk level deduplication, repeated data are removed globally across different clients.

3.4 Disk placement based deduplication

Data deduplication methods are categorized into backward refernce deduplication and forward reference deduplication, based on how data is placed in disks.

- **Backward refernce deduplication :** The recent unnessary data chunks pointer are pointed backward to the previously original data chunks.
- **Forward reference deduplication :** All the previously identical data chunks are pointed to the recent redundant data chunks to maintain entirety.

The Forward reference method provides fastest read performance on the recent redundant data, and it also provides more fragmentation on older data chunks which causes more index and metadata update operation that leads to reduce the system performance. So the most of traditional system were based on backward reference deduplication approach. For eg : VM Servers[6].

4. EXISTING METHODOLOGIES OF DEDUPLICATION

Recently Deduplication motivated to the most of researchers due to its popularity. These researchers works on different aspect of deduplication : Semantic attribute based source deduplication, Hadoop based deduplication[7], GPU based deduplication[8], SSD based deduplication[6], deduplication on storage disk[9], signature based deduplication, live deduplication[10], Image based deduplication[11], Scalable deduplication in VM image[12].

There are number of benefits of data deduplication are considered. Few of them are described below:

- Storage based deduplication [13]
- Network data deduplication [14]
- Virtual servers help from deduplications.

5. EXAMPLES AND PRODUCT REVIEW OF DE-DUPLICATION

There are different storage system which are used for different storage purpose. For eg: MAD2, Extreme Binning, Venti, Duplicate Data Elimination, HYDRAstor[15][16] etc. Some of them are listed below:

- MAD2 : This is network backup deduplication service which works at block level and file level both. To achieve the high performance, it focuses on Hash table based load balancing, hash bucket matrix, bloom filter array, dual cache.
- Extreme Binning : It focuses on file similarity instead of locality. It allows single disk access for blocks lookup per file. This technique arranges similar files into bins and removes duplicate pieces/chunks inside each bin.
- Venti :Venti is a network storage system. It applies similar hash values to find the block contents. It generated a block for larger application and apply write once policy to avoid crash of data. This works in the early stage of storage so it is not suitable for large data application. and one of the major drawback is, system is not scalable.
- Duplicate Data Elimination (DDE): It works in the background. It applies the combination of content hashing. DeDu properly de-duplicates and calculates hash values at the source side before it is transmitted.

There are multiple products are available in market for deduplication. Few of them are listed in following table 2.

	Vendor	**Hardware or software**	**Algorithm used**	**Inline or Offline**	**Source or Target**
1	FalconStor	Both	SHA-1 with optional MD5	Offline	Target
2	HP	Hardware	SHA-1	Inline	Target
3	IBM/Diligent	Software	Custom	Inline	Target
4	Symantec	Software	SHA-1	Inline	Source
5	Copan	Hardware	SHA-1	Offline	Target
6	Data Domain	Hardware	SHA-1	Inline	Target
7	EMC	Hardware	SHA-1 and MD5	Offline	Target

6. EXAMPLES AND PRODUCT REVIEW OF DE-DUPLICATION

There are different storage system which are used for different storage purpose. For eg: MAD2, Extreme Binning, Venti, Duplicate Data Elimination, HYDRAstor[15][16] etc. Some of them are listed below:

- MAD2 : This is network backup deduplication service which works at block level and file level both. To achieve the high performance, it focuses on Hash table based load balancing, hash bucket matrix, bloom filter array, dual cache.
- Extreme Binning : It focuses on file similarity instead of locality. It allows single disk access for blocks lookup per file. This technique arranges similar files into bins and removes duplicate pieces/chunks inside each bin.
- Venti :Venti is a network storage system. It applies similar hash values to find the block contents. It generated a block for larger application and apply write once policy to avoid crash of data. This works in the early stage of storage so it is not suitable for large data application. and one of the major drawback is, system is not scalable.
- Duplicate Data Elimination (DDE): It works in the background. It applies the combination of content hashing. DeDu properly de-duplicates and calculates hash values at the source side before it is transmitted.

There are multiple products are available in market for deduplication. Few of them are listed in following table 2.

7. CONCLUSION

Many researchers have investigated on the efficiency and performance of deduplication. In this paper, surveys of various techniques for deduplication which are previously defined are discussed. Cloud backup architecture can be organized depending on data similarity for the efficiency of deduplication. Offline and Inline deduplication can bring improvement in terms of deduplication efficiency. Source level deduplication performs on client side before data is transmitted. This strategy saves bandwidth, reduces transmission cost, computation overhead, backup window size. The research challenge considered while performing deduplication at source side are reading and writing efficiency.

8. ACKNOWLEDGMENTS

Our thanks to the experts who have contributed towards development of the template.

9. 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

[5] Zhu, B. K. Li and H. Patterson, 2008. Avoiding the 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 Jose, CA., USA., pp: 269-282

[6] Mao, B., H. Jiang, S. Wu, Y. Fu and L. Tian, 2013. AR: SSD assited restore optimization for deduplication-based storage systems. Proceedings of IEEE 7th international conference on networking, Architecture and Storage, June 28-30, 2012, Xiamen, Fujian, pp: 328-337

[7] Sun, Z., J. Shen and J. Yong, 2011. DeDu: Building a deduplication storage system over cloud computing. Proceedings of the 15th International Conference on Computer Supported Cooperative Work in Design, June 8-10, 2011, Lausanne, pp: 348-355.

[8] Suttisirikul, K. and P. Uthayopas, 2012. Accelerating the cloud backup using GPU based data deduplication. Proceedings of the 18th International Conference on Parallel and Distributed Systems, December 17-19, 2012, Singapore, pp: 766-769.

[9] Guo, F. and P. Efstathopoulos, 2011. Building a high-performance deduplication system. Proceedings of the USENIX Annual Technical Conference, June 15-17, 2011, Portland, OR., USA., pp: 25-29.

[10] Lin, I.C. and P.C. Chien, 2012. Data deduplication scheme for cloud storage. Int. J. Comput. Consumer Control, 1: 26-31.

[11] Jin, K. and E.L. Miller, 2009. The effectiveness of deduplication on virtual machine disk images. Proceedings of SYSTOR: The Israeli Experimental Systems Conference, May 2009, Haifa, Israel, pp: 1-12.

[12] Zhang, J., S. Han, J. Wan, B. Zhu, L. Zhou, Y. Ren and W. Zhang, 2013. IM-Dedup: An image management system based on deduplication applied in DWSNs. Int. J. Distrib. Sensor Networks, 10.1155/2013/625070