Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 4, Issue. 9, September 2015, pg.338 – 344

REVIEW ARTICLE

A Review on Backup-up Practices using Deduplication

Mohini Vikhe[1], Jyoti Malhotra[2]

[12]Department of Information Technology, MIT College of Engineering, Pune, India

[1] mohinivikhe@gmail.com; [2] jyoti.malhotra@mitcoe.edu.in

Abstract— Deduplication is a technique of eliminating redundant copies of the data, which stores only unique instance for all the redundant data and creates a pointer to the unique data, stored on media. Now a day's, deduplication is needed to make efficient use of the storage space and to minimize the performance overhead for huge storage systems like e-storage. In day to day life; data is been stored on huge storage systems rather than on hard disk. For this purpose cloud computing and Hadoop are the boom words. As the data on these systems, are frequently updated and quick retrieval is also needed. In this paper, we have reviewed existing Deduplication taxonomies and how to make the process dynamic in nature. As deduplication provides only a single instance of data and pointer to other location for fault tolerance and reliability, replication of the server can be done. For storage efficiency replication can be done after deduplication.

Keywords— "Cloud Computing", " Cloud storage", " HDFS", " Deduplication", "Hadoop Distributed File System".

I. INTRODUCTION

Data deduplication is a technique to reduce storage space by eliminating redundant data in your backup system. Only one copy of the data is maintained on storage media, and duplicate data is replaced with a pointer to the unique data copy. Deduplication technology typically divides data sets in to smaller chunks and uses various hashing algorithms to assign each data chunk a hash identifier, which is compare to previously stored identifiers to determine if the data chunk has already been stored.

It is been studied that with Deduplication, can achieve up to 70% to 90% reduction in capacity of our backups. Deduplication technology offers number of benefits for storage and backup such as lower storage space requirements, more efficient use of disk space, and less data sent across a WAN for remote backups, replication, and disaster recovery. Deduplication can be performed either on source side or on target side. Figure 1 shows the basic idea of deduplication. If figure 1, data is shown with multiple colors, after applying deduplication method only one copy of data is been stored rather than multiple copies.

Data deduplication is a method to reduce storage space by eliminating redundant data from backup system. Single copy of the data is maintained on storage media, and duplicate data is replaced with a pointer to the unique data copy [1]. Deduplication technology typically divides data sets in to smaller chunks and uses hash algorithms to assign each data chunk a hash identifier that is fingerprint, which it compares to previously stored identifiers to verify if the data chunk has already been stored. New fingerprint is stored in database and metadata is also updated for future use.

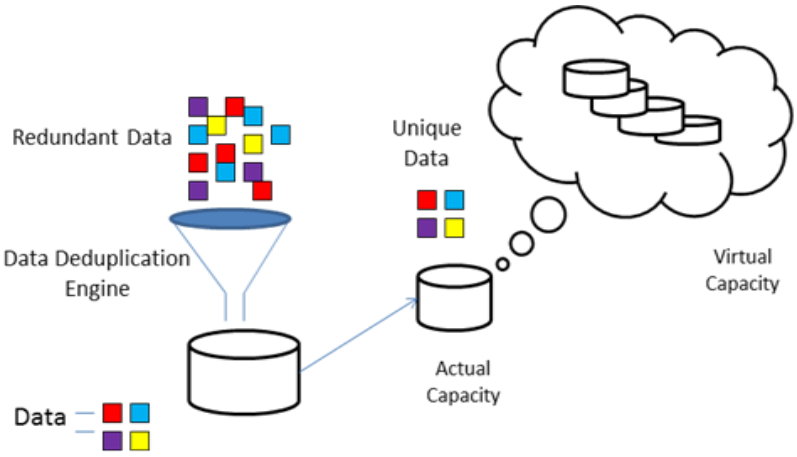

Fig. 1. Deduplication Process

Now a days cloud computing is mostly used for storage as well as also known as for dynamic service provider. Concept of the cloud computing is to give the services as per requirement of the users. It might be software or hardware depending upon the demand. It also aims for resource virtualization. Mostly used in business domains.

Cloud Computing gives one of the simple way to access servers, storage, databases and a broad set of application services over the Internet. Without installing applications, cloud computing allows consumers and businesses to use applications and also allows to access their personal files at any computer with having internet access. Main advantage of this it is much more efficient computing by centralizing data storage, processing and bandwidth. Example of cloud computing is Yahoo, Gmail, or Hotmail etc. Cloud computing is divided into three segments, application, storage and connectivity. This paper focuses on cloud-storage.

As cloud computing provides lot of services so has a huge amount of data and also the access to it. So that purpose deduplication can be applied to it. As per the today's user's needs only deduplication won't provide required throughput and reliability. It has to be integrated with dynamic deduplication.

Cloud storage is storing the data online on cloud. It can also give benefits such as availability, reliability, disaster recovery and reduce storage cost because no need to purchase and maintain expensive hardware. It also provides the security. Main advantage is for files stored in cloud storage are accessible from anywhere and at any time. Main focus is to dynamically deduplicate the cloud storage. The rapid growth of cellular advances, social media and requirements for data analytics has challenged the way of storing and processing the data for many large business entities. To solve the data storage and processing challenges and problems, organizations are starting to deploy large clusters of Apache Hadoop. It is a solution that makes use of parallel processing of large data sets (big data) and creating multiple replications of the data to avoid data loss.

Hadoop Distributed File System (HDFS) is a reliable and scalable storage system. HDFS is a file system, have many readers to a file but only a single writer at a time. The single writer can only update data to the files. Like in disk file systems, files in the HDFS are formed by blocks. HDFS usually replicates blocks to three Data Nodes to ensure the consistency and high availability of data.

In order to explore on the current trends of Deduplication, we have reviewed existing Deduplication techniques. Section –II illustrates this literature survey. Section III compares these techniques. Section –IV mentions the need of dynamism in Deduplication process followed by the conclusion in section V.

II. REVIEW DIFFERENT METHODS FOR DATA DEDUPLICATION

Data deduplication is a technique to remove redundant data either before or after backups. Deduplication reduces both inter file as well as intra file redundancy. Deduplication is a technique of eliminating redundant copies of the data, which stores only unique instance for all the redundant data and creates a pointer to the unique data, stored on media. Deduplication aims to reduce storage space, duplicated data chunks identified and store only one replica of the data in storage. Logical pointers are created for other copies instead of storing redundant data.

Data Deduplication process affects the parameters such as decrease in deduplication ration, improve back up window time, speed up database transactions, and reduce storage space.

Nagapramod et.al [1] focuses on the basic idea of Deduplication and tells us about the deduplication taxonomy: Placement; Timing [Synchronous / In Band or Asynchronous / Out of Band]; Deduplication Algorithms [Whole File Hashing, Sub File/Chunk Hashing].

Here client based deduplication is the deduplication performed on client side rather than on server side, then moved to server. This saves the network bandwidth. Deduplication Appliance consists of In-Band and Out-Of-Band. In-Band deduplication appliance finds duplicate data of arrived one, before writing it to the disk. While, Out-Of-Band performs deduplication after data has been written to the disk. Deduplication can be applied on file-level or sub-file level. File level uses various hash algorithms(SHA-1 or MD5) to obtain whole file hashing fingerprint of the file. Here complete file is matched with another one. Second is sub file hashing it divides the file in sub parts and then find the fingerprint of that. File can be divided into fixed size or variable-size sub-parts.

One of the another architecture for cloud backup services is Semantic-Aware Multi-tiered source de-duplication framework (SAM) [3], which combines global file-level deduplication and local chunk-level de-duplication. It also exploits file semantics (e.g., file locality, file timestamps, file size and file type), to achieve an optimal trade off between the de-duplication efficiency and de-duplication overhead to shorten the backup window. SAM is composed of three key components: (i) Virtual Full Backup (VFB), (ii) Global File-level De-duplication (GFD), and (iii) Local Chunk-level De-duplication (LCD). VFB removes the unchanged files from the backup process based on file timestamps, GFD remove global duplicate files across different clients and LCD removes the local duplicate chunks across similar files within the same client.

SAM architecture has three main parts, are

i. File Agent [Installed on client machine at the time of backup]
ii. Master server [Keep records of each chunk]
iii. Storage Server [Serves backup request from client]

For implementation purpose SAM uses:

- Virtual Full Backup
- Global File level Deduplication
- Local Chunk-level Deduplication

Virtual Full Backup has two phases: Incremental backup based on file's timestamp which leads to no deduplicate overhead. Second one is synthetic full backups. Global file level deduplication performed on Master Server. It compares file hashes of new file with existing stored file hashes. Local chunk level deduplication exploits the redundant data among similar files of same client. To reduce the disk access SAM uses two tire

chunk indexing approach. To identify duplicate chunk need to at least once access the disk. Author has also introduced small files and compressed files to reduce disk access.

Other architecture for cloud backup service is CAB [4] deduplication, here authors aim to focus on reducing the backup and restore time. For scalability they have designed architecture as client side deduplication.

CAB Client, CAB Server and Interface are the three main parts of the CABDedup technique. CABClient helps to process Deduplication at client side. It is composed of two main modules, the Causality-Capture module and Redundancy-Removal module. Causality-Capture module has File Monitor, File List and File Recipe Store, which monitors and captures the causal relationships of all files. Redundancy-Removal module removes unmodified data by using the causality information stored in File List and File Recipe Store with the help of backup and restore.

CABServer handles Deduplication process at the server end; which has File List Remote Store and File Recipe Remote Store as its components. It stores the file lists and file recipes received from CAB-Client that ensures the availability of the causality information updated by CAB-Client if CAB-Client's corrupts. Due to data transmission overheads, the file lists and file recipes stored in CAB-Client do not report to CAB-Server as soon as CABclient is updated; this may affect the performance as updated data may not be available with the Server.

As per client request the load balancer will hand over backup task to one of the deduplicators. This architecture compares the formed Deduplicate with the hash value if exist in metadata server. If not found, hash value is stored in metadata and then the file is backed up to file server, if found no need to store it, just gives pointer to that location.

Andre Oriani et.al [5] focuses on high availability of the backup data. Authors have used the concept of HDFS [Hadoop Distributed File System] and hot standby node as a solution over HDFS concept. It focuses on two main points i) High availability using Avatar Node, ii) Automatic failover mechanism using apache zookeeper. HDFS was used for replication and availability of data. Name node used by HDFS is to communicate between client and data nodes. Backup node plays an important role of deciding the checkpoint for name node. Back up node also stores transactional logs.

Main aim to develop backup node was to obtain high availability; to achieve this, Hot standby node was designed to fight over failover and high availability and it was able to do it in a short period of time. Author also promises that this concept can be used for replication of the backup systems.

In order to modify HDFS to hot standby node authors have implemented following steps:

a. *Extend the state already replicated by backup node* –Here, Replication of block locations is important for a fast failover. It can be achieved by two ways one is if any changes are made to replicas then name node should let know it to Hot standby node, an another is allow data node send their messages to Hot standby node.

b. *Build an automatic failover mechanism*–Here, when name node fails then; to maintain continuity, client with the help of zookeeper will be connected to any other server without affecting present client session. It is highly available to HDFS users and also maintains the data. Till the time name node is recovered zookeeper will handle its sessions. These background processes/changes are not noticed by user. Other way to implement this is by using virtual IP technique. Here failed name node's IP address will be transfer to hot standby node without knowing to client and client can continue the session without failure.

Nowadays we are also approaching towards online storage systems. Widely used tools are Google Drives or Droplet. Droplet[6] is a distributed deduplication storage system which aims high throughput and scalability. It consist of three main components, they are *Single* Metadata server, *Multiple* fingerprinting servers and *Multiple* storage nodes. Metadata server maintains the information related to storage server as well as of fingerprinting server. Fingerprinting server gets the data from client and it divides the data into several blocks and then finds their fingerprints. Each and every block is tagged with its fingerprint, later it is compressed and it is sent to deduplication queue. A process periodically collects the fingerprints from a queue and tries to match its fingerprint with the fingerprints stored on storage server. If match found then discard the data block and if match not found then store data block to storage server. Droplet uses block size of 64KB as fixed size which provides excellence I/O performance and good deduplication. For fingerprinting calculation droplet uses MD5 and SHA-1 techniques. Droplet compresses data blocks on fingerprinting server before sending them to data server, to reduce disk storage network bandwidth.

To summarize, we can quote that - Demystifying Data Deduplication gives the basic idea of Deduplication and where it can be applied and basic algorithms used. Problem of this method is same technique been applied over all types of data which increases backup window time. SAM: A Semantic Aware Multi-Tiered Source De-duplication Framework for Cloud Backup focuses on reduction in backup window time and restores time but disadvantage is it has high restoration time. Another method CAB dedup : A Causality-based Deduplication Performance Booster for Cloud Backup Services its advantage is it improves both the backup and restore performances, but it slow downs client system and if client backup is lost, then it is difficult to recover the same. From Backup to Hot Standby: High Availability for HDFS focuses on transferring HDFS to Hot standby node, even though transferring to hot standby node it was not able to support Fault tolerance.

III. COMPARISON AND ANALYSIS

Table 1.1 summarizes the comparison of various papers reviewed here.

Table 1.1 : Comparison of various Deduplication methods

Sr. No.	Reference Paper	Work Description	Remarks
1	Demystifying Data Deduplication	Gives the basic idea of deduplication and where it can be applied and basic algorithms used.	•Technique is applied over all types of data. •Variable size hash consumes large number of CPU cycles. •It nearly consumes 230% more number of CPU cycles as compared to fixed size hashing. •Increases window backup time.
2	CAB dedup : A Causality-based Deduplication Performance Booster for Cloud Backup Services	Improve both the backup and restore performances.	•Slowdowns client system and if client backup is lost, then difficult to recover. •Reduces backup time to 30%
3	SAM: A Semantic Aware Multi-Tiered Source De-duplication Framework for Cloud Backup	Focuses on reduction in backup window time and restore time.	•Needs high restoration time. •Reduces backup time by an average of nearly 38.7%. •SAM gives the better performance than CAB; as CAB client do not provide latest data to CAB server.

4	Dynamic Data Deduplication in Cloud Storage	Focuses on dynamic features of cloud storage. Uses large number of Deduplicators	•Techniques till now were static, so this was designed. •Processing time increases with an increase in the number of Deduplicators. •Unable to predict at what limit the Deduplicators need to be increased.

IV. NEED OF DYNAMISM IN DEDUPLICATION PROCESS

These days' Backup services are migrating towards Cloud. Waraporn et.al [2] focuses on deduplication that can be applied on dynamic nature of Cloud Computing. Cloud computing is widely used utility for customers to provide the resource on demand. As data storage size is getting increase, deduplication process is applied to the cloud storage to reduce energy consumption and reduce cost for storing large data. Advantage is that it reduces storage space and network bandwidth and maintains fault tolerance with storage efficiency. This paper overcomes the drawback of static architecture and focuses on disaster recovery with replication after deduplication. Author also discusses how to speedup replication time and save bandwidth. Its system design consists of:

- Load balancer: Depending upon the load of deduplicators it shares the load among them.
- Deduplicators: It identifies duplicate data by comparing hash values which are in metadata server.
- Cloud storage: Used to store metadata and file server.
- Redundancy Manager: Keeps track of changing nature of QoS.
- As per the number of times the file or chunk is referred on that level it should be replicated for reliability.

Authors have experimented the architecture for three events namely, file upload, update and delete using CloudSim and HDFS Simulator.

V. CONCLUSION

Understanding the storage demand and available services, a review and analysis of different data deduplication techniques has been done. It was observed that SAM gives the better performance than that of CAB. Research focus can be done on reduction in network bandwidth, high throughput, read/write efficiency, backup window time and size and transmission cost backup process. Lastly we have discussed Dynamic deduplication storage system which stores the data efficiently but creates bottleneck at the time of disk storage.

REFERENCES

[1] Nagapramod Mandagere , Pin Zhou, Mark A Smith, Sandeep uttamchandani "Demystifying Data Deduplication" 2008

[2] Waraporn Leesakul, Paul Townend, Jie Xu "Dynamic Data Deduplication in Cloud Storage" *2014 IEEE 8th International Symposium on Service Oriented System Engineering*

[3] Yujuan Tan, Hong Jiang, Dan Feng, Lei Tian, Zhichao Yan, Guohui Zhou "SAM: A Semantic-AwareMulti-Tiered Source De-duplication Framework for Cloud Backup" *2010 39th International Conference on Parallel Processing*

[4] Yujuan Tan, Hong Jiang, Dan Feng, Lei Tian, Zhichao Yan "CABdedupe: A Causality- based Deduplication Performance Booster for Cloud Backup Services"

[5] Andre Oriani and Islene C. Garcia "From Backup to Hot Standby: High Availability for HDFS" *2012 31st International Symposium on Reliable Distributed Systems*

[6] Yang Zhang, Yongwei Wu and Guangwen Yang "Droplet: a Distributed Solution of Data Deduplication" *2012 ACM/IEEE 13th International Conference on Grid Computing*

[7] Michael Vrable, Stefan Savage, and Geoffrey M. Voelker "Cumulus: Filesystem Backup to the Cloud" *7th USENIX Conference on File and Storage Technologies*

[8] Zhe Sun, Jun Shen, Jianming Young "A novel approach to data deduplication over the engineering oriented cloud system" *2013 Integrated Computer Aided Engineering,*

[9] P. Neelaveni and M. Vijayalakshmi "A Survey on Deduplication in Cloud Storage "Asian Journal of Information Technology13(6): 320-330, 2014.

[10] Jian-ping Luo Xia Li, Min-rong Chen " Hybrid shuffled frog leaping algorithm for energy-efficient dynamic consolidation of virtual machines in cloud data centers " Expert Systems with Applications 2014.

