Dynamic Data Deduplication in Cloud Services

Mohini Ramesh Vikhe, Prof. Dr. Kishor Kinage, Prof. Jyoti Malhotra

mohinivikhe@gmail.com, kishor.kinage@mitcoe.edu.in, jyoti.malhotra@mitcoe.edu.in

Department of Information Technology,
MIT College of Engineering,
Pune, India

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. For this reason, we are applying dynamic data deduplication. As in deduplication; we are providing only a single instance of data and pointer to other location for fault tolerance and reliability, we are replicating the server. We firstly deduplicate, and then replicate the server for storage efficiency.

Keywords— Cloud Computing; Cloud storage; Deduplication; HDFS.

I. INTRODUCTION

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 algorithms to assign each data chunk a hash identifier that is fingerprint, which it compares to previously stored identifiers to determine if the data chunk has already been stored. New fingerprint is stored in database.

We can achieve up to 70% to 90% reduction in capacity of our backups. Dedupe technology offers number of benefits for storage and backup such as lower storage space requirements, more efficient disk space, and less data sent across in every field for remote backups, replication, and disaster recovery. Deduplication can be performed either on source side or on target side.

The figure shows the basic idea of deduplication. As shown in diagram data is shown with multiple colors and redundant data is appeared as multiple copies of data. After applying deduplication method only one copy of data is been stored rather than multiple copies.

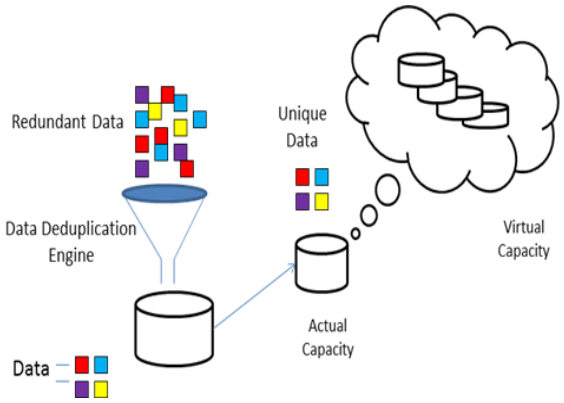

Figure 1: Basic idea of deduplication.

Now a day's cloud computing is mostly used for storage as well as also known as for dynamic service provider [2]. The 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 which are 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 required applications, consumers and businesses use applications and also access their personal files at any computer with having internet access via cloud computing. Main advantage of this it is efficient computing by making data storage, processing and bandwidth as centralized medium for access. Examples of cloud computing are Yahoo email, Gmail, or Hotmail etc. Cloud computing is divided into three segments application, storage and connectivity.

As cloud computing provides lot of services so a huge amount of data should be needed and allow access for this purpose deduplication can be applied to it. As per the today's user's need only deduplication won't provide much more throughput and reliability too. Therefore, dynamic deduplication can be applied to it.

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. The 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.

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 [5]. 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.

This paper is organized as, section II presents related work. Section III summarizes proposed architecture and finally we have summarized in section IV.

II. RELATED WORK

Deduplication has been used for reducing storage capacity with high reliability. It also helps to reduce the cost of storage capacity and reduces CPU time for retrieval and storing huge data. Now-a-days cloud has become the main storage for big data. So deduplication plays vital role in cloud storage.

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 copies of duplicate data instead of storing redundant data.

Dynamic Approach of Deduplication to Cloud Storage:

Cloud computing is a utility for customer to provide the resource on demand. So as data storage size is getting increase, so we are applying deduplication to the cloud storage to reduce energy consumption and reduce cost for storing large data [1].

Advantage:-

It reduces storage space and network bandwidth and maintains fault tolerance with storage efficiency.

Drawback:-

First paper overcomes the drawback of static architecture and focuses on disaster recovery with replication after deduplication so speedup replication time and save bandwidth. As per the number of times the file or chunk is referred on that level it should be replicated for reliability.

Dynamic deduplication design architectural model consistsof :

*Load Balancer:*Handles the client request, deduplicator with less work load at that time gets the work .

Deduplicators: Finds out the duplicate data by comparing with existing hash values which are in metadata server.

Cloud Storage: To store metadata on Metadata server and some files to store on File server.

Redundancy Manager: Usedto identify initial number of copies.

NagapramodMandagere, Pin Zhou, Mark A Smith, Sandeep uttamchandaniproposes the basic idea of deduplication and about the deduplication taxonomy[1].

Here client based deduplication is the deduplication performed on client side rather than on serve side, then transferred to server. Deduplication Appliance consists of In-Band and Out-Of-Band. In In-Band deduplication appliance finds duplicate data of arrived one, before writing it to the disk. In Out-Of-Band performs deduplication after data has been written to the disk. Deduplication algorithms,whole file hashing uses SHA-1 or MD5 hash function to obtain fingerprint of the file. Here complete file is matched with another one. Another algorithm is sub file hashing it divides the file in sub parts and then find the fingerprint of that. It has to different types first is fixed block hashing and second is variable block hashing.

Advantage:-
It saves network bandwidth.

Disadvantage:-
Same method applied to various forms of data.

WarapornLeesakul, Paul Townend, Jie Xu focuses on deduplication that can be applied on dynamic nature of Cloud computing [2]. This paper overcomes the drawback of static architecture and focuses on disaster recovery with replication after deduplication so speedup replication time and save bandwidth.Cloud computing becomes utility for customer to provide the resource on demand facility. So as data storage size is growing rapidly, therefor applying deduplication to the cloud storage will reduce energy consumption and reduce cost for storing large data [2]. System design consist of : Load balancer, Deduplicators, Cloud storage, Redundancy Manager.

As per the number of times the file or chunk is referred on that level it should be replicated for reliability.

Advantage:-
It reduces storage space and network bandwidth and maintains fault tolerance with storage efficiency.

One of the another architecture for cloud backup services is Semantic-Aware Multi-tiered source de-duplication framework (SAM) [3], here Yujuan Tan, Hong Jiang, Dan Feng, Lei Tian, Zhichao Yan, Guohui Zhoucombines 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 tradeoff between the de-duplication efficiency and de-duplication overhead to shorten the backup window.

SAM is composed of three key components such as Virtual Full Backup, Global File-level De-duplication, and Local Chunk-level De-duplication.

Global file level deduplication performed on Master Server. To reduce the disk access SAM uses two tire chunk indexing approach. To identify duplicate chunk need to at least once access the disk. In future to reduce disk accesses SAM has two types of file small files and compressed files.

Other architecture for cloud backup services is CAB [4] deduplication where Yujuan Tan, Hong Jiang, Dan Feng, Lei Tian andZhichao Yanfocuses on reduction in backup time and restore time. For scalability they have designed architecture as client side deduplication.

CABDedup has CAB Client, CAB Server and Interface are the three main parts of the CABDedup technique.

Deduplication process at client side handle by CABClient. It is composed of two main modules, the Causality-Capture module and Redundancy-Removal module.Causality-Capture module has File Monitor, File List and FileRecipe Store, whichmonitors and capture 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.

Deduplication process at server side is handleby CABServer which has File List Remote Store and File Recipe Remote Store as its components. It stores the file lists and file recipes sent 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.

As per client request the load balancer will assign to any one of the deduplicator. This architecture compares generated deduplicate value with the hash value if exist in metadata server. If not found stores in metadata server and then backup to file server, if found no need to store it, just gives pointer to that location.

Waraporn Leesakul, Paul Townend, Jie Xu "Dynamic Data Deduplication in Cloud Storage [2],Yujuan Tan, Hong Jiang, Dan Feng, Lei Tian, Zhichao Yan, Guohui Zhou [3], and Yujuan Tan, Hong Jiang, Dan Feng, Lei Tian, Zhichao Yan [4] focused on storage and backup of data retrieval time factor. Andr´e Oriani and Islene C. Garcia focuses on high availability of the backup data[5]. This paper uses the concept of HDFS [Hadoop Distributed File System] and used 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 here backup node name node used by HDFS to communicate between client and data nodes. Backup node acts as a main role, because it decides checkpoint for name node. Backup node also used to store transactional logs.

Main aim to develop backup node was to obtain high availability but not gained as per the requirement. So, 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. This concept can be used for replication of the backup systems.

In order to modify HDFS to hot standby node it has to implement following two steps:

a. Extend the state already replicated by backup node.
b. Build an automatic failover mechanism.

In first 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.

Yang Zhang, Yongwei Wuand Guangwen Yangsay's droplet is a distributed deduplication storage system aims at high throughput and scalability [6].Droplet consist of three main components, they are*single Metadata server, multiple fingerprinting servers, multiple storage nodes.* Where, Metadata server maintains the information related to storage server as well as of fingerprinting server, when fingerprinting server gets the data from client then it divides the data in blocks and then find out their fingerprints. Each and every block is tagged with its fingerprint then it's been compressed and sends 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 discards 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 IO 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.

III. PROPOSED DESIGN

The system architecture consists of client side deduplication. Hashing done at client side and connects to any of deduplicator depending upon load at that time. Deduplication can be identified by comparing hash values which are already in metadata server. If not found, then it is stored in file server. File reference number will be increased depending upon number of times that file been referred. File with highest reference number will be replicated more number of time for high availability.

The block diagram is shown in Figure 2. The system has following components:

1. Load Balancing and load sharing:
When client provides hash value to the load balancing and sharing it assigns the work to one of the deduplicator depending upon the existing load of deduplicators.

2. Deduplicators:
It identifies duplicate data by matching available fingerprint with existing one which is in metadata server.

3. Cloud storage:
It consists of metadata server, file server and their replicas. Metadata server stores metadata (fingerprint, hash value), file server to store actual data with its replicas.

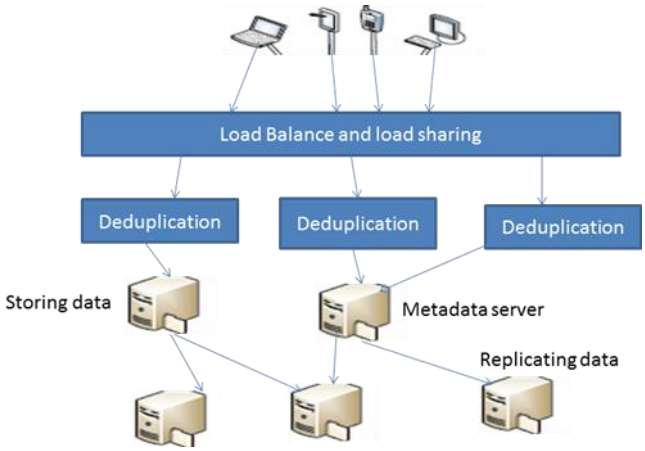

Figure2: Block Diagram

Detailed architecture is as shown in figure 3. It consists of client, metadata server, main storage server, backup of client and replicas of deduplicate data.

Client sends the fingerprint to the metadata servers, which simultaneously removes duplicate chunks locally from its memory to speed up the other operations.

When metadata server receives the client request it tries to match it with fingerprint, which is already in its memory. If not found it will request to main storage server,if found in main server, it will return back to master server. If master server already have then it will not store it, it will just give pointer to new location and file access pointer will be increased, and even if not found from storage server then it will store it in storage server.

After a period of time master server will write the data to backup of client and after specific time period it will backup the data to the main server.

For high availability of data main storage server need to replicate. File's number of replicas will depend on its access time,more the number of accesses more the number of replicas. For replication purpose we are trying to use HDFS system.

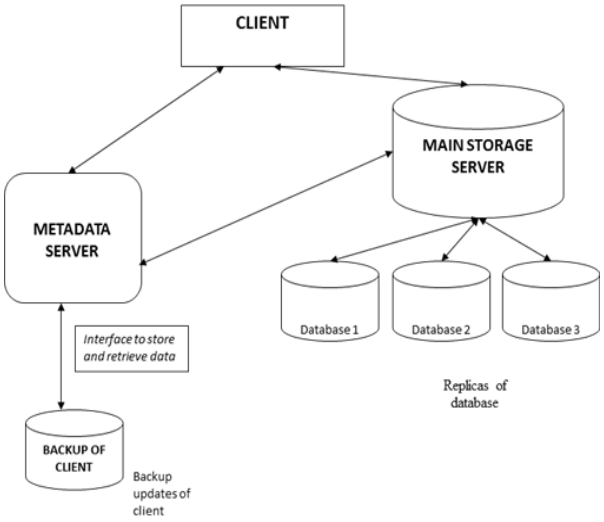

Figure 3: Architecture

IV. CONCLUSION

Deduplication techniques have been applied on cloud storage for improving backup performance. This reduces the system overhead and improves the data transfer efficiency on cloud, as now a day's cloud storage is becoming primary source for storing big data. A review and analysis of different data deduplication techniques has been done. In this paper we focus on deduplication before storing the data on server side in cloud environment and on replicating the storage for high availability.

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] Andr´e 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 Wuand 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.