International Journal of Innovative Research in Advanced Engineering (IJIRAE) ISSN: 2349-2163
Volume 1 Issue 10 (November 2014) www.ijirae.com

# Storage Optimization on cloud using De-duplication

P.B.Mane* N.B. Pokale R.V.Powar
*Computer Engg University of Pune* *University of Pune* *Shivaji University*

***Abstract--Now days cloud computing becomes more popular as it can provide low-cost and on-demand use of vast storage and processing resources. With the explosive growth of online digital information, the cost of storage infrastructure and management cost increases. Therefore it becomes a crucial to reduce the amount of data that is to be transferred, stored, and managed in cloud storage systems. As number of files increases, it leads to the waste of hardware resources and increases complexity of data center, which further degrades the performance of the cloud storage system. So we identified research gap as large number of storage of duplicate files on cloud systems from different users. For this reason, to decrease the workload caused by duplicated files, this paper proposes a new data index technique, which integrates data de-duplication for storage optimization on cloud systems.***

***Keywords: Cloud Storage, Optimization, Fingerprint, Data De-duplication.***

## 1. INTRODUCTION

**C**loud computing services can be classified as either computing or storage. Cloud storage is becoming a very popular research field. One reason is that more and more data-intensive applications are being attracted by the clouds. Cloud storage can provide high scalability, availability, fault tolerance, security, and cost-effective data services.

In a cloud storage environment, data is usually stored and managed in the space provided by the third-party companies. The common storage protocols include two types: NAS and SAN. However, the cloud network manager cannot control the performance of different storage nodes because of the great number of users and devices.

As far as data storage is concerned, although numerous schemes have been presented to improve file chunking and data compression is often overlooked. When the files that are reuploaded to the server may seriously affect the network bandwidth and also degrade efficiency.

In addition, the cloud network covers a great scope and domain and the data written on storage devices by different users might be similar or identical. In addition most users access similar data, operate the same functions, or repeat similar behaviors. Consequently, the system manager can no

longer guarantee the optimal status of each storage node in the cloud system. With the enlargement of the network, data integration bottleneck and waste of resources may occur as the system processes duplicate and redundant data, despite the flexibility and rapidity of the cloud storage system.

In enterprise databases, a large amount of redundant copies of data exist because of full system backups, shared documents etc. Data deduplication aims at reducing these redundant copies of data. When a backup application creates a backup, which is scheduled fortnightly or weekly depending on the criticality of the data, it creates a big file or series of individual files. Backup of these files has to be taken. Every such backup creates a redundant copy of data. Also, transmission of huge volumes of data over the network consumes a lot of network resources.

## 2. RELATED WORK

Q.He et al. [1] talk about various deduplication techniques. The techniques relies on the principle is to maintain only one copy of the duplicate data and a pointer to point to all the duplicate copies. The three types of deduplication can be possible at file level, block level or byte level. The old and new data are compared at byte level and if they match, they are marked as duplicate and pointers are updated.

He et al. [2] discuss various cloud storage techniques. About data deduplication technology, they suggest to retain only the unique FIStance of the data, reducing data storage volumes. Data deduplication engine creates an index of the digital signature for the data segment and the signature of a given repository to identify data blocks. The index provides a pointer to determine whether the data block is already present. In the copy operation, the data deduplication software found in a block of data inserts a link to the original data block index location instead of storing the data block again. If the same block appears more than once, more pointers to the indexing table are generated. Data migration of cloud storage means moving data from one storage system to another which are at different geographical locations. It aims at cooperating and keeping load balance in cloud storage system. The data should be migrated into other cloud storage units and while keeping pointers in the old stored positions intact, or

 

modify and update the index as changes occur. However it may bring overhead to network bandwidth and access bottleneck to concurrent clients.

H.-G. Yang[3] propose a layered view of the cloud storage architecture, which composed of user application layer, apps hosting platform layer, storage management layer and storage resources layer. In this focusing on cloud storage for data-intensive applications. The details of how to organize the nodes in the cloud storage system is also described. Availability is another important aspect of cloud storage that the users very concern. One aim of cloud storage is to provide persistent availability of data service even when one or more I/O node is failure.

M. Ripeanu [4] discussed initial progress towards developing an automated solution to configure a distributed storage system, that is, to enable/disable its various optimizations and configure their parameters with minimal human intervention. To better understand the challenges entailed by automated storage system configuration, is optimization, namely online data compression through similarity detection, in the context of check pointing applications.

S. W. Clyde[5] proposed a novel learning-based fusion technique for rule-based deduplication. This system utilizes the original deduplication rules, thereby making use of the expert domain knowledge that was used to create those rules. Fused deduplication technique achieves high average accuracy, without requiring extensive manual tuning of the deduplication rules. Also demonstrated the efficacy of our fused deduplication technique on biographic text records that conform to the significant real world schema, GJXD.

## 3. PROPOSED WORK

The technique of deduplication over a cloud has been around for few years. The current market products provide excellent deduplication of data for their clients. However, the existing products work in the following manner

1. If a single block of data has many copies, a single copy is maintained and all the other copies are deleted. However, a pointer pointing to the existing copy indicates how many original copies were present. When more than one copy is required, the pointers point to the existing copy and the user gets the block that he was looking for.
2. While sending these blocks over the net, all the copies are sent which makes deduplication not very worthwhile and increases the bandwidth requirement.

We propose a system, where in, while sending these blocks over the network, only fingerprint of the block is sent. On the receiver's side when a user desires for this block, the correct block of data is returned. Thus, reducing the bandwidth and saving the storage costs.

In our propose system uses the Fingerprint index server (FIS) to process cloud storage functions, including file compression, chunk matching, data de-duplication. Therefore, our proposed FIS can manage and optimize the storage nodes according to the client-side transmission conditions so that every storage node can maintain its optimal status and provide suitable resources to clients.

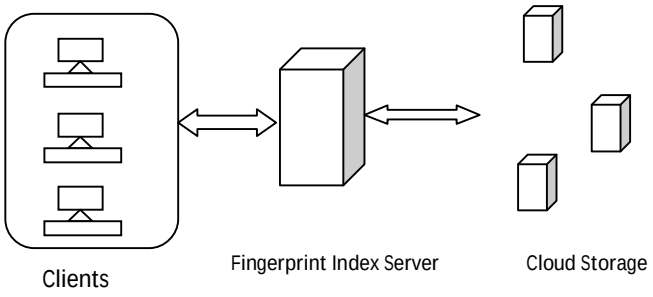


Fig1: System Architecture

*3.1 Fingerprint Index Server (FIS)*

Fingerprint Index Server (FIS) similar to Domain Name System (DNS) structure, manages the cloud data by a complex P2P-like architecture . Although FIS resembles DNS in structure and functions, FIS mainly processes the one-to-many matches of the storage nodes. IP addresses and hash codes. In general, FIS has following functions:

 

- Switching between the fingerprints and their corresponding storage nodes.
- Satisfying client demands for transmission as much as possible.

For file transmission optimization, every FIS has exclusive databases of its own domain, which include the fingerprints and their corresponding storage nodes. However, for WAN cloud network environment, to manage the file system by few FISs will cause great burden on the FIS. Therefore, based on the existing DNS structure, we propose to divide the FISs according to the domain FIS and adopt hierarchical management architecture to reduce the workload of the FISs.

The FIS mainly query and control the data between fingerprints and storage nodes, and coordinate the transmission by feedback control between storage node and clients. FIS records fingerprints and storage nodes of all data chunks. The FIS record only the locations of the fingerprints and manage the storage nodes.

*3.2 De-duplication*

De-duplication technique is adopted in our scheme to scatter and remix the data at local hosts, divide the file into several chunks for uploading, and designate a unique fingerprint to each file by MD5.
Because of its uniqueness, every fingerprint is regarded as the identification and fingerprint of a data chunk. After checking a requested fingerprint, the FIS will make sure whether the file chunk of the same fingerprint exists in the storage space. If not, the system continues the following uploading procedure and assigns tasks to the storage node.

*3.3 FIS Querying Process*

Each domain-based FIS has databases of fingerprints and storage nodes. The database of fingerprints records the fingerprints of different files and their corresponding storage nodes. When a user are looking for specific fingerprints, the FIS queries and confirms if the file already exists in the storage node within the domain before taking the next step. While the clients want to access data, they can use the obtained fingerprints as the index and query the FIS of the upper layer, which searches for the best access node based on the content in the database in case the inefficiency of the access node.

## 4. CONCLUSION

This paper proposes the FIS to process file compression, chunk matching, data de-duplication and file storage. Two major contributions of this paper include the following.
1) By compressing and partitioning the files according to the chunk size of the cloud file system, we can reduce the data duplication rate. The processed files are encoded into the signature by MD5 fingerprint for the FISs to match the file, designate to the storage servers, and provide necessary uploading information for the clients. After downloading and modifying the files, the clients compress and partition the modified chunks only, encode these chunks by MD5 fingerprint and reupload the chunks.
2) FIS receives the feedback of the previous transmissions and adjusted the transmission parameters to attain the optimal performance for the storage nodes.

## REFERENCES

[1] Q. He, Z. Li, and X. Zhang, “Data deduplication techniques,” in *International Conference on Future Information Technology and Management Engineering*, 2010, pp. 431–432.
[2] Z. Li, X. Zhang, and Q. He, “Analysis of the key technology on cloud storage,” in *International Conference on Future Information Technology and Management Engineering*, 2010, pp. 427-428.
[3] Y.-M. Huo, H.-Y. Wang, L.-A. Hu, and H.-G. Yang, “A cloud storage architecture model for data-intensive applications,” in *Proc. Int. Conf.Comput. Manage.*, May 2011, pp. 1–4.
[4] L. B. Costa and M. Ripeanu, “Towards automating the configuration of a distributed storage system,” in *Proc. 11th IEEE/ACM Int. Conf. Grid Comput.*, Oct. 2010, pp. 201–208.
[5] J. Dinerstein, S. Dinerstein, P. K. Egbert, and S. W. Clyde, “Learning based fusion for data deduplication,” in *Proc. 7th Int. Conf. Mach. Learning Appl.*, Dec. 2008, pp. 66–71.
[6] D. R. Mikkilineni and V. Sarathy, “Cloud computing and the lessons from the past,” in *IEEE International Workshops on Enabling Technologies: Infrastructures for Collaborative Enterprises*, Los Altos, CA, 2009, pp. 4–5.
[7]W. Zeng, Y. Zhao, K. Ou, and W. Song, “Research on cloud storage architecture and key technologies,” in *International Conference on Information Sciences*, Seoul, Korea, November 2009, p. 4.
[8] C.-Y. Chen, K.-D. Chang, and H.-C. Chao, “Transaction pattern based anomaly detection algorithm for IP multimedia subsystem, *IEEE Trans. Inform. Forensics Security*, vol. 6, no. 1, pp. 152–161, Mar. 2011.

 