IJISET - International Journal of Innovative Science, Engineering & Technology, Vol. 2 Issue 10, October 2015.
www.ijiset.com
ISSN 2348 – 7968

Study on Dedupe: Data Deduplication Scheme for Cloud Backup of Personal Storage

Ankur Kudale[1], Swapnil Gaikwad[2], Ajay Bankar[3], Datta Patil[4]

[1,2,3,4,] B.E. Students Department of Computer Engineering, SVPM's C.O.E. Malegaon (Bk.), 413115, Savitribai Phule, Pune University, Maharashtra, India

Abstract

The Cloud backup is an important issue since large amount of valuable data has been stored on personal computers. Inescapable challenge facing source deduplication for cloud backup services is the low deduplication efficiency. The proposed Dedupe scheme that improves data deduplication efficiency by exploiting application awareness, and further combines local and global duplicate detection to strike a good balance between cloud storage capacity saving and deduplication time reduction It can be done by using an intelligent chunking scheme. This system is based on an application-awareness with the use of hash function. To improve the efficiency of system with low system overhead on client side it combines the local and global source deduplication with application awareness

Keywords: *Data Deduplication, Chunking Scheme, Application Awareness, Cloud backup Services.*

1. Introduction

According to the present scenario, the backup has become the most essential mechanism for everyone. Backing up files can protect against accidental loss of user data, database corruptions, hardware failures, and even natural disasters. However, the large amount of redundancies which is found in the backups makes the storage of the backups a concern, thus utilizing a large of disk space. Data de-duplication comes as a rescue for the problem of redundancies in the backup. It is a capacity optimization technology that is being used to dramatically improve the storage efficiency. Data de-duplication eliminates the redundant data and stores only unique copy of the data. Here instead of saving the duplicate copy of the data, data de-duplication helps in storing a pointer to the unique copy of the data, thus reducing the storage costs involved in the backups to a large extent. It need not be applied in only backups but also in primary storage, cloud storage or data in flight for replication, such as LAN and WAN transfers. It can help us to manage the data growth, increase efficiency of storage and backup, reduce overall cost of storage, reduce network bandwidth and reduce the operational costs and administrative costs. The five basic steps involved in all of the data deduplication systems are evaluating the data, identify redundancy, create or update reference information, store and/or transmit unique data once and read or reproduce the data. Data de-duplication technology divides the data into smaller chunks and uses an algorithm to assign a unique hash value to each data chunk called fingerprint. The algorithm takes the chunk data as input and produces a cryptographic hash value as the output. The most frequently used hash algorithms are SHA1, MD5. These fingerprints are then stored in an index called chunk index. The data de-duplication system compares every fingerprint with all the fingerprints already stored in the chunk index. If the fingerprint exists in the system, then the duplicate chunk is replaced with a pointer to that chunk. Else the unique chunk is stored in the disk and the new fingerprint is stored in the chunk index for further process.

2. Literature Survey

In paper [2] the cloud computing is a technology which is used to provide resources as a service. There are many services provided by cloud provider. Such as SAAS, IAAS, PAAS. The cloud computing provides the Storage-as-Service which is used to backup the users data into cloud. The Storage-as-a-Service is provided by Storage Service Provider or Cloud Service Provider. This service is provided by Cloud Service Provider which is effective, reliable and cost-effective. The existing backup scheduling provides the reliability by maintaining the same copy of the data twice. The existing backup scheduling provides the reliability and backup speed, but the redundancy of data is not considered. The existing backup scheduling not considers much of the security issues. The limitations of the existing backup scheduling algorithm is improved by proposing a backup scheduling algorithm(IBSD) which aims at reducing redundancy without compromising on availability. The IBSD algorithm reduces redundancy by deduplication techniques. The deduplication is a technique which is used to identify the duplicate data. The de-duplication identifies the duplicate data and eliminates it, by storing only one

ISSN 2348 – 7968

copy of the original data. If the duplicate occurs then the link will be added to the existing data. Also paper [3] tells us that, Data Deduplication describes approach that reduces the storage capacity needed to store data or the data has to be transfer on the network. Source Deduplication is useful in cloud backup that saves network bandwidth and reduces network space Deduplication is the process by breaking up an incoming stream into relatively large segments and deduplication each segment against only a few of the most similar previous segments. To identify similar segments use block index technique The problem is that these schemes traditionally require a full chunk index, which indexes every chunk, in order to determine which chunks have already been stored unfortunately, it is impractical to keep such an index in RAM and a disk based index with one seek per incoming chunk is far too slow. In this paper we describes application based deduplication approach and indexing scheme contains block that preserved caching which maintains the locality of the fingerprint of duplicate content to achieve high hit ratio and to overcome the lookup performance and reduced cost for cloud backup services and increase deduplication efficiency. In paper [4], to improve space utilization and reduce network congestion, cloud backup venders (CBVs) always implement data deduplication in the source and the destination. Towards integrating source and destination data deduplication, we mainly propose two proposals in this area. One of the important things of this is benefit-cost model for users to decide in which degree the deduplication executes in client and in cloud, an d let the data centre to decide how to handle the duplications. This will give better reliability, quality of service etc. Combining caching and prefetching, and the requirements of different cloud backup services, the read performance in the cloud backup systems can be improved.

3. System Architecture

The main purpose of the local and global deduplication scheme is to utilizing not only low overhead but also to utilize high overhead cloud assets to reduce the Computational transparency by using an intelligent data chunking scheme. In this system there will be the adaptive use of the hash function based on the application awareness [1]. To advance the efficiency of the system and low system overhead on client side it combines the local and global source deduplication with application awareness

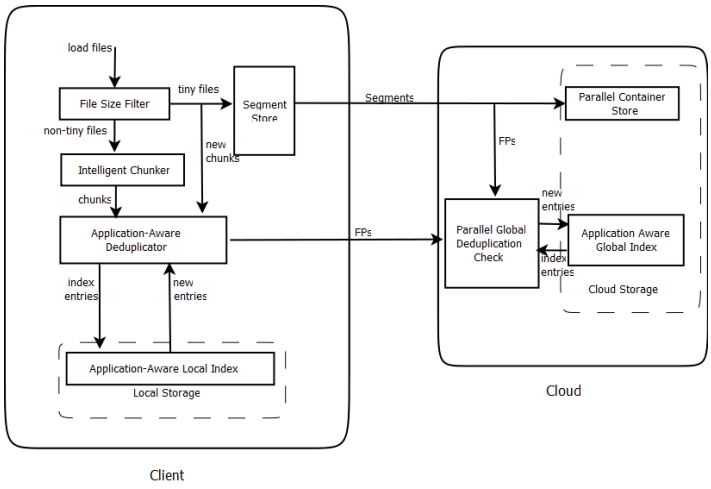

Fig. 1 System Architecture

Proposed system contains five main components:

3.1 File Size Filter for Backup Data Stream

A personal computing device contains the most of the tiny files which holds a negligibly small percentage of the storage capacity. To decrease the data overhead the planned system filters out these tiny files in file size filter before stating the deduplication process. And make the group of all tiny files together into large unit in the segment store and that will be stored as segment in segment store. Efficiency of the data transfer over WAN is increase due to segment store.

3.2 Intelligent Chunker

Data chunking scheme having the great impact on the efficiency of data deduplication. There will be the contrary relationship among the deduplication ratio and the average chunk size. Chunking can be prepared on divide files into three categories depending on file type i.e. Compressed files, static uncompressed files and dynamic uncompressed files. Compressed files are chunked by WFC based chunking like file with extension .mp3, .jar, etc. Static uncompressed files are not editable. These files chunked into fixed size by using SC chunking like file with extension .pdf, .exe, etc. Dynamic uncompressed files are editable. It breaks dynamic uncompressed files variable-sized chunks using CDC based chunking like file format .txt, .doc, .ppt.

ISSN 2348 – 7968

3.3 Application - Aware Deduplicator

After data chunking in intelligent chunker module, data chunks will be deduplicated in the application aware deduplicator by generating chunk fingerprints in the hash engine and detecting duplicate chunks in both the local client and remote cloud. If static chunking then SHA1 is used for fingerprinting. If WFC or CDC then MD5 is used for generating fingerprint. To achieve high deduplication efficiency, the application aware deduplicator first detects duplicate data in application aware local index corresponding to local dataset. And then compares local deduplicated data chunks with all stored cloud dataset looking up fingerprints in application aware global index on cloud side. After that only unique data chunks will be stored in cloud storage with parallel container management.

3.4 Application - Aware Index Structure

It consists of small hash table and application index based on disk indices classified by file type. Index structure is used for speed up the I/O operation. It uses two application aware indices: local index on client side and global index on cloud side. It can achieve the high deduplication throughput by looking up chunk fingerprints concurrently in small indices classified by application type.

3.5 Segment and Container Management

The segment contains the files which is less than predefined size. Dedupe often group deduplicated data from many files and chunk into larger units called segments before these data transferred over WAN. After a segment is sent to the cloud, it will be routed to a storage node in the cloud with its corresponding fingerprints, and be packed into container, a data stream based structure, to keep spatial locality for deduplicated data. A container includes a large number of chunks and their metadata, and it has a size of several MB. An open chunk container is maintained for each incoming backup data stream in storage nodes, appending each new chunk or tiny file to the open container corresponding to the stream it is part of. When a container fills up with a predefined fixed size, a new one is opened up. If a container is not full but needs to be written to disk, it is padded out to its full size. This process uses chunk locality to group chunks likely to be retrieved together so that the data restoration performance will be reasonably good. Supporting deletion of files requires an additional process in the background.

4. Conclusions

By studying all the previously work done on deduplication we summarize that, Dedupe is an application aware local-global source-deduplication scheme for cloud backup in the personal computing environment to improve deduplication efficiency is proposed. Also an intelligent deduplication strategy in Dedupe is designed to exploit file semantics to minimize computational overhead and maximize deduplication effectiveness using application awareness. It combines local deduplication and global deduplication to balance the effectiveness and latency of deduplication.

References

[1]. Y.Fu, H.Jiang, ”Application -Aware Local-Global Source Deduplication for Cloud Backup Services of Personal Storage,” IEEE TRANSACTIONS ON PARALLEL AND DISTRIBUTED SYSTEMS, VOL. 25, NO 5, MAY 2014.

[2] Improved Backup Scheduling With Data Deduplication Techniques For Saas In Cloud 1tamilselvi.T, 2k.Saruladha 1Department of Distributed Computing Systems (CSE), Pondicherry Engineering College, Pondicherry 2Department of computer science and engineering, Pondicherry engineering college, Pondicherry.

[3] A Novel Way of Deduplication Approach for Cloud Backup Services Using Block Index Caching Technique Jyoti Malhotra1 ,Priya Ghyare2 Associate Professor, Dept. of Information Technology, MIT College of Engineering, Pune, India1 PG Student [IT], Dept. of Information Technology, MIT College of Engineering, Pune , India.

[4] Data Deduplication in Cloud Backup Systems Xiongzi Ge, Zhichao Cao CSci 8980, Fall 2013, Final Report Computer Science and Engineering, University of Minnesota, Twin Cities {xiongzi, zcao}@cs.umn.edu.

[5] Hemant Palivela, Chawande Nitin P, Sonule Avinash, Wani Hemant, ”Development of servers in cloud computing to solve issues related to security and backup”, NJ, USA: IEEE Computer Society, 158-163, 2011