International Journal of Emerging Technology and Advanced Engineering
Website: www.ijetae.com (ISSN 2250-2459 (Online), Volume 5, Special Issue 2, May 2015)
International Conference on Advances in Computer and Communication Engineering (ACCE-2015)

A Survey on Aware of Local-Global Cloud Backup Storage for Personal Purpose

Abhirupa Chatterjee[1], Divya. R. Krishnan[2], P. Kalamani[3]

[1,2]UG Scholar, Sri Sairam College Of Engineering, Bangalore. India
[3]Assistant Professor, Sri Sairam College Of Engineering, Bangalore. India

[1]abhirupachatterjie@gmail.com, [2]divyark95@gmail.com, [3]pskalamani@gmail.com

***Abstract*—According to the present scenario, the requirement for data protection in the personal computing environment has increased significantly. This is because the volume and value of digital information is growing rapidly. For protecting the data, it is the needed to have a good backup and recovery plan. But the redundant nature of the backup data makes the storage a concern; hence it is necessary to avoid the redundant data present in the backup. Data de-duplication is one such solution that discovers and removes the redundancies among the data blocks. Actually data de-duplication scheme is motivated on personal storage. The proposed scheme improves data de-duplication efficiency by exploiting application awareness.**

***Index Terms*— Application awareness, Cloud backup service, Chunking schemes, Data redundancy, Data de-duplication, De-duplication efficiency, Hashing algorithm .**

I. INTRODUCTION

Data is the heart of any organization; hence it is necessary to protect it. Now-a-days, the backup has become the most essential mechanism for any organization. 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 can help organizations 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 de-duplication 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 SHA, MD5. These fingerprints are then stored in an index called chunk index. The data de-duplication system compares every finger-print with all the fingerprints already stored in the chunk index. If the fingerprint exists in the system, then the du-plicate 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.

Many personal computing devices rely on a cloud storage environment for data backup. Source de-duplication for cloud backup services is faced by a critical challenge. It has low de-duplication efficiency because of the re-source intensive nature of de-duplication and the limited system resources.

In this paper, an Application-aware Local-Global source de-duplication scheme is proposed that combines local and global duplicate detection to strike a good balance between cloud storage capacity saving and de-duplication time reduction

II. CLOUD STORAGE

Cloud storage is a service model in which data is maintained, managed and backed up remotely and made available to users over a network.

Cloud storage provides users with storage space and make user friendly and timely acquire data, which is foundation of all kinds of cloud applications. The storage cloud provides storage-as-a-service. The organization providing storage cloud uses online interface to upload or download files from a user's desktop to the servers on the cloud. Typical usage of these sites is to take a backup of files and data. Storage cloud exists for all the types of cloud. A cloud storage SLA is a service-level agreement between a cloud storage service provider and a client that specifies details of the service, usually in quantifiable terms.

A. Advantages Of Cloud Storage

Cloud storage has several advantages over traditional data storage. For example, if we store our data on a cloud storage system, we will be able to get that data from any location that has internet access. There is no need to carry around a physical storage device or use the same computer to save and retrieve our information. With the right storage system, we could allow other people to access the data.

III. DATA DE-DUPLICATION

The traditional backup solutions require a rotational schedule of full and incremental backup, which move a significant amount of redundant data every week. Most organizations also create a second copy of this information to be shipped to a secondary site for disaster recovery purposes. Thus aggregating, the costs of traditional backup in terms of bandwidth, storage infrastructure, and time increases the cost of IT organizations for information management. Backing up of redundant files and data increases the backup window size this results in over utilization of Network resources, and require too much additional storage capacity to hold unnecessary backup data.The organizations need solutions to manage this increasing information and data.

Thus de-duplication techniques can reduce your bandwidth requirements; it can improve the data transfer speed and maintain your cloud storage needs including cloud storage fees to a minimum.

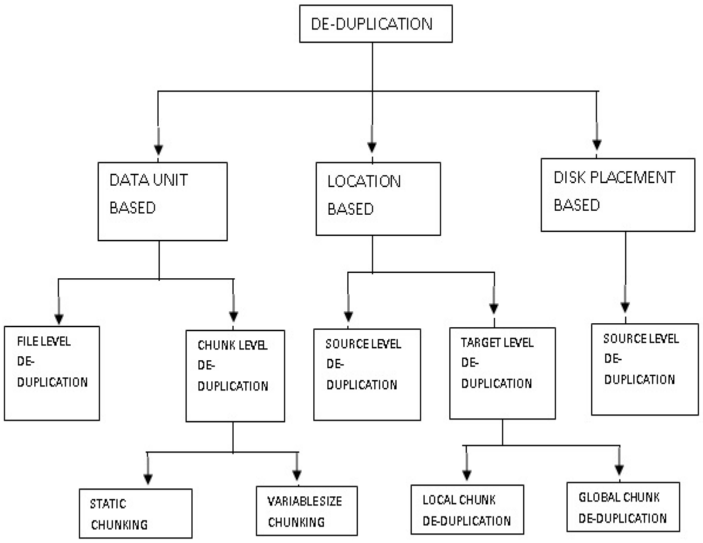

DE-DUPLICATION
DATA UNIT BASED
LOCATION BASED
DISK PLACEMENT BASED
FILE LEVEL DE-DUPLICATION
CHUNK LEVEL DE-DUPLICATION
SOURCE LEVEL DE-DUPLICATION
TARGET LEVEL DE-DUPLICATION
SOURCE LEVEL DE-DUPLICATION
STATIC CHUNKING
VARIABLE SIZE CHUNKING
LOCAL CHUNK DE-DUPLICATION
GLOBAL CHUNK DE-DUPLICATION

Figure 1:Data De-Duplication techniques

(1) Data unit based:

Here Data duplication strategies are basically classified in to File level de-duplication and Block (chunk) level de-duplication. In File level de-duplication only one copy of the file is stored. Two files are identical if they have the same hash value. On the other hand file is fragmented into blocks in Block level De-duplication and one copy of each block is stored. Each block may be fixed (static) or variable size chunk. In fixed size chunks, size of each block is same. In case of variable, size of each chunk is varies.

(2) Location based:

De-duplication can be categorized in to two basic approaches depending on the location where redundant data is to be eliminated [6]. In the target based approach, De-duplication is performed in the Destination storage system. Here client is not aware about strategies in de-duplication

The positive part of this method is storage utilization increases but bandwidth is not saved. The elimination of duplicate data is performed closed to where data is created in source based de-duplication.

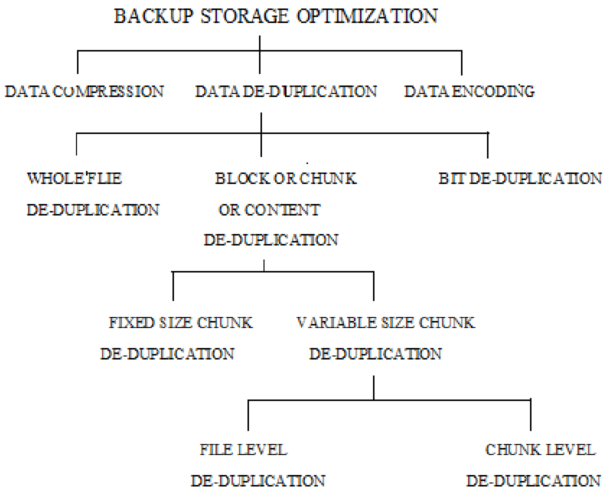
The source de-duplication approach is implemented at the client side. The client software communicates with the backup server by sending hash signatures to check for the existence of files or block. The duplicate are replaced by pointers and the actual duplicate data is never sent over the network.

(3) Disk Placement based de-duplication:

Backward reference de-duplication and Forward reference de-duplication are two major classifications in Disk placement. In backward reference the recent redundant data chunks are associated with pointers that point backward to the older identical data chunks. In case of forward reference de-duplication the recent redundant data chunks are maintained in their entirety and all the old identical data chunks are associated with pointers that point forward to the recent data chunks.

B. De-Duplication Techniques:

The optimization of backup storage technique is shown below. The data de-duplication can operate at the whole file, block (chunk) and bit level.


BACKUP STORAGE OPTIMIZATION
DATA COMPRESSION
DATA DE-DUPLICATION
DATA ENCODING
WHOLE'FLIE DE-DUPLICATION
BLOCK OR CHUNK OR CONTENT DE-DUPLICATION
BIT DE-DUPLICATION
FIXED SIZE CHUNK DE-DUPLICATION
VARIABLE SIZE CHUNK DE-DUPLICATION
FILE LEVEL DE-DUPLICATION
CHUNK LEVEL DE-DUPLICATION

C. De-Duplication Methods

Whole file de-duplication or single instance storage (SIS) finds the hash value for the entire file which is the file index.

If the new incoming file matches with the file index, then it is regarded as duplicate and it is made pointer to existing file index, If the new file is having new file index,then it is upgraded to the storage.Thus only single instance of the file is saved and subsequent copies are replaced with a pointer to the original file.

Block de-duplication divides the files into fixed-size block or variable-size blocks. For a fixed-size chunking, a file is partitioned into fixed-size chunks, for example each block with 8KB or 16KB.In variable-size chunking, a file is partitioned into chunks of different size. Both the fixed size and variable size chunking creates unique ID for each block using a hash algorithm such as MD5 or SHA-1.The unique ID is then compared with a central index. If the ID exists, then that data block has been processed and stored before.

Therefore, only a pointer to the previously stored data needs to be saved. If the ID is new, then the block is unique. The unique ID is added to the index and the unique chunk is stored. Block and Bit de-duplication looks within a file and saves unique iterations of each block or bit. This method makes block and bit de-duplication more efficient.

IV. EXISTING SYSTEM

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-a-service which is used to backup the user's data into cloud. The 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, but the redundancy of the data is not considered. This does not consider much of the security issues. The limitations of the existing backup scheduling algorithm is improved is improved by proposing a backup scheduling algorithm(IBSD) which aims at reducing redundancy without compromising on availability.

The IBSD algorithm reduces redundancy by de-duplication techniques, which is used to identify the duplicate data and eliminates it by storing only one copy of the original data. If the duplicate occurs then the link will be added to the existing data.

Also, the data de-duplication reduces the storage capacity needed to store or the data to be transferred on the network. Source de-duplication is useful in cloud backup that saves network bandwidth and reduces network space.

To identify similar segments we use block index technique. The problem is that these schemes require a full chunk index, which indexes every chunk, inorder to determine which chunks have already been stored, it is impractical to keep such an index in RAM and a disk based index.

In this paper we describe application based de-duplication approach and indexing scheme which contains block that preserves caching and 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 de-duplication efficiency. To improve space utilization and reduce network congestion, cloud backup vendor's (CBVs) always implement data de-duplication in the source and the destination. Towards integrating source and destination, we mainly focus on two proposals.

One of the important things of this is benefit-cost model for users to decide in which degree the de-duplication executes in client and in cloud. This will give better reliability, quality of service etc. Combining, caching and pre-fetching the requirements of different cloud backup services, the read performance in the cloud backup systems can be improved.

V. Local-Global Source De-Duplication

Local source de-duplication only detects redundancy backup data set from the same device at the client side and only sends the unique data chunks to the cloud storage. Local source de-duplication eliminates intra-client redundancy with low duplicate elimination ratio by low-latency .Global source de-duplication performs duplicate check in backup data sets from all clients in the cloud side before data transfer over WAN. It has intra-client and inter-client redundancy with high de-duplication effectiveness by performing high-latency duplication detection on the cloud side.

In the traditional storage applications like file systems and storage hardware, each of the layers contains different kinds of information about the data they manage. Such information in one layer will not be available to any other layers. ADMAD improves redundancy detection by application specific chunking methods that exploit the knowledge about concrete file formats.

All the related and prior work related to de-duplication focus only on the effectiveness of de-duplication. They are designed to remove more redundancy from the data.

Earlier systems have not considered the system over-heads for high efficiency in de-duplication process.

A) Design And Implementation

ALG-dedupe is designed to meet the requirement of de-duplication efficiency with high de-duplication effectiveness and low system overhead. The main idea of ALG-dedupe is (1)exploiting both low-overhead local resources and high-overhead cloud resources to reduce the computational overhead by employing an intelligent data chunking scheme and an adaptive use of hash functions.(2)to mitigate the on-disk index lookup by dividing the full index into small independent and application specific indices in an application aware index structure. It combines local-global source de-duplication with application awareness to improve de-duplication effectiveness with low system overhead on the client side.

B) Architecture

An architectural overview of ALG-dedupe is illustrated below, where tiny files are first filtered out by file size filter for efficiency reasons and backup data streams are broken into chunks by an intelligent chunker using an application aware chunking strategy.

Data chunks from the same type of files are then de-duplicated in the application aware de-duplicator by generating chunk fingerprints in hash engine and performing data redundancy check in trade off between duplicate elimination ratio de-duplication overhead, we de-duplicate compressed files with WFC, separate static uncompressed files into fix-sized chunks by SC with ideal chunk size and break dynamic uncompressed files into variable sized chunks with optimal average chunk size.

C) DE-Duplication Ratio

In the above figure shows that De-duplication comparison of different chunk size from file contains 20 MB for text file and PDF file. As chunk size increases it affect De-duplication ratio in terms of time factor .For improving backup performance and Reduce the system overhead, improve the data transfer. Efficiency on cloud is essential.

So that we use various chunking strategy such as CDC and static Chunking for improving running performance of the system and increase De-duplication ratio. Variation in chunk size affects De-duplication efficiency. Maintaining threshold size of chunk we get better De-duplication ratio pdf as well as text files.

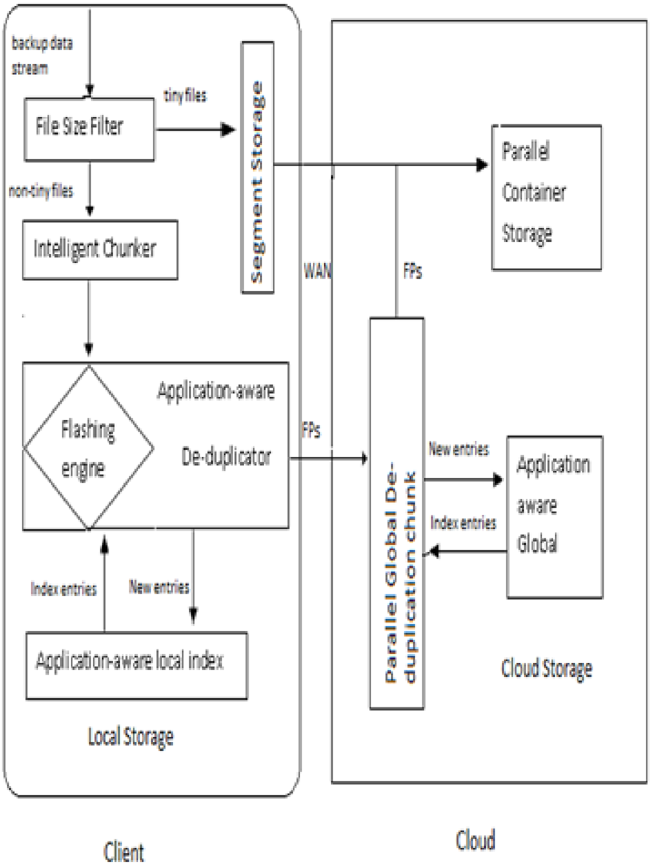


Figure 2: De-duplication comparison of different chunk size

File Size Filter: Most of the files in the PC dataset are tiny files that are less than 10KB in file size, accounting for a negligibly small percentage of the storage capacity. About 60.3 percent of all files are tiny files, accounting for only 1.7 percent of the total storage capacity of the dataset.

To reduce the metadata overhead, ALG-dedupe filters out these tiny files in the file size filter before the de-duplication process and groups data from many tiny files together into larger units of about 1MB each in the segment store to increase the data transfer efficiency over WAN.

Intelligent Data Chunking: The de-duplication efficiency of data chunking scheme depends upon the particular data chunking scheme selected. Depending on whether the file type is static or dynamic, the chunking scheme is selected.

MD5 has been employed in a wide variety of security applications and is also commonly used to check the integrity of files. An MD5 hash is typically expressed as a 32-bit hexadecimal number.

Application Aware Deduplicator: After data chunking in intelligent chunker module, data chunks will be de-duplicated in the application aware de-duplicator by generating chunk fingerprints in the hash engine and detecting duplicate chunks in both the local client and remote cloud. The proposed system will be implemented in java and MYSQL will be used as a database. The general flow of the system is shown in the activity diagram. Initially, the system performs file size filter and chunks the data. Depending on file type the chunking scheme is selected. Static chunking and content defined chunking uses SHA-1 and MD5 hash algorithms respectively to generate a unique identification code for each chunk.

*Sha-1 (Secure Hash Algorithm1) :*SHA-1 is a cryptographic hash function 160-bit hash value. A SHA-1 hash value is typically rendered as a hexadecimal number, 40digits long.

Md5 (Message-Digest 5): It is a widely used cryptographic function with a 128-bit hash value.MD5 has been employed in a wide variety of security applications and is also commonly used to check the integrity of files. An MD5 hash is typically expressed as a 32-bit digit hexadecimal number. In our system we are considering following points: An ALG-De-dupe, an Application aware Local-Global source de-duplication scheme is proposed that not only exploits application awareness, but also combines local and global duplication detection.

The system is proposed to achieve high de-duplication efficiency by reducing the de-duplication latency to as low as the application-aware local de-duplication while saving as much cloud storage cost as the application-aware global de-duplication. The application design is motivated by the systematic de-duplication analysis on personal storage.

The basic idea of ALG-Dedupe is to effectively exploit the application difference and awareness by treating different types of applications independently and adaptively during the local and global duplicate check processes. This will help to significantly improve the de-duplication efficiency and reduce the system overhead.

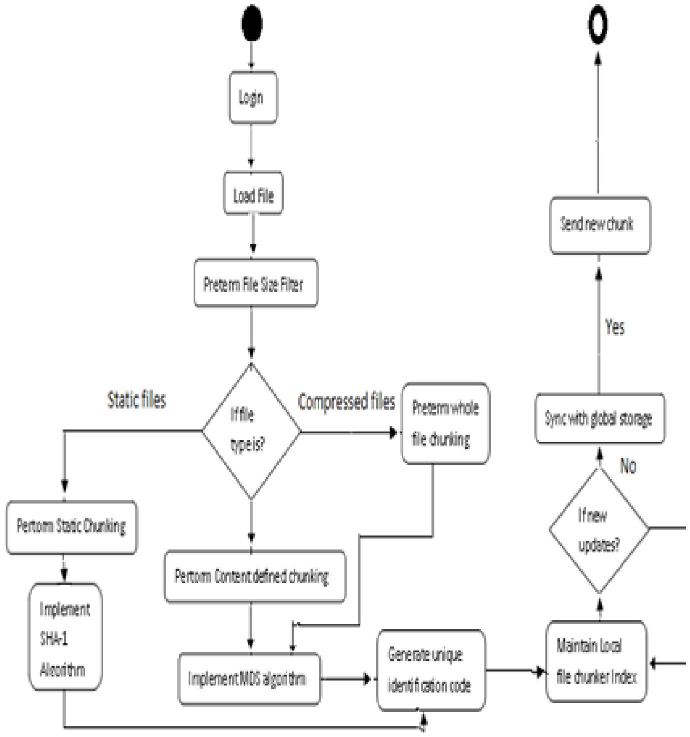


Figure 3: General flow of the proposed system

VI. CONCLUSION

ALG-Dedupe is an application aware local-global source-de-duplication scheme for cloud backup in the personal computing environment to improve de-duplication efficiency is proposed. Also an intelligent de-duplication strategy in ALG-Dedupe is designed to exploit file semantics to minimize computational overhead and maximize de-duplication effectiveness using application awareness.

It combines local de-duplication and global de-duplication to balance the effectiveness and latency of de-duplication. De-duplication technology can accelerate backup efficiency and drive down IT costs. The implementation of De-duplication technique at client side, servers and also data centers will reduce the redundant data to much extent. Also the storage space will be utilized wisely by saving the only the unique data. Thus the availability of data and proper utilization of storage space can be managed with the De-duplication schemes.

REFERENCES

[1] Yinjin Fu, Hong Jiang, Nong Xiao, Lei Tian, Fang Liu, and Lei Xu, “Application-Aware Local-Global Source Deduplication for Cloud Backup Services of Personal Storage”, IEEE, 2013.

[2] Katiyar and J. Weissman, ‘‘ViDeDup: An Applica-tion-Aware Framework for Video De-Duplication,’’ in Proc. 3rd USENIX Workshop Hot-Storage File Syst., 2011.

[3] M. Armbrust, A. Fox, R. Griffith, A.D. Joseph, R. Katz, A. Konwinski, G. Lee, D. Patterson, A. Rabkin, I. Stoi-ca, and M. Zaharia, ‘‘A View of Cloud Computing,’’ ACM, 2010.

[4] Fu Y, Jiang H, Xiao N, et al. Application-Aware Local-Global Source Deduplication for Cloud Backup Services of Personal Storage[J]. 2013.

[5] Mao B, Jiang H, Wu S, et al. SAR: SSD Assisted Restore Optimization for Deduplication-Based Storage Systems in the Cloud[C]//Networking, Architecture and Storage (NAS), 2012 IEEE 7th International Conference on. IEEE,2012: 328-337.

[6] Yinjin Fu, Hong Jiang, Nong Xiao, Lei Tian, Fang Liu,’’ Application-Aware local global Source Deduplication for Cloud Backup Service of personal storage “ IEEE International Conference on Cluster Computinges in the Personal Computing-Environment(2012).

[7] Tan Y, Jiang H, Feng D, et al. CABdedupe: A causality-based deduplication performance booster for cloud backup services[C]//Parallel & Distributed Processing Symposium (IPDPS), 2011 IEEE International. IEEE, 2011: 1266-1277.