ISSN: 2321-7782 (Online)
Volume 4, Issue 1, January 2016
International Journal of Advance Research in Computer Science and Management Studies
Research Article / Survey Paper / Case Study
Available online at: www.ijarcsms.com

Survey on Cloud Backup Services of Personal Storage

Pradeep Sarjerao Jagtap[1]
Computer Department
TSSM's Bhivarabai Sawant College of Engineering and Research
Pune, India

Prof. A. D. Gujar[2]
Computer Department
TSSM's Bhivarabai Sawant College of Engineering and Research
Pune, India

Abstract: In widespread cloud environment cloud services is tremendously growing due to large amount of personal computation data. Deduplication process is used for avoiding the redundant data. A cloud storage environment for data backup in personal computing devices facing various challenge, of source deduplication for the cloud backup services with low deduplication efficiency. Challenges facing in the process of deduplication for cloud backup service are-1)Low deduplication efficiency due to exclusive access to large amount of data and limited system resources of PC based client site.2)Low data transfer efficiency due to transferring deduplicated data from source to backup server are typically small but that can be often across the WAN.

Key words: Cloud computing; Deduplication; chunking scheme; cloud backup; application awareness.

I. INTRODUCTION

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. Cloud computing is more striking a service having a great potential to alter the large part of the IT industry. Cloud computing is the centralized storage for the data and it also provides the online access to various computer services and resources. Cloud computing broadly focuses on maximizing the efficiency of shared resources. Cloud backup for end user's is nothing but an unlimited amount of data storage space which is secure and highly available for backup data from personal computing devices.

Data deduplication an effective technology for eliminating the redundant data in backup data. 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.

The propose approach of application aware Local-Global source deduplication it not only formulate use of application awareness but also combines the local and global duplicate data detection. ALG-Dedupe scheme helps to achieve not only higher deduplication efficiency by reducing deduplication latency but also saves the cloud storage cost. Application awareness adapts different types of applications independently during the local and global duplicate check process, which helps to reduce the system. The proposed system combines local deduplication and global deduplication to balance the effectiveness and latency of duplicate data.

II. LITERATURE SURVEY

The existing source deduplication strategies can be divided into two categories: 1] Local source deduplication 2] Global source deduplication. Local source deduplication [5] only detects redundancy in backup dataset from the same device at the client side and only sends the unique data chunks to the cloud storage. Local deduplication eliminates intra-client redundancy with low duplicate elimination ratio by low-latency client-side duplicate check. Global source deduplication [6] performs duplicate check in backup datasets from all clients in cloud side before data transfer over WAN. Global source deduplication has intra-client and inter-client redundancy with high-latency duplication detection on cloud side.

In paper [4], Cloud4Home enhances data services by combining limited local resources with low latency and powerful Internet resources with high latency. It has local-global source deduplication scheme that eliminates intra-client redundancy at client before suppression inter-client redundancy in the cloud, The scheme can potentially improve deduplication efficiency in cloud backup services to save as much cloud storage space as the global method but at as low latency as the local mechanism.

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 [2] improves redundancy detection by application-specific chunking methods that exploit the knowledge about concrete file formats. ViDeDup [3] is a frame-work for video deduplication based on an application-level view of redundancy at the content level rather than at the byte level.

All the related and prior work related to deduplication focus only on the effectiveness of deduplication. They are designed to remove more redundancy from the data. Earlier systems have not considered the system over-heads for high efficiency in deduplication process.

III. 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.

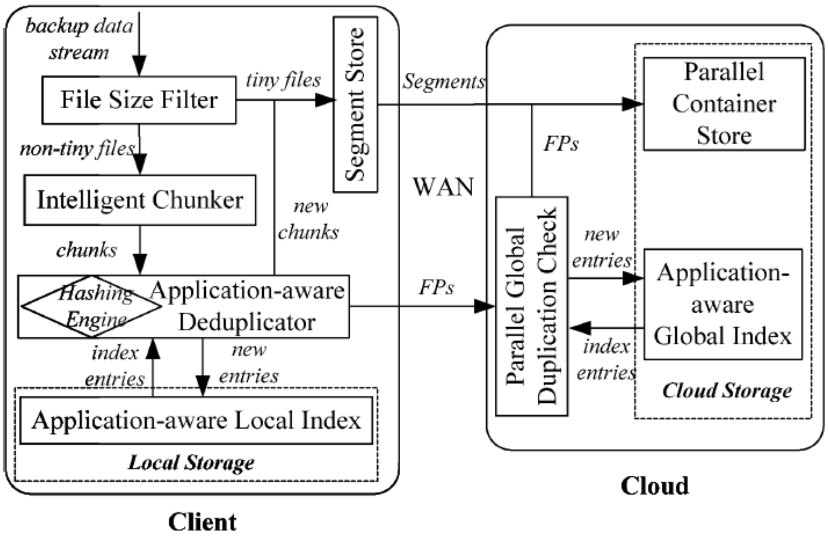

Fig. 1 Application aware local -Global deduplication scheme

The propose system contains four main components:

A. *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.

B. *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 the basis of frequency based or content based. CBC is the stateless chunking algorithm which divides the long stream of data into smaller units by removing duplicates. To achieve a better mean balance between deduplication ratio and deduplication overhead, we deduplicate compressed files with WFC for its low sub-file redundancy.

C. *Applications-Aware Deduplicator*

After performing the data chinking in Chunker, the deduplication of the data chunks will be performed in application aware, which generates the FP in hash engine of the data chunks and detecting replica chunks in both local client and remote cloud. Application aware local and global deduplication strikes the superior stability between computation overhead on client side and hash collision will be avoided to keep data reliability. For performing deduplication on client side and on global cloud it requires two types of application aware indices such as local index on client side and global index on cloud side. ALG Deduplication performs the periodic synchronization to backup application aware index and to protect the data integrity of PC backup datasets.

D. *Segment Management*

To reduce the cost of cloud storage and avoiding higher overhead of network protocol due to small file transfer, Application aware deduplication will group the deduplicate data of various smaller files and chunk into large units that will be known as segments. These group of deduplicate data will be stored in the segment store before transferring these data over WAN.

E. *Container Management*

A container is nothing but the self describing data structure in chunk descriptors for stored chunk. Container is maintained for each arriving backup data stream. These backup data is nothing but the segments send over the cloud and which will be routed to store node over the cloud with its respective fingerprints.

IV. CONCLUSION

The proposed scheme is used for cloud backup in the personal computing environment to improve deduplication efficiency. The proposed scheme minimizes computational over-head by using file semantics and maximizes deduplication effectiveness by using application awareness.

ACKNOWLEDGEMENT

I have taken efforts in this review of deduplication for cloud backup services of personal storage. However, it would not have been possible without the kind support and help of many individuals. I am highly indebted to Prof. A. D. Gujar for his guidance and constant supervision as well as for providing necessary information regarding this approach.

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. C. Liu, Y. Lu, C. Shi, G. Lu, D. Du, and D.-S. Wang, ''ADMAD: Application-Driven Metadata Aware De-Deduplication Archival Storage Systems,'' in Proc. 5th IEEE Int'l Workshop SNAPI I/Os, 2008, pp. 29-35.
3. A. Katiyar and J. Weissman, ''ViDeDup: An Application-Aware Framework for Video De-Duplication,'' in Proc. 3rd USENIX Workshop Hot-Storage File Syst., 2011.
4. S. Kannan, A. Gavrilovska, and K. Schwan, ''Cloud4HomeV Enhancing Data Services with @Home Clouds,'' in Proc. 31st ICDCS, 2011, pp. 539-548.
5. Y. Fu, H. Jiang, N. Xiao, L. Tian, and F. Liu, ''A Dedupe: An Application-Aware Source Deduplication Approach for Cloud Backup Services in the Personal Computing Environment,'' in Proc. 13th IEEE Int'l Conf. CLUSTER Comput., 2011.
6. P. Anderson and L. Zhang, ''Fast and Secure Laptop Backups with Encrypted De-Duplication,'' in Proc. 24th Int'l Conf. LISA, 2010, pp. 29-40.

AUTHOR(S) PROFILE

Pradeep Sarjerao Jagtap, is currently pursuing M.E (Computer) from Department of Computer Engineering, TSSM's Bhivarabai Sawant College of Engineering and Research, Pune, India at Savitribai Phule Pune University. He received his B.E (Information Technology) Degree from Annasaheb Dange College of Engineering & Technology, Ashta, India at Shivaji University Kolhapur. His area of interest is cloud computing.

Anil Gujar, received the M.Tech (IT) degree from the Department of Information Technology, Bharati Vidyapeeth Pune, India. He is currently working as Asst. Professor with Department of Computer Engineering, Bhivarabai Sawant College of Engineering and Research, Pune, India.