e-ISSN : 2349-9745
p-ISSN : 2393-8161

Scientific Journal Impact Factor (SJIF): 1.711

International Journal of Modern Trends in Engineering and Research

www.ijmter.com

A Hybrid Cloud Approach for Secure Authorized De-Duplication

Divya.K[1], Santhosh Kumar.M[2],Sudha.C[3],Ramya Devi.R[4]

[1]*Department of ME-CSE, Srinivasan Engineering College,*

[2]*Assistant Professor/CSE, Srinivasan Engineering College,*

[3]*Department of ME-CSE, Srinivasan Engineering College,*

[4]*Department of ME-CSE, Srinivasan Engineering College,*

Abstract— The cloud backup is used for the personal storage of the people in terms of reducing the mainlining process and managing the structure and storage space managing process. The challenging process is the deduplication process in both the local and global backup de-duplications. In the prior work they only provide the local storage de-duplication or vice versa global storage de-duplication in terms of improving the storage capacity and the processing time. In this paper, the proposed system is called as the ALG- Dedupe. It means the Application aware Local-Global Source De-duplication proposed system to provide the efficient de-duplication process. It can provide the efficient de-duplication process with the low system load, shortened backup window, and increased power efficiency in the user's personal storage. In the proposed system the large data is partitioned into smaller part which is called as chunks of data. Here the data may contain the redundancy it will be avoided before storing into the storage area.

Keywords*:* Data Deduplication, Backup Window,ALG Dedupe,Cloud Backup

I. INTRODUCTION

Cloud computing is the delivery of computing services over the Internet. Cloud services allow individuals and businesses to use software and hardware that are managed by third parties at remote locations. Data deduplication, an effective data compression approach that exploits data redundancy, partitions large data objects into smaller parts, called chunks, represents these chunks by their fingerprints, replaces the duplicate chunks with their fingerprints index lookup, and only transfers or stores the unique chunks for the purpose of communication or storage efficiency. Source Deduplication that eliminates redundant data at the client site is obviously preferred to target deduplication due to the former's ability to significantly reduce the amount of data transferred over wide area network with low communication bandwidth.

A hybrid cloud is a combination of different methods of resource pooling (for example, combining public and community clouds).Cloud services is popular Cloud services are popular because they can reduce the cost and complexity of owning and operating computers and networks. Since cloud users do not have to invest in information technology infrastructure, purchase hardware, or buy software licenses, the benefits are low up-front costs, rapid return on investment, rapid deployment, customization, flexible use, and solutions that can make use of new innovations. Some other benefits to users include scalability, reliability, and efficiency. Scalability means that cloud computing offers unlimited processing and storage capacity. The cloud is reliable in that it enables access to applications and documents anywhere in the world via the Internet. Cloud computing is often considered efficient because it allows organizations to free up resources to focus on innovation and product development.

Another potential benefit is that personal information may be better protected in the cloud. Specifically, cloud computing may improve efforts to build privacy protection into technology from the start and the use of better security mechanisms. ALG-Dedupe outperforms the existing state-of-

the-art source deduplication schemes in terms of backup window, efficiency and cost saving for its high deduplication efficiency and low system overhead. Thus, the basic idea of ALG-Dedupe is to effectively exploit this application difference and awareness by treating different types of applications independently and adaptively during the local and global deduplication.

II. RELATED WORK

First process is the input file selection. Select the input file. Then input file has been split into the data chunks. After the data chunks has been loaded into the database, based on the input file. Then calculate the backup window size. In ALG-Dedupe filters out these tiny files in the file size filter before the deduplication process, and groups data from many tiny files together into larger units of about 1 MB each in the segment store to increase the data transfer efficiency over WAN.

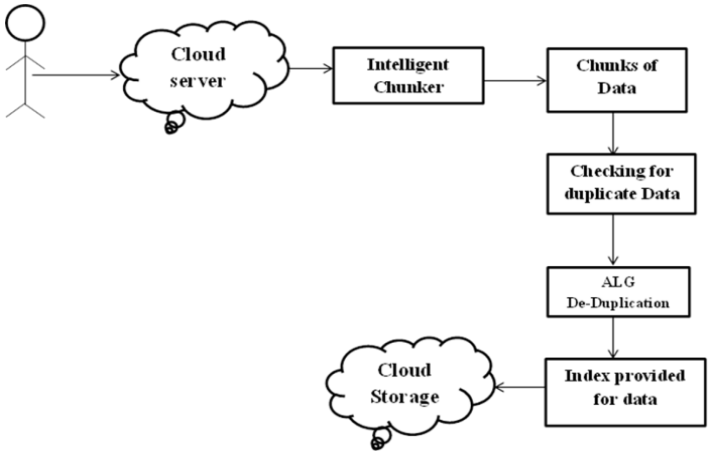

Figure 1 system architecture

An Application aware Local-Global source deduplication scheme that not only exploits application awareness, but also combines local and global duplication detection, to achieve high deduplication efficiency by reducing the deduplication latency to as low as the application-aware local deduplication while saving as much cloud storage cost as the application-aware global deduplication. Most of the files in the PC dataset are tiny files that less than 10 KB in file size, accounting for a negligibly small percentage of the storage capacity. The statistical evidences about 63 percent of all files are tiny files, accounting for only 1.9 percent of the total storage capacity of the dataset. To reduce the metadata overhead. The deduplication efficiency of data chunking scheme among different applications differs. Depending on whether the file type is compressed or whether SC can outperform CDC in deduplication efficiency, The file can be classified into three main categories: compressed, static compressed files, and dynamic uncompressed files. The dynamic files are always editable, while the static files are uneditable in common. To strike a better tradeoff between duplicate elimination ratio and deduplication, An deduplicate 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 chunk size using CDC based on the Rabin finger printing to identify chunk boundaries.

While there are benefits, there are privacy and security concerns too. Data is travelling over the Internet and is stored in remote locations. In addition, cloud providers often serve multiple customers simultaneously. All of this may raise the scale of exposure to possible breaches, both accidental and deliberate. Concerns have been raised by many that cloud computing may lead to "function creep" — uses of data by cloud providers that were not anticipated when the information was originally collected and for which consent has typically not been obtained. Given how inexpensive it is to keep data, there is little incentive to remove the information from the cloud. Some other benefits to users include scalability, reliability, and efficiency. Scalability means that

cloud computing offers unlimited processing and storage capacity. The cloud is reliable in that it enables access to applications and documents anywhere in the world via the Internet. Cloud computing is often considered efficient because it allows organizations to free up resources to focus on innovation and product development.

A hybrid cloud is a combination of different methods of resource pooling (for example, combining public and community clouds). Since cloud users do not have to invest in information technology infrastructure, purchase hardware, or buy software licenses, the benefits are low up-front costs, rapid return on investment, rapid deployment, customization, flexible use, and solutions that can make use of new innovations. Depending on the location where redundant data is eliminated. The deduplication can be categorized into source deduplication that applies data deduplication at the client site and target deduplication that eliminates redundant data at the backup server site. Since data backup for personal computing in the cloud storage environment implies geographic separation between the client and the service provider.

III. SCOPE OF THE PAPER

Present and evaluate a ALG Dedupelication to eliminate the redundant data. This work is focuses on designed to meet the requirement of deduplication efficiency with high deduplication effectiveness and low system overhead. The main idea of ALGDedupe 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 based on application awareness, and 2) to mitigate the on-disk index lookup bottleneck by dividing the full index into small independent and application-specific indices in an application-aware index structure. It combines local-global source deduplication.

Data chunking in intelligent chunker module, data chunks will be deduplicated in the application-aware by generating chunk fingerprints in the hash engine and detecting duplicate chunks in both the local client and remote cloud. An employ extended 12-byte Rabin hash value as chunk fingerprint for local duplicate data detection and a MD5 value for global duplicate detection of compressed files with WFC. In both local and global detection scenarios, a SHA-1 value of chunk serves as chunk fingerprint of SC in static uncompressed files and MD5 value is used as chunk fingerprint of dynamic files since chunk length is another dimension for duplicate detection in CDC-based deduplication. To achieve high efficiency, application aware deduplicator first detects duplicate data in the application-aware index corresponding to the local dataset with low deduplication latency in the PC client, and then compares local deduplicated data chunks with all data stored in the cloud by looking up fingerprints in the application-aware global index on the cloud side for high data reduction ratio. Only the unique data chunks after global duplicate detection are stored in the cloud storage.

An application-aware index structure for ALG-Dedupe is constructed It consists of an in-RAM application index and small hash-table based on-disk indices classified by application type. According to accompanied file type information, the incoming chunk is directed to the chunk index with the same file type. Each entry of the index stores a mapping from the fingerprint(fp) of a chunk or with its length (len) to its container ID(cid).As chunk locality exists in backup data streams a small index cache is allocated in RAM to speedup index lookup by reducing disk I/O operations. The index cache is a key-value structure, and it is constructed by a doubly linked list indexed by a hash table. When the cache is full, fingerprints of those containers that are ineffective in accelerating chunk fingerprint lookup are replaced to make room for future prefetching and caching.

Aggregation of data produces larger files for the cloud storage, which can be beneficial in avoiding high overhead of lower layer network protocols due to small transfer sizes, and in reducing the cost of the cloud storage. Amazon S3, for example, has both a per-request and a per-byte cost when storing a file, which encourages the use of files greater than 100 KB. ALG-Dedupe will often group deduplicated data from many smaller files and chunks into larger units called segments before these data are 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 be packed into container, a data stream based structure, to keep spatial locality for 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. When a container fills up with a predefined fixed size.

Where the former removes duplicate data at the file granularity with low duplicate elimination effectiveness and low computational overhead, while the latter removes the duplicate data at the sub-file (that is, chunk) level with high duplicate elimination effectiveness and high computational overhead. To achieve high effectiveness of deduplication, source chunk-level deduplication has become popular and represents state of the art. However, such fine-grained data deduplication is very expensive in terms of memory and processing especially on resource-constrained clients.

IV. CONCLSION

A class of deduplication systems splits the data stream into data blocks (chunks) and then finds exact duplicates of these blocks. ALG-Dedupe improves the deduplication ratio of AA-Dedupe by further leveraging global deduplication with cloud computing. It achieve high deduplication efficiency by reducing the deduplication latency to as low as possible. The application-aware local deduplication while saving as much cloud storage cost as application-aware global deduplication.

The primary goal of this work is to develop a scheme that, eventually, allows for efficient data deduplication by utilizing data file pattern. The proposed scheme efficiently performs Fixed-length Chunking for Head Section and End Section with overhead.Itcombines local deduplication and global deduplication to balance the effectiveness and latency of deduplication. The proposed application-aware index structure can significantly relieve the disk index lookup bottleneck by dividing a central index into many independent small indices to optimize lookup performance. As future work to encrypt the file and after the chunking process is executed.

REFERENCES

[1] M. Armbrust, A. Fox, R. Griffith, A.D. Joseph, R. Katz,A. Konwinski, G. Lee, D. Patterson, A. Rabkin, I. Stoica, andM. Zaharia,(Apr 2010), ‘‘A View of Cloud Computing,’’Commun. ACM,vol. 53, no. 4, pp. 49-58.
[2]H.Biggar,(Feb 2007),‘‘Dupless:Server-Aided Encryption For De-Duplicated Storage,’’Enterprise Strategy Grp., Milford, MA, USA, White Paper.
[3] C.Liu, Y. Lu, C. Shi, G. Lu, D. Du, and D.S.Wang,(2008),‘‘ADMAD:Application-DrivenMetadata Aware De-Deduplication Archival Storage Systems,’’ in Proc. 5th IEEE Int’l Workshop SNAPI I/Os,pp. 29-35.
[4] Y. Tan,H. Jiang, D. Feng, L. Tian, Z. Yan, and G. Zhou,(2010),‘‘SAM: A Semantic-Aware Multi-Tiered Source De-Duplication Frame-Work for Cloud Backup,’’ in Proc. 39th ICPP,pp. 614-623.
[5] A. Muthitacharoen, B. Chen, and D. Mazie` res,(2001),‘‘A Low- Bandwidth Network File System,’’ in Proc. 18th ACM SOSP, pp. 174-187.
[6] S. Kannan, A. Gavrilovska, and K. Schwan,(2011),‘‘Cloud4HomeV Enhancing Data Services with @Home Clouds,’’ in Proc. 31st ICDC, pp. 539-548.
[7] Maximizing Data Efficiency: Benefits of Global Deduplication-NEC, Irving, TX, USA, NEC White Paper, 2009.
[8] D. Meister and A. Brinkmann,(2009), ‘‘Secure Deduplication with encrypted data for Cloud Storage,’’ in Proc. 2nd Annu. Int’l SYSTOR, pp. 1-8.
[9] D. Bhagwat, K. Eshghi, D.D. Long, and M. Lillibridge,(Sept 2009),‘‘A Reverse Deduplication Storage System Optimized For Reads To Latest Backups,’’ HP Lab., Palo Alto, CA, USA, Tech. Rep. HPL-2009-10R2.
[10] K. Eshghi,(2005),‘‘A Weak Leakage-Resilient Client-Side Deduplication Of Encrypted Data In Cloud Storage,’’ HP Laboratories, Palo Alto, CA,USA, Tech. Rep. HPL-2005-30 (R.1).

[11] B. Zhu, K. Li, and H. Patterson,(Feb 2008), ‘‘Message Locked Encryption And Secure Deduplication,’’ in Proc. 6th USENIX Conf. FAST,pp. 269-282.
[12] M. Lillibridge, K. Eshghi, D. Bhagwat, V. Deolalikar, G. Trezise, and P. Camble,(2009),‘‘Sparse Indexing: Large Scale, Inline Deduplication Using Sampling and Locality,’’ in Proc. 7th USENIX Conf. FAST,pp. 111-123.
[13] P. Anderson and L. Zhang,(2010),‘‘Fast and Secure Laptop Backups With Encrypted De-Duplication,’’ in Proc. 24th Int’l Conf. LISA, pp. 29-40.
[14] F. Liu, L. Tian and L.Xu,(MAY 2014), “Application-Aware Local-Global Source Deduplication for Cloud Backup Services of personal Storage,’’ IEEE Trans.Parallel Distrib.Syst, Vol.25, NO. 5.
[15] P. Shilane, M. Huang, G. Wallace, and W. Hsu,(2012), ‘‘WAN Optimized Replication of Backup Datasets Using Stream-Informed Delta Compression,’’ in Proc. 10th USENIX Conf.FAST,pp. 49-64.
[16] F. Douglis, D. Bhardwaj, H. Qian, and P. Shilane,(Dec 2011),‘‘Content-Aware Load Balancing for Distributed Backup,’’ in Proc. 25th USENIX Conf. LISA, pp. 151-168.

