Scientific Journal of Impact Factor (SJIF): 4.72

e-ISSN (O): 2348-4470
p-ISSN (P): 2348-6406

International Journal of Advance Engineering and Research Development

Volume 4, Issue 11, November -2017

Review On Data Deduplication In Cloud Computing

Kinzal Patel[1], Prof. Kapildev Naina[2]

[1]*M. Tech, Department of Computer Engineering, School of Engineering RK University, Rajkot, India.*
[2]*Assistant Professor, Department of Computer Engineering, School of Engineering RK University, Rajkot, India.*

Abstract — *Data deduplication is a technique to improve the storage utilization. De-duplication technologies can be designed to work on primary storage as well as on secondary storage. However, there is only one copy for each file stored in cloud even if such a file is owned by a huge number of users. As a result, de-duplication system improves storage utilization while reducing reliability. De-duplication with the use of chunking data that is passed through the de-duplication engine is chunked into smaller units and assigned identities using cryptographic hash functions. Data deduplication technique allows the cloud users to manage their cloud storage space effectively by avoiding storage of repeated data's and save bandwidth. The data are finally stored in cloud server. To ensure data confidentiality the data are stored in an encrypted type using Advanced Encryption Standard (AES) algorithm.*

Keywords- *Data-deduplication; Hash Function; AES; Cryptography; Secondary Storage.*

I. INTRODUCTION

Cloud computing provides a low-cost, scalable, location independent infrastructure for data management and storage. Owing to the population of cloud service and the increasing of data volume, more and more people pay attention to economize the capacity of cloud storage than before .Therefore how to utilize the cloud storage capacity well becomes important issue nowadays. Data deduplication is a specialized data compression technique for eliminating duplicate copies of repeating data.

Data de-duplication, is a file system feature that only saves unique data segments to save space. It has been most popular and successful for secondary storage systems (backup and archival) Data de-duplication technology, the basic principle is to filter the data block to find the same data block, and the only instance of a pointer to point to replace. Data de-duplication technology to identify duplicate data, eliminate redundancy and reduce the need to transfer or store the data in the overall capacity. Duplication to detect duplicate data elements, to judge a file, block or bit and another file, block or bit the same. Data de-duplication technology to use mathematics for each data element, algorithms to deal with, and get a unique code called a hash authentication number. Each number is compiled into a list, this list is often referred to as hash index.

II. LITERATURE SURVEY

A number of deduplication systems have been proposed based on various deduplication strategies.

2.1 Message-Locked Encryption and Secure Deduplication

Bellare et al formalized this primitive as message-locked encryption, and explored its application in space efficient secure outsourced storage. Bellare et al showed how to protect data confidentiality by transforming the predictable message into unpredictable message [1].

2.2 A Hybrid Cloud Approach for Secure Authorized Deduplication

Li addressed the key-management issue in block-level deduplication by distributing these keys across multiple servers after encrypting the files [2].

2.3 HASBE: A Hierarchical Attribute-Based Solution for Flexible and Scalable Access Control in Cloud Computing

Z. Wang addressed the HASBE scheme for realizing scalable, flexible, and fine-grained access control in cloud computing. The HASBE scheme seamlessly incorporates a hierarchical structure of system users by applying a delegation algorithm to ASBE. HASBE not only supports compound attributes due to flexible attribute set combinations, but also achieves efficient user revocation because of multiple value assignments of attributes [3].

2.4 Accelerating Restore and Garbage Collection in Deduplication-Based Backup Systems via Exploiting Historical Information

M. Fu et al conclude that the fragmentation decreases the efficiencies of restore and garbage collection in deduplication-based backup systems. It is observed that the fragmentation comes in two categories: sparse containers and out-of-order containers. Sparse containers determine the maximum restore performance of a backup while out-of-order containers determine the required cache size to achieve the maximum restore performance. The ability of HAR to reduce sparse containers facilitates the garbage collection. It is no longer necessary to offline merge sparse containers, which relies on identifying valid chunks. He propose a Container-Marker Algorithm (CMA) that identifies valid containers instead of valid chunks. Since the metadata overhead of CMA is bounded by the number of containers, it is more cost effective than existing reference management approaches whose overhead is bounded by the number of chunks [4].

2.5 Improving Restore Speed for Backup Systems That Use Inline Chunk-Based Deduplication

M. Lillibridge addressed poor restore performance due to chunk fragmentation can be a serious problem for inline, chunk-based deduplicating backup systems: if nothing is done, restore performance can slow down orders of magnitude over the life of a system.

M. Lillibridge conclude that system designers use all available RAM for restoring a stream (including operating system cache space) for a single large forward assembly area and associated buffers. If the number of streams being restored at a time can vary, we recommend using more RAM per stream when fewer streams are being restored. Unless deduplication is at a great premium, at least a small amount of capping should be employed [5].

2.6 Protect Pervasive Social Networking Based on Two Dimensional Trust Levels

Z. Yan introduces the scheme seamlessly incorporates a hybrid trust management framework for PSN by applying ABE. The PSN can be automatically secured since the related cryptographic keys can be automatically managed based on two dimensions of node trust levels, node revocation, and the validity period of the keys [6].

III. Methodology

The data de-duplication technique is used to store single instance of redundant data and eliminates the duplicate data in datacenter. It is used to decrease the size of datacenter and reduce the replication of data that were duplicated on cloud. The de-duplication process helps to remove any block or file that are not unique and store in smaller group of blocks . The basic steps for data de-duplication process are.

- The files are converted into small segments.
- Then new and existing data are checked for redundancy
- Metadata are updated and segments are compressed.
- Duplicate data are deleted and check the data integrity.

3.2 Data-deduplication levels

Deduplication strategy can be categorized into two main strategies as follow, differentiated by the type of basic data Units[9].

3.2.1 File-level deduplication

A file is a data unit when examining the data of duplication, and it typically uses the hash value of the file as its identifier. If two or more files have the same hash value, they are assumed to have the same contents and only one of these files will be stored.

Advantage:

1. If any change is made in a file it makes to save the whole file again in file level deduplication.
2. In file level deduplication indexes are small, and so it takes less time for computational when it identifies the duplicate copies.

3.2.2 Block-level deduplication

This strategy segments a file into several fixed-sized blocks or variable-sized blocks, and computes hash value for each block for examining the duplication blocks.

Advantage:

1. Block level deduplication can eliminate or delete the small redundant chunk of data when compared to whole file.
2. Each and every file system can use same deduplication algorithm in block level deduplication.

VI. CONCLUSION

Cloud computing has reached a maturity that leads it into a productive phase. This means that most of the main issues with cloud computing have been addressed to a degree that clouds have become interesting for full commercial exploitation. This however does not mean that all the problems listed above have actually been solved, only that the according risks can be tolerated to a certain degree.

Managing encrypted data with deduplication is significant in practice for running a secure, dependable, and green cloud storage service, especially for big data processes. To secure the confidentiality of sensitive data during deduplication, the convergent encryption technique is used to encrypt the data before outsourcing.

REFERENCES

[1] M. Bellare, S. Keelveedhi, and T. Ristenpart, “Message-Locked Encryption and Secure Deduplication,” Advances in Cryptology (EUROCRYPT 13), LNCS 7881, 2013, pp. 296–312.

[2] J. Li et al., “A Hybrid Cloud Approach for Secure Authorized Deduplication,” IEEE Trans. Parallel Distributed Systems, vol. 26, no. 5, 2015, pp.1206–1216.

[3] Z. Wan, J. Liu, and R.H. Deng, “HASBE: A Hierarchica Attribute-Based Solution for Flexible and Scalable Access Control in Cloud Computing,” IEEE Trans. Information Forensics and Security, vol. 7, no. 2, 2012, pp. 743–754.

[4] M. Fu et al., “Accelerating Restore and Garbage Collection in Deduplication-Based Backup Systems via Exploiting Historical Information,” Proc. Usenix Ann. Technical Conf., 2014, pp. 181–192.

[5] M. Lillibridge, K. Eshghi, and D. Bhagwat, “Improving Restore Speed for Backup Systems That Use Inline Chunk-Based Deduplication,” *Proc. 11th Usenix Conf. File and Storage Technologies*, 2013, pp. 183–198.

[6] Z. Yan and M.J. Wang, “Protect Pervasive Social Networking Based on Two Dimensional Trust Levels,” *IEEE Systems J.*, Sept. 2014, pp. 1–12; doi: 10.1109/JSYST.2014.2347259.

[7] Iuon –Chang Lin, Po-chingChien ,”Data Deduplication Scheme for Cloud Storage” International Journal of Computer and Control(IJ3C),Vol1,No.2(2012).

[8] Z. Yan, W. Ding, and H. Zhu, “Manage Encrypted Data Storage with Deduplication in Cloud,” Proc. Int’l Conf. Algorithms and Architectures for Parallel Processing (ICA3PP), 2015, pp. 547–561.

[9] Priyadharsini, Dhamodran, Kavitha,”A Survey On Deduplication In Cloud Computing”, IJCSMC, Vol. 3, Issue. 11, November 2014, pg.149 – 155.

[10] P.Neelaveni,M.VijayaLakshmi,”A Survey On Deduplication in Cloud Storage”,Asian journal Of Information Technology 13(6):320-330©Medwell journals,2014.