ISSN(Online): 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 7, Issue 12, December 2018

Survey Paper on Efficient Data Storage Management with Deduplication in Cloud Computing

Soudagar Shinde, Manoj Thakur, Shriram Ghotekar, DhirajNavale.

B. E Students, Department of Computer Science Engineering, Sinhgad Academy Of Engineering , Kondhwa, Pune, India.

ABSTRACT*:* Cloud storage as one of the most important cloud computing services helps cloud users overcome the bottleneck of limited resources and expand storage without upgrading their devices. To ensure the safety and privacy of cloud users, data is always outsourced in encrypted form. However, encrypted data could generate a lot of storage waste in the cloud and complicate the exchange of data between authorized users. We are still facing challenges in storage and management encrypted data with deduplication. Traditional deduplication schemes always focus on specific application scenarios, where deduplication is completely controlled by data owners or servers in the cloud. They cannot flexibly satisfy the different requests from data owners based on the level of sensitivity of the data. In this paper, scheme that flexibly offers both deduplication management and access control at the same time through multiple cloud service providers (CSPs). We evaluate your performance with security analysis, comparison and implementation. The results show its safety, effectiveness and efficiency towards a possible practical use

KEYWORDS: Data Deduplication, Cloud Computing, Access Control, Storage Management

I. INTRODUCTION

Although the storage system in the cloud has been adopted mostly does not meet some important emerging needs, such as the ability to verify the integrity of files in the cloud by customers in the cloud and the detection of duplicate files on servers in the cloud. We report both problems below. These servers in the cloud can free customers from the heavy burden of storage management and maintenance. The biggest difference between cloud storage and traditional internal storage is that data is transferred over the Internet and stored in an uncertain domain, which is not under the control of customers, which inevitably raises major concerns about your data integrity. These concerns stem from the fact that cloud storage is affected by security threats both outside and inside the cloud, and servers in the uncontrolled cloud can passively hide some episodes of customer data loss to maintain its reputation What is more serious is that to save money and space, cloud servers can even exclude an active and deliberate data file that we only have access to and belong to a common customer. Given the large size of outsourced data files and the limited capacity of customer resources, the first problem is widespread so the customer can perform integrity checks effectively, even without a local copy of the data file. Cloud computing is computing in which large groups of remote servers are networked to allow centralized data storage and online access to services or IT resources.

With cloud computing, large groups of resources can be connected via a private or public network. In the public cloud, services (that is, applications and storage space) are available for general use on the Internet. A private cloud is a virtualized data center that operates within a firewall. Cloud computing provides computing and storage resources on the Internet. The increasing amount of data is stored in the cloud, and users with specific privileges share it, which defines special rights to access stored data. Managing the exponential growth of a growing volume of data has become a critical challenge. According to the IDC 2014 cloud report, companies in India are gradually moving from the legacy of premise to different forms of cloud. As the process is gradual, it began during the migration of some cloud application workloads. To perform scalable management of data stored in cloud computing, deduplication has been a

ISSN(Online): 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 7, Issue 12, December 2018

well-known technique that has become more popular recently. Deduplication is a specialized data compression technique that reduces storage space and charges bandwidth in cloud storage. In deduplication, only a single instance of data is actually on the server and the redundant data is replaced with a pointer to the copy of the unique data. Deduplication can occur at the file or block level. From the user's point of view, security and privacy issues arise, as data is susceptible to internal and external attacks. We must properly apply the confidentiality, integrity verification and access control mechanisms of both attacks. Deduplication does not work with traditional cryptography. The user encrypts their files with their own individual encryption key, a different encryption text may also appear for identical files. Therefore, traditional cryptography is incompatible with data duplication. Converged encryption is a widely used technique for combining storage savings with deduplication to ensure confidentiality. In converged encryption, data copy is encrypted with a key derived from the data hash. This converging key is used to encrypt and decrypt a copy of data. After key generation and data encryption, users keep keys and send encrypted text to the cloud. Because cryptography is deterministic, copies of identical data will generate the same convergent key and the same encrypted text. This allows the cloud to duplicate encrypted texts. Cryptographic texts can only be decrypted by the owners of the corresponding data with their converging keys. Differential authorization duplication control is an authorized duplication elimination technique in which each user is granted a set of privileges during system initialization. This privilege set specifies what types of users can perform duplicate checks and access files.

A. MOTIVATION

- Data deduplication has been widely deployed in storage systems for space savings, the fingerprint-based deduplication approaches have an inherent drawback: they often fail to detect the similar chunks that are largely identical except for a few modified bytes, because their secure hash digest will be totally different even only one byte of a data chunk was change.
- It becomes a big challenge when applying data deduplication to storage datasets and workloads that have frequently modified data, which demands an effective and efficient way to eliminate redundancy among frequently modified and thus similar data.
- One of the main challenges facing the application of deduplication systems

II. RELATED WORK

Literature survey is the most important step in any kind of research. Before start developing we need to study the previous papers of our domain which we are working and on the basis of study we can predict or generate the drawback and start working with the reference of previous papers.

In this section, we briefly review the related work on Data Deduplication and their different techniques.

M. Lillibridge, K. Eshghi, and D. Bhagwat [2013] represents the improvement in recovery speed for backup systems that use online block-based deduplication. The slow recovery due to the fragmentation of the pieces is a serious problem faced by the one-piece data deduplication systems: the recovery speeds for the most recent backup can eliminate orders of magnitude during the life cycle of a system We have studied three techniques: increase the size of the cache, limit the containers and use a direct assembly area to solve this problem. Container limitation is a time-consuming task that reduces fragmentation of fragments at the cost of losing part of deduplication, while using a direct assembly area is a new technique of recovering and caching in the process of recovery that takes advantage of the perfect knowledge of future access to the fragments available when restoring a backup to reduce the amount of RAM needed for a certain level of caching upon recovery [1].

D. Meister, J. Kaiser, and A. Brinkmann [2013] have represented Block locality caching for data deduplication. The author propose a novel approach, called Block Locality Cache (BLC), that captures the previous backup run significantly better than existing approaches and also always uses up-to-date locality information and which is, therefore, less prone to aging.We evaluate the approach using a trace-based simulation of multiple real-world backup datasets. The simulation compares the Block Locality Cache with the approach of Zhu et al. [37] and provides a

ISSN(Online): 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 7, Issue 12, December 2018

detailed analysis of the behavior and IO pattern. Furthermore, a prototype implementation is used to validate the simulation [2].

D. T. Meyer and W. J. Bolosky [2012] has represents A study of practical Deduplication. We collected file system content data from 857 desktop computers at Microsoft over a span of 4 weeks. We analyzed the data to determine the relative efficacy of data deduplication, particularly considering whole-file versus block-level elimination of redundancy. We found that whole-file deduplication achieves about three quarters of the space savings of the most aggressive block-level deduplication for storage of live file systems, and 87% of the savings for backup images. We also studied file fragmentation finding that it is not prevalent, and updated prior file system metadata studies, finding that the distribution of file sizes continues to skew toward very large unstructured files [3].

V. Tarasov, A. Mudrankit, W. Buik, P. Shilane, G. Kuenning, and E. Zadok [2012] have represents Generating realistic datasets for deduplication analysis.The author developed a generic model of file system changes based on properties measured on terabytes of real, diverse storage systems. Our model plugs into a generic framework for emulating file system changes. Building on observations from specific environments, the model can generate an initial file system followed by ongoing modifications that emulate the distribution of duplicates and file sizes, realistic changes to existing files, and file system growth [4].

P. Shilane, M. Huang, G. Wallace, and W. Hsu,[2012] has discover WAN optimized replication of backup datasets using stream-informed delta compression. Replicating data off site is critical for disaster recovery reasons, but the current approach of transferring tapes is cumbersome and error prone. Replicating across a wide area network (WAN) is a promising alternative, but fast network connections are expensive or impractical in many remote locations, so improved compression is needed to make WAN replication truly practical. We present a new technique for replicating backup datasets across a WAN that not only eliminates duplicate regions of files (deduplication) but also compresses similar regions of files with delta compression, which is available as a feature of EMC Data Domain systems [5].

III. EXISTING APPROACHES

Existing solutions for deduplication suffer from brute-force attacks. They cannot friendly support data access control and revocation at the same time. Most existing solutions cannot ensure reliability, security and privacy with sound performance. First data holders may not be always online or available for each a management, which could come storage delay. Second deduplication could become too complicated in the term of communication and computation to involve data holder into deduplication process. Third, it may intrude the privacy of data holder in a process of discovering duplicated data. Forth a data holder may have no idea how to issue data access right or deduplication key to users in some situation when it does not know other data holders due to data suffer distribution. Therefore, CSP cannot cooperate with data holders on data storage deduplication in many situations.

IV. PROPOSED APPROACHES

In this paper, we propose a scheme trusted on data ownership challenge and encryption to manage encrypted data storage with deduplication. We aim to solve the issue of deduplication in the situation where the data holder it not available or difficult to get involved. Meanwhile, the performance of data deduplication in our scheme is not influenced by the size of data. We motivate to save cloud storage and preserve the privacy of data holders by proposing scheme to manage encrypted data storage with deduplication. We prove the security and assess the performance of the proposed scheme through analysis and simulation. The results show its efficiency, effectiveness and applicability.

ISSN(Online): 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 7, Issue 12, December 2018

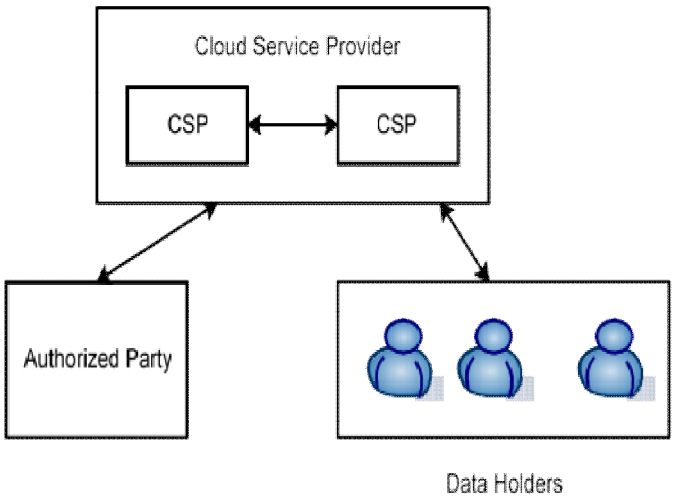

Fig. System Architecture

V. CONCLUSION

Data deduplication is important and significant in the practice of data storage in the cloud, in particular for the management of big data filing. In this paper, we proposed a heterogeneous data storage management scheme, which offers flexible data deduplication in the cloud and access control. Our schema can be adapted to different scenarios and application requests and offers cost-effective management of big data storage across multiple CSPs. Data deduplication and access control can be achieved with different security requirements.Security analysis, comparison with existing work and implementation-based performance evaluation have shown that our scheme is safe, advanced and efficient.

REFERENCES

[1] D. Meister, J. Kaiser, and A. Brinkmann, "Block locality caching for data deduplication," in Proc. 6th Int. Syst. Storage Conf., 2013, pp. 1–12.
[2] M. Lillibridge, K. Eshghi, and D. Bhagwat, "Improving restore speed for backup systems that use inline chunk-based deduplication," in Proc. 11th USENIX Conf. File Storage Technol, Feb. 2013, pp. 183–197.
[3] V. Tarasov, A. Mudrankit, W. Buik, P. Shilane, G. Kuenning, and E. Zadok, "Generating realistic datasets for deduplication analysis," in Proc. USENIX Conf. Annu. Tech. Conf., Jun. 2012, pp. 261–272.
[4] D. T. Meyer and W. J. Bolosky, "A study of practical deduplication," ACM Trans. Storage, vol. 7, no. 4, p. 14, 2012.
[5] G. Wallace, F. Douglis, H. Qian, P. Shilane, S. Smaldone, M. Chamness, and W. Hsu, "Characteristics of backup workloads in production systems," in Proc. 10th USENIX Conf. File Storage Technol., Feb.2012,pp.33–48.