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

Shubham Bhandari.
ME Student, Department of Computer Engineering,
JSPM'S Imperial college of Engineering and Research, Wagholi Pune.

S. T. Waghmode.
Professor, Department of Computer Engineering,
JSPM'S Imperial college of Engineering and Research, Wagholi Pune.

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 security 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 storing and managing 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 of data owners based on the level of data sensitivity. In this paper, we propose a heterogeneous data storage management scheme that flexibly offers both deduplication management and access control at the same time across 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- Cloud Computing, Data Deduplication, Access Control, Storage Management.

I. INTRODUCTION

Even though cloud storage system has been mostly adopted, it fails to accommodate some important emerging needs such as the capability of auditing integrity of cloud files by cloud clients and detecting duplicated files by cloud servers. We disclose both problems below. These cloud server is able to relieve clients from the bulky burden of storage management and maintenance. The most difference of cloud storage from traditional in-house storage is that the data is transferred via Internet and stored in an uncertain domain, not under control of the clients at all, which inevitably raises clients great concerns on the integrity of their data. These concerns originate from the fact that the cloud storage is affected to security threats from both outside and inside of the cloud, and the uncontrolled cloud servers may passively hide some data loss incidents from the clients to maintain their reputation. What is more serious is that for saving money and space, the cloud servers might even actively and deliberately discard barely accessed data files belonging to an ordinary client. Considering the large size of the outsourced data files and the clients constrained resource capabilities, the first problem is generalized as how can the client efficiently perform regularly integrity verifications even without the local copy of data file. Cloud computing is computing in which large groups of remote servers are networked to allow centralized data storage and online access to computer services or resources.

Cloud computing, large pools of resources can be connected through private or public network. In public cloud, services (i.e. applications and storage) are available for general use over the internet. A private cloud is a virtualized data center that operates within a firewall. Cloud computing provides computation and storage resources on the Internet. Increasing amount of data is being stored in the cloud and it is shared by users with specified privileges, which defines special rights to access stored data. Managing the exponential growth of ever-increasing volume of data has become a critical challenge. According to IDC cloud report 2014, companies in India are making a gradual move from on premise legacy to different forms of cloud. While the process is gradual, it has started by migrating certain application workloads to cloud. To make scalable management of stored data in cloud computing, de-duplication has been well known technique which becomes more popular recently. De-duplication is a specialized data compression technique, which reduce storage space and upload bandwidth in cloud storage. In de-duplication, only one unique instance of the data is actually on the server and redundant data is replaced with a pointer to the unique data copy. Deduplication can take place either at file level or block level. From the user perspective, security and privacy concerns are arise as data are susceptible to both insider and outsider attack. We must properly enforce confidentiality, integrity checking, and access control mechanisms both attacks.

De-duplication does not work with traditional encryption. User encrypts their files with their individual encryption key, different cipher text would emerge even for identical files. Thus, traditional encryption is incompatible with data de duplication. Convergent encryption is a widely used technique to combine the storage saving of de-duplication to enforce confidentiality. In convergent encryption, the data copy is encrypted under a key derived by hashing the data itself. This convergent key is used for encrypt and decrypt a data copy. After key generation and data encryption, users retain the keys and send the cipher text to the cloud. Since encryption is deterministic, identical data copies will generate the same convergent key and the same cipher text. This allows the cloud to perform de duplication on the cipher texts. The cipher texts can only be decrypted by the corresponding data owners with their convergent keys. Differential authorization duplicate check is an authorized de-duplication technique where each user is issued a set of privileges during system initialization. This set of privileges specifies that which kind of users is allowed to perform duplicate check and access the files.

II. Related Work

This paper[1] developed Characteristics of backup workloads in production systems. The author presents a complete characterization of backup workloads by analyzing statistics and content metadata collected from a large set of EMC Data Domain backup systems in production use. This analysis is complete (it covers the statistics of over 10,000 systems) and in depth (it uses detailed traces of the metadata of different production systems that store almost 700TB of backup data). We compared these systems with a detailed study of Microsoft's primary storage systems and demonstrated that back-up storage differs significantly from the primary storage workload in terms of data quantities and capacity requirements, as well as the amount of data storage capacity redundancy within the data. These properties offer unique challenges and opportunities when designing a disk-based file system for backup workloads.

This paper[2] developed Primary data deduplication-large scale study and system designThe author presents a large-scale study of primary data deduplication and uses the results to guide the design of a new primary data deduplication system implemented in the Windows Server 2012 operating system. The file data were analyzed by 15 servers of globally distributed files that host data for over 2000 users in a large multinational company. The results are used to achieve a fragmentation and compression approach that maximizes deduplication savings by minimizing the metadata generated and producing a uniform distribution of the portion size. Deduplication processing resizing with data size is achieved by a frugal hash index of RAM and data partitioning, so that memory, CPU and disk search resources remain available to meet the main workload of the IO service.

Redundancy [3] elimination within large collections of files. Propose a new storage reduction scheme that reduces data size with comparable efficiency to the most expensive techniques, but at a cost comparable to the fastest but least effective. The scheme, called REBL (Block Level Redundancy Elimination), exploits the advantages of compression, deletion of duplicate blocks and delta encoding to eliminate a wide spectrum of redundant data in a scalable and efficient way. REBL generally encodes more compactly than compression (up to a factor of 14) and a combination of compression and suppression of duplicates (up to a factor of 6.7). REBL is also coded similarly to a technique based on delta encoding, which significantly reduces the overall space in a case. In addition, REBL uses super fingerprint, a technique that reduces the data needed to identify similar blocks by drastically reducing the computational requirements of the matching blocks: it converts the comparisons of O (n2) into searches of hash tables. As a result, the use of super fingerprints to avoid enumerating the corresponding data objects decreases the calculation in the REBL resemblance phase of a couple of orders of magnitude.

Encrypted Data [4] Storage with De-duplication Approach on Twin Cloud. The data and the private cloud where the token generation will be generated for each file. Before uploading the data or file to the public cloud, the client will send the file to the private cloud for token generation, which is unique to each file. Private clouds generate a hash and token and send the token to the client. The token and hashes are kept in the private cloud itself, so that whenever the next token generation file arrives, the private clone can refer to the same token. Once the client gets the token for a given file, the public cloud looks for the token similar if it exists or not. If the public cloud token exists, it will return a pointer to the existing file, otherwise it will send a message to load a file. A system that achieves confidentiality and allows block-level deduplication at the same time. Before uploading the data or file to the public cloud, the client will send the file to the private cloud for token generation, which is unique to each file. The private cloud generates a hash and token and sends them to the client. The token and the hash are kept in the private cloud itself so that whenever the next token generation file arrives, the private clone can refer to the same token.

In the proposed system [5], we are getting data deduplication by providing data evidence from the data owner. This test is used when the file is uploaded. Each file uploaded to the cloud is also limited by a set of privileges to specify the type of users who can perform duplicate verification and access the files. New duplication constructs compatible with authorized duplicate verification in the cloud hybrid architecture where the private cloud server generates duplicate file verification keys. The proposed system includes a data owner test, so it will help implement better security issues in cloud computing.

The author proposes[7] a new approach, called Block Locality Cache (BLC), which captures the previous backup execution significantly better than existing approaches and always uses up-to-date information about the location and is therefore less prone to aging. We evaluated the approach using a simulation based on the detection of multiple sets of real backup data. The simulation compares the Block Locality Cache with the approach of Zhu et al. and provides a detailed analysis of the behavior and the IO pattern. In addition, a prototype implementation is used to validate the simulation.

We collect data [8] from the file system content of 857 desktop computers in Microsoft for a period of 4 weeks. We analyze the data to determine the relative efficiency of data deduplication, especially considering the elimination of complete file redundancy against blocks. We have found that full file deduplication reaches about three quarters of the space savings of more aggressive block deduplication for live file system storage and 87 of backup image savings. We also investigated file fragmentation and found that it does not prevail, and we have updated previous studies on file system metadata, and we have found that file size distribution continues to affect very large unstructured files.

The author[9] has developed a generic model of file system changes based on properties measured in terabytes of real and different storage systems. Our model connects to a generic framework to emulate changes in the file system. Based on observations from specific environments, the model can generate an initial file system followed by continuous changes that emulate the distribution of duplicates and file sizes, realistic changes to existing files and file system growth.

This paper[10] discovered the optimized WAN replication of backup data sets using delta compression reported by the stream Off-site data replication is critical for disaster recovery reasons, but the current tape transfer approach is cumbersome and error prone. Replication in a wide area network (WAN) is a promising alternative, but fast network connections are expensive or impractical in many remote locations, so better compression is needed to make WAN replication very practical. We present a new technique for replicating backup data sets through a WAN that not only removes duplicate file regions (deduplication) but also compresses similar file regions with delta compression, which is available as a feature of EMC Data Domain systems."

III. EXISTING SYSTEM

Existing solutions for deduplication suffer from many 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 SYSTEM

In this paper, Author propose a confidence scheme in the challenge of data ownership and cryptography to manage the storage of encrypted data with deduplication. Our goal is to solve the problem of deduplication in the situation where the data owner is not available or it is difficult to get involved. Meanwhile, the data size does not affect the performance of data deduplication in our schema. Author are motivated to save space in the cloud and to preserve the privacy of data owners by proposing a scheme to manage the storage of encrypted data with deduplication. Author test safety and evaluate the performance of the proposed scheme through analysis and simulation. The results show its efficiency, effectiveness and applicability.

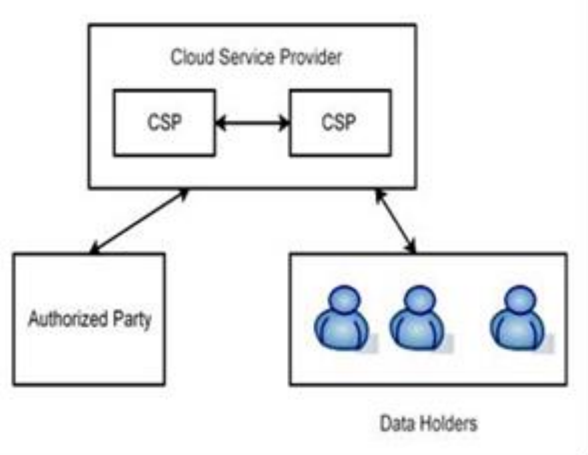

Fig. System Architecture

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 require-ments. Security analysis, comparison with existing work and implementation-based performance evaluation have shown that our scheme is safe, advanced and efficient.

VI. 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.

[6] El-Shimi, R. Kalach, A. Kumar, A. Ottean, J. Li, and S. Sengupta, "Primary data deduplication-large scale study and system design," in Proc. Conf. USENIX Annu. Tech. Conf., Jun. 2012, pp.285–296.

[7] P. Shilane, M. Huang, G. Wallace, and W. Hsu, "WAN optimized replication of backup datasets using stream-informed delta compression," in Proc. 10th USENIX Conf. File Storage Technol.,Feb.2012,pp.49–64.

[8] P. Kulkarni, F. Douglis, J. D. LaVoie, and J. M. Tracey, "Redundancy elimination within large collections of files," in Proc.USENIXAnnu.Tech.Conf. Jun.2012, pp.59–72.

[9] Jin Li, Yan Kit Li, Xiaofeng Chen, Patrick P. C. Lee, Wenjing Lou "A Hybrid Cloud Approach for Secure Authorized De-duplication" IEEE Transactions on Parallel and Distributed Systems: PP Year 2014.

[10] Shweta D. Pochhi, Prof. Pradnya V. Kasture "Encrypted Data Storage with De-duplication Approach on Twin Cloud " International Journal of Innovative Research in Computer and Communication Engineering