



IJIRCCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798




INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 11, Issue 6, June 2023

ISSN INTERNATIONAL
STANDARD
SERIAL
NUMBER
INDIA

Impact Factor: 8.379

 9940 572 462

 6381 907 438

 ijircce@gmail.com

 www.ijircce.com

Cloud Based Improved File Handling and Duplication Removal Using CHF

Mr.Yogesh Jatharam Choudhari, Prof.Prasad Bhosale

Department of Computer Engineering, Trinity College of Engineering and Research, Pune, India

ABSTRACT– For cloud storage providers, data deduplication, a method for reducing redundant data by maintaining only one duplicate of a file, saves a significant amount of storage and bandwidth. Using cloud storage services, people and corporations can outsource data storage to distant servers. To increase the effectiveness of IT resources, deduplication has become a commonly used technology in cloud data centres. The duplication check can be used by an attacker to determine whether a file (such as a pay stub with a certain name and salary amount) has already been stored (by someone else), revealing the user's private information. Due to information leakage, weak attack resistance, high computational cost, and other factors, the majority of existing solutions that rely on the aid of a trustworthy key server (KS) are weak and have limitations. The system as a whole disintegrates if the reliable KS fails, creating a single point of failure. Data duplication in the current system has security problems. Therefore, we are using double encryption and hash code creation to resolve these concerns. It combines access control with cloud data deduplication. Because only authorised data holders can receive the symmetric keys required for data decryption, encrypted data can be accessed safely. Extensive performance testing and analysis revealed that our method is highly effective for huge data deduplication under the given security architecture. We assess its performance using in-depth research and computer simulations. The outcomes demonstrate the scheme's improved efficacy and efficiency for possible practical application, particularly for huge data deduplication in cloud storage.

KEYWORDS– Cloud Computing, Data deduplication, 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. Author 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 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.

II. RELATED WORK

G. Wallace, F. Douglis, H. Qian, P. Shilane, S. Smaldone, M. Chamness, and W. Hsu has developed Characteristics of backup workloads in production systems. By examining statistics and content metadata gathered from a sizable number of EMC Data Domain backup systems in use, the author provides an in-depth analysis of backup workloads. This investigation uses meticulous traces of the metadata of several production systems that hold about 700TB of backup data, making it thorough (it covers the statistics of over 10,000 systems). The author compared these systems with a thorough examination of Microsoft's primary storage systems and showed how the workloads for backup storage and primary storage are very different in terms of data volumes, capacity needs, and the quantity of data storage capacity. Inconsistency in the data. When creating a disk-based file system for backup workloads, these characteristics present both distinct obstacles and opportunities[1].

A. El-Shimi, R. Kalach, A. Kumar, A. Ottean, J. Li, and S. Sengupta have developed Primary data deduplication-large scale study and system design. The main data deduplication system designed by the author is integrated into the Windows Server 2012 operating system. The author presents a comprehensive research of primary data deduplication and uses the findings to inform design decisions. 15 servers that host files that are worldwide spread and house the data for more than 2000 users in a huge multinational organisation analysed the file data. By reducing the amount of metadata created and generating a consistent distribution of portion size, the results are used to achieve a fragmentation and compression technique that maximises deduplication savings. A thrifty RAM hash index and data partitioning are used to achieve deduplication processing scaling with data size such that memory, CPU, and disc search resources are still available to handle the IO service's primary workload [2] .

P. Kulkarni, F. Douglis, J. D. LaVoie, and J. M. Tracey, "Redundancy elimination within large collections of files". Make a novel storage reduction plan that reduces data size as effectively as the priciest ways while costing about the same as the quickest but least efficient. The technique, known as REBL (Block Level Redundancy Elimination), makes use of the benefits of compression, the removal of duplicate blocks, and delta encoding to efficiently and effectively remove a variety of redundant data. Generally speaking, REBL encrypts data more efficiently than compression (up to a factor of 14) and a compression and duplicate-suppression combo (up to a factor of 6.7). A method based on delta encoding, which considerably decreases the overall space in a case, is comparable to how REBL is also coded. Additionally, REBL makes use of a technique called super fingerprint, which transforms comparisons of $O(n^2)$ into hash table searches, hence lowering the amount of data required to find blocks that are similar to one another. As a result, the calculation in the REBL resemblance phase is reduced by a couple of orders of magnitude when super fingerprints are used to avoid counting the corresponding data objects [3].

Shweta D. Pochhi, Prof. Pradnya V. Kasture have represents "Encrypted Data Storage with De-duplication Approach on Twin Cloud. the information and the private cloud where each file's token will be produced. The client will transmit the file

to the private cloud for token generation, which is specific to each file, before uploading the data or file to the public cloud. A hash and token are generated by private clouds, and the token is sent to the client. So that the private clone can use the same token everytime the next token generation file is received, the token and hashes are stored in the private cloud itself. As soon as the client receives the token for a particular file, the public cloud checks to see if the token already exists or not. A pointer to the existing file will be returned if the public cloud token already exists; otherwise, a message to load a file will be sent. a technology that permits block-level deduplication while also achieving confidentiality. The client will transmit the file to the private cloud for token generation, which is specific to each file, before uploading the data or file to the public cloud. A hash and token are created by the private cloud and sent to the client. In order for the private clone to use the same token whenever the next token generation file is received, the token and hash are saved in the private cloud itself[4].

Jin Li, Yan Kit Li, Xiaofeng Chen, Patrick P. C. Lee, Wenjing Lou have developed A Hybrid Cloud Approach for Secure Authorized De-duplication. Data deduplication is achieved in the suggested method by supplying data proof from the data owner. When the file is uploaded, this test is run. A set of privileges that limit the kinds of users who can do duplication verification and access the files are also applied to each file that is uploaded to the cloud. In a cloud hybrid architecture where the private cloud server creates duplicate file verification keys, new duplication structures are compatible with authorised duplicate verification. The suggested system has a data owner test, which will make cloud computing security challenges more effectively implemented[5].

M. Lillibridge, K. Eshghi, and D. Bhagwat shows the increase in recovery time for block-based online deduplication backup solutions. Data deduplication systems in one piece confront a severe issue with the delayed recovery caused by the fragmentation of the parts: the recovery rates for the most recent backup might decrease orders of magnitude throughout the course of a system. To address this issue, the author has researched three solutions: enlarging the cache, restricting the number of containers, and using a direct assembly area[6].

D. Meister, J. Kaiser, and A. Brinkmann represented the places where data deduplication took place. Block Locality Cache (BLC), the novel method the author suggests, captures the previous backup execution substantially better than existing methods and always uses the most recent information about the location, making it less susceptible to ageing. The method was assessed by the author using a simulation based on the discovery of numerous actual backup data sets. The simulation contrasts Zhu et al.'s method with the Block Locality Cache and offers a thorough examination of the behaviour and IO pattern. Additionally, the simulation is validated using a prototype implementation [7].

D. T. Meyer and W. J. Boloskyhas represents A study of practical Deduplication. For a period of 4 weeks, the author collected data from the file system content of 857 desktop PCs in Microsoft. In order to assess the relative effectiveness of data deduplication, the author analyses the data, taking into account the removal of whole file redundancy versus blocks. Full file deduplication, according to the author, saves 87% more space for backup images and around 75% more space for live file system storage than more aggressive block deduplication. The author also looked at file fragmentation and discovered that it does not occur. The author also updated earlier studies on file system metadata and discovered that very large unstructured files are still affected by file size distribution[8].

V. Tarasov, A. Mudrankit, W. Buik, P. Shilane, G. Kuenning, and E. Zadok having are a way of creating data sets that are realistic enough for the deduplication analysis. Based on characteristics measured in terabytes of actual data and a variety of storage systems, the author has created a general model of file system changes. To simulate changes to the file system, our model is connected to a general framework. The model may create an initial file system and then continuously update it to simulate the distribution of duplicates and file sizes, realistic changes to existing files, and file system growth based on observations from certain situations[9].

P. Shilane, M. Huang, G. Wallace, and W. Hsu utilising the delta compression supplied by the stream, it was possible to optimise WAN replication of backup data sets. For disaster recovery purposes, offsite data replication is essential, but the present tape transfer method is inefficient and error-prone. A viable option is replication across a wide area network (WAN), however in many remote sites, fast network

connections are expensive or unfeasible. Therefore, improved compression is required to make WAN replication highly practical. Using a feature of EMC Data Domain systems, the authors offer a new method for replicating backup data sets via a WAN that not only deduplicates file regions (deduplication), but also compresses related file regions with delta compression[10].

Proposed system

In this study, the author proposes a confidence system to handle the storage of encrypted data with deduplication while addressing the issue of data ownership. Our objective is to address the deduplication issue in scenarios where the data owner is unavailable or unable to participate. In the meantime, the performance of data deduplication in our schema is unaffected by the amount of the data. Authors propose a method to control the storage of encrypted data using deduplication in order to conserve cloud storage capacity and protect data owners' privacy. Through analysis and simulation, the author assesses the performance of the suggested plan and tests its safety. The outcomes demonstrate its applicability, efficacy, and efficiency.

A. System Architecture

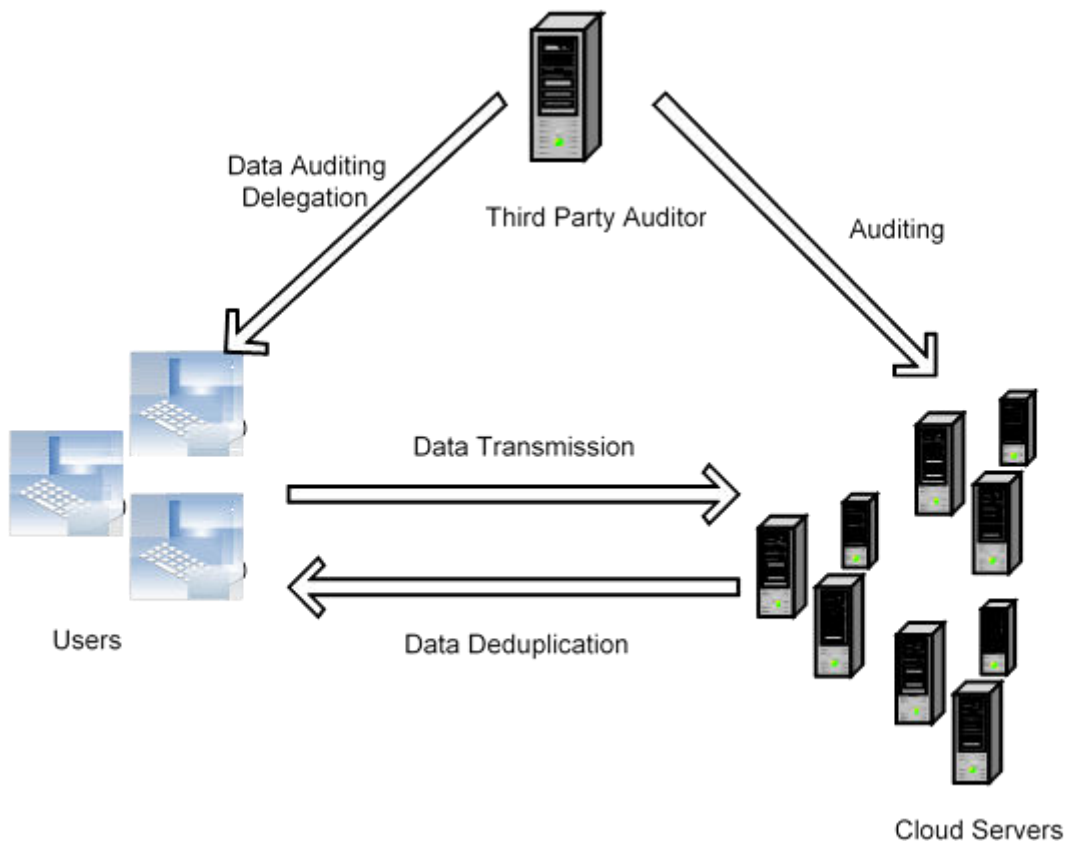


Fig. System Architecture

Cloud Servers

The cloud computing framework is an information-technology (IT) concept that enables all kinds of popular classes of configurable tools to be available easily and often via the Internet (such as computer networks, databases, processing,



applications, and services). Cloud computing provides consumers and businesses with varying computational capabilities, the option of storing data in private servers, or on a third party database in a data center, thereby increasing the efficiency and consistency of data access systems.

Data Deduplication

Duplication of data occurs when the data owner attempts to archive the same information already contained in the CSP. The CSP should check it by comparing tokens. In the case of a successful match, the CSP should contact the deduplication system supplying the data holder with the token and public key.

Storage Management

Some storage management techniques, including virtualized storage, deduplication along with compression, allow businesses to use current data storage efficiently. The advantages of these approaches include reduced cost, the specific spending on resources for processing facilities as well as continuing operational costs for the repair of these systems. Network and equipment management is also simplified through the majority of storage management techniques. In order to save time and money, the number of IT employees required for maintaining their storage systems can even be reduced by businesses and the total cost of storage in turn can be reduced. The quality of the data center can also be enhanced by data management. The compression and engineering, for instance, will increase the distribution and automation of storage assets to various applications.

B. Algorithms

1. AES Algorithm for Encryption.

AES (advanced encryption standard). It is symmetric algorithm. It used to convert plain text into cipher text. The need for coming with this algo is weakness in DES. The 56 bit key of des is no longer safe against attacks based on exhaustive key searches and 64-bit block also consider as weak. AES was to be used 128-bit block with 128-bit keys.

Rijndael was founder. In this drop we are using it to encrypt the data owner file.

Input:

128_bit / 192 bit / 256 bit input (0, 1)

Secret key (128_bit) + plain text (128_bit).

Process:

10/12/14-rounds for-128_bit / 192 bit / 256 bit input

Xor state block (i/p)

Final round: 10, 12, 14

Each round consists: sub byte, shift byte, mix columns, add round key.

Output:

cipher text (128 bit)

2. MD5 (Message-Digest Algorithm)

The MD5 message-digest algorithm is a widely used cryptographic hash function producing a 128-bit (16-byte) hash value, typically expressed in text format as a 32 digit hexadecimal number. MD5 has been utilized in a wide variety of cryptographic applications, and is also commonly used to verify data integrity.

Steps:

1. A message digest algorithm is a hash function that takes a bit sequence of any length and produces a bit sequence of a fixed small length.
2. The output of a message digest is considered as a digital signature of the input data.
3. MD5 is a message digest algorithm producing 128 bits of data.
4. It uses constants derived to trigonometric Sine function.
5. It loops through the original message in blocks of 512 bits, with 4 rounds of operations for each block, and 16 operations in each round.
6. Most modern programming languages provides MD5 algorithm as built-in functions



III. RESULTS AND DISCUSSION

In our experimental setup, in table 1, find out different file upload and time required for time for uploading that file. In our experimental setup, in our system first is uploading file size and time for that file.

Sr.No	File Size(Kb)	Time(ms)
1	10351	226
2	17541	500
3	8500	140

Table1: File Uploading Time and Size

IV. CONCLUSION

In the practise of data storage in the cloud, data deduplication is crucial and relevant, especially for the management of huge data filing. The author of this work developed a method for managing heterogeneous data storage that provides access control and customizable data deduplication in the cloud. Our schema provides efficient handling of massive data storage over numerous SPs and can be customised to varied scenarios and application needs. Data deduplication can be accomplished with various levels of security. Our plan is secure, cutting-edge, and effective, as demonstrated by security analysis, comparison with prior work, and implementation-based performance evaluation.

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. Dougliis, 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. Dougliis, J. D. LaVoie, and J. M. Tracey, "Redundancy elimination within large collections of files," in Proc. USENIX Annu. 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



Impact Factor: 8.379



INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

 9940 572 462  6381 907 438  ijircce@gmail.com



www.ijircce.com

Scan to save the contact details