IJIRSET
International Journal of Innovative Research in
SCIENCE | ENGINEERING | TECHNOLOGY

e-ISSN: 2319-8753 | p-ISSN: 2347-6710

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

Volume 12, Issue 6, June 2023

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.423

9940 572 462 | 6381 907 438 | ijirset@gmail.com | www.ijirset.com

International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)

|e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 6, June 2023 ||

| DOI:10.15680/IJIRSET.2023.1206079 |

Survey on 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-Access control, cloud computing, data deduplication.

I. INTRODUCTION

By rearranging different Internet resources, cloud computing offers a fresh approach to service delivery. Data storage is an important and well-liked cloud service. The data is often saved encrypted in order to protect the data subjects' privacy. Data deduplication in the cloud, which is essential for processing and preserving enormous volumes of data there, is put to significant difficulties by encrypted data. Traditional deduplication techniques cannot function on encrypted data. Data deduplication suffers from inadequate security in current encryption techniques. They can't survive restrictions on and cancellation of access to data. Consequently, some of them. They are simple to put into practise. In this article, we suggest using challenge properties and a redefining proxy to deduplicate encrypted data that has been stored in the cloud. Access control and cloud data deduplication are integrated. Based on thorough analytics and computer simulations, we evaluate their performance. The outcomes demonstrate the highest possible distribution of efficiency and efficiency schemes, particularly for big data deduplication storage in the cloud.

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].

|e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 6, June 2023 ||

| DOI:10.15680/IJIRSET.2023.1206079 |

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 (n2) 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].

|e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 6, June 2023 ||

| DOI:10.15680/IJIRSET.2023.1206079 |

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].

III. EXISTING SYSTEM APPROACH

Brute-force assaults are a problem for deduplication methods currently in use. They cannot simultaneously enable friendly data access control and revocation. Most current systems cannot guarantee performance while also maintaining dependability, security, and privacy. The first data holders could not always be online or accessible for every management, which might result in a storage delay. To include data holders in the deduplication process, second deduplication may become excessively complex in terms of communication and processing. Third, when looking for duplicate data, it might invade the privacy of the data owner. In some circumstances where it does not know other data holders because to data suffer distribution, a data holder may not know how to grant data access rights or a deduplication key to users. As a result, CSP frequently finds it impossible to work with data owners on data storage deduplication. on PRE and the ownership dilemma. Even when the data owner is offline, our system can flexibly allow data updating and exchange with deduplication. Because only authorised data holders can receive the symmetric keys needed for data decryption, encrypted data can be accessed safely. Our scheme is efficient and secure within the outlined security paradigm and is ideally suited for deduplication, according to extensive performance analysis and testing. The outcomes of our computer simulations further demonstrated how viable our plan is.

IV. CONCLUSION

For a successful cloud storage service, discuss how to manage encrypted data with deduplication for huge data storage. In this paper, a method for managing encrypted massive data in the cloud with deduplication based on ownership challenge and PRE is proposed. Even when the data holders are offline, this approach can adaptably provide data updating and exchange with deduplication. The symmetric keys required to safely access encrypted data can only be utilised by authorised data holders. The performance analysis and graphs demonstrated that, using the big data deduplication security paradigm as defined, the system is secure. The variations in the graph further displayed the outcomes of the computer simulations. The optimisation of their design and execution for cloud large data storage is a task for the future. To make sure CSP operates as anticipated when managing deduplication investigating auditable computation. Many academics from other professions have been drawn to the topic of cloud computing, but much work needs to be done before cloud computing technology is widely used. Propose a huge data deduplication for the future. The unique aspect of the system is that it can concentrate on personalising search based on user feedback sessions as well as recommendations based on user point of interest.

|e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 6, June 2023 ||

| DOI:10.15680/IJIRSET.2023.1206079 |

REFERENCES

[1] M. Bellare, S. Keelveedhi, and T. Ristenpart, "DupLESS: Server aided encryption for deduplicated storage," in Proc. 22nd USENIX Conf. Secur., 2013, pp. 179–194.
[2] J. Li, Y. K. Li, X. F. Chen, P. P. C. Lee, and W. J. Lou, "A hybrid cloud approach for secure authorized deduplication," IEEE Trans. Parallel Distrib. Syst., vol. 26, no. 5, pp. 1206–1216, May 2015, doi:10.1109/TPDS.2014.2318320.
[3] P. Meye, P. Raipin, F. Tronel, and E. Anceaume, "A secure two phase data deduplication scheme," in Proc. HPCC/CSS/ICESS, 2014, pp. 802–809, doi:10.1109/HPCC.2014.134.
[4] Z. Yan, X. Y. Li, M. J. Wang, and A. V. Vasilakos, "Flexible data access control based on trust and reputation in cloud computing," IEEE Trans. Cloud Comput., vol. PP, no. 99, Aug. 2015, doi:10.1109/ TCC.2015.2469662, Art. no. 1.
[5] P. Puzio, R. Molva, M. Onen, and S. Loureiro, "ClouDedup: Secure deduplication with encrypted data for cloud storage," in Proc. IEEE Int. Cof. Cloud Comput. Technol. Sci., 2013, pp. 363–370, doi:10.1109/CloudCom.2013.54.
[6] C. Y. Liu, X. J. Liu, and L. Wan, "Policy-based deduplication in secure cloud storage," in Proc. Trustworthy Comput. Serv., 2013, pp. 250–262, doi:10.1007/978-3-642-35795-4_32.
[7] Z. Sun, J. Shen, and J. M. Yong, "DeDu: Building a deduplication storage system over cloud computing," in Proc. IEEE Int. Conf. Comput. Supported Cooperative Work Des., 2011, pp. 348–355, doi:10.1109/CSCWD.2011.5960097.
[8] Z. C. Wen, J. M. Luo, H. J. Chen, J. X. Meng, X. Li, and J. Li, "A verifiable data deduplication scheme in cloud computing," in Proc. Int. Conf. Intell. Netw.Collaborative Syst., 2014, pp. 85–90, doi:10.1109/INCoS.2014.111.

INNO SPACE
SJIF Scientific Journal Impact Factor

Impact Factor: 8.423

doi cross ref

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

Scan to save the contact details