IJIRCCE

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

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 9, Issue 10, October 2021

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 7.542

International Journal of Innovative Research in Computer and Communication Engineering

-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.542

|| Volume 9, Issue 10, October 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0910008 |

Survey on Deduplication of Encrypted Big Data in Cloud

K.Vinitha[1], Dr.P.Thirumoorthy[2]

PG Scholar, Department of Computer Science and Engineering, Nandha Engineering College, Erode, Tamil Nadu, India[1]

Professor, Department of Computer Science and Engineering, Nandha Engineering College, Erode, Tamil Nadu, India[2]

ABSTRACT: Cloud storage service is an efficient solution for increasing storage needs of organizations and individuals. In ordertoensuresecurityanddataprivacy,usersmayencrypttheir data before uploading to cloud. In such cases same or different users may up-load same data in encrypted form and it results in theexistenceof duplicated data.Datadeduplicationisatechnique designed to identify and eliminate redundant data. When users are uploading data in encrypted form, then deduplication is a challenging problem. This project proposes a deduplication system in cloud storage that supports big data deduplication also. The system is based on symmetric encryption and re-encryption. The system aims at paragraph-level deduplication of text data files. A token will be generated for each paragraph in a file before uploading to cloud. When a new request for uploading is received, system checks the existence of duplicate files using these tokens. If the file is already exists in the cloud, then system rejects uploading, else system will perform deduplication using re-encryption algorithm.

KEYWORDS: Cloud Computing, Data De-duplication, Re-encryption.

I. INTRODUCTION

Cloud storage deliver resources to users over internet. Clouds which are larger in sizes have functions which is distributed over more than one location from central servers. Clouds are of different types. Some are restricted to a single company or organization and it is called enterprise clouds. Some are available to more than one organizations and they are called public clouds and some are combination of both public and private clouds and they are called hybrid clouds. Amazon AWS is the largest among public cloud. High-capacitynetworks,economicalcomputersandstorage services results in the rapid growth of cloud computing. Cloud computing mainly aims at making the resource sharing economical. By linking together resources over network cloud computing delivers wide collection of information sources to users. Cloud computing provides data processing services to users by offering data storage, computing and so forth. It offers a large collection of data sources by combining resources over a network. It should possess some attributes namely, adaptability, flexibility and fail resistance.

1.1 CLOUD MODELS

There are three types of models present in cloud computing which are given as follows:

Public Cloud Model: The public cloud model is defined as a cloud infrastructure which is managed by an organization providing third-party service. This is available as a service over the internet for both individual users and software companies/ organizations. This model's main advantage is that it is very large in scale. With limited configurations and security protection, the users in this model share the same infrastructure pool as provided by the service provider.

PrivateCloud Model: The private cloud model is defined as a cloud computing infrastructure exclusively developed by a given company for each project or software. This requires a policy of permission to host cloud applications to enforce system security and control. In addition to being generated for each specific project, an external party or supplier also provide the cloud service.

Hybrid Cloud Model: The hybrid cloud model is defined as a cloud computing infrastructure that combines both public and private cloud models' advantageous factors. This is done using separate algorithms used to switch between the two infrastructures.

-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.542

|| Volume 9, Issue 10, October 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0910008 |

1.2 CLOUD COMPUTING MODELS

Infrastructure as a service(IaaS) allows users to use their storage or computational units remotely to access the given network. It does soon a demand-based basis whenever the service is required by the user. E.g: Microsoft Azure, Amazon Web Service.

Platform as a Service (PaaS) enables users to quickly and easily create web applications with permissions to provide a substitute for the purchase and maintenance of the system's software and infrastructure. Eg: GoogleApp Engine.

Software as a service (SaaS) enables users to obtain an application license for any user, either as an ondemand service or through Internet subscription. In a simple way, it can be rented for use in a pay-as-you-go way instead of buying the required software. Example: Sales force, Cisco WebEx.

1.3. CLOUD COMPUTING TOOLS

Cloud services across a network are used as efficient, organizational-based business solutions. Various cloud computing tools, such as Eucalyptus, Open Nebula, Nimbus, Open stack, etc., are available where they all have different deployment strategies.

Cloud computing load balancing is defined as the process of distributing workload and computing resources within a networked cloud computing environment. It enables an organization to manage applications or workload demands on a task-by-task basis, by allocating resources on the networks between the various computers or through servers.

1.4. DEDUPLICATION OF ENCRYPTED BIG DATA:

Cloud computing offers a new way of service provision by re-arranging various resources over the Internet. The most important and popular cloud service is data storage. In order to preserve the privacy of data holders, data are often stored in cloud in an encrypted form. However, encrypted data introduce new challenges for cloud data deduplication, which becomes crucial for big data storage and processing in cloud. Traditional deduplication schemes cannot work on encrypted data. Existing solutions of encrypted data deduplication suffer from security weakness. They cannot flexibly support data access control and revocation. Therefore, few of them can be readily deployed in practice. In this paper, we propose a scheme to deduplicate encrypted data stored in cloud based on ownership challenge and proxy re-encryption. It integrates cloud data deduplication with access control.

II. LITERATURE SURVEY

[1] Huijun Wu, Chen Wang, Yinjin Fu, Kai Lu, Liming Zhu" presents an existing primary deduplication techniques either use inline caching to exploit locality in primary workloads or use post-processing deduplication to avoid the negative impact on I/O performance. However, neither of them works well in the cloud servers running multiple services for the following two reasons: Firstly, the temporal locality of duplicate data writes varies among primary storage workloads, which makes it challenging to efficiently allocate the inline cache space and achieve a good deduplication ratio. Secondly, the post-processing deduplication does not eliminate duplicate I/O operations that write to the same logical block address as it is performed after duplicate blocks have been written. A hybrid deduplication mechanism is promising to deal with these problems. Inline fingerprint caching is essential to achieving efficient hybrid deduplication. We present a detailed analysis of the limitations of using existing caching algorithms in primary deduplication in the cloud.

[2] Jianwei Yin, Yan Tang, Shuiguang Deng, Bangpeng Zheng and Albert Y. Zomaya, presents a cloud storage service vendors, balancing the client-perceived IO performance and the selfperceived space cost is always one of the standing challenges. Enabling deduplication decreases the storage space cost, whereas the IO performance will be somewhat affected due to extra processing overhead and data fragmentation. First, we propose a novel notation of Dedup-SLA (deduplication-oriented service level agreement). Second, MUSE adopts multi-tiered deduplication that orchestrates several combinational forms of deduplication into multiple tiers with varied "deduplication strength". Third, we implement a mechanism called dynamic deduplication regulation (DDR) to adjust the deduplication behavior during runtime. MUSE's deduplication behavior is periodically switched between tiers according to the predefined Dedup-SLA and instant system status.

-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.542

|| Volume 9, Issue 10, October 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0910008 |

[3] UrsNiesen, purposesdeduplication finds and removes long-range data duplicates. It is commonly used in cloud and enterprise server settings and has been successfully applied to primary, backup, and archival storage. Despite its practical importance as a source-coding technique, its analysis from the point of view of information theory is missing. This paper provides such an information-theoretic analysis of data deduplication. It intro- duces a new source model adapted to the deduplication setting. It formalizes the two standard fixed-length and variable-length deduplication schemes, and it introduces a novel multi-chunk deduplication scheme. It then provides an analysis of these three deduplication variants, emphasizing the importance of boundary synchronization between source blocks and deduplication chunks.

[4] Xue Yang, RongxingLu,Jun Shao, Xiaohu Tang, and Ali A.Ghorbani, presented the study which aims at presents acloud storage as one of the most important services of cloud computing significantly facilitates cloud users to outsource their data to the cloud for storage and share them with authorized users. In cloud storage, secure deduplication has been widely investigated as it can eliminate the redundancy over the encrypted data to reduce storage space and communication overhead. Regarding the security and privacy, many existing secure deduplication schemes generally focus on achieving the following properties: data confidentiality, tag consistency, access control and resistance to brute-force attacks. However, as far as we know, none of them can achieve these four requirements at the same time. To overcome this shortcoming, in this paper, we propose an efficient secure deduplication scheme that supports user-defined access control.

[5] Haoran Yuan, Xiaofeng Chen, Jin Li, Tao Jiang, Jianfeng Wang, and Robert H. Deng, propose a secure data deduplication scheme with efficient re-encryption based on the convergent all-or-nothing transform (CAONT) and randomly sampled bits from the Bloom filter. Due to the intrinsic property of one-way hash function, our scheme can resist the stub-reserved attack and guarantee the data privacy of data owners' sensitive data. Moreover, instead of re-encrypting the entire package, data owners are only required to re-encrypt a small part of it through the CAONT, thereby effectively reducing the computation overhead of the system. Finally, security analysis and experimental results show that our scheme is secure and efficient in re-encryption.

[6] Chia-Mu Yu, Sarada Prasad Gochhayat, Mauro Conti, and Chun-Shien Lu,presents a cloud storage services enable individuals and organizations to outsource data storage to remote servers. Cloud storage providers generally adopt data deduplication, a technique for eliminating redundant data by keeping only a single copy of a file, thus saving a considerable amount of storage and bandwidth. We propose ZEUS (zero-knowledge deduplication response) framework. We develop ZEUS and ZEUS+, two privacy-aware deduplication protocols: ZEUS provides weaker privacy guarantees while being more efficient in the communication cost, while ZEUS+ guarantees stronger privacy properties, at an increased communication cost.

[7] Hui Cui, Robert H. Deng, Yingjiu Li, and Guowei Wu, present an attribute-based storage system with secure deduplication in a hybrid cloud setting, where a private cloud is responsible for duplicate detection and a public cloud manages the storage. Compared with the prior data deduplication systems, our system has two advantages. Firstly, it can be used to confidentially share data with users by specifying access policies rather than sharing decryption keys. Secondly, it achieves the standard notion of semantic security for data confidentiality while existing systems only achieve it by defining a weaker security notion.

[8] Shunrong Jiang, Tao Jiang and Liangmin Wang, presents a data deduplication has been widely used in cloud storage to reduce storage space and communication overhead by eliminating redundant data and storing only one copy for themSpecially, our scheme supports both cross-user filelevel and inside-user block-level data deduplication. During the file-level deduplication, we construct a new PoW scheme to ensure the tag consistency and achieve the mutual ownership verification. Moreover, we design a lazy update strategy to achieve efficient ownership management. For inside-user block-level deduplication, the user-aided key is used to realize convergent key management and reduce the key storage space.

[9]Zheng Yan, Lifang Zhang, Wenxiu Ding, and QinghuaZheng, propose a heterogeneous data storage management scheme, which flexibly offers both deduplication management and access control at the same time across multiple Cloud Service Providers (CSPs). We evaluate its performance with security analysis, comparison and implementation. The results show its security, effectiveness and efficiency towards potential practical usage.Cloud storage as one of the

-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.542

|| Volume 9, Issue 10, October 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0910008 |

most important services of cloud computing helps cloud users break the bottleneck of restricted resources and expand their storage without upgrading their devices.

[10] Yinjin Fu, Nong Xiao, Hong Jiang, Guyu Hu, and Weiwei Chen,presented the study which aims at presents deduplication has become a widely deployed technology in cloud data centers to improve IT resources efficiency. However, traditional techniques face a great challenge in big data deduplication to strike a sensible tradeoff between the conflicting goals of scalable deduplication throughput and high duplicate elimination ratio. We propose App Dedupe, an application-aware scalable inline distributed deduplication framework in cloud environment, to meet this challenge by exploiting application awareness, data similarity and locality to optimize distributed deduplication with inter-node two-tiered data routing and intra-node application-aware deduplication

[11] NahlahAslam K.P., proposes a deduplication system in cloud storage that supports big data deduplication also Thesytem is based on symmetric encryption and re-encryption. The system aims at paragraph-level deduplication of text data files. A token will be generated for each paragraph in a file before uploading to cloud. When a new request for uploading is received, system checks the existence of duplicate files using these tokens. If the file is already exists in the cloud, then system rejects uploading, else system will perform deduplication using reencryption algorithm.

[12] Zheng YAN, Wenxiu DING, Xixun YU, Haiqi ZHU, DENG, Robert H.,propose a scheme to deduplicate encrypted data stored in cloud based on ownership challenge and proxy re-encryption. It integrates cloud data deduplication with access control. We evaluate its performance based on extensive analysis and computer simulations. The results show the superior efficiency and effectiveness of the scheme for potential practical deployment, especially for big data deduplication in cloud storage.

[13] ShengmeiLuo, Guangyan Zhang, Chengwen Wu, Samee U. Khan,andKeqin Li, present the Boafft, a cloud storage system with distributed deduplication. Firstly, the Boafft uses an efficient data routing algorithm based on data similarity that reduces the network overhead by quickly identifying the storage location. Secondly, the Boafft maintains an in-memory similarity indexing in each data server that helps avoid a large number of random disk reads and writes, which in turn accelerates local data deduplication. Thirdly, the Boafft constructs hot fingerprint cache in each data server based on access frequency, so as to improve the data deduplication ratio. Moreover, the Boafft makes better usage of the storage space, with higher read/write bandwidth and good load balance.

[14] A.G.Ashmita, presents a cloud computing one of the services is cloud storage where data is remotely maintained, managed and backed up. Due to increase in data exponentially day to day, issues related to the storage space, data confidentiality and volume of search space complexities increases. To resolve these issues, the proposed model aims to, address the demand of storing the data redundantly by means of efficient de-duplication technique and also to protect the confidentiality of sensitive data while supporting de-duplication. Monitoring the activities on top of the storage environment in datalake to provide security to the storage nodes. As the storage nodes are geographically distributed the prime focus is on Optimal data storage and retrieval storage management and data security.

[15] Youngjoo Shin, Dongyoung Koo, Joobeom Yun and Junbeom Hur Member,extend server-aided encryption to a decentralized setting that consists of multiple KSs. The key idea of our proposed scheme is to construct an inter-KS deduplication algorithm, by which a cloud storage service provider can perform deduplication over ciphertexts from different KSs within a tenant or across tenants. This way, our scheme simultaneously offers flexibility of KS management and crosstenantdeduplication over encrypted data. The novelty of the approach is using a decentralized architecture that does not require any centralized entities for the coordination or pre-sharing of secrets among KSs

-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.542

|| Volume 9, Issue 10, October 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0910008 |

III.COMPARATIVE ANALYSIS

S.No	Title	Techniques & Mechanisms	Parameter Analysis	Tools	Future Work
1	A Differentiated Caching Mechanism to Enable Primary Storage Deduplication in Clouds Hybrid.	Hybrid deduplication mechanism, post processing technique, cache replacement algorithm.	Cache management, storage workloads.	Cloud FTP, Virtual machine.	To improve the disk capacity.
2	MUSE A Multi-tierd and SLA-Driven Deduplication Framework for Cloud Storage Systems.	Dynamic Deduplication Regulation(DDR), Context-based rewriting algorithm(CBR).	Decreases the storage space cost, critical control, throughput.	Object Storage Cluster(OSC),Intel Xeon E5620 CPU,SATA hard disk.	Utilize advanced machine learning techniques to promote the current DDR mechanism.
3	An InformationTheoretic Analysis of Deduplication	Encoding algorithm, multichunk.	Reduce storage requirements in data center.	Virtual machine disk, Appendix C.	Analyzing the effect of distributions on the deduplication rate.
4	Achieving Efficient Secure Deduplication with User-Defined Access Control in Cloud.	Brute force attack, efficient secure deduplication scheme, hybrid cloud.	Reduce storage cost.	Proxy.	Extensive performance evaluation on file-level deduplication and chunk-level Deduplication.
5	Secure Cloud Data Deduplication with Efficient Re-encryption.	Stub-reserved attack Encryption algorithm.	Increasing the capacity of storage space.	Memopal.	Computational cost, communication overhead and storage cost.
6	Privacy Aware Data Deduplication for Side Channel in Cloud Storage.	ZEUS(zero-knowledge deduplication response) framework.	Saving amount of storage and bandwidth, reducingboth cost and complexity.	Proxy, Obfuscation.	To improve capability of eliminating data deduplication.
7	Attribute-Based Storage Supporting Secure Deduplication of Encrypted Data in Cloud.	Hybrid cloud, tag matching algorithm.	Increase the storage space and security.	Charm 0.43,Python 3.4,Ubuntu 16.04,Intel core i5.42100 CPU.	Reduce the storage.
8	Secure and Efficient Cloud Data Deduplication with Ownership Management.	PoW schema: verification algorithm + generation algorithm, Convergent encryption.	To reduce the update frequency and computation overhead, saving the bandwidth.	SATA hard disk.	Improve the capacity of storage.

-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.542

|| Volume 9, Issue 10, October 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0910008 |

9	Heterogeneous Data Storage Management with Deduplication in Cloud Computing	Heterogeneous data storage management scheme.	Security, efficiency towards potential.	C++,MySQL 5.5.46, Ubuntu,Intel Xeon CPU E5-2670.	Improve the performance of our schema towards practical deployment.
10	Application-Aware Big Data Deduplication in Cloud Environment.	Routing algorithm, AppDedupe.	To maintain high efficiency.	Virtual machine, Linux kernel.	In future, save the network bandwidth during data transfer.
11	Data Deduplication with Encrypted Big Data Management in Cloud Computing medias such as audio, video and images	Hashing algorithm.	Removing redundant data, Energy consumption.	Java using Eclipse IDE, MySQL.	In future, this system can be expanded to manage the storage with deduplication when there is any updation or deletion of files and implementing Map Reduce algorithms to reduce execution time.
12	Deduplication on encrypted big data in cloud	Encrypted data store in cloud.	Cloud storage space, efficiency.	Intel core i53337U CPU, Ubuntu V13.10, MySQL, Dualcore processor.	Future work includes Optimizing design and implementation for practical deployment and verifiable computation to ensure that CSP behaves as expected in deduplication management.
13	Boafft Distributed Deduplication for Big Data Storage in the Cloud	Boafft, Routing algorithm.	Better usage of storage space, load balance.	Intel Xeon E52620,Ubuntu,Jdk 1.7	To improve the network bandwidth.
14	Data De-Duplication on Encrypted Data Lake in Cloud Environment	Encrypted data lake scheme, AppDedupe, Avalanche effect.	Reduces the latency, increasing throughput, storage allocation.	Intel (R) Core TM Processor.	To improve the performance of this schema.
15	Decentralized Serveraided Encryption for Secure Deduplication in Cloud Storage	Server-aided encryption scheme.	High deduplication efficiency and scalability for storage services.	Intel Core i7-4770, Intel Xeon E5-2676, Ubuntu 14.04 LTS	In future, reduces the data security problems.

-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.542

|| Volume 9, Issue 10, October 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0910008 |

IV. CONCLUSION AND FUTURE ENHANCEMENT

In a secure and well organized cloud storage, removing redundant data has got much importance. Particularly in case of big data storage, removing redundant information will greatly improve utilization of space. Day by day need for cloud services for storing large amount of data is increasing and as a result need for efficient deduplication technique is also arising. Huge data load on storage systems are emphasizing the focus on development of novice techniques to remove duplicate data. With the evolution of data deduplication and its various techniques, cloud computing has potential to remove unnecessary duplicates. Deduplication is an important technique for reducing storage cost, bandwidth and energy consumption.

In Future, this system can be expanded to manage the storage with deduplication when there is any updation or deletion of files and implementing Map Reduce algorithms to reduce execution time. Also the system can be expanded to perform deduplication on other Medias such as audio, video and images.

REFERENCES

[1] Huijun Wu, Chen Wang, Yinjin Fu, Kai Lu, Liming Zhu, IEEE Transactions on Parallel and Distributed Systems,2018.
[2] Jianwei Yin, Yan Tang, Shuiguang Deng, Bangpeng Zheng and Albert Y. Zomaya, "MUSE: A Multi-tierd and SLA-Driven Deduplication Framework for Cloud Storage systems" , IEEE Transactions on Computers, MAY 2018
[3] UrsNiesen, "An Information-Theoretic Analysis of Deduplication", IEEE TRANSACTIONS ON INFORMATION THEORY, VOL. 65, NO. 9, SEPTEMBER 2019
[4] Xue Yang, RongxingLu,Jun Shao, Xiaohu Tang, and Ali A.Ghorbani, "Achieving Efficient Secure Deduplication with User-Defined Access Control in Cloud", IEEE Transactions on Dependable and Secure Computing,2020.
[5] Haoran Yuan, Xiaofeng Chen, Jin Li, Tao Jiang, Jianfeng Wang, and Robert H. Deng, "Secure Cloud Data Deduplication with Efficient Re-encryption", IEEE Transactions on Services Computing,2019.
[6] Chia-Mu Yu, Sarada Prasad Gochhayat, Mauro Conti, and Chun-Shien Lu, "Privacy Aware Data Deduplication for Side Channel in Cloud Storage", IEEE Transactions on Cloud Computing, VOL. 14, NO. 8, AUGUST 2015.
[7] Hui Cui, Robert H. Deng, Yingjiu Li, and Guowei Wu, "Attribute-Based Storage Supporting Secure Deduplication of Encrypted Data in Cloud", IEEE Transactions on Big Data,2017.
[8] Shunrong Jiang, Tao Jiang and Liangmin Wang, "Secure and Efficient Cloud Data DeduplicationwithOwnership Management", IEEE Transactions on Services Computing,2017.
[9]Zheng Yan, Lifang Zhang, Wenxiu Ding, and QinghuaZheng, "Heterogeneous Data Storage Management with Deduplication in Cloud Computing", IEEE Transactions on Big Data,2017.
[10] Yinjin Fu, Nong Xiao, Hong Jiang, Guyu Hu, and Weiwei Chen, "Application-Aware Big Data Deduplication in Cloud Environment", IEEE Transactions on Cloud Computing,2017.
[11] NahlahAslam K.P., "Data Deduplication with Encrypted Big Data
Management in Cloud Computing", IEEE Xplore ISBN: 978-1-7281-1261-9,2019.
[12] Zheng YAN, Wenxiu DING, Xixun YU, Haiqi ZHU, DENG, Robert H., "Deduplication on encrypted big data in cloud", IEEE Transactions on Big Data,2016.
Page 20 of 20
[13] ShengmeiLuo, Guangyan Zhang, Chengwen Wu, Samee U. Khan,andKeqin Li, "Boafft: Distributed Deduplication for Big Data Storage in the Cloud", IEEE TRANSACTIONS ON CLOUD COMPUTING, VOL. 61, NO. 11, JANUARY 2015.
[14] A.G.Ashmita, "Data De-Duplication on Encrypted Data Lake in Cloud Environment", International Journal of Engineering Research & Technology (IJERT),Vol. 7 Issue 03,ISSN: 22780181, MARCH 2018.
[15] Youngjoo Shin, Dongyoung Koo, Joobeom Yun and Junbeom Hur Member, "Decentralized Server-aided Encryption for Secure Deduplication in Cloud Storage", IEEE Transactions onServices Computing,2017.

INNO SPACE
SJIF Scientific Journal Impact Factor
Impact Factor: 7.542

doi cross ref

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर NISCAIR

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