e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:06/Issue:11/November-2024 Impact Factor- 8.187 www.irjmets.com

CLOUD SIM FRAMEWORK UTILIZATION ON DATA DUPLICATION

Aditya Bhairawkar*1, Jay Devare*2, Nilam Asawale*3,
Anjali Dhawale*4, Prof. Dr. Vaishali Thorat*5

*1,2,3,4,5Department Of Computer Engineering, D Y Patil University, D Y Patil College of Engineering Ambi, Pune, Maharashtra, India.

DOI : https://www.doi.org/10.56726/IRJMETS63719

ABSTRACT

ACloudSim is a widely adopted simulation framework designed to facilitate the modelling and evaluation of cloud computing environments and services. This study explores the utilization of CloudSim in addressing data duplication challenges within cloud storage systems. By simulating various cloud architectures and data management strategies, the research analyses the impact of data duplication on resource utilization, performance, and cost-efficiency. We demonstrate how CloudSim can be leveraged to model different scenarios involving data replication strategies, providing insights into optimal configurations for minimizing redundancy while maintaining data availability. The findings contribute to the understanding of effective data management practices in cloud computing, aiding cloud service providers in enhancing their operational efficiency. As the world has seen exceptional movement throughout the most recent decade, there is an unusual development in the wrongdoing rate and besides the amount of law breakers is extending at an upsetting rate, this leads toward an uncommon stress over the security issues. The individualistic characters of the human face can be isolated by face affirmation. Face affirmation is a clear and deft biometric development. Face distinguishing proof and affirmation is the development which is used to perceive a person from a video or picture. In this system, we can recognize and see the characters of the criminals in a video move got from a camera ceaselessly. Criminal records generally contain individual nuances and the photograph of the hoodlum. Thusly, we can use this photograph close by his nuances. The video got from the perception camera are changed over into diagrams. Right when a face is recognized in a packaging, it is pre-dealt with and a while later it goes through feature extraction. The components of the dealt with consistent picture are differentiated and the features of taken care of pictures which are taken care of in the criminal informational collection. Accepting that a match is found, a caution message close by the live region of the criminal would be delivered off the power. So, this system will be incredibly useful for the police division to recognize the criminal through video got from camera consistently. In this paper Haar Cascade Algorithm is used for face affirmation.

Keywords: CloudSim Framework, Data Duplication, Data Management, Storage Efficiency, Resource Allocation.

I. INTRODUCTION

Data deduplication is a computer technique that eliminates multiple copies of repeated data. When used successfully, this method can improve storage utilization and save capital by requiring less storage media to meet storage capacity requirements. Data replication is a technique that reduces storage overhead by eliminating duplicate data. This method ensures that only one distinct instance of data is stored on a storage medium such as disk, flash memory, or tape. The pointer to a single copy of the data is used instead of the excess data blocks. The duplication of data and increasing backup are similar, as they only copy data that has changed since the last backup. Data deduplication is one of the techniques used to solve the problem repetitive data. Deduplication technology is usually used in cloud servers to reduce the space on the servers. To prevent unauthorized access to data and creation of duplicate data in the cloud, an encryption method is used to encrypt the data before storing it on the cloud server. Cloud storage usually contains business critical data and processes; therefore, high security is the only solution to maintain a strong trust relationship between cloud users and cloud service providers. Therefore, to overcome the security threats, this article proposes several cloud storage solutions. Thus, common forms of data storage such as user specific files and databases are separated and stored in different cloud storages (e.g., Cloud A and Cloud B). Data deduplication allows users to reduce redundant data and manage backup processes more efficiently while realizing the benefits of more efficient backups, cost savings, and load balancing.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:06/Issue:11/November-2024 Impact Factor- 8.187 www.irjmets.com

II. LITERATURE SURVEY

Title	Eliminating Redundancy in cloud	A secure data duplication system	Big Data Analytics in cloud	Data Recovery & backup Management	Secure and Efficient Deduplication
Year	2015	2020	2022	2023 June	2023 Dec
Abstract	Cloud computing improves deployment speed and storage efficiency with data deduplication, granting authorized users secure access through unique tokens and encryption.	The paper introduces a method combining Convergent Encryption and Modified Elliptic Curve Cryptography for secure and efficient data deduplication in cloud-fog environments.	The paper examines the importance of Big Data and Cloud Computing, emphasizing Big Data Analytics and a case study on Google's BigQuery for scalable data analysis across various sectors.	The paper highlights cloud computing's benefits for secure data storage and recovery, focusing on Google Photos and Google Drive while addressing data loss and technological threats.	The paper introduces a secure deduplication protocol that enhances data privacy and efficiency while managing client ownership and ensuring secrecy.
Keyword	Convergent encryption, De-duplication, duplicate check, hybrid cloud.	Convergent encryption (CE), Modified elliptic curve cryptography (MECC), Edge computing, Integrated cloud and fog networks, Hash tree. Secure hash algorithm (SHA)	Big data, Analytics, BigQuery, Cloud computing	Cloud Computing, Data Recovery, Secure Storage, Backup Managemen, Disaster Recovery	Deduplication; cloud storage; data sharing; message-locked encryption; dynamic ownership update
Introduction	Cloud computing offers cost-effective data storage with deduplication for efficiency and convergent key encryption to ensure secure, authorized access in a hybrid environment.	The IoT's data growth demands cloud and fog computing, with fog handling local processing to reduce delays before cloud storage.	The paper highlights the explosive growth of digital information, projected to rise from 44 zettabytes in 2020 to 163 zettabytes by 2025 due to technological advancements and IoT devices.	The paper highlights Google Cloud's secure and cost-effective data storage solutions, focusing on encryption, data retention policies, and recovery benefits.	The paper examines the generation of vast client data through digital technology and its analysis for personalized services, emphasizing the efficiency of cloud services in data management and access.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:06/Issue:11/November-2024 Impact Factor- 8.187 www.irjmets.com

Advantage	1. Storage Efficiency: Data deduplication reduces duplicate data, optimizing storage costs. 2. Security: Unique tokens and private keys ensure that only authorized users can access sensitive information.	The method improves storage efficiency and security by reducing data redundancy while encrypting data in cloud-fog environments.	Big Data Analytics enables organizations to make informed, data-driven decisions, improving efficiency and outcomes across various sectors.	Google Cloud offers secure, cost-effective data storage and backup solutions with strong encryption and flexible retention policies.	The proposed secure deduplication protocol reduces computational costs for clients while enhancing data privacy and efficiently managing ownership.
Disadvantages	Cloud data deduplication poses security risks to sensitive information and increases management complexity for storage providers.	A potential disadvantage is the increased computational complexity due to dual encryption, which may impact system performance and processing speed.	A disadvantage of Big Data Analytics is the potential for data privacy concerns and security risks associated with handling large volumes of sensitive information.	A disadvantage of Google Cloud is the potential concern over data privacy and the reliance on internet connectivity for accessing stored information.	A disadvantage of the proposed secure deduplication protocol is the reliance on complete trust in the cryptographic server for data security and management.
Future Scope	The future of cloud data deduplication involves advanced encryption for security, AI-driven management, and improved scalability to accommodate growing data volumes.	Future scope includes optimizing dual encryption for performance, integrating machine learning for enhanced deduplication, and applying the method to edge computing and blockchain technologies.	Future developments in Big Data Analytics will focus on real-time processing, advanced machine learning, and stronger data privacy measures.	Future scope for Google Cloud includes enhancing data privacy, integrating advanced AI tools, and expanding hybrid cloud capabilities for greater flexibility.	Scope would be interesting to incorporate efficient key revocation techniques into our proposed protocol. This integration would lead to a substantial enhancement in the ownership update.

III. PROBLEM STATEMENT

In cloud computing environments, data duplication poses significant challenges to storage efficiency, data management, and cost optimization. Redundant data blocks across distributed storage resources can increase storage costs, lead to inefficient data retrieval, and potentially degrade system performance. Effective solutions to identify, manage, and minimize data duplication are essential to optimize resource utilization and reduce operational costs in cloud systems.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:06/Issue:11/November-2024 Impact Factor- 8.187 www.irjmets.com

IV. PROPOSED SYSTEM

The proposed system for utilizing the CloudSim Framework to address data duplication issues is designed to efficiently manage, simulate, and evaluate the performance of cloud environments while focusing on the duplication of data storage across multiple virtual machines and datacentres. In modern cloud computing, data redundancy is essential to ensure data availability, reliability, and fault tolerance, yet excessive data duplication leads to increased storage costs, unnecessary use of resources, and higher energy consumption. This system integrates data deduplication algorithms into the CloudSim Framework, allowing for simulation and analysis of cloud environments with an emphasis on storage optimization. The CloudSim Framework, a widely adopted toolkit for modelling cloud infrastructures, enables detailed simulation of complex cloud scenarios. By incorporating deduplication mechanisms, the system aims to identify and remove redundant data blocks, thus reducing storage requirements and improving resource allocation efficiency.

In the simulation setup, virtual machines (VMs) and data centres are configured to handle different data sets, where the system continually monitors data for potential duplication across the cloud environment. Advanced hash-based or content-based algorithms, such as SHA-256 or Rabin fingerprinting, are applied to uniquely identify data chunks, ensuring that only unique data is stored while duplicate copies are referenced, not replicated. This process optimizes storage space, reduces latency, and minimizes I/O load on storage devices. Furthermore, CloudSim's inherent flexibility allows for real-time simulation adjustments, enabling researchers to experiment with different deduplication algorithms, varying data sizes, and diverse cloud configurations. By evaluating metrics like data storage savings, resource utilization, and processing times, the system provides critical insights into how deduplication can enhance the performance and scalability of cloud infrastructures. Ultimately, this proposed system allows for a comprehensive assessment of data deduplication's benefits within cloud environments, paving the way for more sustainable and cost-effective cloud storage solutions.

Expanding on the proposed system, this approach goes beyond merely reducing storage costs; it aims to optimize the entire cloud computing infrastructure by addressing redundancy at multiple levels, including storage, data transmission, and computational efficiency. When cloud storage solutions experience high levels of duplication, resources such as CPU, memory, and bandwidth can be heavily burdened, affecting performance and user satisfaction. The proposed system leverages the CloudSim Framework not only to simulate and detect redundancy but also to implement different deduplication strategies tailored to specific use cases, such as file-level, block-level, and byte-level deduplication. Each approach has unique benefits: file-level deduplication is faster and simpler, while block-level and byte-level deduplication offer higher granularity, thus greater storage savings.

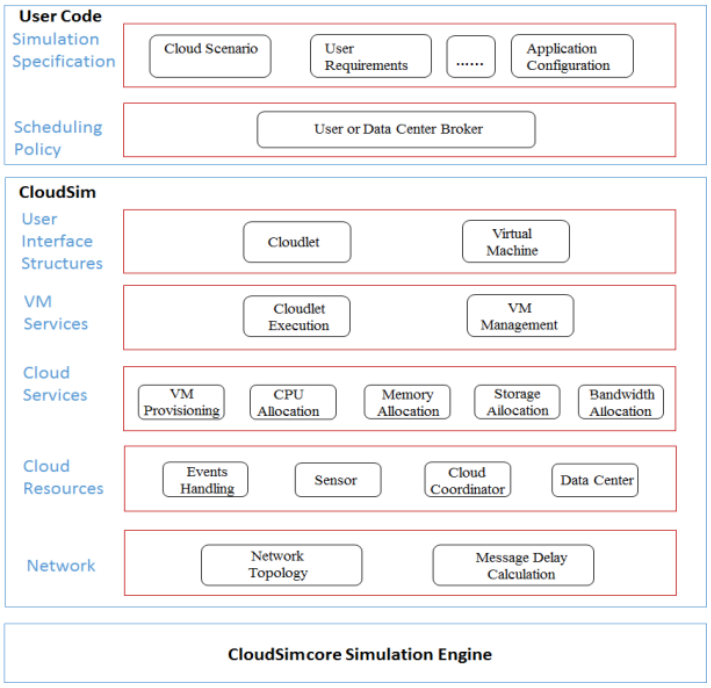

Fig1. Cloud Sim Architecture

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:06/Issue:11/November-2024 Impact Factor- 8.187 www.irjmets.com

The integration of the deduplication mechanism within CloudSim allows for a versatile testing environment, where different deduplication configurations and policies can be simulated under various workloads and conditions. For instance, data from different applications and users can be analysed to measure the effect of deduplication on network load and overall performance across distributed datacentres.

The deduplication module can also incorporate compression techniques to further reduce the storage footprint of unique data, adding another layer of efficiency. By compressing only unique data blocks rather than duplicates, the system achieves a fine balance between compression efficiency and processing overhead.

The proposed system also takes into account the energy consumption factor. Storage and redundant data processing lead to significant energy expenditure in cloud infrastructures. By removing redundant data, the system indirectly reduces the energy needed to store, retrieve, and maintain data across multiple datacentres. This aspect is crucial for cloud providers aiming to meet sustainability goals. Through CloudSim, researchers can simulate various cloud setups under different deduplication techniques to measure energy savings, providing insights into the environmental impact of their cloud strategies.

Moreover, the system includes mechanisms to track and validate the integrity of deduplicated data, ensuring no data loss or corruption occurs during the deduplication process. By leveraging hash values or digital fingerprints for each data block, the system verifies the integrity of data when stored and retrieved. In a real-world setting, this feature ensures data security and reliability, especially for sensitive or regulated data.

The proposed system thus offers a robust platform for simulating, testing, and validating data deduplication strategies within cloud infrastructures. By integrating deduplication at various stages of data storage, transmission, and processing, cloud providers can achieve a more cost-effective and sustainable solution. With CloudSim's extensive capabilities, this system not only reduces data redundancy but also paves the way for optimized resource utilization, lower operational costs, and an overall more efficient cloud environment. Ultimately, this system could provide a blueprint for implementing data deduplication across large-scale cloud deployments, fostering greater innovation and sustainability in cloud computing.

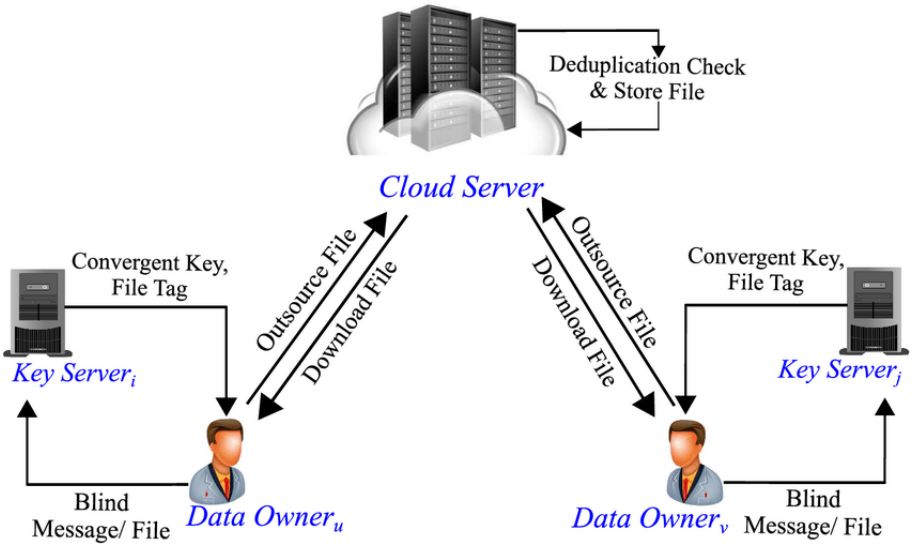

Fig2. Working of Project

Building on the system's capabilities, this proposed approach to data deduplication within CloudSim extends to dynamic data management techniques, incorporating real-time data analytics and workload predictions to continuously improve deduplication efficiency. As cloud environments handle vast, often unpredictable workloads, the system is designed to dynamically adapt deduplication strategies based on current data patterns and usage trends. For example, during periods of high data inflow, the system can prioritize more aggressive deduplication methods to handle temporary storage pressures, while shifting to less intensive methods during low-traffic periods to optimize processing overhead.

A major feature of this system is its ability to model distributed deduplication across multiple datacentres. In a typical cloud setup, data is often replicated across locations to improve fault tolerance and ensure high availability. However, this redundancy across geographically distributed datacentres can lead to unnecessary

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:06/Issue:11/November-2024 Impact Factor- 8.187 www.irjmets.com

duplication at a global scale. The proposed system's deduplication module is capable of simulating cross-datacentre deduplication, where unique data is identified and shared across multiple sites rather than replicated independently. By doing so, this system not only conserves local storage but also optimizes inter-datacenter bandwidth, reducing costs associated with data transfer across regions. The system's integration within CloudSim enables simulations of cross-regional deduplication under different network configurations, providing insights into how global-scale deduplication strategies affect latency, bandwidth utilization, and data availability.

Another core component is the system's support for hybrid deduplication models. In many cloud settings, a hybrid cloud model – combining private and public clouds – is used to manage different data sensitivity levels and workloads. The system can simulate deduplication across hybrid cloud models, allowing administrators to explore deduplication across both public and private environments. For example, less sensitive data in a public cloud can undergo more aggressive deduplication, while critical data in private cloud environments can use milder, integrity-focused deduplication strategies. This approach ensures that deduplication does not compromise data privacy or integrity while optimizing storage across hybrid architectures.

The system also considers deduplication's impact on disaster recovery (DR) and data backup. In cloud environments, DR strategies often involve data replication across primary and secondary locations to safeguard against data loss. However, direct deduplication can interfere with DR objectives if not managed carefully. By simulating DR scenarios within CloudSim, the system can implement selective deduplication that balances storage optimization with data recovery requirements. For instance, rather than deduplicating all data, only certain non-critical files or infrequently accessed data could be deduplicated, ensuring that critical files are always readily available for recovery.

Furthermore, the system is designed to handle multi-tenancy, a common characteristic of cloud environments where multiple users or organizations share resources. Multi-tenant deduplication, particularly within public clouds, presents unique challenges, as data privacy and security are paramount. The proposed system incorporates tenant-aware deduplication that prevents data from being shared across tenants, addressing privacy concerns while still minimizing duplication within each tenant's allocated resources. By assigning unique identifiers to each tenant's data, the system ensures deduplication is confined to data owned by the same tenant, aligning with regulatory and compliance requirements.

In terms of user interface and accessibility, the system offers a dashboard for monitoring deduplication performance metrics in real-time. Metrics such as deduplication ratios, storage savings, CPU usage, and network traffic can be visualized, providing administrators with valuable insights into the ongoing efficiency of deduplication. Additionally, administrators can test different deduplication configurations through the CloudSim Framework's dashboard, adjusting parameters to simulate various scenarios, such as high data loads, security settings, and backup frequencies. This interface not only provides actionable data but also enables experimentation, allowing cloud architects to test and identify the most effective deduplication strategies for their unique needs.

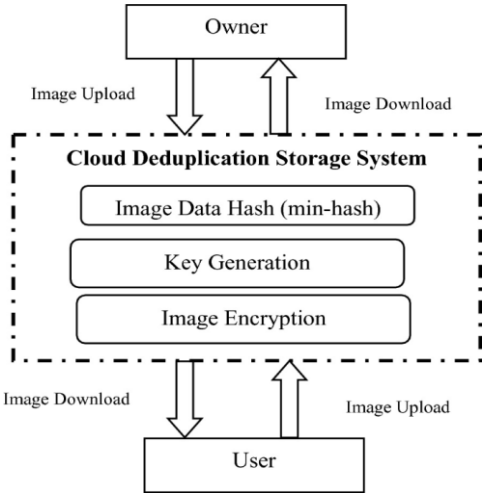

Fig3. An Efficient Secure Cloud Image Deduplication with Weighted Min-Hash Algorithm

V. RESULT

CloudSim is a widely used framework in cloud computing research, particularly for modelling, simulating, and analysing cloud computing environments. It allows researchers to assess various aspects of cloud infrastructure, including data duplication, in a controlled environment.

When applied to data duplication in a parallel (para) processing environment, Cloud Sim enables the simulation of scenarios where data is duplicated across multiple virtual machines (VMs) or data centres. This is especially useful in analysing the efficiency of resource utilization, network bandwidth, and storage costs.

Using CloudSim for data duplication studies in parallel processing environments allows researchers to optimize cloud performance and data management strategies without incurring the high costs and risks associated with testing on real infrastructure.

VI. CONCLUSION

The utilization of the CloudSim Framework for addressing data duplication demonstrates significant potential in enhancing cloud computing efficiency and resource management. By leveraging CloudSim's simulation capabilities, researchers and developers can model, test, and optimize data deduplication processes without deploying them in a live environment, reducing time and costs. This framework allows for the analysis of various deduplication algorithms under different network conditions and workloads, offering insights into performance trade-offs and resource savings. Overall, CloudSim serves as an effective tool for assessing and improving data duplication strategies, contributing to more scalable and optimized cloud environments.

VII. REFERENCES

[1] Muhammad Zulkifl Hasan, Muhammad Zunnurain Hussain, and Nadeem Sarwar, "Data Recovery and Backup Management: A Cloud Computing Impact," Faculty of Computer Science and Information Technology University of Central Punjab Lahore, Pakistan, June 2023.

[2] M. Bellare, S. Keelveedhi, and T. Ristenpart, "Dupless: Server-aided encryption for Deduplicated Storage," in Proc. 22nd USENIX Conf. Sec. Symp., 179-194 (2013).

[3] H. Wang, "Identity-Based Distributed Provable Data Possession in Multicloud Storage," IEEE, 8(2), 328-340 (2015).

[4] J. M. Bohli, N. Gruschka, M. Jensen, L. L. Iacono, and N. Marnau, "Security and Privacy-Enhancing Multicloud Architectures," 10(4), 212-224 (2013).

[5] N. Yager and A. Amin, "Fingerprint Verification Based on Minutiae Features: A Review," Pattern Anal. Apple., 10, 94-113 (2004); Feb 14, 7(1), pp. 94-113.

[6] A Study on Authorized Deduplication Techniques in Cloud Computing, Int. J. Adv. Res. Computer Engg. Technol. (IJARCET), 3(12) (2014).

[7] Data Deduplication in Cloud Computing Systems, International Workshop on Cloud Computing and Information Security (CCIS) (2013).

[8] Mira Lee and Minhye Seo, “Secure and Efficient Deduplication for Cloud Storage with Dynamic Ownership Management,” Appl. Sci. 2023, 13, 13270.

[9] P. Anderson and L. Zhang, "Fast and Secure Laptop Backups with Encrypted DeDuplication," in Proc. 24th Int. Conf. Large Installation Syst. Admin., 29-40 (2010).

[10] Berisha et al., "Big Data Analytics in Cloud Computing: An Overview," Journal of Cloud Computing https://doi.org/10.1186/s13677-022-00301-.

[11] P. G. et al., "A Secure Data Deduplication System for Integrated Cloud-Edge Networks," Journal of Cloud Computing: Advances, Systems and Applications (2020) 9:61 https://doi.org/10.1186/s13677-020-00214-.

[12] Cloud Computing Security: From Single to Multi-Clouds Using Digital Signature, Int. J. Engg. Technol., Manage. Appl. Sci. www.ijetmas.com 2(6), (2014).

[13] Search Engine Market Share Worldwide. Available online: https://gs.statcounter.com/search-engine-market-share#monthly-202201-202212-bar (accessed on 15 October 2023).

[14] Ng, W.K.; Wen, Y.; Zhu, H. Private data deduplication protocols in cloud storage. In Proceedings of the

27th Annual ACM Symposium on Applied Computing, Trento, Italy, 26–30 March 2012; pp. 441–446. Appl. Sci. 2023, 13, 13270 22 of 22

[15] Dutch, M. Understanding data deduplication ratios. In Proceedings of the SNIA Data Management Forum, Orlando, FL, USA, 74. 5. 6. 7. 8. 9. 10. 11. 12. 13. 14. 15. 16. 17. 18. 19. 20. 21. 22. 23. 24. April 2008; Volume 7.

[16] Douceur, J.R.; Adya, A.; Bolosky, W.J.; Simon, P.; Theimer, M. Reclaiming space from duplicate files in a serverless distributed file system. In Proceedings of the 22nd International Conference on Distributed Computing Systems, Vienna, Austria, 2–5 July 2002; pp. 617–624.

[17] Ali, M.; Dhamotharan, R.; Khan, E.; Khan, S.U.; Vasilakos, A.V.; Li, K.; Zomaya, A.Y. SeDaSC: Secure data sharing in clouds. IEEE Syst. J. 2015, 11, 395–404. [CrossRef]

[18] Hur, J.; Koo, D.; Shin, Y.; Kang, K. Secure data deduplication with dynamic ownership management in cloud storage. IEEE Trans. Knowl. Data Eng. 2016, 28, 3113–3125. [CrossRef]

[19] Areed, M.F.; Rashed, M.M.; Fayez, N.; Abdelhay, E.H. Modified SeDaSc system for efficient data sharing in the cloud. Concurr. Comput. Pract. Exp. 2021, 33, e6377. [CrossRef]

[20] Keelveedhi, S.; Bellare, M.; Ristenpart, T. DupLESS: Server-Aided encryption for deduplicated storage. In Proceedings of the 22nd USENIX Security Symposium (USENIX Security 13), Washington, DC, USA, 14–16 August 2013; pp. 179–194.

[21] Bellare, M.; Keelveedhi, S.; Ristenpart, T. Message-locked encryption and secure deduplication. In Advances in Cryptology—EUROCRYPT 2013, Proceedings of the Annual International Conference on the Theory and Applications of Cryptographic Techniques, Athens, Greece, 26–30 May 2013; Springer: Berlin/Heidelberg, Germany, 2013; pp. 296–312.

[22] Puzio, P.; Molva, R.; Önen, M.; Loureiro, S. ClouDedup: Secure deduplication with encrypted data for cloud storage. In Proceedings of the 2013 IEEE 5th International Conference on Cloud Computing Technology and Science, Bristol, UK, 2–5 December 2013; Volume 1, pp. 363–370.

[23] Scanlon, M. Battling the digital forensic backlog through data deduplication. In Proceedings of the 2016 Sixth International Conference on Innovative Computing Technology (INTECH), Dublin, Ireland, 24–26 August 2016; pp. 10–14.

[24] Kim, D.; Song, S.; Choi, B.Y.; Kim, D.; Song, S.; Choi, B.Y. HEDS: Hybrid Email Deduplication System. In Data Deduplication for Data Optimization for Storage and Network Systems; Springer: Cham, Switzerland, 2017; pp. 79–96.

[25] Shin, Y.; Koo, D.; Yun, J.; Hur, J. Decentralized server-aided encryption for secure deduplication in cloud storage. IEEE Trans. Serv. Comput. 2017, 13, 1021–1033. [CrossRef]

[26] Yuan, H.; Chen, X.; Wang, J.; Yuan, J.; Yan, H.; Susilo, W. Blockchain-based public auditing and secure deduplication with fair arbitration. Inf. Sci. 2020, 541, 409–425. [CrossRef]

[27] Ma, X.; Yang, W.; Zhu, Y.; Bai, Z. A Secure and Efficient Data Deduplication Scheme with Dynamic Ownership Management in Cloud Computing. In Proceedings of the 2022 IEEE International Performance, Computing, and Communications Conference (IPCCC), Austin, TX, USA, 11–13 November 2022; pp. 194–201.