Grenze International Journal of Engineering and Technology, June Issue

Data Auditing in Cloud Computing: A Survey

Sumit Sharma[1], Prasanna Dwivedi[2] and Swapnil Kaushal[3]
[1-2]Department of CSE, GL Bajaj Institute of Technology and Management, Greater Noida, Uttar Pradesh
Email: asyncsunlight@gmail.com, prasannadwivedi12@gmail.com
[3]Assistant professor, JIMS Engineering Management Technical Campus, Greater Noida, Uttar Pradesh
Email: swapnilkaushal5@gmail.com

***Abstract*— The use of data replication as a technique for improving performance and addressing various issues in cloud storage systems, such as availability, stability, security, bandwidth, and data access response times, has been discussed in this article. A thorough assessment and There has been a classification of cutting-edge data replication techniques in a number of current cloud computing technologies. Data de-duplication techniques, data auditing techniques, and data management strategies have been classified as replication techniques and each scheme has been studied in-depth, including its key characteristics, type, implementation location, evaluation instruments, benefits, and disadvantages. The article concludes with a comparison analysis of the different replication strategies and a discussion of unresolved problems and future research challenges in the area of cloud data replication. Overall, a significant contribution to the field of cloud data replication has been offered by this study, which can be useful for researchers and practitioners working on cloud storage systems.**

***Index Terms*— Cloud Computing, Data Auditing, Data Deduplication, Cloud Storage Security, Data Privacy and Data Replication.**

I. INTRODUCTION

Your Large-scale data processing is becoming more and more necessary across a range of industries, including business, industry, and healthcare. Researchers are using a variety of data processing approaches and paradigms to satisfy these criteria. Distributed cloud computing is one of these new paradigms. It offers consumers a virtualized pool of computer resources they may use as needed, including storage, software, memory, processing power, and services.

Software-as-a-Service (SaaS), and Infrastructure-as-a-Service (IaaS), Platform-as-a-Service (PaaS), are the three models used to classify cloud services. While PaaS allows clients to use applications on the cloud provider's infrastructure regardless of the operating system they are using, IaaS allows end users to access applications that have been placed on the provider's infrastructure.

Researchers have been introducing cutting-edge technologies including mobile edge computing (MEC), mobile cloud computing (MCC), fog computing, edge computing, to lessen the processing burden. Offloading is a technique to move the resource-intensive portion of the data or code to these nearby settings or eventually cloud servers to manage the resources' utilization properly [8]. These technologies make it possible.

Numerous various data compression methods, including data type-based, application-based, coding-based, and data quality-based methods, have been introduced in the literature to reduce storage size.

Grenze ID: 01.GIJET.12.2.240_20

De-duplication, often known as smart compression, is one method for removing redundant data from several files. This strategy is useful when a lot of data is accessible. Data duplication is detected using hash values in the de-duplication process, and chunks can be fitted to ensure data integrity.
A promising method for storing data blocks across several cloud servers is cloud storage, to sum up. The burden of linked storage concerns can be lessened by businesses or even by individuals, making cloud storage one of the key cloud computing services that allows authorised users to access stored data from multiple locations.
The basic goal of data replication, which is increasingly used today, is to handle a vast volume of data in a dispersed manner. This method improves data availability while lowering access latency and increasing access speed. The right number of data replicas and their placement for efficient system job execution are two essential issues that arise in the context of cloud storage clusters. Keeping numerous rigid duplicates does not address the first issue because it is neither practical nor cost-effective. The standard data replication strategy used by distributed data storage systems like the Google File System (GFS), Amazon S3 and Hadoop Distributed File System (HDFS) is three replicas [8]. In relation to the latter issue, load balancing is necessary to match user expectations, which necessitates a thorough system analysis. In cloud-based systems, which are used globally, many levels of replication are frequently used, particularly between clusters, servers, data centers, or even between cloud services. Optimistic solutions (also known as lock-based solutions) and pessimistic solutions are common approaches for maintaining consistency amongst an object's replicas. Although a pessimistic method offers higher consistency, it has drawbacks such decreased performance, limited availability, and poor scaling. The best approach, however, relies on conditional consistency and permits a little divergence in the replication process to maintain availability and performance. In connection to the latter problem, a detailed system analysis is required for load balancing in order to meet user expectations. Many layers of replication are regularly employed in cloud-based systems, which are widely used, especially amongst clusters, servers, data centers, and even cloud services. Pessimistic solutions and optimistic solutions are typical strategies for preserving consistency across an object's clones. Optimistic solutions are often referred to as lock-based solutions. A pessimistic approach has lower performance, less availability, and poor scaling, but it also provides stronger consistency. To preserve availability and performance, the optimal strategy, however, relies on conditional consistency and allows for some divergence in the replication process. It is frequently used in applications that exchange vast amounts of data and needs to be accessed from many locations across various geographies. As a result, even if one of the sites hosting its copy fails, the required data file can still be delivered from other locations. By granting requests from the closest sites to maintain the desired file replica, data replication is able to accomplish its objectives. Services, virtual machines, tasks, and data replications can all be done using replication solutions. This study focuses on data replication as one of these solutions. In this study, the state-of-the-art data replication techniques are thoroughly surveyed and categorized from the cloud computing literature. In order to increase the availability and dependability of cloud-based systems and lower the costs of consistency, communication and data storage, it describes the presentation style and uses replication techniques. The replication techniques' key characteristics, benefits, and drawbacks are highlighted in a thorough comparative study. Additionally, future unresolved or inadequately addressed research issues to enhance data replication strategies in the cloud system are mentioned.

II. Preliminary Metrics of Data Auditing

The first step of auditing, called preliminary data auditing, is to look at the data to evaluate if the information is pertinent enough for more extensive research. Before doing an extensive auditing, this step includes checking the data for inconsistencies, errors and unusual things that might need fixed using suitable methodology. Data profiling is about collecting metadata to understand the data's structure, content, and quality is often part of this process. Other chores may include changing, normalizing, and cleaning of the data to make sure it is correct, complete, and consistent with the standards required. The goals of this auditing is to find problems that could pose a threat to the data's reliability and also to prepare it for a more detailed analysis. The appropriateness of the data for further applications is increased by this, resulting in a higher overall efficiency.

A. Data Integrity

Any system that lies under data storage such as cloud storage system is ought to have data integrity as a crucial need. This ensures the accuracy, fullness, and alignment throughout its life span. Methods like replication, data check-summing, data encryption, and other similar methods can be taken into consideration for the maintenance of data integrity. A distinctive value is created for each data block, when checking data, and that unique value is paralleled to the checksum value, the instant data is accessed. Encryption data require converting them into a

coded format which can only be accessed through a key. Data corruption is detected as well as corrected via replication, which imbibes creation of numerous copies of the same.

B. Data Privacy

Another key factor for data auditing is data privacy. The prevention of the provision of unauthorized access or revelation of sensitive data lies under this. Various procedures such as data masking, data encryption, and restricted access to unauthorized guests can ensure the preservation of data privacy. Data masking demands transforming the data to hide sensitive information while ensuring access for authorized guests. The transformation of data into coded format which is key accessible is known as data encryption. Limited access to the data is provided in accordance with users' roles and permissions through access control.

C. Data Security

The process of averting unsanctioned access, theft, or impairment to the data is known as data security. Approaches like firewalls, access control, and data encryption can be taken into consideration to preserve data security.

Transformation of the data into an encrypted and structured code format, which can only be accessed via a key, is known as data encryption. In accordance with the users' roles and permissions, access control entails limiting access to the data. Inbound and outbound network traffics are kept in track and managed via firewalls per pre-established security standards.

D. Data Compliance

Ensuring that the said data abides by all the applicable laws, regulations, and known industry standards is called data compliance. Methods, including data encryption, access control, and data retention regulations, can be taken into consideration for data compliance maintenance.

Transformation of said data into encoded format which is accessible only through a defined key is known as data encryption. Limited access is provisioned in accordance with the roles and permissions of the guests through access control. Approaches for data retention determine for how long the said data must be commissioned and when it must be discarded.

E. Audit Logs

Audit logs are documentation of all events and activities related to data storage, processing, and access. Audit logs are useful for tracking and analyzing data access and patterns of usage, emphasizing the potential security risks and compliance violations, as well as troubleshooting the errors in the provided system. To certify the integrity and accessibility, audit logs are backed up timely and stored in secured locations.

F. Audit Trails

An audit trail is a sequential record of data access and modifications. Audit trails can be utilized for tracing the origin of data changes, modifications in data tracking, and detection of potential breaches. Audit trails could be brought about by both the users themselves or automatically via the system.

G. Monitoring and Reporting

The process of monitoring and reporting involves tracking and analyzing data access and usage patterns, as well as generating reports to identify potential security risks and compliance violations. It is important to consider monitoring and reporting as a means of detecting unusual data access and usage patterns, identifying potential data breaches, and troubleshooting system errors.

H. Cost

Expenditure must be taken into consideration while doing data auditing. Data auditing can be an exorbitant operation that demands specific hardware and software. Data auditing expenses are variable contingent on volume of data, the convolution of the system, and the desired levels of security and adherence.

I. Bandwidth

Another important factor in data auditing is the bandwidth. Data audit could potentially generate a substantial amount of network traffic flow, which surely would impact system's availability and performance. Limitations on the bandwidth shall also put affects on the speed and accuracy of data auditing. Therefore, it is vital to consider the bandwidth requirements and limitations when designing a data auditing system.

III. RELATED WORK

Serial no	Type of auditing	year	Place of Auditing	Performances metrics	Evaluation tools	Advantages	Dis-advantages
1	Private	2020	Client-Side	Cost Time	python	Suitable system Model	Not comprehensive
2	Public	2020	Server-side	Integrity	C	Suitable Security Model	High Overhead
3	Public	2019	Third party-side	Availability Cost	C	Suitable Organization	Low efficiency High Complexity
4	Public	2019	Client-Side	Bandwidth Size	java	Easy to design model	Poor result
5	Public	2020	Client-Side	Bandwidth Size	python	Embedded security	Lack of privacy High overhead
6	Public	2018	Client-Side	Time Security	C++	Powerful organization	High Complexity
7	Public	2019	Client-Side	Bandwidth Security	C++	Suitable system Model	Low efficiency

Figure 1. Comparison of data auditing schemes within the cloud

A. Dwivedi et al

In the research paper from cloud computing environment suggests a distributed and lax auditing strategy for outsourced data. The proposed plan streamlines the audit process, triggering it only as necessary to minimize the complexities and overhead inherent in traditional methods. Several measures assess the performance of the proposed scheme, whose design and implementation are described. According to the findings provided, The presented plan would surely lower expenses for the scheme and enhances auditing efficiency. Potential uses of the suggested technique in cloud computing environments are also highlighted in the study. [1]

B. Yu et al

An Inventive tactic for replicated data auditing in the cloud infrastructure using Identity-Based Cryptography (IBC) and dynamic auditing techniques are presented in this paper. Suggested strategy optimizes a dynamic auditing protocol to unravel any unauthorized changes are fabricated to replicated data stored in that cloud. The protocol also enables efficient auditing of large-scale cloud infrastructure by minimizing communication overhead and by reducing the numbers of audit requests. The authors simulation experiments and a performance evaluation to illustrate the effectiveness and efficiency of the presented scheme. Finally, the paper provides us with a useful contribution towards guaranteeing data security and data integrity in cloud computing environments. [2]

C. Wang, H et al

For public cloud storage auditing, the study suggests a blockchain-based fair payment smart contract to guarantee the objectivity and fairness of the auditing procedure. The suggested system automates the payment procedure and ensures that auditors are paid properly for their job. The authors present a brand-new method for auditing public cloud storage that combines Merkle tree and homomorphic encryption. By minimising the quantity of data exchanged during the auditing process and enhancing the privacy of user data, the suggested approach enables effective auditing. Through simulated tests, the authors assess the suggested scheme's performance and show how efficient and successful it is. Overall, the research makes a significant contribution to assuring fairness and data integrity in public cloud storage auditing. [3]

D. Hema et al

The Distributed Storage Hash Algorithm or DSHA proposed for file-based data matching in cloud computing settings. Because of the issue with repetitive files, the author signifies the issue of redundant storage in cloud

systems and advances DSHA as a resolution. Duplicate files can be caught and eliminated using DSHA's hash feature, which curates an unusual identifier for every file. A widespread storage architecture is also a key for suggested algorithm, ensuring effective and reliable data storage across a number of cloud servers. Through a series of experimentation and stimulations, the author's assessment of DSHA's and comparison to other replication methods is already in work. The results portray that DSHA defeats the pre-existing approaches in the speed of processing and storage efficacy. The research leads to a general useful contribution in enhancement of the working and reliability of data storage in cloud computing systems. [4]

E. Pooranian et al

The paper "LEVER: Secure De-duplicated Cloud Storage with Encrypted Two-Party Interactions in Cyber-Physical Systems" presents a new system called LEVER, which is a secure cloud storage solution. The system utilizes de-duplication techniques to reduce storage space, while ensuring data integrity and confidentiality. To certify the safety of client data in encrypted format against unauthorized access, LEVER implements an innovative two-party interaction protocol between the client and the cloud storage provider. This protocol also enables the client to verify that the cloud storage provider has executed its storage operations correctly. In order to assess LEVER's security and effectiveness, authors conducted experiments and simulation using real-world datasets. This resulted that LEVER is an effective and reliable cloud storage solution capable of resisting various attacks, including replay attacks, man-in-the-middle attacks, and data theft. In overall, the paper makes a remarkable contribution to the concern of dependable cloud storage, especially regarding cyber-physical systems, where data integrity and confidentiality are very crucial. [5]

F. Zhou et al

The paper proposes a similarity-aware encrypted deduplication scheme with flexible access control in cloud computing environments. The proposed scheme aims to achieve efficient and secure data deduplication while preserving data privacy and enabling flexible access control. The authors introduce a novel approach that utilizes the similarity of data blocks to improve the efficiency of deduplication while preserving data confidentiality through encryption. The scheme also allows for flexible access control, enabling users to control the sharing of their data with different entities. The authors evaluate the performance of the proposed scheme through simulation experiments, demonstrating its efficiency and effectiveness. The results show that the proposed scheme outperforms existing deduplication schemes in terms of storage efficiency and processing time while preserving data confidentiality and enabling flexible access control. Overall, the paper presents a valuable contribution towards improving the efficiency and security of data storage in cloud computing environments. [6]

G. Li et al

The paper proposes a proof of ownership based client-side encrypted deduplication (CSED) proposition for cloud storage. This method enables successful and guarded data deduplication while pointing the data privacy breaches and ensuring client-phase management. The authors suggest a progressive approach that entails use of proof of ownership so the clients could exercise their ownership that they possess over the data with the disclosure of data itself, managing data privacy. The plan incorporates an innovative client-side deduplication approach that enables data deduplication to be safe and fruitful. Using simulated tests, assessment of the performance of the presented plan is done, and it shows the order and success of this method. The results indicate how the said method may outshine many pre-existing deduplication techniques in terms of speed of data processing and storage efficiency of the data, while still directing to safeguarding the data privacy and ensuring client-phase authority on it. In conclusion, the paper shows a remarkable contribution in improvising the efficacy and dependability of data storage in cloud computing environments while prevent data breaches and partaking client-phase control. [7]

III. CONCLUSIONS

It is our common knowledge that cloud computing facilitates businesses to store more data on remote servers than on local ones, doing away with the need for ongoing server maintenance and other computing difficulties. Organizations can generate a large number of data files and achieve data accessibility without the need for local storage thanks to remote data storage, which enables authorized consumers to access data from geo-located nodes or nodal centers.

This research analysed and contrasted several potential strategies for data replication cloud systems. It was then revealed how the articles that had been examined had been divided into three categories: replication auditing, replication management, and replication de-duplication. Some problems with data replication were taken into

consideration while evaluating various systems based on crucial elements including the used type, place, class, evaluation tools, and their advantages and limitations.
Security, multi-environment replication, replication placement, replication migration, replication schemes, and fault tolerance, are only a few of the difficulties that arise throughout the data replicating process in cloud systems. These issues were taken into account in the existing gap and discussed to provide a summary.

REFERENCES

[1] Dwivedi, A.K., Kumar, N., Pathela, M.: Distributed and lazy auditing of outsourced data. In: International conference on distributed computing and internet technology (pp. 364–379). Springer, Cham (2020)
[2] Yu, H., Cai, Y., Sinnott, R.O., Yang, Z.: ID-based dynamic replicated data auditing for the cloud. Concurr. Comput. 31(11), e5051 (2019)
[3] Wang, H., Qin, H., Zhao, M., Wei, X., Shen, H., Susilo, W.: Blockchain-based fair payment smart contract for public cloud storage auditing. Inf. Sci. 519, 348–362 (2020)
[4] Hema, S., Kangaiammal, A. (2019) Distributed storage hash algorithm (DSHA) for file-based deduplication in cloud computing. In: International conference on computer networks and inventive communication technologies (pp. 572–581). Springer, Cham (2019)
[5] Pooranian, Z., Shojafar, M., Garg, S., Taheri, R., Tafazolli, R.: LEVER: secure deduplicated cloud storage with encrypted two-party interactions in cyber-physical systems. IEEE Trans. Ind. Inf. (2020).
[6] Zhou, Y., Feng, D., Hua, Y., Xia, W., Fu, M., Huang, F., Zhang, Y.: A similarity-aware encrypted deduplication scheme with flexible access control in the cloud. Futur. Gener. Comput. Syst. 84, 177–189 (2018)
[7] Li, S., Xu, C., Zhang, Y.: CSED: client-side encrypted deduplication scheme based on proofs of ownership for cloud storage. J. Inf. Secur. Appl. 46, 250–258 (2019)
[8] S. Slimani, A. Hamou-Lhadj, and J. Mabrouki, "Metadata replications and request balancing for cost-effective data replication in cloud systems," Cluster Computing, vol. 24, no. 3, pp. 2545–2579, Sep. 2021, doi: 10.1007/s10586-021-03283-7.