JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Ensuring Data Security and Storage Efficiency in the Cloud with Deduplication

Mrs. SK. Sofia[1] | T. Sai Sandeep[2] | Y. Chandra Sekhar[3] | SK. Tazeem Ali Shah[4] | Y. Muni Chandu [5] | P. Vivek[6]

[1]Assistant Professor, [2,3,4,5,6] B.Tech Scholars
Narayana Engineering College, Department of Computer Science and Engineering, Nellore, Andhra Pradesh, India

Abstract: Information redundancies in cloud capacity can be effectively eliminated through information deduplication, which also lowers client transfer speed requirements. Nevertheless, the majority of previous schemes that rely on the provision of support from a trustworthy key server (KS) are helpless and constrained because of the difficulties they face in terms of data discovery, weak protection against attacks, enormous computing cost, etc. Specifically, the entire framework fails (single-point-of-failure) if the trustworthy KS fails. In this work, we present a Secure and Productive data Deduplication scheme (SED) in a Cloud capacity architecture that provides global services through cooperation with several clouds. Furthermore, SED promotes active information sharing and upgrading without the offer of support from the reliable KS. Besides, SED can overcome the single-point-of-failure that commonly happens in the classic cloud capacity framework. Concurring to the hypothetical investigations, our SED guarantees the semantic security in the irregular prophet demonstrate and has solid anti-attack capacity such as the brute-force assault resistance and the conniving assault resistance. Other than, SED can viably kill information redundancies with moo computational complexity and communication and capacity overhead. The effectiveness and usefulness of SED progresses the convenience in client-side.

Key words: Cloud Storage, Data Deduplication, Data Security, Storage Efficiency, Redundancy Elimination, Encryption, Data Integrity, Cloud Computing, Data Management, Storage Optimization.

1. INTRODUCTION

A "pay-as-you-go" model is used to provide large-scale information capacity and benefits access through cloud capacity. However, some of the extra data in the cloud has really wasted and consumed capacity resources. An effective technique to recognize and eliminate redundant information is information deduplication. After that, the data is moved and stored, effectively creating a single replica. In this sense, the information deduplication innovation can reduce the client-side requirement for transmission capacity and significantly increase the server-side's effectiveness in terms of space usage. It is being widely used in many cloud computing services to increase customer engagement and conserve resources.

The traditional information deduplication diagram and its modifications, in which the system consists of clients, a cloud capacity supplier (CSP), and a key server (KS), ensure security based on the reliable KS. Even worse, these traditional plans might still work because of "platform lock-in" and single-point-of-failure problems. The cloud capacity framework malfunctions and information outsourcing protocols cannot be carried out if the trustworthy KS is insufficient. Recently, a collaborative cloud computing framework known as the Cloud computing framework was developed to fully understand the aforementioned problems. Cloud computing is organized around customers and many cloud service providers offering various services. In order to provide computing services, clients can communicate with any of these clouds as they work together without the need for a trusted key. It is evident that cloud computing can provide efficient cross-cloud services and meet the demands of globally acceptable cloud services through multilateral cooperation across various clouds. It can also be integrated into the decentralized structure. Cloud computing has garnered significant attention from academia and business.

Large-scale data breaches that affect billions of individual records are not uncommon. In order to ensure information secrecy in cloud capacity frameworks, the outsourced data are typically asked to be jumbled. That being said, finding and removing the replicated duplicates in the ciphertext region is a challenge. Because the ciphertexts of the same plaintext that different clients have scrambled using traditional encryption algorithms differ from one another. It has been suggested that concurrent encryption (CE) and its variants be used to really do jumbled deduplication. Information is jumbled in these plans using keys that are derived from the information itself. In other words, the storyline is helpless and the mysterious key is deterministic.

2. LITERATURE STUDY

Many deduplication strategies have currently been put forth. One of the main methods for ensuring information security in deduplication is convergent encryption, which protects outsourced data from malicious and unreliable CSPs. A primitive known as message-locked encryption (MLE) conspire was formalized by Bellare et al. At that point, some modifications based on Bellare's work were suggested. Regardless, these MLE-based schemes were up against a lot of potential risks because the keys used to jumble data are extracted from the data itself. For bounded message disseminations, Abadi et al. designed a fully randomized plot and a deterministic scrambled conspiracy based on the non-degenerate fluently computable bilinear outline. Li and colleagues demonstrated a coordinated effort to achieve robust key management in deduplication.

Jiang et al then seemed to have a safe plot for deduplication using the randomized tag. Nevertheless, they did not outline the requirements for information revamping. After then, Hur et al. examined the administration of energetic proprietorships for safe deduplication, and each client was required to keep the keys discovered along a double tree of all the keys in order to achieve deduplication. Based on mystery sharing ideas, Li et al. presented secure deduplication with key administration. Following that, a decentralized server-aided encryption for deduplication was planned by Shin et al. However, it necessitated a lot of intuition between the clients and KS, which provided opportunities for the attackers to obtain important information from the exchange. Secure deduplication for multiserver-aided was suggested by Miao et al.

Quick and efficient content-defined chunking was described by Xia et al. to achieve fine-grained information deduplication. In order to reduce layer reestablish overhead and deduplicate layers for space sparing, Zhao et al. presented a deduplication scheme based on the Docker registry concept. Analysts have recently focused on the use of a few emerging ideas, such blockchain, for deduplication in various applications.

3. PROPOSED SYSTEM

By making information deduplication in cloud capacity faster, more secure, and easier to manage, the suggested framework advances the process. This system uses sophisticated algorithms to quickly identify and remove duplicate data, saving storage space and lowering expenses. Additionally, it includes updated encryption protocols to ensure that all data is safe and accessible only to authorized users. The system is designed to manage large datasets efficiently without slowing down, and it includes tools that are easy to use for easier management and observation.

Additionally, this framework overcomes the shortcomings of the current frameworks by combining far superior error detection and correction tools to prevent data anomalies and mishaps. It uses an adaptive architecture to grow with the amount of information that is available, ensuring continued efficacy and performance. The suggested architecture now offers a more stable and safe way to manage data in the cloud thanks to these improvements. Furthermore, the proposed work will include the implementation of a feedback mechanism to gather input from students on their experiences with the grievance handling system. This feedback will be used to continuously refine and enhance the system, ensuring that it remains responsive to the needs and preferences of its users.

To sum up, this suggested method enhances cloud data deduplication by making it quicker, safer, and simpler to administer. While data is protected by new encryption standards, it uses sophisticated algorithms to remove duplicate data, saving space and money. Large datasets are handled by the system with efficiency, and it comes with easy-to-use tools for management. Moreover, it has an adaptive design to sustain performance as data accumulates and better mistake detection and correction to avert data problems. A feedback mechanism will also be put in place so that the system can be improved over time depending on user experiences. In general, this framework provides cloud data management in a more dependable and safe manner.

4. METHODOLOGY

Our methodology focuses on using multiple important strategies to improve the security and efficiency of data deduplication in cloud storage. In order to maximize resource use and enable scalable access to cloud capacity while reducing superfluous data storage, we first adopt a pay-as-you-go strategy. By identifying and removing redundant data using sophisticated deduplication techniques, only unique information is kept, saving money and storage space.

Security is crucial, and it is attained by using strong encryption techniques. To preserve privacy, data is encrypted at the client-side before being uploaded to the cloud. Access control techniques like role-based access control and multi-factor authentication are used to prevent unwanted access, while secure key management guarantees that encryption keys are managed efficiently.

We use a collaborative cloud computing system to solve conventional vulnerabilities related to centralized key servers. This method removes reliance on a single trustworthy party by enabling cooperation between several cloud service providers. By allowing customers to use services from any supplier inside the framework, flexibility and resistance to single points of failure are increased.

Information privacy must be maintained in light of the frequency of data breaches. Our approach incorporates related variations of convergent encryption (CE) to provide secure deduplication even in encrypted data. CE ensures data privacy while enabling efficient deduplication by generating consistent ciphertexts for identical plaintexts across multiple users using data-derived keys.

Reliable and accurate information retrieval is ensured by these technologies, which continuously monitor and correct any irregularities in stored data. In addition, easy-to-use management tools are incorporated for monitoring and administration, and a feedback mechanism is in place to collect user input and improve system performance.

Our methodology, in its entirety, provides a thorough approach to safe and effective data deduplication in cloud storage. We guarantee optimal resource utilization, data security, and system dependability in contemporary cloud computing environments by combining sophisticated deduplication algorithms, robust encryption, collaborative cloud frameworks management.

5. ANALYSIS

Data deduplication in cloud storage is crucial for optimizing resource utilization and enhancing data security. By identifying and eliminating redundant data, organizations can significantly reduce storage costs and improve overall efficiency. One of the primary benefits of deduplication is its ability to minimize storage footprint. Instead of storing multiple copies of identical data, deduplication ensures that only one instance is retained, thereby conserving storage space. This is particularly advantageous in cloud environments where storage costs can escalate with increasing data volumes.

Another crucial component of data deduplication in cloud storage is security. It is crucial to guarantee the integrity and confidentiality of data, particularly when working with sensitive data. Robust encryption techniques are essential for deduplication procedures in order to safeguard data while it's in transit and at rest. In order to guarantee that only encrypted data is sent to the cloud, client-side encryption is usually used to encrypt data before it leaves the user's premises. Encryption keys must be kept safe, and secure key management procedures must stop unwanted access to private data.

Reducing storage needs is just one aspect of efficient data deduplication; other factors include processing times and data transport optimization. Sophisticated deduplication algorithms are engineered to effectively detect redundant data segments in extensive datasets. Hash-based fingerprinting and content-defined chunking are two techniques that allow for quick comparison and removal of unnecessary material. By lowering the amount of data that has to be processed and transported, this not only conserves storage space but also improves data retrieval rates.

Another important factor to take into account in cloud storage setups is scalability. The deduplication system has to scale smoothly to accommodate growing workloads without sacrificing performance as data quantities increase. Parallel computing, distributed processing, and scalable architectures are frequently used to make sure deduplication procedures are effective and adaptable to organizational requirements.

In cloud storage, data deduplication conserves space and maintains data security. Organizations may save storage expenses and improve system performance by identifying and eliminating duplicate files. This is particularly helpful in cloud environments where large data storage might be costly. By ensuring that only one duplicate of the same data is retained, deduplication guarantees that just one copy is kept, freeing up space for other purposes.

Furthermore, deduplication systems' dependability is crucial. Cloud storage providers are essential to organizations in order to sustain high service availability and data durability. Fault-tolerant designs and redundancy in storage help reduce the risk of hardware failures and network outages, guaranteeing uninterrupted access to deduplicated data. System dependability is further increased by regular data integrity checks and error correction procedures, which identify and fix any discrepancies in recorded data.

In the context of data deduplication, security is crucial. Data security and privacy are crucial, particularly when handling sensitive information. To do this, data is safeguarded throughout transmission to and storage in the cloud using robust encryption techniques. Only encrypted data is transferred since it is typically encrypted on the user's end before being sent to the cloud. Maintaining safe encryption key management is also crucial to preventing unwanted access to the data.

Another important factor is scalability. The deduplication system must be able to handle increasing amounts of data without experiencing any lag. The system scales effectively with the use of methods like distributed processing and parallel computing. This indicates that even with a rise in data volume, the deduplication process stays quick and efficient, guaranteeing the system can handle the expanding demands of the company.

To sum up, safe and effective data deduplication in cloud storage is necessary for cost savings, improved security, data management optimization, and regulatory compliance. Organizations may efficiently manage and safeguard their data assets in cloud settings by using strong encryption, scalable architectures, effective deduplication methods, and stringent data integrity checks. Ongoing developments in deduplication technology will be critical to enhancing data security, dependability, and efficiency for enterprises of all kinds as cloud computing continues to develop.

6. CONCLUSION

In conclusion, no matter how many people upload a file, cloud storage with secure and effective data deduplication ensures that only one duplicate of the content is kept. This is a powerful way to manage data. By removing redundant files, this procedure lowers the amount of storage space needed, improving the effectiveness and economy of the system.

The increased security that this system provides is one of its main advantages. Potential security concerns are decreased by eliminating the requirement for a trusted Key Server (KS) by utilizing the Secure and Efficient Deduplication (SED) technique. This technique guarantees the confidentiality and security of data during the deduplication process, in conjunction with convergent encryption and access control. An additional degree of security is added when users must utilize encryption and decryption keys in order to access their data.

All file transactions, including who submitted the file, what it is called, and actions like uploading, searching, and downloading, are carefully logged by the system. Every action's date and time are also recorded. Transparency is provided by this thorough tracking, which also facilitates improved file management. It makes sure the system is utilized effectively and securely by enabling users and administrators to keep an eye on how data is used and accessed.

All things considered, cloud storage's safe and effective data deduplication provides a solid option for handling massive data volumes. It guarantees thorough file transaction tracking, improves security using cutting-edge encryption techniques, and lowers storage expenses by getting rid of duplicate files. This system offers a safe and effective method of managing and accessing data in the cloud, in addition to saving space.

In conclusion, by fusing efficiency and security, this method of cloud storage tackles the main issues surrounding data management. It makes sure that information is kept secure from unwanted access and saved to take full advantage of the space that is available. Maintaining comprehensive documentation of file transfers and utilization provides significant insights for enhancing both system efficiency and user satisfaction. Because of this, cloud storage becomes a more dependable, safe, and affordable option for people and businesses alike.

End User must register with details like name, email, password, date of birth, mobile number, and address. After click on register. If the user already registered then they can click on login. After login successful the End User can perform their activities.

After login the End User menu you can see the options like search files it means you can search which file you want and view files in cloud and you can download files from the cloud by requesting the keys from the data owner and from the cloud after request successful you can download file from the cloud.

After login successful the Data Owner menu shows these options like upload means you can upload a text file into cloud, update the file means you can change the contents of the file , delete the file you have unnecessary information, you can view files whatever you have uploaded into the cloud and there is secret key permissions like if any user requested to download a file you can give access request.

After login the Cloud it shows the above menu options like view end users details and authorize and view data owner details and also you can authorize it means authorize people can have access. And also you can give the permissions for the end users to download the files from the cloud.

In the above cloud login page you can see the uploaded files into the cloud with their addresses it means the file name, data owner name and date and time of the uploaded file.

Performing the deduplication means This is done by comparing the contents of files and keeping just one copy. If a new file matches an existing one, then it simply shows the file is already uploaded with contents and it is duplicated.

We measure how often a file is used. Every time someone looks for the file, search count will be increased. Every time someone opens or views the file, The count will be updated. Every time someone downloads the file, count the count will be updated in chart.

The above figure shows that how much time it takes to upload the file into cloud. It shows the time in (milli seconds) for every file which are uploaded in cloud. For Example: java2, java3, TEMP…etc files are uploaded in cloud and it shows the time in the chart.

7. REFERENCES

[1] P. Christen, "A review of indexing methods for deduplication and scalable record linkage," IEEE Transactions on Knowledge and Data Engineering, vol. 24, no. 9, pp. 1537–1555, 2012.

[2] G. Jia, G. Han, J. J. P. C. Rodrigues, J. Lloret, andW. Li, "Sync memory partitioning and deduplication to enhance cloud computing performance," IEEE Transactions on Cloud Computing, vol. 7, no. 2, pp. 357–368, 2019.

[3] W. Xia, X. Zou, H. Jiang, Y. Zhou, C. Liu, D. Feng, Y. Hua, Y. Hu, and Y. Zhang, "The development of quick content-defined chunking for storage systems based on data deduplication," IEEE Transactions on Parallel and Distributed Systems, vol. 31, no. 9, pp. 2017–2031, 2020.

[4] J. Li, J. Li, D. Xie, and Z. Cai, "Cloud deduplication and secure auditing," IEEE Transactions on Computers, vol. 65, no. 8, pp. 2386–2396, 2016.

[5] L. Liu, Y. Zhang, and X. Li, "Keyd: Integrating identity-based broadcast encryption with secure key deduplication," IEEE Transactions on Cloud Computing, pp. 1–1, 2018.

[6] J. Ni, K. Zhang, Y. Yu, X. Lin, and X. S. Shen, "By using fog computing to provide job allocation and safe deduplication for mobile crowdsensing," IEEE Transactions on Dependable and Secure Computing, pp. 1–1, 2018.