IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 8, August 2024

DOI: 10.17148/IJARCCE.2024.13838

A Survey on the Ranking and Deduplication Strategies for Cloud Storage Monitoring

Jayashree G M[1], A M Prasad[2]

Student, M.Tech, Department of CSE, Dayananda Sagar College of Engineering, Bengaluru, India[1]

Assistant Professor, Department of CSE, Dayananda Sagar College of Engineering, Bengaluru, India[2]

Abstract: Cloud computing has become an essential technology in today's cutthroat business environment. It cannot be avoided. Consumers can easily access recycled data, which reduces overall access times and results in high-quality goods, but it still requires more store space, leading to increased storage costs. Cloud Storage Monitoring (CSM) is an IaaS storage utilisation tracking system that uses multi-tenancy data to watch and evaluate access patterns in order to identify data size, frequency of access, future usage, and data recycling in the cloud. Every document receives a ranking, which also predicts future patterns of access. To free up more space for later usage, this gives customers dashboard options where they may decide whether to upload, download, or keep stuff and remove unnecessary files stored in the cloud. Java technology, encryption, decryption keys, deduplication ideas, compression, and decompression methods are all used in the implementation of this project. The frequency distribution algorithm provides a greater storage improvement than conventional techniques. Additionally, it forecasts when archives will become available in the future and upholds the access balance.

Keywords: cloud computing, cloud storage, data deduplication

I. INTRODUCTION

The field of cloud computing has evolved over the last 10 years, which has resulted in the outsourcing of data storage to cloud services. Because it lessens the work required to maintain and handle massive data, this trend is advantageous. However, there are concerns about attaining data de-duplication in the cloud while still preserving storage efficiency due to the unreliability of the outsourced storage cloud. In this paper, we examine the challenge of concurrently de-duplicating the cloud-based data and preserving auditing integrity. Our specific objective is to detect duplicate data in the cloud, and we suggested a secure method to achieve this. [1] Numerous deduplication systems have been proposed, each with specific deduplication strategies—such as file-level or block-level deduplication—chosen according to the amount of deduplication being done. These systems fall into two categories: inline deduplication systems and post- process deduplication systems. While deduplication is done after storage in the latter, it is done before in the former. In particular, the growing interest that data deduplication techniques are gaining in academic and industry circles is noteworthy, especially considering the growing acceptance of cloud storage. In the information-driven world of today, this method is now necessary to handle the increasing amount of data.

II. RELATED WORK

IoT data processing, storage, and management are all supported by cloud computing; nevertheless, storage efficiency is decreased when encrypted duplicate data is stored. Current deduplication systems lack flexibility for safe data access control and contain security flaws. a secure access control system that uses attribute-based encryption (ABE) to deduplicate encrypted data. Performance assessments show the scheme's scalability, efficacy, and efficiency, which qualify it for real-world implementation[2]. a secure deduplication method based on proof of power (PoW) and essential sharing.

The original uploader sends the material to authorised drug dealers after encrypting it with a randomly selected key (CK). Only previously for indistinguishable data is the key stored. With the use of a thickness policy and deduplication checks on plaintexts, our technique can cypher indistinguishable data with a small number of possessors. In our security paradigm, this method is more efficient and safe. Future research will focus on improving block-position deduplication efficacy and decreasing reliance on the Index Garçon (IS) in order to mitigate failures brought on by data block border disruption[3].

DOI: 10.17148/IJARCCE.2024.13838

The paper presents a novel deduplication approach for encrypted cloud storage with the goal of increasing efficiency and security. By employing randomised convergent encryption and safe ownership group key distribution, it addresses issues with existing methods such as preventing data leaks and handling dynamic ownership changes. This strategy significantly improves deduplication for encrypted data stored in cloud storage by guaranteeing low computational overhead and data access control[4]. The Cloud Storage Monitoring (CSM) system ranks lines based on frequency and fashionability, increasing the amount of available storage space in IaaS environments. Train rankings are assessed using a vaticination algorithm, and lines are either relocated or archived as a result. Deduplication is used to eliminate identical lines. Simulated experiments utilising lines ranging from 0.11 MB to 1.00 MB demonstrated a mean decrease of 10.91 in storage space along with 3.8 GB of DUD. In the future, the CSM system can be expanded to PaaS and SaaS environments, offering an efficient data storage solution[5].

In order to improve efficiency and preserve redundancy for failure tolerance, the research presents a dynamic deduplication strategy for Pall Storehouse. In order to improve storage efficacy and reliability, it overcomes the drawbacks of static deduplication techniques by steadfastly adapting to shifting data operations and access patterns in dynamic environments[6]. Using software-defined storage (SDS) technologies, this work built a heterogeneous cloud storage system to achieve load balance through uniform data delivery. The platform includes three types of open-source SDS software and simulates several public cloud storage scenarios. Due to hardware constraints, it currently exclusively employs open-source back-end storage. Future advancements will include hybrid cloud environments and enhancements to security, availability, and user interface components[7].

This study looks at data deduplication techniques for cloud storage, highlighting the ways in which they might improve storage efficiency and reduce costs. Among other data types, it covers many deduplication taxonomies for photo, video, and text data. The study's conclusion offers recommendations for enhancing deduplication methods while outlining challenges and potential paths forward[8]. By employing deduplication and train access pattern ranking, the suggested storehouse optimisation system (SoS) lowers the pall storehouse application. Trials on ten lines with sizes ranging from 0.10 MB to 1.00 MB revealed 11.91% smaller storehouses and 4.28% lower bandwidth usage as compared to existing systems, which reduced IaaS expenses. With the fewest possible detentions, the SoS system permits efficient access. Future developments will focus on handling unshaped data in the cloud and security with translated lines[9].

Cloud computing resource management issues brought on by bus-scaling can be resolved by suggesting an autonomous vaccination suite. By selecting appropriate time-series vaticination algorithms based on incoming workload patterns, it seeks to improve vaticination delicacy. Following an empirical validation, a theoretical discussion of threat minimisation principles and their benefits for vaticination delicacy is presented.

A tone-adaptive vaticination suite that automatically chooses the fashionable vaticination algorithm for the workload was designed as a result of the findings[10]. a dynamic data replication strategy for cloud computing that maximises replica placement and selection to speed up data access. By enabling users to create, manage, and update copies, the system increases the accessibility of data. There are two primary stages to it: locating and creating replicas using a catalogue and index, and ascertaining the availability of destination space. When the system is used in the Eucalyptus cloud environment, it decreases delay times, access fees, and shared bandwidth usage while also increasing resource availability. Findings demonstrate that the suggested algorithm performs more accessible than current techniques[11].

III. CLOUD STORAGE SCHEMA

A cloud solution structure that includes both "on-premises" and "cloud storage" resources is provided by the cloud architecture. Among the components of the software are geolocation services, as well as middleware. having an accessible parameter externally, and that connects the interfaces, as seen in Figure 1.

The front end is used by the client. It possesses the client-side interfaces and programs required to access platforms for cloud computing. Web servers, like the Internet Tablets, thin and fat clients, Internet Explorer, Firefox, and Chrome, and The front end is made up of mobile devices. The back end is used by the service provider. It is in charge of all the resources required to provide cloud computing services. An enormous amount of servers, virtual computers, data storage, and traffic deployment strategies, management approaches, and all security precautions are incorporated.

DOI: 10.17148/IJARCCE.2024.13838

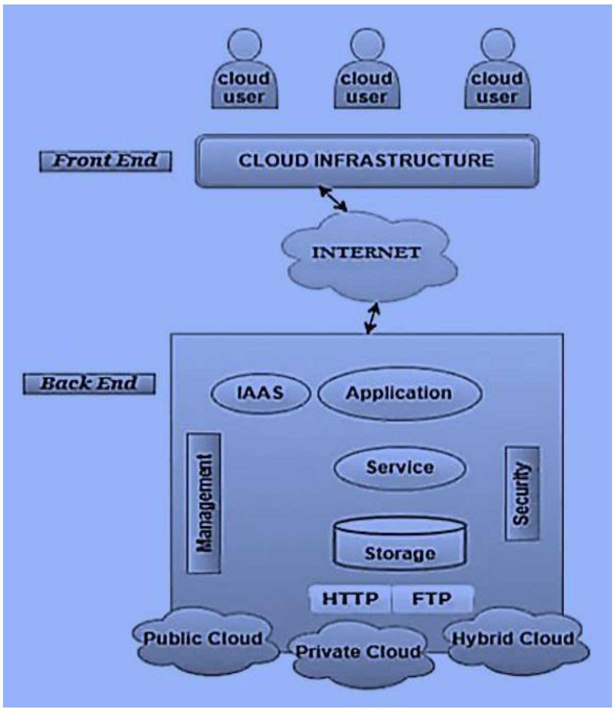

Fig. 1 Cloud Storage Architecture

A. Infrastructure

This software layer is offered by cloud storage providers as a way to link various cloud users together so they can use the Internet to access cloud IaaS storage services. This layer uses a single authentication and authorisation process for sign-on.Techniques for confirming and authenticating individuals with their credentials. Java AWT/Swing will be utilised in our project for the tables, graphs, and forms that users interact with the cloud storage monitoring system through an interface. A variety of components are available in rich GUI components to help designers create user interfaces that are responsive and easy to use and understand. Permits significant changes to GUI elements to satisfy certain user requirements.

B. Management and Security

The management layer is responsible for managing data movement throughout the network in addition to monitoring and validating the data of individual cloud users with regard to data architecture and operations like data synchronisation, data splitting, and storage, taking into account the distribution of content among storage sites. Infrastructure, as well as other backend security challenges. network and to ensure replication, backup, and consistency of applications, services, runtime clouds, storage, and data. Controlled and arranged with help from management. Basic Products for storage and storage virtualisation make up the majority of this.

Quantity of cloud storage available. Storage virtualisation maps storage devices with varying storage volumes and heterogeneous hardware. Using a single allotment of storage space, hence enabling the creation of a shared platform . Storage technology that has been virtualised offers applications availability, security, and scalability. Networks, virtual machines, and virtual hardware produced by IaaS are a few examples. Users can operate on any OS or application with the help of Infrastructure as a Service (IaaS), which offers customised infrastructure. Security protocols and monitoring systems, like MySQL 5.0 servers for storage management, are part of cloud computing.

C. Frameworks and Operational Models

Using protocols like FTP and HTTP for remote server uploading and downloading, as well as size and response times, the study looks at client and cloud server security. Large files and entire directories can be transferred over networks using the File Transfer Protocol (FTP). Passwords and usernames are needed for both passive and active modes. Most access to and transfers of websites and web-connected services occur over HTTP, which is used to enable communication between clients and by providing access to cloud services and APIs for cloud-based apps. Furthermore, in internet programs for downloading and uploading files to the transport of data. Each request a client submits to a server is handled as a separate transaction, such as JSON, XML, Videos, images, and HTML.

D. The Complexity of Integrating Cloud Storage

Companies are increasingly adopting cloud storage services for data management due to their ease of use, large volume, and quick response times. However, challenges remain in best practices and optimizing storage features, particularly for enterprises with significant on-site storage needs. Compared to internal storage, public cloud security is less secure, and IT managers are frequently uneasy managing sensitive data in public settings. According to 62% of Cloudian polls, security concerns are a frequent problem. IOPs and user storage capacity affect cloud storage costs. Costs can be decreased by optimising pay-as-you-grow and deduplication choices. Public clouds are used by most businesses to minimise storage expenses. Interoperability with hybrid cloud IaaS storage principles presents issues for many organisations, especially when it comes to important on-premise applications. Nearly 20% of businesses struggle with vendor lock-in, which makes data transfer expensive and difficult[12].

IV. IMPLICATIONS OF OUTCOME

A. System for Monitoring Cloud Storage(CMS)

A prediction- and rating-based system can be proposed with the following design, which aims to handle cloud storage duplication. Find out how frequently the access pattern occurs. Identify redundant files on cloud storage. Build a storage system that works well. Increase the efficiency of the system. Make searching more enjoyable. Avoid duplicate files in the future. To reduce cloud storage duplication, a prediction- and grade-based approach is advised. This establishes the access pattern's frequency. Provide a forecast for file access. Find the duplicate files stored in the cloud. builds a storage mechanism that works well. Increase system efficiency. Improve the way you search. Avoid more duplicate files. The CSM system ranks the files based on how popular and frequently they are accessed. This is the recommended study endeavour [13]. The CMS system generates a ranking dashboard to optimize storage capacity and availability by removing duplicate data and ensuring every file is available.

B. The Process

The CMS module can check for duplicate files and can create a link for uploads with file sizes less than 100 bytes. Files are arranged according to rank and encrypted using a strong algorithm. The module also compresses and stores files for storage, ensuring they are not used by multiple clients. The CMS sends upload requests, storage nodes save data, check memory consumption, respond, send content, delete files, and refresh memory after responses and download requests.

C. Public Cloud Services

Open-source public cloud services provide cost-effective, scalable, and flexible file storage solutions without licensing fees. They allow users to choose from various providers without being tied to a single provider, allowing customization to meet unique business requirements. Public availability of source code enhances security and facilitates community detection. Open-source solutions support sustainability, minimize carbon footprints, and maximize resource use.

D. Data Deduplication and Compression

Data de-duplication or single-instance techniques eliminate redundant data on cloud servers, reducing host space requirements. Three main forms include compression, single-instance storage , and comparing files to remove duplicates. Backups are crucial before de-duplication techniques, and compression is often used as a long-term technique. The ZIP compression standard can be utilized for proposal for compressing and decompressing data. Two techniques are used: compress, which compresses byte arrays, and decompress, which decomposes data from a byte array. The compressed data is then moved using less bandwidth and stored in cloud storage.

E. Public and Private Keys

Cloud storage systems use public and private keys for encryption and decryption, ensuring data integrity. Public keys establish digital signatures and authenticate senders. Key management enforces access control, while public and private key pairs support multi-tenant security. Public key encryption enables secure file sharing, teamwork, and effective key distribution. Automated key management systems reduce administrative overhead and improve data confidentiality. Public and private keys also enhance data protection standards.

F. Effective Storage

Depending on convenience, we can establish three to five servers or nodes with storage capacity ranging from two to five gigabytes using deduplication ranking and encryption/decryption algorithms or keys. This results in an effective storage system that is affordable and scalable, which is vital given that storage requirements are known to be expensive and security-required. Protocols and deployment models can be used to establish a connection to either a public or private cloud service, and uploaded data can be securely stored.

DOI: 10.17148/IJARCCE.2024.13838

V. CONCLUSION

In an IaaS-Cloud system, the Cloud Storage Monitoring (CSM) solution is recommended to increase storage space availability. The frequency of files is graded and quantified. The files are ranked according to their frequency and popularity. A prediction algorithm assesses the file's rating. Depending on their classification, the files are either transferred or archived. The de-duplication method eliminates duplicate files. The CSM method generates an average de-duplication by comparing the average reduction to the "without using CSM" technique. Consequently, the suggested CSM system offers a productive way to store data. This system can be improved further in the future to support further cloud computing series, such as SaaS and PaaS.

REFERENCES

[1]. J. Gantz, D. Reinsel, The digital universe in 2020: big data, bigger digital shadows, and biggest growth in the far east, http://www.emc.com/collateral/ analyst-reports/idc-the-digital-universein-2020.pdf, Dec. 2012.

[2]. Z. Yan, M. Wang, Y. Li and A. V. Vasilakos, “Encrypted Data Management with Deduplication in Cloud Computing,” in IEEE Cloud Computing, vol. 3, no. 2, pp. 28-35, Mar.- Apr. 2016, doi: 10.1109/MCC.2016.2

[3]. Liang Wang, Baocang Wang, Wei Song, Zhili Zhang, A key-sharing based secure deduplication scheme in cloud storage, Information Sciences, Volume 504, 2019, Pages 48-60, ISSN 0020-0255,https://doi.org/10.1016/j.ins.2019.07.058.

[4]. J. Hur, D. Koo, Y. Shin and K. Kang, “Secure Data Deduplication with Dynamic Ownership Management in Cloud Storage,” in IEEE Transactions on Knowledge and Data Engineering, vol. 28, no. 11, pp. 3113-3125, 1 Nov. 2016, doi: 10.1109/TKDE.2016.2580139.

[5]. Devarajan, A & Muthu T, Sudalai. (2019). Cloud Storage Monitoring System analyzing through File Access Pattern. 1-6. 10.1109/ICCIDS.2019.8862113..

[6]. W. Leesakul, P. Townend and J. Xu, "Dynamic Data Deduplication in Cloud Storage," 2014 IEEE 8th International Symposium on Service Oriented System Engineering, Oxford, UK, 2014, pp. 320-325, doi: 10.1109/SOSE.2014.46.

[7]. C. -T. Yang, S. -T. Chen, W. -H. Cheng, Y. -W. Chan and E. Kristiani, "A Heterogeneous Cloud Storage Platform With Uniform Data Distribution by Software-Defined Storage Technologies," in IEEE Access, vol. 7, pp. 147672-147682, 2019, doi: 10.1109/ACCESS.2019.2946962.

[8]. Kaur, R., Chana, I. & Bhattacharya, J. Data deduplication techniques for efficient cloud storage management: a systematic review. J Supercomput 74, 2035–2085 (2018). https://doi.org/10.1007/s11227-017-2210-8 A System for Monitoring the Cloud Storage with Deduplication Ranking Dept. of CSE, DSCE 2023-24 Page 21

[9]. A. Augustus Devarajan and T. Sudalai Muthu, "Enhanced Storage optimization System (SoS) for IaaS Cloud Storage," 2020 Fourth International Conference on Inventive Systems and Control (ICISC), Coimbatore, India, 2020, pp. 756-760, doi: 10.1109/ICISC47916.2020.9171182.

[10]. An Autonomic Prediction Suite For Cloud Resource Provisioning, Nikravesh, A.Y., Ajila, S.A. & Lung, CH.. J Cloud Comp 6, 3 (2017). https://doi.org/10.1186/s13677-017-0073-4

[11]. Rajalakshmi, A. & Vijayakumar, D. & Srinivasagan, K. (2014). An improved dynamic data replica selection and placement in cloud. 2014 International Conference on Recent Trends in Information Technology, ICRTIT 2014. 1-6. 10.1109/ICRTIT.2014.6996180.

[12]. Sathish N, Ranjana P, “Secure remote access fleet entry management system using UHF band RFID”, Advances in Intelligent Systems and Computing, vol. 216, pp. 141-149, 2014.

[13]. T. S. Muthu, R. Vadivel, A. Ramesh and G. Vasanth, "A novel protocol for secure data storage in Data Grid environment," Trendz in Information Sciences & Computing (TISC2010), Chennai, 2010, pp. 125-130.