iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 **Issue:** VI **Month of publication:** June 2025

DOI: https://doi.org/10.22214/ijraset.2025.71988

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue VI June 2025- Available at www.ijraset.com

Reducing Latency and Storage Costs in Cloud Applications Through Advanced Data Management

Janakiraman S[1], Narmadha K[2]
[1]*Assistant Professor,* [2]*II MCA, Department of Master of Computer Applications, Er. Perumal Manimekalai College of Engineering, Hosur*

Abstract: ***Junk files, including outdated backups, redundant document versions, and orphaned objects, accumulate in cloud storage, leading to inefficiencies in data retrieval, increased latency, and higher storage costs. As cloud applications grow in scale, managing and optimizing storage resources becomes crucial for maintaining performance and reducing operational overhead. The problem of unnecessary files taking up valuable space is especially critical in cloud environments where efficient resource management is essential for smooth operations. This project proposes a solution to optimize cloud data management by integrating automated cleanup, structured data lifecycle management, and advanced deduplication techniques. Regex algorithms will drive the cleanup process, identifying and eliminating obsolete files regularly to ensure that only relevant data is stored. Additionally, the Data Life Cycle Guard Scheme provides a framework for managing data according to predefined compliance rules, improving overall data governance and integrity. These measures aim to streamline data processes and maintain the efficiency of cloud applications. Fuzzy Matching techniques will further enhance the deduplication process, improving accuracy in identifying and removing duplicate files, thus optimizing storage space. By automating the identification of unnecessary files and improving data lifecycle management, this system helps reduce storage costs, minimize latency, and ensure that cloud applications run more efficiently. The solution is designed to set new standards in cloud data management, optimizing resource utilization and ensuring long-term sustainability for cloud-based environments***

I. INTRODUCTION

An enterprise cloud brings together private, public, and distributed clouds in a unified IT environment. It offers a centralized control point. From there, businesses can manage enterprise cloud applications and infrastructure in any cloud. An enterprise cloud provides businesses with a seamless, consistent, and high-performance experience. Enterprise cloud computing is the process of using virtualized IT resources such as external servers, processing power, data storage capacity, databases, developer tools, and networking infrastructure by companies and organizations. Enterprise cloud solutions help organizations optimize their operations and cut costs.The cloud computing framework provides an optimal environment for faster, safer and cheaper delivery of IT services within an enterprise. The enterprise architecture and cloud computing model form the skeleton and blueprint that gives form to the digital side of your organization

II. EXISTING ALGORITHM

In a traditional system context, the management of junk files, outdated backups, and orphaned objects typically involves manual intervention and relies on established processes within an organization's on-premises infrastructure. Here are some characteristics and approaches associated with handling these issues in a traditional system:

1) Manual Cleanup Processes: Addressing junk files, outdated backups, and orphaned objects often requires manual cleanup processes. IT administrators or system operators manually identify and remove unnecessary files periodically.
2) Scheduled Maintenance: Cleanup activities are scheduled as part of routine system maintenance. Regular intervals, such as weekly or monthly, may be designated for reviewing and purging unwanted files from the system.
3) Backup Rotation Policies: Traditional systems typically follow backup rotation policies, where older backups are systematically replaced or deleted to make room for new backups. This process helps manage storage space and ensures data recoverability.
4) File Auditing and Tracking: Some traditional systems implement file auditing and tracking mechanisms to monitor file activities. However, the identification of orphaned objects

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue VI June 2025- Available at www.ijraset.com

III. DISADVANTAGES

1) Manual data cleanup processes are time-consuming and error-prone.
2) Lack of structured data lifecycle management leads to inefficient storage practices.
3) Limited or absent deduplication results in redundant storage and increased costs.
4) Manual reporting is labor-intensive and prone to inaccuracies.
5) Difficulty in scaling to accommodate growing data volumes and evolving storage needs.
6) Increased security risks due to inadequate access controls and monitoring mechanisms.

IV. PROPOSED SYSTEM

The proposed system aims to address the challenges of inefficient cloud data management, storage clutter, and data integrity. Here's an overview of the proposed system:
The CloudClean system is to streamline cloud data management processes, reduce storage costs, and ensure data integrity through the implementation of automated cleanup, robust data lifecycle management, and efficient deduplication strategies.

A. Automated Cleanup

Utilizing Regex algorithms, the system automates the identification and removal of obsolete files or data regularly. This ensures that unnecessary files, including junk and temporary files, are systematically cleaned up, optimizing storage space and minimizing clutter.

B. Data Life Cycle Guard Scheme

The system implements a structured framework, the Data Life Cycle Guard Scheme, to govern the entire lifecycle of data within the cloud environment. This framework defines policies and rules for data creation, storage, usage, and deletion, ensuring compliance with regulations and industry standards.

C. Deduplication Strategies

Leveraging Fuzzy Matching algorithms, the system implements efficient deduplication strategies to identify and eliminate duplicate files or data chunks during the data upload process. By removing redundant copies of data, the system optimizes storage space and resource utilization.
The proposed system aims to optimize resource utilization, improve performance, and align with compliance standards. Through regular identification and elimination of obsolete files, structured data lifecycle management, and accurate deduplication techniques, the system sets new standards in cloud data management, contributing to enhanced storage efficiency and the seamless operation of cloud applications. CloudClean offers a comprehensive solution to address the challenges of cloud storage clutter and inefficiency, empowering organizations to effectively manage their cloud data, reduce costs, and ensure data integrity and compliance.

V. MODULES DESCRIPTION

A. Cloud Service Provider Web App

The design and development of the Cloud Service IaaS Provider Web App will leverage Python, Flask, MySQL, Wampserver, Pandas, Matplotlib, NumPy, and Bootstrap. Python and Flask will be utilized for backend development, MySQL for database management, and Wampserver for local development environment setup. Pandas, Matplotlib, and NumPy will facilitate data processing and visualization tasks, while Bootstrap will enhance the frontend with responsive design elements. It serves as the primary interface for users to interact with the Infrastructure as a Service (IaaS) cloud service provider's platform. It encompasses a range of functionalities and features designed to facilitate the provisioning, management, and monitoring of cloud resources. Users can easily provision virtual machines, storage volumes, and network resources through the web application. This includes selecting instance types, specifying storage configurations, and defining network settings. Once deployed, users can efficiently manage their resources, such as starting/stopping virtual machines, attaching/detaching storage volumes, and configuring network security groups. The web application provides robust monitoring and analytics tools, allowing users to monitor the performance and health of their cloud resources in real-time. This includes metrics on CPU utilization, memory usage, network traffic, and storage capacity, enabling proactive management and optimization of resources. Security is paramount, and the web application offers features for managing security settings and access control policies.

This includes configuring firewall rules, setting up identity and access management (IAM) policies, and enabling multi-factor authentication for user accounts, ensuring the confidentiality and integrity of data. Billing and cost management functionalities are also integrated into the web application, allowing users to track resource consumption, estimate costs, and set budget alerts to prevent unexpected charges. Additionally, the web application provides APIs and SDKs for programmatically accessing and managing cloud resources, enabling automation and integration with third-party tools. With customizable user interfaces, users can tailor their experience to suit their preferences and workflows. Furthermore, access to comprehensive documentation, tutorials, and support resources ensures users can effectively onboard, troubleshoot, and optimize their use of cloud services. This module empowers users to leverage the benefits of Infrastructure as a Service for their IT infrastructure needs with ease and efficiency.

B. Cloud User Interface

1) Data Owner

The Data Owner, as a key user of the Cloud User Interface, has access to a comprehensive set of functionalities:

- Register/Login: Data Owners can register for an account or log in using their credentials to access the platform.
- Add and Manage Data: Data Owners can upload and manage their data assets within the cloud storage system, organizing them into folders or categories for easy access and management.
- Set Expiry Date for Each Data: Data Owners have the capability to set expiry dates for their data, ensuring compliance with data retention policies and facilitating automatic data cleanup.
- Add and Manage Data User: Data Owners can add and manage Data Users, granting them access to specific datasets based on their roles and permissions.
- Set Access Control for Users: Data Owners have granular control over access permissions, defining who can view, edit, or delete specific datasets.
- Receive Bill/Pay Bill: Data Owners receive billing statements for their usage of cloud resources and can conveniently make payments through the user interface.
- View Deduplication Report: Data Owners can access reports detailing the results of deduplication processes, providing insights into storage optimization efforts.
- View Junk File or Temp File Deletion Report: Data Owners receive reports on the deletion of junk files or temporary files, ensuring efficient cleanup of unnecessary data.
- Receive Expiry Alert: Data Owners receive alerts when data reaches its expiry date, prompting them to review and renew if necessary.
- View Expired Data: Data Owners can view a list of expired data assets and take appropriate actions, such as archiving or deleting them.
- View Data Log History: Data Owners have access to a log history that records all user actions and system events related to their data assets, providing transparency and accountability.

2) Data User

Data Users, as consumers of data within the cloud environment, have simplified functionalities:

- Login: Data Users can log in to the Cloud User Interface using their credentials to access the data shared with them.
- Access Data: Once logged in, Data Users can access the data assets shared with them by Data Owners, allowing them to view, download, or interact with the data as permitted by the access control settings.

C. Data Access

The Data Access Module, encompassing upload, access, and download functionalities, forms the backbone of interactions with data stored in the cloud environment. This module is designed to facilitate seamless and secure data operations, ensuring efficient data management while upholding data integrity and security standards.

1) Upload Functionality

Users can upload data to the cloud storage infrastructure through various interfaces, including web applications, APIs, or command-line tools. The module manages the upload process, ensuring that data is securely transmitted and stored in the designated storage locations. Upload operations may involve transferring individual files, batches of files, or entire datasets, with mechanisms in place to handle large volumes of data efficiently. Data upload can also include metadata tagging and validation to enhance data organization and integrity.

2) Access Functionality

Authorized users and applications can access data stored in the cloud environment using designated access points and authentication mechanisms. The module verifies user credentials and permissions to grant appropriate access privileges based on predefined roles and policies. Access operations encompass querying, retrieving, and manipulating data objects as required by users. This includes reading data for analysis, updating records, or executing transactions within applications.

3) Download Functionality

Users have the capability to download data from the cloud storage infrastructure to local or remote environments for further processing or analysis. The module facilitates secure and efficient data transfer, ensuring that downloaded data retains its integrity and confidentiality. Download operations may involve retrieving individual files, datasets, or specific subsets of data based on user-defined criteria. Users can choose from various download options and formats to suit their requirements, with the module supporting standard protocols for data transfer. In summary, the Data Access Module serves as a comprehensive solution for managing data interactions within the cloud environment, offering robust upload, access, and download functionalities while prioritizing security, integrity, and performance. By facilitating seamless data operations and enforcing stringent access controls, the module empowers users to leverage cloud-stored data effectively for their applications and analytics needs.

D. Data Life Cycle Guard

The Data Life Cycle Guard Scheme establishes a structured framework to govern the entire lifecycle of data within the cloud environment. It defines policies and rules for data creation, storage, usage, and deletion, ensuring compliance with regulations and industry standards. This module encompasses various functionalities, including defining data creation policies to ensure consistency and standardization, allocating storage based on classification and access requirements, and implementing access control policies for data security. Additionally, it establishes rules for data retention and expiry, facilitating automatic cleanup processes and compliance with retention regulations. Procedures for data deletion and archiving are outlined, along with mechanisms for audit and compliance monitoring to track data-related activities and ensure adherence to policies. Automated enforcement mechanisms are implemented to enforce policy compliance, with provisions for continuous improvement and adaptation based on evolving business needs and regulatory requirements. In essence, the Data Life Cycle Guard Scheme module provides a comprehensive framework for managing data throughout its lifecycle, promoting compliance, efficiency, and security within the cloud environment.

E. Data Deduplication

The Data Deduplication using Fuzzy Matching Module is designed to optimize storage efficiency by identifying and eliminating duplicate data during the process of uploading data to the Cloud Service Provider IaaS Web App. It employs fuzzy matching algorithms to compare data objects and identify similarities, enabling the removal of redundant copies and conserving storage space.

- Data Upload Process: The module integrates seamlessly with the data upload process within the Cloud Service Provider IaaS Web App. As users upload new data to the cloud environment, the module intercepts and analyzes the incoming data objects for duplicate content.
- Fuzzy Matching Algorithms: Fuzzy matching algorithms are utilized to compare the content of incoming data objects with existing data within the cloud environment. These algorithms employ similarity measures to identify data objects that exhibit partial matches or similarities, even in the presence of variations or discrepancies.
- Duplicate Identification: The module systematically compares incoming data objects with existing data within the cloud environment using fuzzy matching algorithms. If a data object is found to be similar or nearly identical to an existing data object above the defined similarity threshold, it is flagged as a duplicate.
- Duplicate Removal: Upon identifying duplicates, the module initiates the removal process to eliminate redundant copies of data objects from the cloud storage environment. This ensures that only unique instances of data are retained, conserving storage space and improving data organization.
- Handling Variations and Discrepancies: Fuzzy matching algorithms are robust and capable of handling variations and discrepancies in data, such as spelling errors, typos, or formatting differences. This allows the module to accurately identify duplicates even in cases where data objects exhibit minor differences.
- Logging and Reporting: The module maintains logs of duplicate identification and removal activities, including details of data objects analyzed, similarities detected, and duplicates removed. This logging provides administrators with visibility into the deduplication process and its outcomes.

Through the Data Deduplication using Fuzzy Matching Module, redundant copies of data are systematically identified and eliminated during the data upload process to the Cloud Service Provider IaaS Web App, optimizing storage efficiency and promoting a streamlined data environment.

F. Automated Clean Up

The Automated Clean Up Regex Algorithm Module is specifically designed to target and remove junk or temporary files that are created while accessing data within the cloud storage environment. It employs Regex algorithms to systematically identify and eliminate these unnecessary files, enhancing storage efficiency and maintaining a clean and organized data environment.

- Pattern Definition: The module begins by defining Regex patterns that match criteria indicative of junk or temporary files. These patterns may include file names, extensions, prefixes, or specific content characteristics associated with temporary data.
- File Scanning: The Regex algorithm systematically scans the cloud storage environment, examining each file to determine if it matches any of the predefined patterns. This scanning process covers all files within the specified scope, ensuring comprehensive coverage.
- Pattern Matching: When a file is scanned, the Regex algorithm applies the predefined patterns to the file's metadata or content. If a match is found between the file's characteristics and any of the patterns, the file is flagged as a potential junk or temporary file.
- Flagging for Removal: Files that match the predefined patterns are flagged for removal during the cleanup process. The algorithm keeps track of flagged files, storing information about their location, metadata, and reasons for potential obsolescence.
- Cleanup Execution: Once the scanning process is complete, the cleanup process is executed based on the flagged files. The algorithm systematically removes flagged files from the cloud storage environment, ensuring that only relevant and necessary files remain.
- Verification and Confirmation: After the cleanup process is executed, the algorithm verifies the removal of flagged files and confirms that they have been successfully deleted from the storage environment. This verification step ensures the effectiveness and accuracy of the cleanup process.
- Logging and Reporting: Throughout the workflow, the Regex algorithm logs details of its activities, including files scanned, matches found, files flagged for removal, and cleanup actions taken. This logging enables administrators to track the progress of the cleanup process and review its outcomes.
- Iterative Improvement: The module supports iterative improvement, allowing for continuous refinement of Regex patterns and scanning parameters based on feedback and performance evaluation. Administrators can fine-tune the algorithm to optimize its effectiveness in identifying and removing junk or temporary files over time.

Through the Automated Clean Up Regex Algorithm Module, unnecessary junk or temporary files created while accessing data within the cloud storage environment are systematically identified and removed, promoting storage efficiency and maintaining a clutter-free data environment.

G. Notification

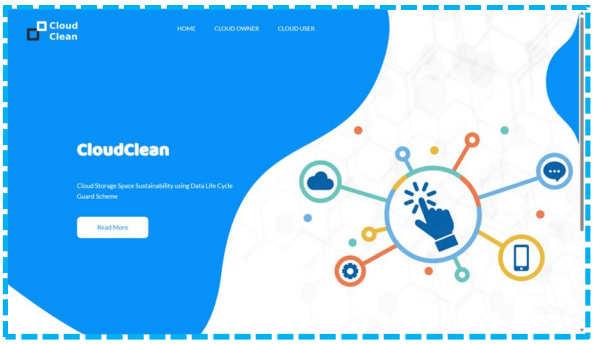
The Notification module is essential within the cloud environment, ensuring users are promptly informed about critical events, updates, or actions concerning their data and system activities. It encompasses functionalities aimed at keeping users engaged and informed, promoting transparency and facilitating quick responses to important events. Notifications are triggered based on predefined events or conditions, such as data upload/download completion, system maintenance schedules, storage quota thresholds, or security alerts. These notifications are delivered through various channels, including email, SMS, in-app notifications, and push notifications to mobile devices, allowing users to customize their preferred notification channels for accessibility. Additionally, users can personalize notifications according to their preferences, specifying factors like frequency, content, and urgency levels to tailor their notification experience. Notifications are prioritized based on the severity and urgency of the event or action being communicated, ensuring that critical alerts receive prompt attention from users. Some notifications may include actionable elements, enabling users to take immediate steps or responses directly from the notification interface, such as confirming alerts or performing specific actions to address an issue. The module maintains a history of notifications sent to users, including timestamps, senders, and content details, providing users with a reference for past events and actions.

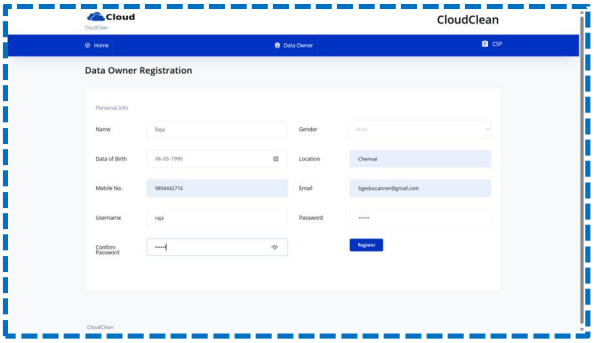
Notifications seamlessly integrate with monitoring systems and event management platforms to automate alert generation and dissemination, proactively notifying users of system status changes or anomalies for prompt response. All notification activities are logged for compliance and audit trail management, ensuring accountability and transparency in communication processes.

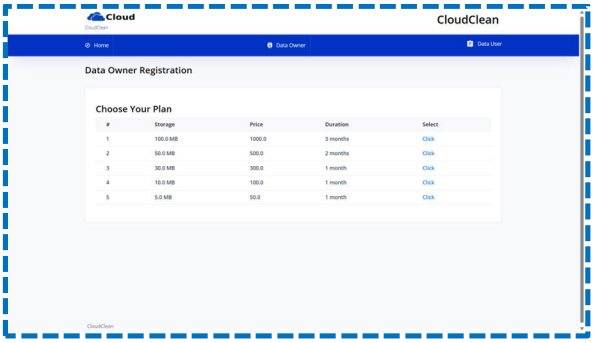
H. Logging and Reporting

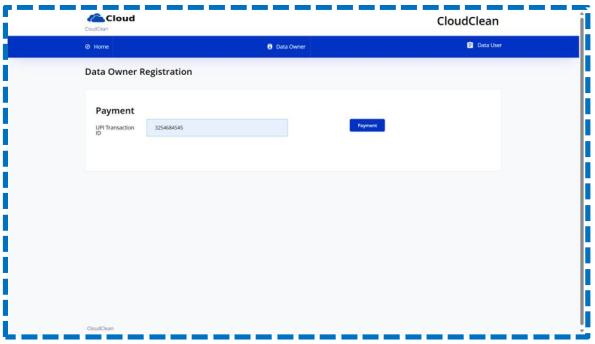
The Logging and Reporting module in the cloud environment is vital for recording and analyzing system activities. It securely stores log data, aggregates information from various sources, and offers real-time monitoring capabilities for proactive response to critical events. Administrators can efficiently search and query log data, receive alerts for important events, and generate reports to gain insights into system performance and compliance. This module ensures transparency, accountability, and security by creating compliance reports and audit trails, demonstrating adherence to regulatory standards and internal policies.

VI. SCREENSHOTS









VII. CONCLUSION

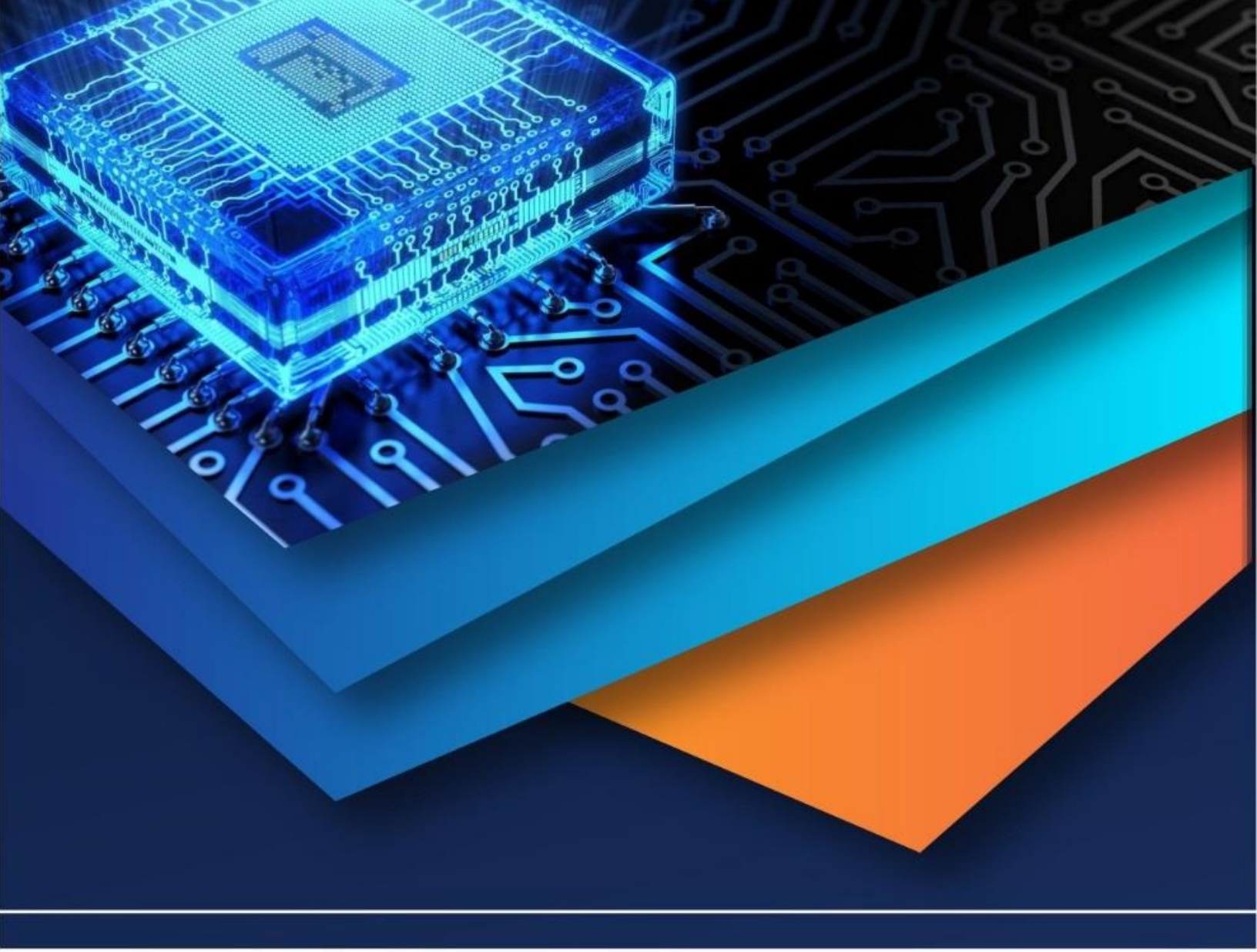
In conclusion, the project represents a significant advancement in cloud data management, addressing the limitations of traditional systems and offering a comprehensive solution to optimize storage efficiency, reduce costs, and ensure data integrity. Through the implementation of automated cleanup processes, robust data lifecycle management, and efficient deduplication strategies, it streamlines data management workflows and mitigates the challenges associated with junk files, outdated backups, and orphaned objects. By leveraging automation tools and algorithms, this project simplifies the process of identifying and removing unnecessary files, ensuring that cloud storage remains clutter-free and optimized for performance. The integration of a structured data lifecycle management framework facilitates adherence to compliance standards and regulatory requirements, while also enabling efficient resource allocation and data retention practices. Furthermore, the incorporation of deduplication algorithms enhances storage efficiency by identifying and eliminating redundant data, thereby reducing storage costs and optimizing resource utilization. This not only improves the overall performance of cloud applications but also contributes to environmental sustainability by minimizing the carbon footprint associated with excess data storage. Additionally, it offers advanced monitoring and reporting capabilities, allowing users to track storage usage, data access patterns, and compliance metrics in real-time. This proactive approach to data management enables organizations to identify potential issues early and take corrective actions to maintain data integrity and security. In summary, the project represents a paradigm shift in cloud data management, offering a holistic approach to address the complexities and challenges of modern data environments. By combining automation, intelligent algorithms, and proactive monitoring, it empowers organizations to optimize their cloud storage resources, improve operational efficiency, and drive innovation in the digital age.

REFERENCES

Journal References

[1] J. Qiu et al., "Light-Dedup: A Light-weight Inline Deduplication Framework for Non-Volatile Memory File Systems", Proceedings of the USENIX Annual Technical Conference (USENIX ATC), 2023.
[2] M. Song, Z. Hua, Y. Zheng, T. Xiang and X. Jia, "FCDedup: A two-level deduplication system for encrypted data in fog computing", IEEE Trans. Parallel Distrib. Syst., vol. 34, no. 10, pp. 2642-2656, Jul. 2023.
[3] A. Makris, I. Kontopoulos, E. Psomakelis, S. N. Xyalis, T. Theodoropoulos and K. Tserpes, "Performance analysis of storage systems in edge computing infrastructures", Appl. Sci., vol. 12, no. 17, pp. 8923, 2022.
[4] G. Cheng, D. Guo, L. Luo, J. Xia and S. Gu, "LOFS: A lightweight online file storage strategy for effective data deduplication at network edge", IEEE Trans. Parallel Distrib. Syst., vol. 33, no. 10, pp. 2263-2276, Oct. 2022.
[5] C. Tian, H. Liu, X. Liao and H. Jin, "UCat: Heterogeneous memory management for unikernels", Frontiers Comput. Sci., vol. 17, no. 1, pp. 171204-171215, 2022.
[6] C. Deng, Q. Chen, X. Zou, E. Xu, B. Tang and W. Xia, "imDedup: A lossless deduplication scheme to eliminate fine-grained redundancy among images", Proc. IEEE Int. Conf. Data Eng., pp. 1071-1084, 2022.
[7] D. Yang, H. Liu, H. Jin and Y. Zhang, "HMvisor: Dynamic hybrid memory management for virtual machines", Sci. China Inf. Sci., vol. 64, no. 9, pp. 192-16, 2021.
[8] C. Ji et al., "Pattern-guided file compression with user-experience enhancement for log-structured file system on mobile devices", Proc. 19th USENIX Conf. File Storage Technol. (FAST), pp. 127-140, 2021.
[9] J. Li, Z. Yang, Y. Ren, P. P. Lee and X. Zhang, "Balancing storage efficiency and data confidentiality with tunable encrypted deduplication", Proc. 15th Eur. Conf. Comput. Syst., pp. 1-15, 2020.
[10] J. Kosińska and K. Zieliński, "Autonomic management framework for cloud-native applications", J. Grid Comput., vol. 18, no. 4, pp. 779-796, Dec. 2020.
[11] S. Li and T. LAN, "HotDedup: Managing hot data storage at network edge through optimal distributed deduplication", Proc. IEEE Conf. Comput. Commun. pp. 247-256, Jul. 2020.
[12] Y. Tan et al., "Improving the Performance of Deduplication-based Storage Cache via Content-Driven Cache Management Methods", IEEE Transactions on Parallel and Distributed Systems (TPDS), 2020.

[13] Y. Zhang et al., "Finesse: Fine-grained Feature Locality based Fast Resemblance Detection for Post-Deduplication Delta Compression", Proceeding of the USENIX Conference on File and Storage Technologies (FAST), 2019.
[14] Q. Yang et al., "SmartDedup: Optimizing Deduplication for Resourceconstrained Devices", Proceedings of the USENIX Annual Technical Conference (USENIX ATC), 2019.
[15] A. Nicolaescu, O. Ascigil and I. Psaras, "Edge data repositories - The design of a store-process-send system at the edge", Proc. ACM CoNEXT Workshop Emerg. Netw. Comput. Paradigms, pp. 41-47, 2019.
[16] C. Wang, Q. Wei, J. Yang, C. Chen, Y. Yang and M. Xue, "NV-Dedup: High-performance inline deduplication for non-volatile memory", IEEE Trans. Comput., vol. 67, no. 5, pp. 658-671, May 2018.
[17] W. Xia et al., "FastCDC: A fast and efficient content-defined chunking approach for data deduplication", Proc. USENIX Annu. Tech. Conf., pp. 101-114, 2016.
[18] M. R. Mesbahi et al., "Highly Reliable Architecture Using the 80/20 Rule in Cloud Computing Datacenters", Future Generation Computer Systems (FGCS), 2017.
[19] M. Fu et al., "Design tradeoffs for data deduplication performance in backup workloads", Proc. 13th USENIX Conf. File Storage Technol., pp. 331-344, 2015.
[20] B. Mao, H. Jiang, S. Wu and L. Tian, "POD: Performance oriented I/O deduplication for primary storage systems in the cloud", Proc. IEEE Int. Parallel Distrib. Process. Symp., pp. 767-776, 2014.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY