ISSN: 2395-7852

International Journal of Advanced Research in Arts, Science, Engineering & Management

Volume 10, Issue 3, May 2023

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 6.551

+91 9940572462 +91 9940572462 ijarasem@gmail.com www.ijarasem.com

International Journal of Advanced Research in Arts, Science, Engineering & Management (IJARASEM)

| ISSN: 2395-7852 | www.ijarasem.com | Impact Factor: 6.551 |Bimonthly, Peer Reviewed & Referred Journal|

| Volume 10, Issue 3, May 2023 |

Cloud Data Deduplication System Using Per File Parity and File Name Interpreter

Mr.R.KARTHIKEYAN.,[1] Mr.S.NANTHAKUMAR.,[2]

Assistant Professor, Department of Master of Computer Application, Gnanamani College of Technology, Namakkal Tamilnadu, India[1]

PG Scholar, Department of Master of Computer Application, Gnanamani College of Technology, Namakkal, Tamilnadu, India[2]

ABSTRACT: Cloud storage has become an integral part of modern data management, providing convenient and scalable storage solutions. However, the rapid growth of data stored in the cloud has led to challenges in storage efficiency and data redundancy. Data deduplication techniques have emerged as a promising approach to address these challenges by eliminating duplicate copies of data.This paper presents a novel cloud data deduplication system that utilizes per-file parity and a file name interpreter to enhance the deduplication process. This information is then used to optimize the deduplication process by selectively eliminating redundant data while preserving essential file versions.Additionally, the system incorporates per-file parity, a technique that generates parity information for each file independently. Unlike traditional block-level parity, per-file parity allows for granular deduplication at the file level. By calculating and storing parity information on a per-file basis, the system can quickly identify and eliminate duplicate files, further enhancing storage efficiency.Experimental results demonstrate the effectiveness of the proposed system in terms of deduplication ratio, storage savings, and processing time. The system achieves significant reductions in storage requirements while maintaining data integrity and minimizing processing overhead.

KEYWORDS: Deduplication, Essential technique for optimizing storage, File name interpreter, cloud storage environment, Client-side and server-side, Large-scale cloud storage environments.

I. INTRODUCTION

In today's digital era, the exponential growth of data has become a significant challenge for organizations in terms of storage, management, and cost. Cloud storage systems have emerged as a viable solution to address these challenges by providing scalable and cost-effective storage options. However, the efficient utilization of cloud storage resources remains a critical concern.Cloud data deduplication systems play a crucial role in optimizing storage utilization by identifying and eliminating duplicate data across multiple files or within the same file. This process involves identifying redundant data blocks and replacing them with references to a single instance, thereby reducing storage requirements and improving overall system performance.This paper introduces a novel cloud data deduplication system that leverages per-file parity and file name interpreter techniques to enhance the deduplication process. The per-file parity technique involves calculating parity information for each file individually, enabling efficient recovery and integrity checking. The file name interpreter technique utilizes advanced algorithms to interpret the file names and extract valuable metadata, facilitating more accurate and efficient deduplication.The proposed system aims to address some of the limitations of existing deduplication approaches, such as the high computational overhead and storage requirements associated with global deduplication schemes. By employing per-file parity and file name interpreter techniques, the system can achieve significant storage savings while minimizing the impact on performance and resource consumption.Additionally, the system incorporates robust data integrity mechanisms to ensure the reliability and consistency of the deduplicated data. This includes checksum calculations, error detection, and error correction capabilities, providing an extra layer of protection against data corruption or loss.Overall, this paper presents a comprehensive approach to cloud data deduplication, combining per-file parity and file name interpreter techniques to optimize storage utilization, improve performance, and enhance data integrity. The subsequent sections will delve into the technical details and experimental evaluations of the proposed system, demonstrating its effectiveness in real-world scenarios.

II. EXISTING SYSTEM

Cloud data reduplication is a technique used to eliminate redundant data within a cloud storage environment. It aims to reduce storage costs and optimize data transfer by identifying and storing only unique data blocks, eliminating duplicates across multiple files or versions of files.
In a typical cloud data deduplication system, several methods are commonly employed:

1. **Chunk-based Deduplication:** Data is divided into fixed-size or variable-size chunks. Each chunk is then fingerprinted using hashing algorithms such as MD5 or SHA-1. Duplicate chunks are identified by comparing their hash values, and only unique chunks are stored in the cloud storage system.
2. **Delta Encoding:** Instead of dividing data into fixed-size chunks, delta encoding breaks data into variable-size chunks based on changes within the file. It stores the differences (delta) between chunks, rather than the entire chunk. This approach is useful when dealing with large files that have only small changes between versions.
3. **Content-Defined Chunking:** This technique breaks data into variable-size chunks based on the content rather than predefined boundaries. It uses algorithms like Rabin, Rolling Hash, or Similarity Digests to identify natural chunk boundaries within the data.
4. **Metadata Management:** In addition to deduplicating data blocks, metadata management is crucial for efficient deduplication. Metadata includes file attributes, such as file names, sizes, timestamps, and ownership information. It helps in identifying duplicate files and managing the deduplication process.

III. PROPOSED SYSTEM

We propose Per- File Parity (PFP) to improve the reliability of deduplication-based storage systems. PFP computes the parity for every N chunks, where N is a configurable parameter, or for a whole file. When a disk failure is detected, the generated parity chunks can be used to recover from the read errors and failed data chunks by intra-file recovery. On the other hand, when several errors occur in a parity group of parity and these failed data chunks each have reference counts of greater than 1, PFP can recover these failed data chunks by inter-file recovery by leveraging the parity groups of the unaffected referencing files.

The reliability results show that PFP can tolerate much more data chunk failures and guarantee file availability upon multiple data chunk failures. Moreover, the failure-injection based evaluations show that the PFP scheme can tolerate hundreds of concurrent chunk errors without file loss for data sets with high deduplication ratios. The evaluations also show that PFP is highly cost effective in terms of file-loss-tolerance/redundancy measure by an average of 52.2% and 197.5% over the DTR and RCR schemes, respectively. On the other hand, the performance assessment shows that PFP's significant reliability gain comes at an acceptable performance cost of an average of 5.7% performance degradation to the deduplication-based storage system.

The data stored in the cloud can be retrieved and the integrity of these data can be ensured. It was based on pseudo random function and BLS signature, a private remote data integrity auditing scheme and a public remote data integrity auditing scheme. To protect the data privacy, a privacy-preserving remote data integrity auditing scheme with a random masking technique has been used. To reduce the burden of signature generation we designed a remote data integrity auditing scheme based on the in distinguish ability obfuscation technique. A Third Party Medium (TPM) is designed a light-weight remote data integrity auditing scheme. In this scheme, the TPM helps user generate signatures on the condition that data privacy can be protected. The data sharing is an important application to protect the identity privacy of user. At the same time data has been saved in to fog server as a temporary data which reduces the risks of data loss or data damage in the cloud server. By using the fog and cloud servers data can be recovered from cloud or the fog server in an efficient way.

IV. SYSTEM OVERVIEW

A cloud data deduplication system using per-file parity and file name interpreter is designed to optimize storage efficiency by identifying and eliminating duplicate data across multiple files stored in the cloud. Here is an overview of the system components and how they work together:

1. **Cloud Storage:** The system utilizes cloud storage services to store the files and data. This can be a public cloud provider like Amazon S3, Google Cloud Storage, or Microsoft Azure Blob Storage.
2. **File Upload:** Users upload files to the cloud storage through the deduplication system. Each file is identified by a unique file name.
3. **File Name Interpreter:** The file name interpreter component analyzes the file names to extract relevant information. This can include metadata such as the file's creation date, type, size, or any other information that

can assist in identifying duplicates. The interpreter may use various techniques like parsing naming conventions or leveraging file metadata to derive this information.

4. **Deduplication Engine:** The deduplication engine is responsible for identifying duplicate files based on their content. It compares the files' data and applies algorithms to detect similarities or identical segments. The engine calculates and stores the file's parity information, which represents the unique data segments within the file.
5. **Per-File Parity:** The system uses per-file parity, which means that instead of creating a single global index for all files, each file has its own parity information. This approach allows for efficient deduplication within a file while maintaining data integrity and minimizing the impact of data corruption or loss.
6. **Deduplication Index:** The deduplication index maintains a record of the unique data segments across all files in the cloud storage. It stores the parity information for each file, allowing the system to quickly identify duplicate segments during the deduplication process.
7. **Deduplication Process:** When a file is uploaded, the system first checks the deduplication index to determine if any duplicate segments exist. If duplicates are found, the system only stores the unique segments and updates the file's parity information. If the file is entirely duplicate, it may be skipped altogether.
8. **Storage Optimization:** By eliminating duplicate data, the system reduces storage requirements, as only unique data segments are stored in the cloud. This optimization can be particularly beneficial when dealing with large files or datasets with significant redundancy.
9. **Retrieval and Reconstruction:** When a user requests a file for retrieval, the system reconstructs the file by retrieving the unique data segments and using the parity information stored for that file. This ensures that the original file can be reconstructed accurately, even though it may be distributed across multiple storage blocks.

Data Deduplication, often called Dedup for short, is a feature that can help reduce the impact of redundant data on storage costs. When enabled, Data Deduplication optimizes free space on a volume by examining the data on the volume by looking for duplicated portions on the volume. Duplicated portions of the volume's dataset are stored once and are (optionally) compressed for additional savings. Data Deduplication optimizes redundancies without compromising data fidelity or integrity. More information about how Data Deduplication works can be found in the 'How does Data Deduplication work?' section of the Understanding Data Deduplication page.

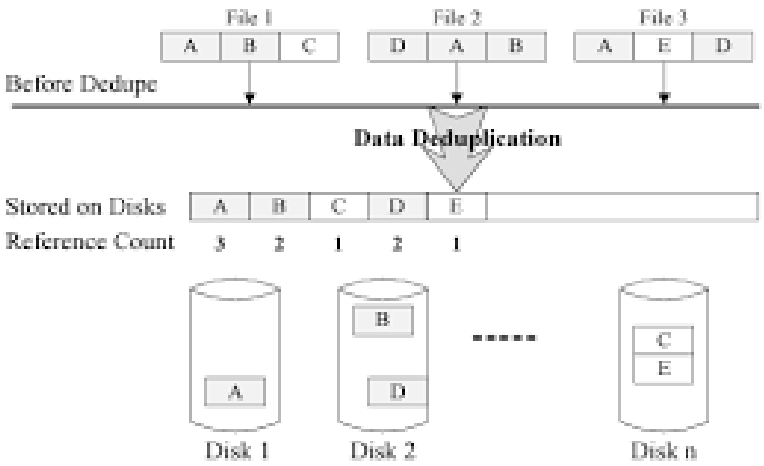


Fig.4.1 Data Deduplication

Data Deduplication helps storage administrators reduce costs that are associated with duplicated data. Large datasets often have a lot of duplication, which increases the costs of storing the data. For example:

- User file shares may have many copies of the same or similar files.
- Virtualization guests might be almost identical from VM-to-VM.
- Backup snapshots might have minor differences from day to day.

The space savings that you can gain from Data Deduplication depend on the dataset or workload on the volume. Datasets that have high duplication could see optimization rates of up to 95%, or a 20x reduction in storage utilization.

General purpose file servers: General purpose file servers are general use file servers that might contain any of the following types of shares:

- Team shares
- User home folders

| ISSN: 2395-7852 | www.ijarasem.com | Impact Factor: 6.551 |Bimonthly, Peer Reviewed & Referred Journal|

| Volume 10, Issue 3, May 2023 |

- Work folders
- Software development shares

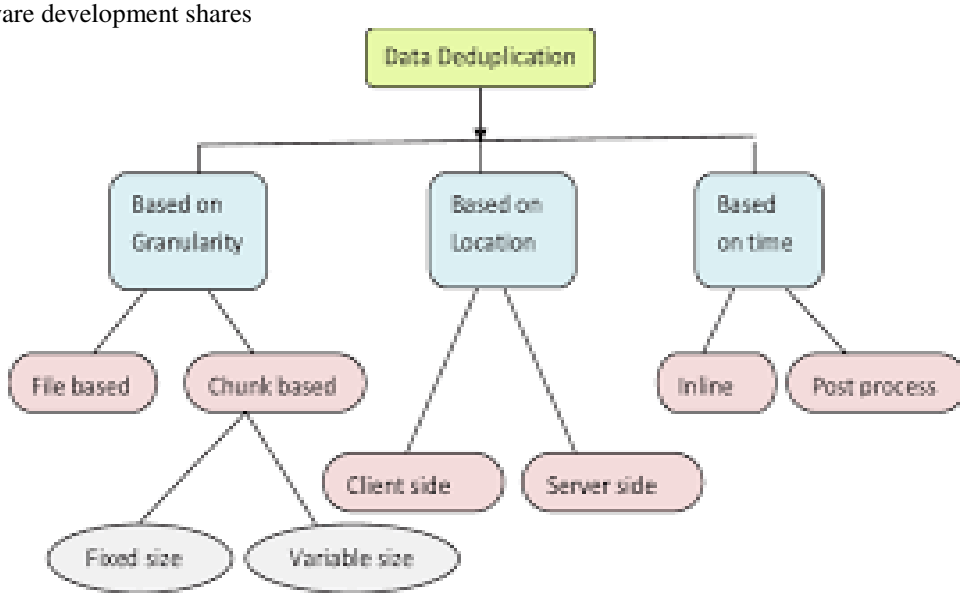


Fig. 4.2. File Servers

General purpose file servers are a good candidate for Data Deduplication because multiple users tend to have many copies or versions of the same file. Software development shares benefit from Data Deduplication because many binaries remain essentially unchanged from build to build.

Virtual Desktop Infrastructure (VDI) deployments: VDI servers, such as Remote Desktop Services, provide a lightweight option for organizations to provision desktops to users. There are many reasons for an organization to rely on such technology:

- Application deployment: You can quickly deploy applications across your enterprise. This is especially useful when you have applications that are frequently updated, infrequently used, or difficult to manage.
- Application consolidation: When you install and run applications from a set of centrally managed virtual machines, you eliminate the need to update applications on client computers. This option also reduces the amount of network bandwidth that is required to access applications.
- Remote Access: Users can access enterprise applications from devices such as home computers, kiosks, low-powered hardware, and operating systems other than Windows.
- Branch office access: VDI deployments can provide better application performance for branch office workers who need access to centralized data stores. Data-intensive applications sometimes do not have client/server protocols that are optimized for low-speed connections.

VDI deployments are great candidates for Data Deduplication because the virtual hard disks that drive the remote desktops for users are essentially identical. Additionally, Data Deduplication can help with the so-called *VDI boot storm*, which is the drop in storage performance when many users simultaneously sign in to their desktops to start the day.

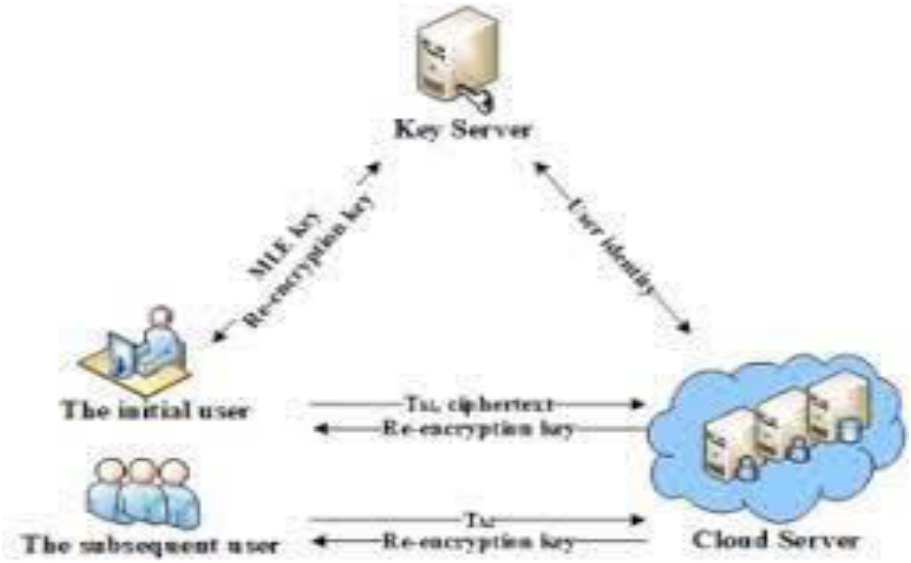


Fig. 4.3 Microsoft Data Protection Manager

Backup targets, such as virtualized backup applications: Backup applications, such as Microsoft Data Protection Manager (DPM), are excellent candidates for Data Deduplication because of the significant duplication between backup snapshots.

V. SYSTEM IMPLEMENTATION

- **PHP (Hypertext Preprocessor):**

PHP is a server-side scripting language used for developing dynamic web applications. It allows you to generate dynamic HTML content, interact with databases, handle forms, perform file operations, and more. PHP code is executed on the server, and the resulting HTML is sent to the client's browser.

- **WAMP:**

WAMP is an acronym that stands for Windows, Apache, MySQL, and PHP. It's a software stack which means installing WAMP installs Apache, MySQL, and PHP on your operating system (Windows in the case of WAMP). Even though you can install them separately, they are usually bundled up, and for a good reason too.

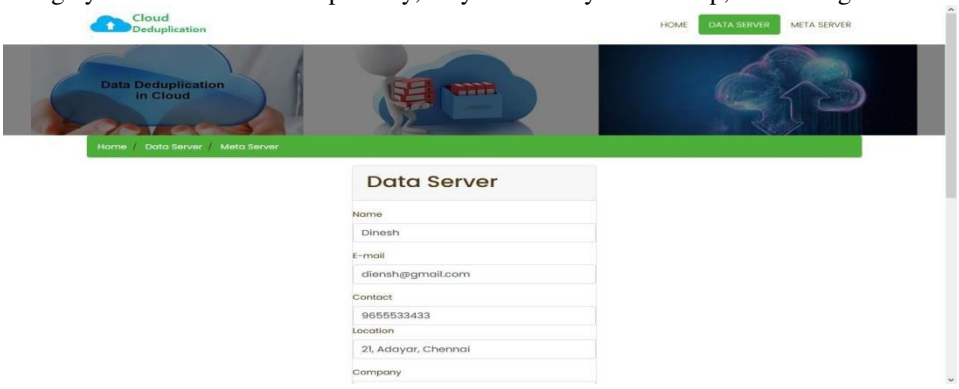


Fig.5.1 DATA SERVER

FILE UPLOADING

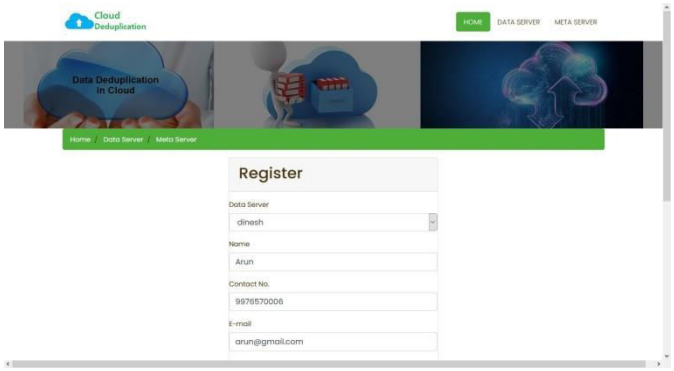


Fig. A. REGISTER PAGE

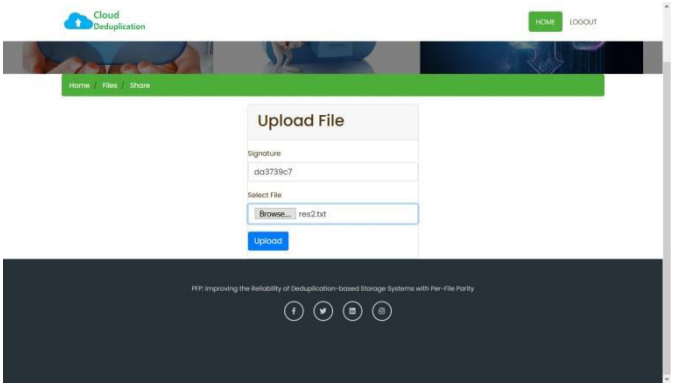


Fig. B. FILE UPLOADING

FILE SHARING

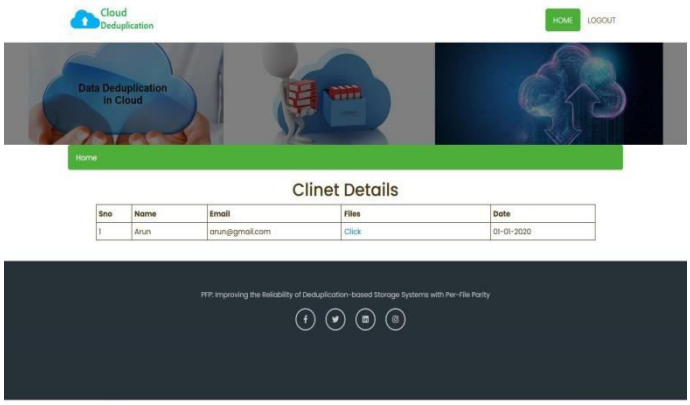


Fig. C.CLINT DETAILS

Fig. D. FILE SHARING

In the future, there are several enhancements that could be applied to a cloud data deduplication system using per-file parity and a file name interpreter. These enhancements could improve the efficiency, reliability, and flexibility of the system. Here are a few potential future enhancements:

1. Intelligent File Name Interpreter: The file name interpreter could be enhanced with advanced natural language processing (NLP) techniques and machine learning algorithms. This would enable the system to extract more meaningful metadata from file names, such as file type, content description, and relevant keywords. By understanding the semantics of file names, the deduplication system could make more accurate decisions about duplicate files and optimize storage resources accordingly.

2. Enhanced Deduplication Algorithms: Deduplication algorithms could be improved to handle more complex scenarios. For example, instead of relying solely on per-file parity, the system could utilize content-aware chunking algorithms that break files into smaller, variable-sized chunks based on their content rather than fixed block sizes. This approach would enhance deduplication efficiency by identifying similar content within and across files more effectively.

3. Hybrid Deduplication Techniques: Combining different deduplication techniques can lead to better results. Hybrid deduplication combines the benefits of both inline and post-processing deduplication. Inline deduplication eliminates duplicate data as it enters the system, while post-processing deduplication identifies and eliminates duplicates after the data has been stored. By utilizing a combination of these techniques, the system can achieve higher deduplication ratios and reduce the impact on system performance.

VI. CONCLUSION

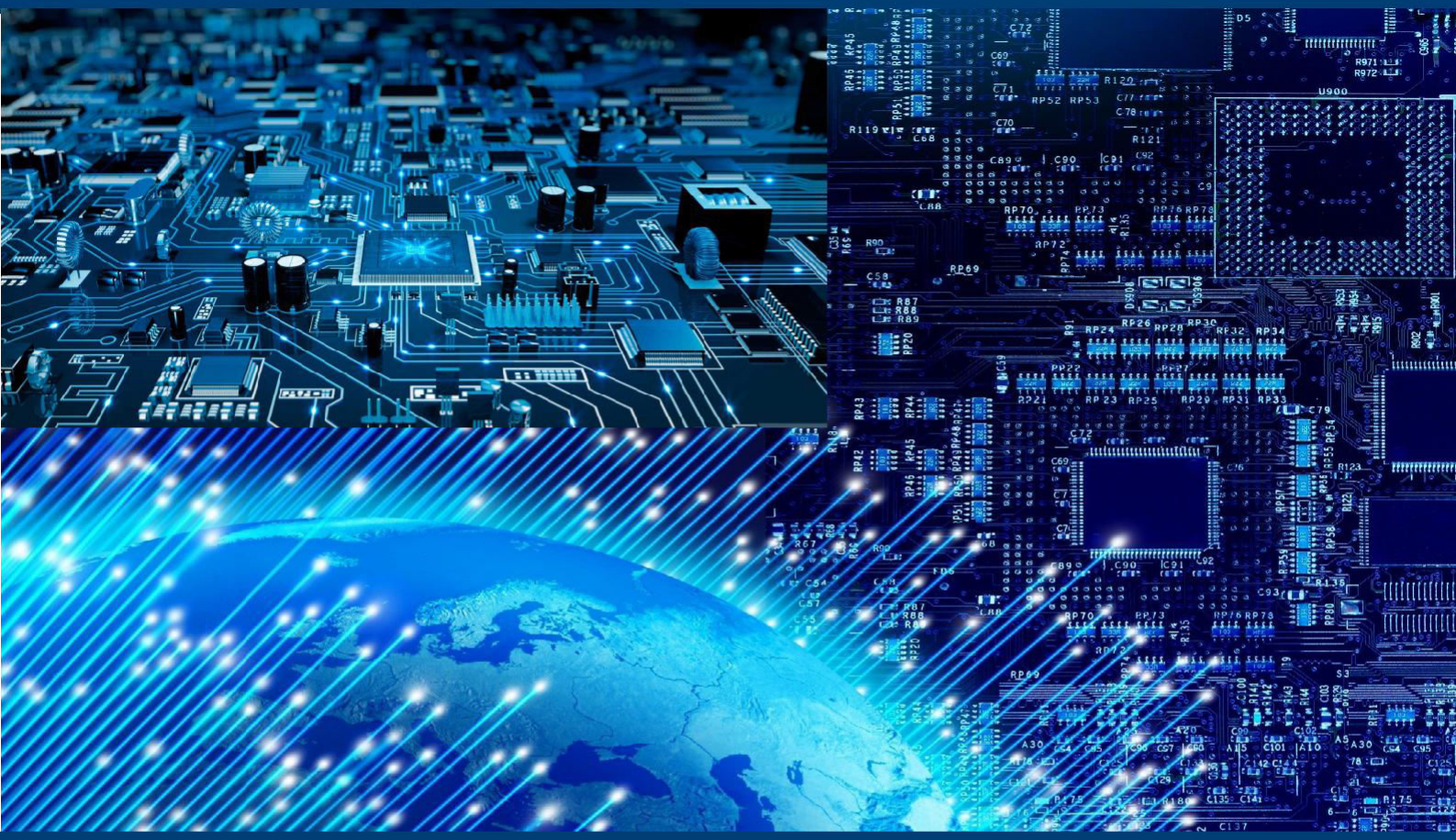
The cloud data reduplication system using per file parity and file name interpreter offers several benefits and features that make it a valuable tool for efficient data storage and retrieval in cloud environments. Overall, the cloud data

deduplication system utilizing per file parity and file name interpreter combines storage optimization, data integrity, and efficient data retrieval capabilities. By reducing storage costs, enhancing data reliability, and streamlining file management processes, this system offers a comprehensive solution for effective data management in cloud environments.

REFERENCES

1. R.Karthikeyan, & et all ”Biometric for Mobile Security” in the international journal of Engineering Science & Computing, Volume7,Issue6, June 2017, ISSN(0):2361-3361,PP No.:13552-13555.
2. R.Karthikeyan, & et all ”Data Mining on Parallel Database Systems” in the international journal of Engineering Science & Computing, Volume7,Issue7, July 2017, ISSN(0):2361-3361,PP No.:13922-13927.
3. R.Karthikeyan, & et all ”Ant Colony System for Graph Coloring Problem” in the international journal of Engineering Science & Computing, Volume7,Issue7, July 2017, ISSN(0):2361-3361,PP No.:14120-14125.
4. R.Karthikeyan, & et all ”Classification of Peer –To- Peer Architectures and Applications” in the international journal of Engineering Science & Computing, Volume7,Issue8, Aug 2017, ISSN(0):2361-3361,PP No.:14394-14397.
5. R.Karthikeyan, & et all ”Mobile Banking Services” in the international journal of Engineering Science & Computing, Volume7,Issue7, July 2017, ISSN(0):2361-3361,PP No.:14357-14361.
6. R.Karthikeyan, & et all ”Neural Networks for Shortest Path Computation and Routing in Computer Networks” in the international journal of Engineering and Techniques, Volume 3 Issue 4, Aug 2017, ISSN:2395-1303,PP No.:86-91.
7. R.Karthikeyan, & et all ”An Sight into Virtual Techniques Private Networks & IP Tunneling” in the international journal of Engineering and Techniques, Volume 3 Issue 4, Aug 2017, ISSN:2395-1303,PP No.:129-133.
8. R.Karthikeyan, & et all “Routing Approaches in Mobile Ad-hoc Networks” in the International Journal of Research in Engineering Technology, Volume 2 Issue 5, Aug 2017, ISSN:2455-1341, Pg No.:1-7.
9. R.Karthikeyan, & et all “Big data Analytics Using Support Vector Machine Algorithm” in the International Journal of Innovative Research in Computer and Communication Engineering, Volume 6 Issue 9, Aug 2018, ISSN:2320 - 9798, Pg No.:7589 -7594.
10. R.Karthikeyan, & et all “Data Security of Network Communication Using Distributed Firewall in WSN ” in the International Journal of Innovative Research in Computer and Communication Engineering, Volume 6 Issue 7, July 2018, ISSN:2320 - 9798, Pg No.:6733 - 6737.
11. R.Karthikeyan, & et all “An Internet of Things Using Automation Detection with Wireless Sensor Network” in the International Journal of Innovative Research in Computer and Communication Engineering, Volume 6 Issue 9, September 2018, ISSN:2320 - 9798, Pg No.:7595 - 7599.
12. R.Karthikeyan, & et all “Entrepreneurship and Modernization Mechanism in Internet of Things” in the International Journal of Innovative Research in Computer and Communication Engineering, Volume 7 Issue 2, Feb 2019, ISSN:2320 - 9798, Pg No.:887 - 892.
13. R.Karthikeyan & et all “Efficient Methodology and Applications of Dynamic Heterogeneous Grid Computing” in the International Journal of Innovative Research in Computer and Communication Engineering, Volume 7 Issue 2, Feb 2019, ISSN:2320 - 9798, Pg No.:1125 -1128.
14. R.Karthikeyan & et all“Entrepreneurship and Modernization Mechanism in Internet of Things” in the International Journal of Innovative Research in Computer and Communication Engineering, Volume 7 Issue 2, Feb 2019, ISSN:2320 - 9798, Pg No.:887– 892.
15. R.Karthikeyan & etall“Efficient Methodology for Emerging and Trending of Big Data Based Applications” in the International Journal of Innovative Research in Computer and Communication Engineering, Volume 7 Issue 2, Feb 2019, ISSN:2320 - 9798, Pg No.:1246– 1249.
16. R.Karthikeyan & et all “Importance of Green Computing In Digital World” in the International Journal of Innovative Research in Computer and Communication Engineering, Volume 8 Issue 2, Feb 2020, ISSN:2320 - 9798, Pg No.:14 – 19.
17. R.Karthikeyan & et all “Fifth Generation Wireless Technology” in the International Journal of Engineering and Technology, Volume 6 Issue 2, Feb 2020, ISSN:2395–1303.
18. R.Karthikeyan & et all “Incorporation of Edge Computing through Cloud Computing Technology” in the International Research l Journal of Engineering and Technology, Volume 7 Issue 9, Sep 2020 ,p. ISSN:2395–0056, e. ISSN:2395–0072.
19. R.Karthikeyan & et all “Zigbee Based Technology Appliance In Wireless Network” in the International Journal of Advance Research and Innovative Ideas in Education, e.ISSN:2395 - 4396, Volume:6 Issue: 5 , Sep. 2020. Pg.No: 453 – 458, Paper Id:12695.

20. R.Karthikeyan & et all "Automatic Electric Metering System Using GSM" in the International Journal of Innovative Research in Management, Engineering and Technology, ISSN: 2456 - 0448, Volume:6 Issue: 3 , Mar. 2021. Pg.No: 07 – 13.
21. R.Karthikeyan & et all "Enhanced the Digital Divide Sensors on 5D Digitization" in the International Journal of Innovative Research in Computer and Communication Engineering, e-ISSN: 2320 – 9801, p-ISSN: 2320 - 9798, Volume:9 Issue: 4 , Apr. 2021. Pg.No: 1976 – 1981.
22. R.Karthikeyan & et all "Crop Yield Prediction Based On Indian Agriculture Using Machine Learning" in the International Journal Of Engineering and Techniques, ISSN: 2395-1303, Volume:8 Issue: 4 , July. 2022. Pg.No: 11 – 22.
23. R.Karthikeyan & et all "A Blockchain Approach to Ensuring Provenance to Outsourced Cloud Data in A Sharing Ecosystem" in the International Journal Of Multidisciplinary Research In Science, Engineering and Technology, ISSN: 2584-7219, Volume: 5 Issue: 7, July. 2022. Pg.No: 1740 – 1744.
24. R.Karthikeyan & et all "College Bus Transport Management Web Application" in the International Journal Of Multidisciplinary Research In Science, Engineering and Technology, ISSN: 2582-7219, Volume: 6 Issue: 6, June. 2023. Pg.No: 1619 – 1625.
25. R.Karthikeyan & et all "Face Recognition Based Attendance System" in the International Journal of Innovative Research in Computer and Communication Engineering, ISSN: e 2320-9801, Volume: 11 Issue: 6, June. 2023. Pg.No: 8710 – 8717.
26. R.Karthikeyan & et all "Face Recognition Based Attendance System" in the International Journal of Innovative Research in Science, Engineering and Technology, ISSN: e 2319-8753, Volume: 12 Issue: 6, June. 2023.



International Journal of Advanced Research in Arts, Science, Engineering & Management (IJARASEM)

| Mobile No: +91-9940572462 | Whatsapp: +91-9940572462 | ijarasem@gmail.com |