IJIRSET
International Journal of Innovative Research in
SCIENCE | ENGINEERING | TECHNOLOGY

e-ISSN: 2319-8753 | p-ISSN: 2347-6710


# INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

## IN SCIENCE | ENGINEERING | TECHNOLOGY

Volume 12, Issue 10, October 2023

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.423

9940 572 462 | 6381 907 438 | ijirset@gmail.com | www.ijirset.com

International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 10, October 2023 ||

| DOI:10.15680/IJIRSET.2023.1210017 |

# Heterogeneous Data Storage Management with Deduplication in Cloud Computing

**Venkatesh D, Dr. Brahmadesam Viswanathan Krishna**

P.G. Student, Department of Computer Science and Engineering, KCG College of Technology, Chennai, TN, India

Professor, Department of Computer Science and Engineering, KCG College of Technology, Chennai, TN, India

**ABSTRACT**: Cloud computing offers extensive storage capabilities without the need for physical upgrades, making it a vital service. The confidentiality and integrity of data kept in the cloud must be safeguarded encryption is commonly employed. However, this encryption can lead to inefficient storage utilization and complex data sharing procedures. In this essay, we outline the implementation and evaluation of a novel approach for managing data storage while integrating deduplication in a heterogeneous cloud environment. Existing deduplication methods tend to be tailored to specific use cases and lack adaptability concerning data sensitivity. Our proposed solution offers a flexible framework that combines Managing deduplication and access across many Cloud Service Providers (CSPs). We conducted a comprehensive assessment of its performance, encompassing security analyses, comparative evaluations against existing methods, and practical implementation. The findings of our study showcase the robustness, efficiency, and practicality of our approach, reinforcing its suitability for real-world applications. Our research contributes to the advancement of encrypted data storage solutions in cloud computing, aligning with the diverse requirements of cloud users.

**KEYWORDS**: Deduplication, Cloud Service Provider (CSP), Cloud Computing

## I. INTRODUCTION

The way that we store & access data has been revolutionized by cloud computing. It provides users the capability to harness vast storage resources without the need for physical infrastructure upgrades. This transformative technology has become a cornerstone of modern computing, enabling individuals and organizations to scale their data storage needs efficiently. However, this convenience comes with its own set of challenges, particularly in the realm of data security and storage efficiency. The paramount concern when it comes to cloud storage is the protection of sensitive information. Users rightly demand stringent safeguards to protect the privacy and accuracy of their data, especially given the shared nature of cloud environments. Consequently, data is often encrypted before being stored in the cloud, an essential step to thwart unauthorized access. While encryption provides robust security, it introduces a conundrum: encrypted data is inherently resistant to deduplication, a process that eliminates redundant copies of data to optimize storage efficiency. Consequently, cloud users often grapple with inefficient utilization of their cloud storage resources.

Deduplication is a well-established technique in the field of data storage management. It involves the identification and elimination of duplicate data blocks, ensuring that only a single copy is retained, thus optimizing storage space and reducing the associated costs. In traditional data storage systems, deduplication has proven to be an effective strategy for improving storage efficiency. However, its application in the context of encrypted data within the cloud presents unique challenges. Existing deduplication methods are often tailored to specific use cases and lack the flexibility to adapt to varying levels of data sensitivity. This limitation hampers their ability to cater to the diverse demands of cloud users who store a wide array of data types, ranging from highly sensitive corporate documents to less critical personal files. Moreover, cloud computing itself is a heterogeneous environment, with users often leveraging multiple Cloud Service Providers (CSPs) to meet their varied needs. Coordinating deduplication across these different providers while maintaining data security and access control further complicates the storage management landscape.

In light of these challenges, our research endeavors to address the intersection of encrypted data storage and deduplication within the cloud. We suggest a fresh strategy that not only addresses the difficulties of deduplication in a cloud environment but also introduces a dynamic access control mechanism. This method aims to enhance storage efficiency while ensuring data security across multiple CSPs. Our study encompasses a comprehensive evaluation, including rigorous security analysis,

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 10, October 2023 ||

| DOI:10.15680/IJIRSET.2023.1210017 |

comparative assessments against existing deduplication strategies, and practical implementations to demonstrate the viability of our approach.

The results of our work in the area of cloud management of storage are presented in this paperoffering a promising solution that combines deduplication and access control in a heterogeneous cloud environment. Our research bridges the gap between data security and storage efficiency, ultimately striving to provide cloud users with a versatile and secure solution for managing their data in the ever-expanding cloud computing landscape.

## II. RELATED WORK

Deduplication of data is a popular subject that has been studied extensively by numerous scholars. Because of this, a variety of data deduplication techniques have been proposed in the scientific literature, including [1], whereby the researchers created a six-step deduplication design that employs record linkage to detect duplicate information. The framework covers data preparation, attribute matching using sorted neighbourhood, decision mode construction, and clustering. The current issue of the lack of labelled data for massive data deduplication has been brought up by the creators of this work, which makes it challenging to evaluate the model's performance. This problem was also raised in [2], where writers advise employing learning by doing as a substitute for the lack of tagged data in deduplication. The writers have gotten the best outcomes withan F-score for structured datasets of 98,4%. However, an inferior score of 52% was obtained for filthy datasets. As seen in [2], Deep learning was also used to resolve data duplication, where the authors design a binary classification technique for safety engineering utilizing fuzzy string-matching algorithms. The proposed approach is based on a binary Convolutional Neural Networks (CNN) classification and a string similarity-based classifier. The above-mentioned research has greatly aided entity resolution and duplicate identification. However, as far as of accuracy & execution time, these methods are not suitable for datasets with a big size.

The aforementioned methodologies failed to tackle the specific problems raised in a large-scale data setting, in spite of the data set volume concern. Big data has, in fact, given rise to new difficulties, including the breadth of data types, the variety of information sources, and the data's speed and dependability [3] [4]. Recent research, including [5], where the authors carried out a study on indexing methods for large-scale data deduplication, specifically addressed these concerns. According to the experiments, Sorted neighborhood indexing is the most difficult indexing technique for large datasets. Additionally, Christoph ides et al. conducted a thorough analysis of every approach for entity resolution that is currently in use in [6]. Clustering, which is matching, block processing, & blocking were all covered in their overview of the several entity resolution methods used in large data. The difficulty of the deduplication methods' deteriorating performance with time has also been brought up by the authors. In [7] and [8], Abd ElGhafar et al. presented an entity resolution approach for massive data that makes use of hashing TF and Jaccard similarity. The method was used to demonstrate the effects of various Natural Language Processing (NLP) strategies on entity resolution in seven scenarios. With 1 million records in the dataset, this method achieves an accuracy of 91%. A deduplication technique that permits reducing the necessary amount of pairwise comparisons in order to handle duplicate information in the web of data has been proposed by Efthymiou et al. in [9].

The opposite direction recurrent neural networks (RNNs) and Long Short-Term Memory (LSTM) are the foundation of the novel Stacked Dedupe Learning to resolve entities approach developed by the authors of [10] using deep learning. The study does not, however, demonstrate how huge data affects the effectiveness and precision of the model. Recent research has also focused on data block for big data, as seen in [11], where the researchers defined progressive blocking (PB), which was able to identify 93% of duplicates in the first third of the time it took to execute. A multi-phase blocking technique for identifying duplicates in huge data has also been proposed in [12]. Although the reduction of data has been extensively studied in the literature, massive data deduplication challenges have received very little attention. Aside from thatthe majority of current methods only addresses the duplicate discovery stage and do not address data deduplication in its entirety. Furthermore, as the data evolve over time, the model's prediction ability typically deteriorates.

To the greatest extent of our understanding, no deduplication strategies have been developed to deal with the model serving's deteriorating accuracy. Therefore, our goal in this study is to present a three-part end-to-end large-scale Data Deduplication system.

•Establishing a learning plan for deduplication to ensure improved deduplication performance and accuracy.

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 10, October 2023 ||

| DOI:10.15680/IJIRSET.2023.1210017 |

- Creating a more thorough framework for large data deduplication that includes all big data deduplication processes are handled in five steps: data preparation, data labeling, duplicate identification, data cleaning, and model monitoring via continual retraining.
- Putting forth a cutting-edge framework that performs better than the current approaches for large-scale data deduplication that utilize a semi-supervised learning approach. In the following section, we underline the importance of duplication in big data.

## III. METHODOLOGY

### A. *Cloud Computing:*

Cloud computing stands as a transformative technological paradigm, offering a spectrum of computing services via the internet, commonly referred to as "the cloud". This encompasses servers, storage, databases, networking, software, analytics, and more. Its advantages are manifold, including accelerated innovation, resource flexibility, and the economies of scale. Diverging from traditional IT models with physical data centers, cloud computing allows organizations to access and utilize computing resources from suppliers of pay-as-you-go cloud services.
On-demand self-service, which enables users to independently handle resources without direct involvement, is one of the cloud computing model's key features. It offers extensive network connectivity, allowing services to be accessible from various devices, fostering convenience and accessibility. Resource pooling employs multi-tenant models to share resources among multiple customers while maintaining rigorous isolation of data and applications. Rapid elasticity enables resources to scale swiftly, accommodating fluctuating workloads, a trait often referred to as "elasticity". Additionally, cloud usage is meticulously metered, with users billed according to their resource consumption, proving cost-effective.

The three primary categories into which cloud computing services can be separated are infrastructure as a service, also known as IaaS, platforms as an assistance, or PaaS, and software at a service, or SaaS. Users have control over operating systems & apps thanks to the virtualized computing resources provided by IaaS. PaaS offers a higher-level platform encompassing development tools, databases, and application hosting environments. SaaS delivers comprehensive software applications over the internet, with the provider managing all aspects, including maintenance and updates.

A public cloud, a private cloud, cloud hybridization, & multi-cloud deployment options are available. Public cloud resources are made available to everyone and are shared by many users. Private cloud infrastructure offers more control and privacy because it is only dedicated to one company. A dynamic & scalable environment is created by combining public & private clouds in hybrid clouds. Multi-cloud employs multiple providers for distinct workloads, leveraging each provider's unique strengths and mitigating vendor lock-in. These cloud models and deployment options offer diverse solutions to cater to varying organizational needs and preferences.


*Fig2:Architecture of Cloud Computing*

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 10, October 2023 ||

| DOI:10.15680/IJIRSET.2023.1210017 |

### B. *File Deduplication*

Data deduplication in computing is a crucial method for removing duplicate copies of data that repeats itself, improving storage effectiveness. This reduction in duplicate data not only optimizes storage utilization but also lowers capital expenditure by minimizing the required storage media. Additionally, it can be applied to network data transfers, diminishing the volume of bytes that need to be transmitted.

This process involves identifying and storing unique, contiguous data chunks, comparing them to existing data, and replacing redundant chunks with a small reference. Given the frequency with which the same byte patterns can occur, the reduction in stored or transferred data can be substantial.

NetApp employs deduplication as a zero data-loss technology, utilizing both inline and background processes to maximize savings without interfering with client operations. The performance impact is minimal as it operates in a separate efficiency domain, independent of client activities. Deduplication savings persist as data moves between locations, ensuring efficiency in replication, backups, and transitions across different cloud environments.

### C. *MD5 Algorithm:*

The secure hashing technique MD5 accepts messages of any length and transforms them into messages of a predetermined length of sixteen bits. The MD5 algorithm is another name for the message-digest technique. MD5 was developed as an improvement on MD4 with more complex security goals. The output for the digestion size (MD5) is still 128 bits. In 1991, Ronald Rivets developed MD5.

### D. *Use of MD5 Algorithm:*

- It is used for file authentication.
- It is used in online applications to improve security. Secure user credentials, for instance, etc.
- This solution allows us to save our login information in a 128-bit format.

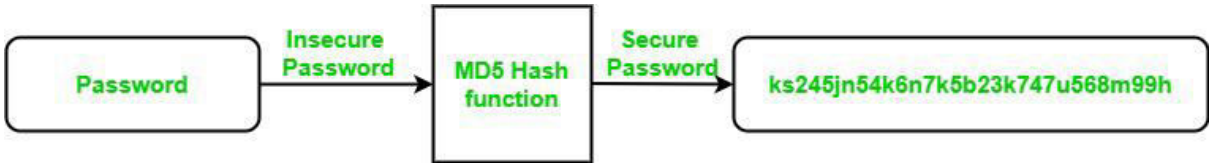


**Fig 3: MD5 Algorithm**

- By encoding a string as a 128-bit fingerprint, an MD5 hash of any length can be created. An identical string will always result in the same 128-bit hash value when encoded using the MD5 method. MD5 hashes are widely used with shorter strings when storing passwords, credit card a number, & various other secret information in database like the popular MySQL. This tool makes it simple and quick to produce an MD5 hash from a short string of up to 256 characters.

- MD5 hashes are also used to guarantee the veracity of data in files. Since the MD5 hash technique always produces the same output for the same input, users can verify that the file is intact and undamaged by comparing a hash of the original document with a recently created hash of a file being exported.

- An MD5 hash is not used for encryption. It just serves to identify the data that was given. However, as it is a one-way deal, it is very challenging to decipher the original string from an MD5 hash.

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 10, October 2023 ||

| DOI:10.15680/IJIRSET.2023.1210017 |

### *E. Local storage:*

Local storage encompasses data directly linked to a computing device or accessible via a local network. This includes JavaScript web storage for saving data on a user's device. It's stored on servers, usually at a business's premises. In contrast, cloud storage disperses data across multiple servers, offering accessibility from anywhere. Cloud storage is also more cost-effective and responsive compared to local storage.

### *F. Public Storage / Thirdy Party Storage*

Businesses and customers can easily rent storage space for their electronic information through public cloud storage. It is divided into two groups in this article: file storage for collaboration and sharing and storage for commercial applications. User-friendly file repositories are available through services from organisations including Dropbox, Google, Microsoft, & others. They serve people and businesses of all sizes. Storage as a Service (SaaS), a service designed for business applications and supported by major IT companies like Microsoft, Amazon, & Google, supports the development of software, data analytics, databases, & enterprise workloads. This article addresses this category of free cloud storage in particular. The following traits of free cloud storage are typical:

- Simple deployment and self-provisioning capability
- Rich APIs for highly automated management
- Dynamic performance scaling and capacity scaling
- Flexible storage options with access to files, objects, and blocks of storage

### *G. File Splitting and Merging*

- File split is the process of dividing a large file into multiple files so that each person can work on a different file. This increases effectiveness and facilitates work. In essence, it involves breaking up your code into different files as opposed to having it all in one. Using the results of any number of group variables, the Split file process divides the data files into separate sections for examination. Cases are grouped by each grouping variable inside the categories of the parent variable if you choose multiple grouping variables. For instance, cases will be categorised by minority category within every gender category if you choose gender as the primary grouping variable & minority as the second.
- Merging files is the complementary process to file splitting. It involves combining individual files or datasets into a unified whole. This consolidation is particularly useful when you want to aggregate the work or data from multiple sources or contributors back into a single, coherent file. Merging allows you to reconstruct the original large file or dataset by bringing together the smaller files that were previously split. The merge file operation typically requires some criteria or keys that align the records from different files correctly. These criteria ensure that the data is combined accurately and that the relationships between the records are preserved. Once merged, the resulting file can be used for various purposes, such as comprehensive analysis, reporting, or sharing the consolidated information with others.

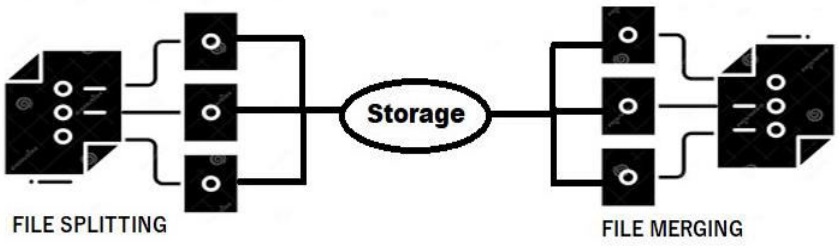


**Fig 4: File Splitting and Merging Process**

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 10, October 2023 ||

| DOI:10.15680/IJIRSET.2023.1210017 |

## IV. PROPOSED WORK

Cloud computing has made storing large amounts of data easier, but ensuring data security and efficient storage can be challenging. Encrypting data for protection can create redundancy issues. Our proposed system addresses this problem by breaking it into manageable parts or modules, simplifying testing and maintenance, system architecture of proposed system is shown in Fig 5.

To enhance the efficiency and manageability of implementing Cloud Computing Heterogeneous Storage of Data Management with Deduplication, the system is divided into distinct modules for easier testing, integration, and code maintenance. This proposed system consists of four main modules:

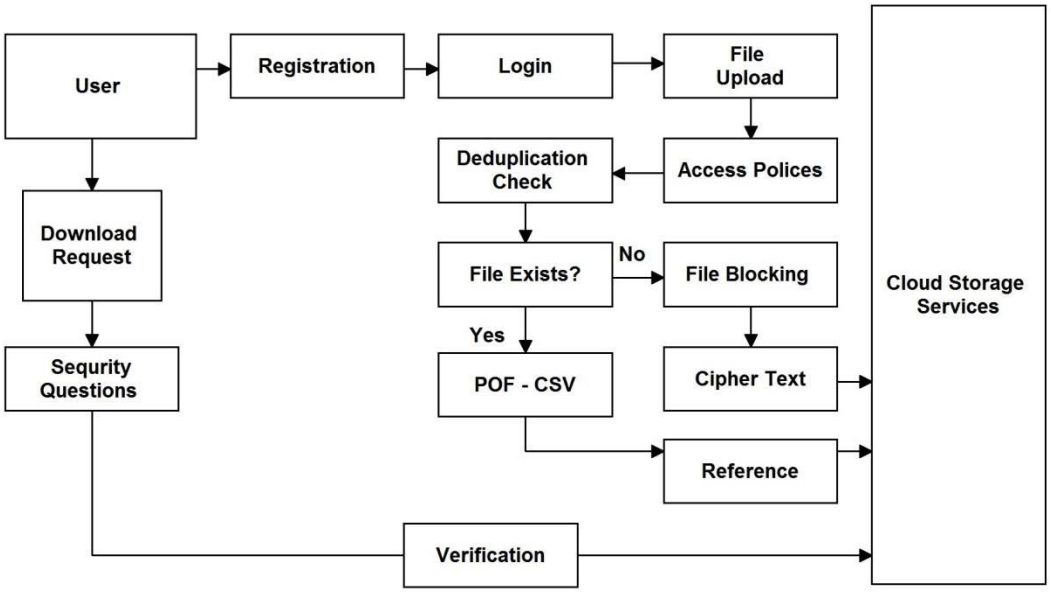


*Fig 5: System Architecture*

**Module 1: Cloud User Authentication**

This module deals with the initial registration process for cloud service providers (CSPs). Owners go through a registration process where they provide their personal information. The server stores this information in its database. Following registration, users can log in to access the cloud service provider's features.

**Module 2: File Upload and Comparison**

In this module, data owners create their accounts within the public cloud and upload files to cloud storage. We introduce the Provable Ownership of the File (POF) scheme for this purpose. When a data owner uploads a file, a unique hash key is generated using the MD5 algorithm. This hash key is distinct for each uploaded file. If another data owner attempts to upload the same file, the system prevents duplication by replacing it with a reference ID through a mapping of indices. Additionally, the module checks for the physical presence of the file by both data owners.

**Module 3: Set Access Policy for File**

This module allows users to select a file and upload it to storage within the Cloud Distributed File System (CDFS). The system generates a signature for each file, which is then divided into multiple blocks. Each block is assigned a unique signature with a key. A 128-bit hash value is generated using the MD5 message-digest method, Computing. Commonly expressed as a 32-digit hexadecimal value in text format, ensuring that files with the same content are deduplicated. Subsequently, convergent keys are generated for each split block, facilitating storage in a CSV file that includes information such as filename, file path, blocks, username, password, and block keys.

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 10, October 2023 ||

| DOI:10.15680/IJIRSET.2023.1210017 |

**Module 4: File Download Request and Handling**

In this module, data owners can download files from the cloud service provider. If they cannot locate the desired file, they have the option to request its download from a different cloud service provider. The system then verifies the file's presence and responds to the data owner accordingly.

## V. RESULT & DISCUSSION

The performance evaluation outcomes are presented in Table 1, which shows the efficiency and cost-effectiveness of our proposed system. Fig 6 shows the pictorial view of system performance.

The table 1 illustrates the comparison between actual storage needs and the storage requirements after employing MapReduce with deduplication in both local and third-party cloud environments. Notably, in the local storage scenario, we achieved a substantial storage reduction of 34.62%, while in the third-party cloud environment, an impressive 44.44% reduction was observed. These findings clearly demonstrate the effectiveness of our deduplication approach in optimizing storage utilization in heterogeneous cloud environments.

**Table 1: Performance Evaluation**

| Cloud | Actual Storage (MB) | Map Reduce Storage (MB) | Map Reduce Storage % | Saved % |
|---|---|---|---|---|
| Local | 520 | 340 | 65.38 | 34.62 |
| Third Party | 450 | 250 | 55.56 | 44.44 |

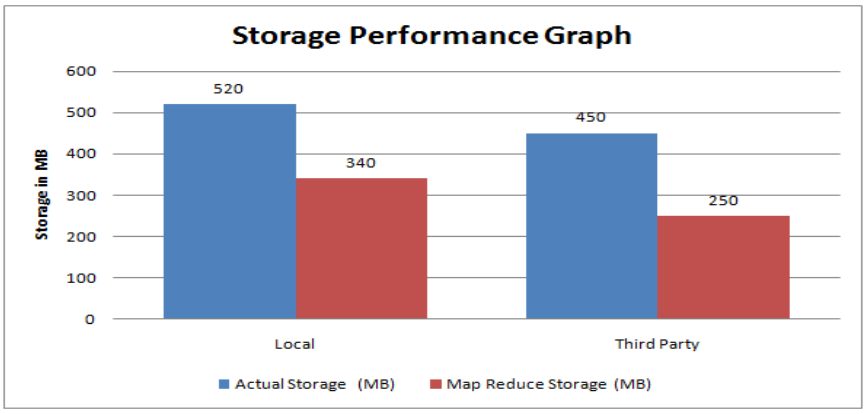


**Fig 6: Comparison Graph for Storage Performance**

## VI. CONCLUSION

In conclusion, our research addresses a critical aspect of cloud computing, emphasizing the significance of efficient storage management in a secure and dynamic environment. The integration of deduplication and access management among several Cloud Service Providers (CSPs) marks a notable advancement, offering adaptability and enhanced data sensitivity handling. Our comprehensive evaluation affirms the robustness and efficiency of this approach, validating its suitability for practical deployment. By mitigating the challenges posed by encryption-induced storage inefficiencies, our solution aligns with the evolving needs of cloud users. This contribution not only enhances data confidentiality and integrity but also fosters more streamlined and accessible cloud storage solutions, ultimately bolstering the broader efficacy of cloud computing services.

Future work in this domain could focus on several key areas. Further refinement and optimization of the proposed framework could be pursued, with a specific emphasis on enhancing scalability to accommodate even larger and more complex cloud environments. Additionally, exploring strategies to dynamically adjust deduplication policies based on evolving data sensitivity requirements would be valuable. Finally, examining the potential environmental impact of deduplication

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 10, October 2023 ||

| DOI:10.15680/IJIRSET.2023.1210017 |

strategies, particularly in terms of energy consumption and carbon footprint reduction, would align with growing concerns for sustainable cloud computing practices.

## REFERENCES

1. O. Azeroual, M. Jha, A. Nikiforova, K. Sha, M. Alsmirat, and S. Jha, 'A Record Linkage-Based Data Deduplication Framework with DataCleaner Extension', Multimodal Technol. Interact., vol. 6, no. 4, Art. no. 4, Apr. 2022, doi: 10.3390/mti6040027.
2. Simonini, Giovanni; Saccani, Henrique; Gagliardelli, Luca; Zecchini, Luca; Benevento, Domenico; Bergamaschi, Sonia. The Case for Multitask Active Learning Entity Resolution / (2021).
3. M. Pikies and J. Ali, 'Analysis and safety engineering of fuzzy string matching algorithms', ISA Trans., vol. 113, pp. 1–8, Jul. 2021, doi: 10.1016/j.isatra.2020.10.014.
4. Y. Gahi and I. El Alaoui, 'Machine Learning and Deep Learning Models for Big Data Issues', in Machine Intelligence and Big Data Analytics for Cybersecurity Applications, Y. Maleh, M. Shojafar, M. Alazab, and Y. Baddi, Eds. Cham: Springer International Publishing, 2021, pp. 29–49. doi: 10.1007/978-3-030-57024-8_2.
5. I. El Alaoui, Y. Gahi, R. Messoussi, A. Todoskoff, and A. Kobi, 'Big Data Analytics: A Comparison of Tools and Applications', in Innovations in Smart Cities and Applications, Cham, 2018, pp. 587–601. doi: 10.1007/978-3-319-74500-8_54.
6. S. Yeddula And K. Lakshmaiah, 'Investigation Of Techniques For Efficient & Accurate Indexing For Scalable Record Linkage & Deduplication', Int. J. Comput. Commun. Technol., Vol. 6, No. 1, Sep. 2020, Doi: 10.47893/Ijcct.2015.1275.
7. X. Chen, 'Towards Efficient and Effective Entity Resolution for HighVolume and Variable Data', p. 167, 2020, doi: 10.25673/35204
8. El-Ghafar, R.M., El-Bastawissy, A.H., Nasr, E.S., Gheith, M.H., & Independent Researcher, C.E. (2021). An Effective Entity Resolution Approach for Big Data. International Journal of Innovative Technology and Exploring Engineering.
9. V. Efthymiou, K. Stefanidis, and V. Christophides, 'Big data entity resolution: From highly to somehow similar entity descriptions in the Web', in 2015 IEEE International Conference on Big Data (Big Data), Oct. 2015, pp. 401–410. doi: 10.1109/BigData.2015.7363781.
10. A. Ngueilbaye, H. Wang, D. A. Mahamat, and I. A. Elgendy, 'SDLER: stacked dedupe learning for entity resolution in big data era', J. Supercomput., vol. 77, no. 10, pp. 10959–10983, Oct. 2021, doi: 10.1007/s11227-021-03710-x.
11. T. Papenbrock, A. Heise, and F. Naumann, 'Progressive Duplicate Detection', IEEE Trans. Knowl. Data Eng., vol. 27, no. 5, pp. 1316– 1329, May 2015, doi: 10.1109/TKDE.2014.2359666.
12. El-Ghafar, R. M. A., El-Bastawissy, A. H., Nasr, E. S., &Gheith, M. H. (2020). An Efficient Multi-Phase Blocking Strategy for Entity Resolution in Big Data. In International Journal of Innovative Technology and Exploring Engineering (Vol. 9, Issue 9, pp. 254–263).

INNO SPACE
SJIF Scientific Journal Impact Factor

Impact Factor: 8.423

doi cross ref

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर NISCAIR

# INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

## IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details