DATA QUALITY IN IOT OPEN DATASETS: A METHODOLOGICAL REVIEW

S.N.B.M. Isa[1], N.A. Emran[1]

[1]Fakulti Teknologi Maklumat dan Komunikasi,
Universiti Teknikal Malaysia Melaka, Hang Tuah Jaya, 76100 Durian Tunggal, Melaka, Malaysia.

Corresponding Author's Email: [1]nurulakmar@utem.edu.my

Article History: Received 18 April 2024 ; Revised 30 April 2024; Accepted May 2024

ABSTRACT: The growth of the Internet of Things (IoT) has significantly increased data generated by connected devices, leading to challenges in data duplication that threaten data quality and reliability. The purpose of this study is to assess and thoroughly examine the quality of open-source IoT datasets, focusing on the occurrence and impact of duplicate data. By employing a Systematic Literature Review (SLR) and a literature-based comparative analysis, we reviewed and compared existing techniques for detecting these issues. Our findings reveal that while various methods have been proposed, there remains a lack of standardized approaches specifically designed for the unique characteristics of IoT environments. The study concludes by highlighting the need for more reliable and scalable solutions that are capable of handling the diverse and dynamic nature of IoT data, also offering insights into future research directions.

KEYWORDS: *IoT Data Quality, Open-Source Datasets, Data Duplication, Systematic Literature Review (SLR) and Comparative Study*

1.0 INTRODUCTION

The Internet of Things (IoT) has transformed how we interact with technology, enabling devices from industrial sensors to home appliances to communicate and share data autonomously [1], [2]. This vast network generates large amounts of data, driving automation and smarter decision-making, but also raises challenges like data

duplication [3].
With IoT devices expected to grow from 15.9 billion in 2023 to 32.1 billion by 2030 [46], managing and ensuring data quality is critical. Data duplication, in particular, impacts the reliability and usefulness of IoT datasets, especially in open-source datasets [4]. Despite its importance, systematic studies on data duplication in open-source IoT datasets are limited.

This review addresses the gap by exploring current studies and methods for detecting data duplication in IoT data. It evaluates detection approaches, highlights the impact of duplication, and stresses the need for better solutions.

Key contributions include analyzing data duplication in open-source IoT datasets to improve data quality and reliability. The review also provides practical insights for researchers, offering data management strategies that can enhance IoT technologies across various domains. Ultimately, this review serves as a guide for future research in IoT data quality management.

2.0 LITERATURE REVIEW

2.1 IoT data types and characteristics

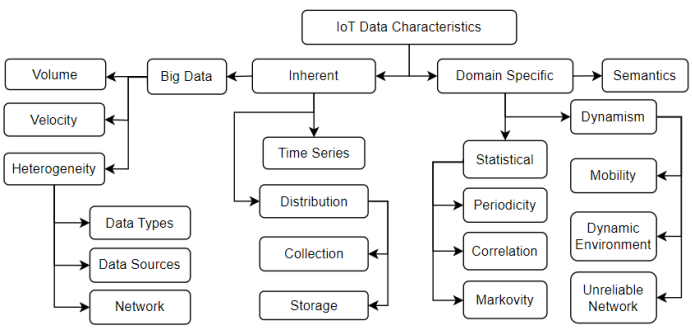

Figure 1: IoT Data Characteristics

IoT data has distinct characteristics that set it apart from traditional big data. [5] classify these into Inherent and Domain-Specific Features. Inherent features include high volume, velocity, and heterogeneity, as IoT data comes from various types and sources. [6] highlight that IoT devices produce structured, semi-structured, and unstructured data, while formats range from numerical to multimedia [7].
Domain-specific features like mobility, dynamic conditions, and

unreliable networks reflect IoT's constantly changing environment. Additionally, IoT data shows statistical properties such as periodicity and correlation, with continuous streams generated by sensor devices, forming time series data [8], [9].

2.2 Open-Source Datasets

Open-source IoT datasets play a crucial role in IoT research and development, providing accessible data for analysis and innovation [10]. [11] highlight their value in offering realistic traffic patterns, network performance, and user demand scenarios. However, the quality of these datasets is a critical concern. [12] and [13] identify main quality issues such as outliers, duplicates, anomalies, and missing values due to massive amount of data by IoT devices, that often leads to noise, uncertainty, and incompleteness which can undermine data reliability and result in poor decision-making [14].

2.3 Data Duplication

Data duplication is a major issue in IoT datasets, varying across different domains. [15] define duplication in big data as redundant entries referring to the same real-world entity. In edge computing, [16] describe duplicates as adjacent records with identical timestamps. Similarly, in environmental datasets, [17] identify redundancy as multiple items sharing the same timestamp.

In the IoT context, data duplication takes on specific forms, reflecting the unique characteristics of IoT data. Table 1 shows the categories of duplicate data in IoT and its example.

Table 1: Categories of duplicate data in IoT

Category	Definition & Example
Exact Duplicates	Records identical in all aspects, including data values and timestamps, can be from multiple sensors, thereby skewing data accuracy and reliability [18].
Example	Location: 1, Sensor: A, Temperature: 25.0°C, Humidity: 60%, Time: 12:00 PM Location: 1, Sensor: A, Temperature: 25.0°C, Humidity: 60%, Time: 12:00 PM
Near Duplicates	Records with minor variations but representing the same real-world entity or data points with slightly different content, where retaining any one does not impact the analysis outcome [19].
Example	Location: 2, Sensor: B, Temperature: 24.8°C, Humidity: 58%, Time: 12:05 PM Location: 2, Sensor: C, Temperature: 25.1°C, Humidity: 57%, Time: 12:05 PM
Spatial Duplicates	Data from multiple sensors in close proximity reporting similar information, potentially causing inaccuracies in data analysis [20].

Example	Location: Park, Coordinates: (40.7128, -74.0060), Sensor: D, Temperature: 23.0°C, Humidity: 55%, Time: 12:10 PM Location: Park, Coordinates: (40.7129, -74.0061), Sensor: E, Temperature: 23.0°C, Humidity: 55%, Time: 12:10 PM
Temporal Duplicates	Nearly identical data is reported over short time intervals by the same sensor, collected at different time points, which can occur in time-series databases and impact the accuracy of data analysis [21], [22].
Example	Location: 3, Sensor: F, Temperature: 26.0°C, Humidity: 62%, Time: 12:15:01 PM Location: 3, Sensor: F, Temperature: 26.0°C, Humidity: 62%, Time: 12:15:02 PM
Semantic Duplicates	Data points representing the same information but expressed differently [23]), occur due to variations in data formats, naming conventions, or contextual usage of terms across different datasets [24].
Example	Location: 4, Sensor: G, Temperature: 22.0°C, Humidity: 65%, Time: 12:20 PM Location: 4, Sensor: H, Temperature: 71.6°F, Humidity: 65%, Time: 12:20 PM

The impact of data duplication is broad, affecting data quality, consistency, analysis, and decision-making. The following table summarizes the key impacts of data duplication across different domains.

Table 2: Summary of key impacts of data duplication

Impact	Descriptions
Data Quality and Consistency	•Challenges in updating/ modifying across every instance •Propagation of errors and misreporting of information •Erroneous results in analysis •Negative impact on data utilization and decision quality [3], [25], [26], [27]
Analysis and Decision-Making	•Difficult to identify recent and reliable data across multiple instances •Anomalies and unexpected query results •Distort analysis results and inflated statistical measures lead to inaccurate conclusions •Affecting reliability of data-driven decision [3], [28], [29]
Resource Utilization and Efficiency	•Utilizes high storage space •Higher energy consumption and additional communication costs •Inefficiencies [30], [31]
Genetic and Biological Implications	•Selection advantage in changing environments •Increased genetic diversity and appearance of new functions •Evolution of viruses (like HIV) and complex traits (such as fertility) [32], [33], [34]
Signal Processing	• Add redundant information to prevent channel failures and packet losses, allowing for error detection and correction to reconstruct the received signal with minimal distortion [35]

Despite these challenges, addressing data duplication through deduplication techniques offers numerous benefits across various domains. The following table summarizes the benefits of eliminating data duplication.

Table 3: Summary of benefits of data deduplication

Category	Description

Cost	Helps lower storage costs & costs associated with maintaining duplicated data [30], [36], [37], [38]
Storage	Effective data storage, keeps only useful data in storage, saving storage space [3], [38]
Data Management	•efficient data migrations, reducing the volume of data to be transferred [38] •allow easy updating and cross-referencing [27] •efficient data updating, error-free query processing [3]
Performance	•improves backup and restore time, improving performance and efficiency [38] •reduces data volume, enhancing overall system's performance [30]
Miscellaneous	•facilitating effective disaster recovery by replicating data after removing redundant data [38] •keep precise info to help medical professionals focus on only critical readings [30]

The unique characteristics of IoT data, combined with the prevalence of data quality issues in open-source datasets, underscore the need for effective strategies to manage and mitigate data duplication. This paper will compare existing methods that allow researchers and practitioners to make informed decisions about which approaches are suitable for specific IoT applications and data characteristics.

3.0 METHODOLOGY

This study employs a Systematic Literature Review (SLR) to analyze research on data duplication detection in IoT open-source datasets. Searches were conducted in IEEE Xplore, ACM Digital Library, Scopus, and MDPI, utilizing keywords such as "Internet of Things" OR "IoT," AND ("data quality" OR "data integrity"), AND ("open-source dataset"), AND ("duplicate data" OR "data redundancy"). This yielded 110 studies, with 95 unique records remaining after duplicate removal. The screening process involved two stages: an initial review of titles and abstracts that excluded 30 irrelevant papers, followed by a full-text review of 65 studies. Ultimately, 47 studies relevant to data duplication in IoT datasets were selected.

Data was extracted using a standardized form, focusing on study details, methods for handling duplicates, and key findings. This information was synthesized to identify themes, trends, and gaps in research. A comparative analysis evaluated methods for detecting and managing data duplication, categorizing them by strengths and limitations to provide insights into their effectiveness in IoT environments.

4.0 RESULT AND DISCUSSION

The analysis of the literature reveals three primary categories of duplicate data detection methods: hash-based methods, content-aware methods, and hybrid/advanced approaches. Each category has its strengths and limitations, making them suitable for different scenarios within IoT data management.

4.1 Hash-based Methods

Hash-based methods calculate unique hash values or fingerprints for each data chunk or file. [39] explain this process in their study for cloud storage, while [37] used SHA-256 (Python code) to remove duplicate files, keeping only the latest copy.

Advantages:

- Quick identification of identical data
- Low computational cost
- Efficient for exact duplicate detection

Limitations:

- Ineffective for detecting near or semantically similar data
- Vulnerable to hash collisions with weaker hash functions

These methods are ideal for exact duplicate detection, such as in backup systems or storage optimization [38]. However, [40] note that these methods may not be effective in detecting similar duplicates or near-duplicates that have minor differences in content.

4.2 Content-aware Methods

Content-aware methods analyze data content to detect similarities, employing techniques from fields like computer vision, NLP, and machine learning [26], [40], [41]. [26] proposed the Multidimensional Similarity Redundancy Detection Algorithm (MSRD), that combines numerical similarity, literal similarity, and semantic similarity for comprehensive duplicate detection. [41] developed a two-stage detection approach utilizing locality-sensitive hashing (LSH) to efficiently select candidate pairs in RDF data and perform similarity analysis on the selected pairs to identify duplicates. [40] introduced the CompoundEyes method, which detects near-duplicate videos by utilizing multiple features and classifiers in parallel.

Advantages:

- Effective for near-duplicates and semantically similar data
- Comprehensive similarity calculations
- Suitable for complex data types and structures

Limitations:

- Computationally intensive
- Requires significant processing power and resources
- Accuracy can be influenced by the quality of data

Content-aware methods excel in detecting near-duplicates or semantically similar data, that ideal in plagiarism detection, copyright infringement detection, or data integration [41], [42].

4.3 Hybrid and Advanced Approaches

Hybrid and advanced approaches combine multiple techniques or incorporate additional processing steps to enhance data redundancy detection accuracy and efficiency. These methods often aim to address specific challenges or requirements, such as handling frequently modified data, dealing with incomplete or missing data, or optimizing performance for large-scale datasets.

[43] proposed the Deduplication-Aware Resemblance Detection and Elimination (DARE) scheme, using duplicate-adjacency (DupAdj), an improved super-feature approach, followed by delta compression to remove redundancy. [44] developed a Deduplication Framework with Metric Functional Dependencies (MFDs) and a weighting scheme for term frequency. [45] introduced the Duplicate Detection within Incomplete Datasets (DDID) method, which selects attributes for generating sort keys for clustering and comparison strings for matching records and Hot Deck imputation to compensate for missing values.

Advantages:

- Handles complex challenges (e.g., modified or incomplete data)
- Potentially more robust and accurate detection
- Suitable for complex scenarios with specific requirements

Limitations:

- Complex to implement
- Requires additional computational resources and expertise
- May be overkill for simpler duplicate detection needs

The effectiveness of these methods depends on the specific use case. For example, the DARE scheme by [43] may be effective for backup or archiving systems that deal with frequently changing data, while the Deduplication Framework proposed by [44] can be effective in scenarios where numeric data or term frequencies play a crucial role in detecting duplicates.

Table 4: Summary of key impacts of data duplication

Category	Advantages	Disadvantages	Suitable Scenarios	References
Hash-based Methods	- Quick identification of identical data - Low computational cost - Efficient for exact duplicate detection	- Ineffective for detecting near-duplicates - Vulnerable to hash collisions with weaker hash functions	- File deduplication in backup systems - Storage optimization	[37], [38], [39], [40]
Content-aware Methods	- Effective for detecting near-duplicates and semantically similar data - Comprehensive similarity calculations	- Computationally intensive - Requires significant processing power and resources - Accuracy influenced by quality of data	- Plagiarism detection - Copyright infringement detection - Data integration tasks	[26], [40], [42], [45]
Hybrid/ Advanced Methods	- Addresses specific challenges (e.g., frequently modified data, incomplete data) - Robust, accurate detection	- Complex to implement - Requires additional computational resources and specialized knowledge	- Scenarios with specific requirements or challenges - Applications needing robust and accurate duplicate	[41], [43], [44]

The choice of duplicate detection method in IoT depends on the application's needs. Hash-based methods work well for storage optimization, while content-aware or hybrid methods are better for detailed duplicate detection, such as data integration or handling sensor data from multiple sources. Hybrid and advanced methods are increasingly popular as they balance computational efficiency with detection accuracy, addressing challenges in frequently modified or incomplete datasets. As IoT technologies evolve, ensuring data quality must also progress. Future research should aim for scalable, adaptive methods that uphold accuracy as IoT datasets expand in size and complexity.

5.0 CONCLUSIONS

This review paper highlights the multilayered nature of data duplication across multiple domains, emphasizing its impact in IoT environments. The analysis reveals a spectrum of duplicate detection methods, from basic hash-based techniques to advanced content-aware and hybrid approaches. Each method presents unique advantages and limitations, highlighting the need for context-specific solutions to manage data redundancy. Developing adaptive and scalable detection methods is crucial for future research, aiming to balance computational efficiency with detection accuracy in increasingly complex data ecosystems as IoT technologies evolve.

ACKNOWLEDGMENTS

This work was conducted independently and did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

REFERENCES

[1] V. D. Gowda, V. N. R. Bandaru, A. Y. Begum, D. Palanikkumar, and A. C. Jadhav, "Internet of Things (IoT): Definitions, Components, Characteristics and Applications," in *Current Overview on Science and Technology Research Vol. 8*, 2022. doi: 10.9734/bpi/costr/v8/3535c.

[2] K. Rajamohan, S. Rangasamy, N. A. Pinto, B. E. Manoj, D. Mukherjee, and J. Shukla, "IoVST: Internet of vehicles and smart traffic - Architecture, applications, and challenges," in *Handbook of Research on Machine Learning-Enabled IoT for Smart Applications Across Industries*, 2023. doi: 10.4018/978-1-6684-8785-3.ch015.

[3] T. T. Nguyen, T. T. Huynh, M. T. Pham, T. D. Hoang, T. T. Nguyen, and Q. V. H. Nguyen, "Validating functional redundancy with mixed generative adversarial networks," *Knowl Based Syst*, vol. 264, p. 110342, Mar. 2023, doi: 10.1016/J.KNOSYS.2023.110342.

[4] T. Mansouri, M. Reza, S. Moghadam, F. Monshizadeh, and A. Zareravasan, "IoT Data Quality Issues and Potential Solutions: A Literature Review," *Comput J*, vol. 66, no. 3, pp. 615–625, Mar. 2023, doi: https://doi.org/10.48550/arXiv.2103.13303.

[5] N. Zubair, N. A, K. Hebbar, and Y. Simmhan, "Characterizing IoT Data and its Quality for Use," Jun. 2019, [Online]. Available: http://arxiv.org/abs/1906.10497

[6] A. Singh and S. Mahapatra, "Network-based applications of multimedia big data computing in IoT environment," in *Intelligent Systems Reference Library*, vol. 163, 2020. doi: 10.1007/978-981-13-8759-3_17.

[7] S. Vongsingthong and S. Smanchat, "A Review of Data Management in Internet of Things," *KKU Research Journal*, vol. 20, no. 2, 2015.

[8] A. Chowdhury, A. Pal, A. Raut, and M. Kumar, "KIHCDP: An Incremental Hierarchical Clustering Approach for IoT Data Using Dirichlet Process," *IEEE Access*, vol. 12, pp. 56019–56032, 2024, doi: 10.1109/ACCESS.2024.3385628.

[9] D. Puschmann, P. Barnaghi, and R. Tafazolli, "Adaptive Clustering for Dynamic IoT Data Streams," *IEEE Internet Things J*, vol. 4, no. 1, 2017, doi: 10.1109/JIOT.2016.2618909.

[10] D. Salian, "Usability of Open Data," in *Open-Source Horizons - Challenges and Opportunities for Collaboration and Innovation*, 2023. doi: 10.5772/intechopen.1003269.

[11] H. Shahid, M. Angel Vázquez, L. Reynaud, F. Parzysz, and M. Shaat, "Open Datasets for AI-Enabled Radio Resource Control in Non-Terrestrial Networks,"

2024. doi: 2404.12813.

[12] Y. Bertrand, R. Van Belle, J. De Weerdt, and E. Serral, "Defining Data Quality Issues in Process Mining with IoT Data," in *Lecture Notes in Business Information Processing*, 2023. doi: 10.1007/978-3-031-27815-0_31.

[13] J. N. Kabi, C. W. Maina, and E. T. Mharakurwa, *Anomaly Detection in IoT Data*. IEEE / 2023 IST-Africa Conference (IST-Africa), 2023.

[14] J. Byabazaire, G. M. P. O'Hare, R. Collier, and D. Delaney, "Dynamic Data Source Selection: A Case of Weather Stations for IoT Applications," in *2022 IEEE 8th World Forum on Internet of Things, WF-IoT 2022*, 2022. doi: 10.1109/WF-IoT54382.2022.10152030.

[15] E. Widad, E. Saida, and Y. Gahi, "Quality Anomaly Detection Using Predictive Techniques: An Extensive Big Data Quality Framework for Reliable Data Analysis," *IEEE Access*, vol. 11, pp. 103306–103318, 2023, doi: 10.1109/ACCESS.2023.3317354.

[16] S. Tverdal *et al.*, "Edge-based Data Profiling and Repair as a Service for IoT," in *ACM International Conference Proceeding Series*, Association for Computing Machinery, Nov. 2023, pp. 17–24. doi: 10.1145/3627050.3627065.

[17] J. H. Buelvas, D. Múnera, and N. Gaviria, "DQ-MAN: A tool for multi-dimensional data quality analysis in IoT-based air quality monitoring systems," *Internet of Things*, vol. 22, p. 100769, Jul. 2023, doi: 10.1016/J.IOT.2023.100769.

[18] W. P. Jiang, B. Wu, Z. Jiang, and S. B. Yang, "Cloning Vulnerability Detection in Driver Layer of IoT Devices," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2020. doi: 10.1007/978-3-030-41579-2_6.

[19] Y. Gao, L. Chen, J. Han, G. Wu, and S. Liu, "Similarity-based deduplication and secure auditing in IoT decentralized storage," *Journal of Systems Architecture*, vol. 142, 2023, doi: 10.1016/j.sysarc.2023.102961.

[20] H. Li, H. Lu, C. S. Jensen, B. Tang, and M. A. Cheema, "Spatial Data Quality in the Internet of Things: Management, Exploitation, and Prospects," *ACM Comput Surv*, vol. 55, no. 3, 2022, doi: 10.1145/3498338.

[21] S. An-Dong and Z. Fang, "Research on Open Source Solutions of Data Collection for Industrial Internet of Things," in *Proceedings - 2021 7th International Symposium on Mechatronics and Industrial Informatics, ISMII 2021*, 2021. doi: 10.1109/ISMII52409.2021.00045.

[22] C. Wang *et al.*, "Apache IoTDB: Time-series Database for Internet of Things," *Proceedings of the VLDB Endowment*, vol. 13, no. 12, 2020, doi: 10.14778/3415478.3415504.

[23] S. Aydin and M. N. Aydin, "Semantic and syntactic interoperability for agricultural open-data platforms in the context of IoT using crop-specific trait ontologies," *Applied Sciences (Switzerland)*, vol. 10, no. 13, 2020, doi: 10.3390/app10134460.

[24] S. Muralidharan, B. Yoo, and H. Ko, "Designing a semantic digital twin model for IoT," in *Digest of Technical Papers - IEEE International Conference on Consumer Electronics*, 2020. doi: 10.1109/ICCE46568.2020.9043088.

[25] X. Ding, H. Wang, G. Li, H. Li, Y. Li, and Y. Liu, "IoT data cleaning techniques: A survey," *Intelligent and Converged Networks*, vol. 3, no. 4, pp. 325–339, Dec. 2022, doi: 10.23919/ICN.2022.0026.

[26] Y. Long, H. Li, Z. Wan, and P. Tian, "Data Redundancy Detection Algorithm based on Multidimensional Similarity," in *Proceedings - 2023 International Conference on Frontiers of Robotics and Software Engineering, FRSE 2023*, Institute of Electrical and Electronics Engineers Inc., 2023, pp. 180–187. doi: 10.1109/FRSE58934.2023.00032.

[27] T. Searle, Z. Ibrahim, J. Teo, and R. Dobson, "Estimating redundancy in clinical text," *J Biomed Inform*, vol. 124, p. 103938, Dec. 2021, doi: 10.1016/J.JBI.2021.103938.

[28] D. Firmani, M. Mecella, M. Scannapieco, and C. Batini, "On the Meaningfulness of 'Big Data Quality' (Invited Paper)," *Data Sci Eng*, vol. 1, no. 1, pp. 6–20, Mar. 2016, doi: 10.1007/s41019-015-0004-7.

[29] E. Frank, "Machine Learning Models for Data Quality Assessment," 2024.

[30] M. N. U. Khan, W. Cao, Z. Tang, A. Ullah, and W. Pan, "Energy-Efficient De-Duplication Mechanism for Healthcare Data Aggregation in IoT," *Future Internet*, vol. 16, no. 2, Feb. 2024, doi: 10.3390/fi16020066.

[31] W. Zheng, Y. Li, X. Wu, and J. Cheng, "Duplicate Bug Report detection using Named Entity Recognition," *Knowl Based Syst*, vol. 284, p. 111258, Jan. 2024, doi: 10.1016/J.KNOSYS.2023.111258.

[32] S. Fouchécourt *et al.*, "Expanding duplication of the testis PHD Finger Protein 7 (PHF7) gene in the chicken genome," *Genomics*, vol. 114, no. 4, p. 110411, Jul. 2022, doi: 10.1016/J.YGENO.2022.110411.

[33] M. J. H. Girard-Madoux *et al.*, "The immunological functions of the Appendix: An example of redundancy?," 2018. doi: 10.1016/j.smim.2018.02.005.

[34] M. Müller and D. Sauter, "The more the merrier? Gene duplications in the coevolution of primate lentiviruses with their hosts," *Curr Opin Virol*, vol. 62, p. 101350, Oct. 2023, doi: 10.1016/J.COVIRO.2023.101350.

[35] R. Ma and F. Labeau, "A family of fast index and redundancy assignments for error resilient multiple description coding," *Signal Process Image Commun*, vol. 27, no. 6, pp. 612–624, Jul. 2012, doi: 10.1016/J.IMAGE.2012.01.020.

[36] A. H. Adhab and N. A. Hussien, "Techniques of Data Deduplication for Cloud Storage: A Review," *International Journal of Engineering Research and Advanced Technology (ijerat)*, vol. 8, no. 4, pp. 7–18, 2022, doi: 10.31695/IJERAT.2022.8.4.2.

[37] H. Deshingkar *et al.*, "Data Deduplication Using Python," in *2023 7th International Conference On Computing, Communication, Control And Automation, ICCUBEA 2023*, Institute of Electrical and Electronics Engineers Inc., 2023. doi: 10.1109/ICCUBEA58933.2023.10391968.

[38] M. U. Tahir, M. R. Naqvi, S. K. Shahzad, and M. W. Iqbal, "Resolving Data De-Duplication issues on Cloud," in *2020 International Conference on Engineering and Emerging Technologies (ICEET), Lahore, Pakistan*, Lahore, Pakistan: IEEE, 2020, pp. 1–5. doi: 10.1109/ICEET48479.2020.9048214.

[39] K. Vijayalakshmi and V. Jayalakshmi, "Analysis on data deduplication techniques of storage of big data in cloud," in *Proceedings - 5th International Conference on Computing Methodologies and Communication, ICCMC 2021*, Institute of Electrical and Electronics Engineers Inc., Apr. 2021, pp. 976–983. doi: 10.1109/ICCMC51019.2021.9418445.

[40] Y. Chen, D. Li, Y. Hua, and W. He, "Effective and Efficient Content Redundancy Detection of Web Videos," *IEEE Trans Big Data*, vol. 7, no. 1, pp. 187–198, May 2019, doi: 10.1109/tbdata.2019.2913674.

[41] Y. Chen, D. Li, L. Yan, and Z. Ma, "Two-stage Detection of Semantic Redundancies in RDF Data," *Journal of Web Engineering*, vol. 21, no. 8, pp. 2313–2337, 2022, doi: 10.13052/jwe1540-9589.2184.

[42] L. Lu and P. Wang, "Duplication Detection in News Articles Based on Big Data," in *2019 IEEE 4th International Conference on Cloud Computing and Big Data Analysis (ICCCBDA), Chengdu, China*, Chengdu, China,: IEEE, 2019, pp. 15–19. doi: 10.1109/ICCCBDA.2019.8725674.

[43] W. Xia, H. Jiang, D. Feng, and L. Tian, "DARE: A Deduplication-Aware Resemblance Detection and Elimination Scheme for Data Reduction with Low Overheads," *IEEE Transactions on Computers*, vol. 65, no. 6, pp. 1692–1705, Jun. 2016, doi: 10.1109/TC.2015.2456015.

[44] Y. Huang and F. Chiang, "Refining Duplicate Detection for Improved Data Quality," in *TDDL/MDQual/Futurity@TPDL*, CEUR-WS.org, 2017. Accessed: May 12, 2024. [Online]. Available: https://ceur-ws.org/Vol-2038/paper3.pdf

[45] A. Ali, N. A. Emran, and S. A. Asmai, "Missing values compensation in duplicates detection using hot deck method," *J Big Data*, vol. 8, no. 1, Dec. 2021, doi: 10.1186/s40537-021-00502-1.

[46] Statista. (2024). *Number of Internet of Things (IoT) connected devices worldwide from 2019 to 2023, with forecasts from 2022 to 2030*. [Online]. Available: https://www.statista.com/statistics/1183457/iot-connected-devices-worldwide/