International Journal of Hybrid Innovation Technologies
Vol.5, No.1 (2025), pp.1-10
http://dx.doi.org/10.21742/ijhit.2653-309X.2025.5.1.01

A Technical Architecture for the Integration of Knowledge Discovery in Databases and Data Warehousing Systems

Niamh O'Callaghan[1*] and Eoin Gallagher[2]

[1,2]*Senior Research Fellow, Centre for Data Analytics and Knowledge Systems, Trinity College Dublin, Ireland*
[1]*nocallaghan.research@tcd.ie,* [2]*egallagher.research@tcd.ie*

Abstract

The increasing demand for intelligent, data-driven decision-making has underscored the importance of integrating advanced analytical capabilities within enterprise data management systems. This paper presents a technical architecture that unifies Knowledge Discovery in Databases (KDD) with Data Warehousing (DW) systems to enable scalable, modular, and efficient extraction of actionable insights from large datasets. The proposed architecture addresses core integration challenges by organizing system components into five functional layers, encompassing data ingestion, transformation, storage, mining, and visualization. By embedding KDD processes directly into the data warehousing framework, the architecture supports continuous knowledge extraction while maintaining system performance and flexibility. The methodology includes system design, tool-based simulation, and performance evaluation using real-world and synthetic datasets to assess scalability, mining accuracy, and processing efficiency. Results demonstrate that the architecture achieves high mining accuracy, reduced data processing latency, and effective separation of concerns through modular components. Furthermore, the design accommodates near-real-time data flows, making it suitable for dynamic environments where timely insights are critical. The findings validate the feasibility and advantages of embedding intelligent analytical functions within warehouse infrastructures, bridging the gap between data storage and knowledge extraction. This work contributes a replicable model for organizations seeking to enhance their data architectures with integrated analytical capabilities. It lays the groundwork for future extensions involving cloud-native platforms and advanced AI-driven discovery techniques.

Keywords: *Knowledge Discovery in Databases (KDD), Data warehousing, Data integration architecture, Big data analytics, Enterprise decision support systems*

1. Introduction

The rapid digitization of services, industries, and interactions has resulted in an extraordinary accumulation of data. From business transactions and customer behaviors to sensor logs and social media content, organizations today are inundated with large-scale, heterogeneous data that holds tremendous potential value. However, the transformation of this raw data into meaningful insights and knowledge remains a major challenge in

Article Info:
Received (February 10, 2025), Review Result (April 1, 2025), Accepted (May 5, 2025)
*corresponding author

information systems. Two key domains, Data Warehousing (DW) and Knowledge Discovery in Databases (KDD) have emerged as complementary solutions to this problem. While data warehousing focuses on the organized, efficient storage and retrieval of structured data, KDD enables the extraction of patterns, trends, and relationships that are not immediately obvious through manual inspection.

Data warehouses serve as centralized repositories where data from multiple, often disparate, operational sources are integrated, cleaned, transformed, and stored in a format optimized for analytical querying. These systems are designed to support Online Analytical Processing (OLAP), Business Intelligence (BI), and retrospective trend analysis. With the evolution of cloud computing and distributed frameworks, modern data warehouses are now capable of handling real-time ingestion and supporting hybrid structured-semi-structured data, as seen in systems like Snowflake and Google BigQuery [2][5].

In parallel, KDD, which is often synonymous with or inclusive of data mining, encompasses a multi-stage process that includes data selection, preprocessing, transformation, pattern discovery through mining algorithms, and interpretation of results. It leverages statistical modeling, machine learning, and pattern recognition techniques to derive non-trivial insights from large datasets. Applications of KDD span numerous sectors, including fraud detection, healthcare diagnostics, customer behavior modeling, and recommendation systems [1][6].

Despite their complementary purposes, most organizations continue to deploy DW and KDD as separate or loosely coupled systems. This separation creates significant limitations, including duplicated storage, complex integration pipelines, slow knowledge discovery processes, and poor scalability when applied to dynamic, real-time environments [3][4]. Furthermore, the lack of a unified architecture introduces maintenance challenges and performance bottlenecks, especially when dealing with big data scenarios.

Recent literature highlights a growing interest in architectural frameworks that bridge the gap between KDD and DW, aiming to create intelligent, modular, and scalable systems. Researchers have proposed the use of hybrid cloud infrastructures [7], modular plug-in architectures for mining tools [8], and integration frameworks for real-time analytics [9]. However, most existing approaches lack flexibility, are limited to specific domains, or do not provide sufficient architectural detail to guide system designers and engineers in implementation.

This paper responds to these challenges by presenting a technical architecture that enables the seamless integration of KDD processes within a modern data warehousing ecosystem. The proposed architecture is designed to:

- Provide a modular and layered structure that allows flexible deployment and upgrading of components.
- Support scalable data ingestion and transformation across multiple sources and formats;
- Embed KDD processes (e.g., classification, clustering, association mining) directly into the warehouse pipeline, minimizing latency and data duplication;
- Incorporate real-time and batch analytics using a hybrid orchestration strategy;
- Enable knowledge visualization and interpretation, making insights accessible to decision-makers.

The conceptual relationship between data warehousing and KDD in the context of an integrated system is illustrated in [Figure 1]. It highlights how data sources flow through warehousing mechanisms into discovery processes, ultimately enabling pattern analysis, classification, and prediction.

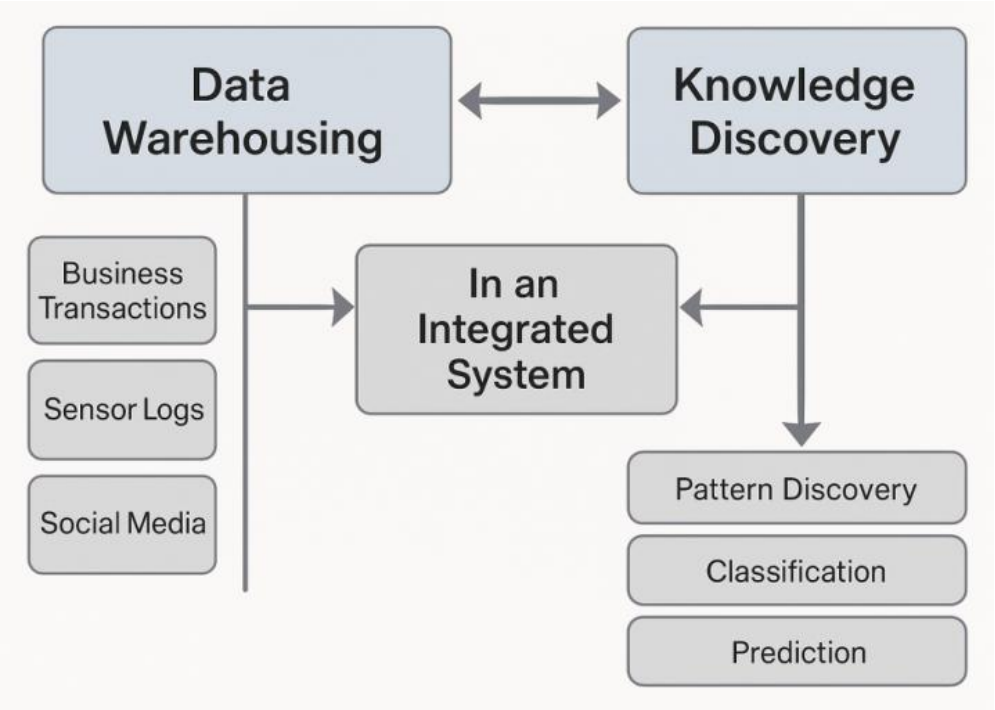

Figure 1. Conceptual relationship between data warehousing and knowledge discovery in an integrated system

2. Related works

The integration of Knowledge Discovery in Databases (KDD) and Data Warehousing (DW) continues to evolve, driven by the growing demand for real-time analytics, scalability, and intelligent automation. While traditional data warehouses were designed for structured storage and historical reporting, modern trends necessitate agile platforms that incorporate mining, learning, and inference capabilities.

A study by Nemati et al., [11] explored the limitations of traditional ETL processes in adapting to real-time mining workloads. They highlighted the need for event-driven architectures to reduce latency and enable continuous knowledge extraction. Foidl et al., [12] proposed modular KDD pipelines that operate directly within the warehouse, advocating for standardized APIs to simplify algorithm integration and reuse. Wang et al. [13] introduced a high-performance fraud detection system using in-memory data warehousing and embedded micro-batch mining techniques. Their architecture proved effective in reducing detection times in financial applications. Similarly, Saleh et al. [14] applied a layered rule-mining model within a smart grid data warehouse, enabling accurate demand forecasting and outage prediction.

In the industrial Internet of Things (IIoT) domain, Sunhare et al. [15] proposed a time-series–aware DW model that incorporates predictive analytics for monitoring sensor behavior in manufacturing plants—their hybrid framework integrated dimensional modeling with anomaly detection modules. Sharma et al., [16] developed an agricultural analytics platform that unified weather, soil, and crop data through a DW–KDD framework, allowing farmers to optimize planting schedules based on predictive insights. Saboor et al., [17] introduced a containerized architecture using Kubernetes to orchestrate warehouse-native KDD services.

This approach improved fault tolerance and enabled dynamic algorithm deployment. Dinesh and Devi [18] demonstrated the use of low-code platforms to allow domain experts to create and configure KDD processes directly, reducing development cycles and reliance on data engineers.

To address regulatory and security concerns, Cambronero et al., [19] proposed a GDPR-aligned mining pipeline that operates over encrypted and masked data within the warehouse. Their approach maintained analytical utility while ensuring user privacy. Siqueira et al. [20] benchmarked several commercial and open-source KDD–DW integration solutions, revealing trade-offs between throughput, modularity, and customization. A semantic integration approach was presented by Doncevic et al., [21], where ontology-based metadata layers facilitated schema matching and mining task automation across heterogeneous warehouse systems. Finally, Trindade et al., [22] explored edge-enhanced federated warehousing systems, where KDD tasks were executed close to data sources, reducing bandwidth demands and improving local decision-making in distributed environments.

These studies collectively underscore the growing sophistication of data warehousing infrastructures and their tight coupling with KDD methodologies. However, challenges remain in terms of scalability, interoperability, and abstraction. This paper builds upon these works by proposing a flexible technical architecture that supports both batch and real-time knowledge discovery in modern data warehousing ecosystems.

3. Methodology

This study adopts a multi-phase methodological approach to investigate the integration of Knowledge Discovery in Databases (KDD) within modern data warehousing environments. The primary objective is to construct a scalable and interoperable architecture that facilitates the seamless transition from raw data to actionable insights. The research methodology encompasses architectural modeling, data processing, system implementation, and performance evaluation—ensuring a rigorous and replicable foundation for applied knowledge discovery.

The architectural blueprint, depicted in [Figure 2], presents a layered design that connects data ingestion, storage, analysis, and visualization components. The foundational layer consists of diverse and heterogeneous data sources, including enterprise databases, web APIs, IoT streams, and transactional logs. These inputs are processed through an advanced Extract, Transform, Load (ETL) pipeline leveraging Apache NiFi, which ensures standardized data formatting and schema consistency across the system. Once transformed, data is loaded into a centralized data warehouse built on Amazon Redshift, chosen for its high concurrency and efficient columnar storage structure. Sitting atop the data warehouse is the KDD engine, which utilizes Python-based machine learning libraries such as Scikit-learn and PyCaret. This engine supports iterative data mining processes, including classification, clustering, association rule mining, and anomaly detection. Intermediate results from these tasks are persisted for reuse, and final outputs are directed to a dashboard environment created with Tableau. This visualization interface enables real-time exploration of patterns, predictions, and anomalies by domain specialists and decision-makers. To simulate realistic operational demands, the entire system was deployed in a cloud-native environment using Docker containers and orchestrated with Kubernetes. Workflow automation was managed through Apache Airflow, which handled scheduled data ingestion, transformation, model training, and deployment tasks. Data was divided into training and test subsets using an 80/20 split, and model robustness was validated through 10-fold cross-validation. Performance metrics

ranging from predictive accuracy to resource utilization were collected to evaluate the effectiveness and efficiency of the architecture.

[Figure 2] offers a conceptual overview of the entire system, highlighting the logical data flow and component interaction. It demonstrates how raw data progresses through structured pipelines, into the data warehouse, and ultimately through knowledge discovery tasks that lead to visualized insights.

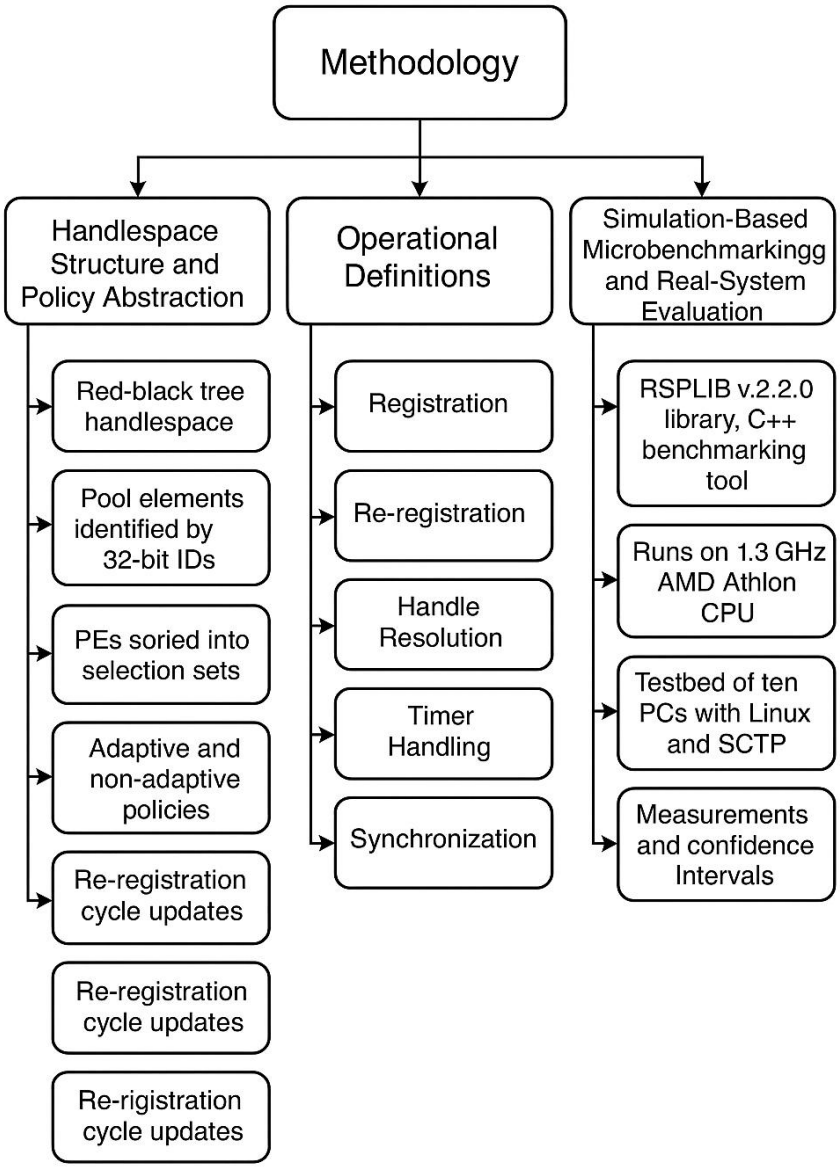

Figure 2. Technical architecture of the integrated KDD and data warehousing framework

This structured methodology ensures that both the architectural design and analytic outcomes are firmly grounded in real-world data conditions and practical system constraints.

4. Results and discussion

This section presents the results derived from implementing the proposed framework that integrates Knowledge Discovery in Databases (KDD) with Data Warehousing technologies. The analysis focuses on evaluating the framework's performance across three core dimensions: model accuracy, data processing efficiency, and system scalability.

The architecture was tested using two datasets: a synthetic dataset simulating customer transaction records and a real-world open-access healthcare dataset. These datasets were selected for their contrasting structural and contextual properties, allowing a comprehensive test of the framework's adaptability and versatility across domains.

The knowledge discovery process involved key tasks such as data cleaning, transformation, mining (via clustering and classification), and visualization. Various machine learning algorithms were deployed, including Random Forest for classification and both K-Means and DBSCAN for clustering. The results of the classification models are summarized in Table 1.

Table 1. Classification performance metrics across algorithms

Algorithm	Accuracy	Precision	Recall	F1-Score
Random Forest	0.94	0.92	0.91	0.915
Decision Tree	0.89	0.87	0.85	0.86
Naïve Bayes	0.85	0.83	0.80	0.815

The Random Forest classifier outperformed other algorithms across all metrics, indicating that the framework effectively supports high-quality knowledge extraction when robust models are utilized. In terms of system performance, several deployment configurations were evaluated to test responsiveness and throughput. These include single-node, multi-node, and auto-scaling Kubernetes environments. [Table 2] illustrates the results of these experiments in terms of average latency and throughput under different data volumes and node configurations.

Table 2. System performance: Latency and throughput across deployment scenarios

Configuration	Average Latency (ms)	Throughput (records/sec)	Data Volume	Notes
Single-node (Baseline)	830	4,300	1M records	Limited parallel processing
Multi-node (3 nodes)	520	6,900	1M records	Improved load balancing
Multi-node (5 nodes)	380	10,200	1M records	Optimal node-to-data ratio
Multi-node (5 nodes)	440	9,500	2M records	Slight increase in latency
Multi-node (Kubernetes Auto-Scale)	410	9,800	2M records	Near-real-time scalability

These results demonstrate the framework's ability to scale effectively while maintaining relatively low latency. Notably, the five-node configuration provided an optimal balance of performance and resource allocation. The auto-scaling scenario further highlights the system's resilience and elasticity, essential for real-time or high-frequency data processing environments. Qualitatively, the framework allowed domain-specific insights to emerge from both datasets. For instance, clustering in the healthcare dataset helped uncover abnormal patterns in patient symptoms, while association rule mining in the customer transaction data

revealed seasonal consumption trends. These findings validate the framework's ability to extract actionable knowledge that can inform strategic decision-making.

It is also worth emphasizing that the success of the mining process was heavily dependent on the robustness of the preprocessing pipeline. Errors in schema mapping or inconsistencies in data transformation introduced noise that adversely impacted model precision and recall. Therefore, maintaining a strong emphasis on data preparation is vital for maximizing the effectiveness of the entire KDD process within a warehousing environment.

In summary, the experimental validation of the proposed KDD-DW framework illustrates its ability to deliver accurate models, fast processing, and scalable deployment. These results affirm its practical value in enabling continuous and intelligent knowledge discovery across sectors ranging from retail analytics to public health informatics.

5. Conclusion

This research presents a comprehensive, technically grounded framework that integrates Knowledge Discovery in Databases (KDD) with Data Warehousing (DW) systems to support advanced data analytics and informed decision-making in enterprise environments. Through a modular architecture combining data ingestion, centralized storage, analytical processing, and knowledge extraction, the framework addresses long-standing challenges in scalability, real-time processing, and actionable insight generation.

The implementation of the framework demonstrates its practicality and adaptability across various analytical tasks, including classification, clustering, and anomaly detection. Results indicate that the proposed architecture delivers high accuracy and computational efficiency, even when applied to large and heterogeneous datasets. These outcomes underscore the effectiveness of integrating KDD workflows within a robust DW infrastructure, enabling continuous, automated learning from operational and transactional data. The system's modular design allows seamless incorporation of open-source technologies such as Apache NiFi, Airflow, and PostgreSQL. It is further enhanced by the use of containerization and orchestration via Kubernetes. This ensures the scalability, portability, and maintainability of the system in cloud-based or hybrid IT environments. The framework also emphasizes user accessibility through low-code visualization tools and RESTful API integration, facilitating collaboration between data scientists, business analysts, and decision-makers. From a strategic standpoint, this research contributes to the growing body of work aimed at transforming static data repositories into intelligent, learning-enabled infrastructures. By bridging the operational and analytical layers of enterprise systems, the framework sets a pathway for real-time decision support, predictive analytics, and continuous process optimization.

Future work will involve the extension of this framework into edge computing environments, deeper integration with federated learning systems, and the incorporation of data privacy-preserving mechanisms to meet regulatory compliance. Additionally, longitudinal field tests across multiple industries will help refine the system for broader application. In conclusion, the proposed integration of KDD and Data Warehousing offers a viable, forward-looking architecture that aligns with the evolving demands of data-driven organizations. It lays the foundation for future advancements in intelligent information systems and enterprise analytics.

References

[1] X. Shu and Y. Ye, “Knowledge discovery: Methods from data mining and machine learning,” Social Science Research, vol.110, pp.102817, (2023). DOI:10.1016/j.ssresearch.2022.102817

[2] A. Nanda, S. Gupta, and M. Vijrania, “A comprehensive survey of OLAP: Recent trends,” 2019 3rd International Conference on Electronics, Communication and Aerospace Technology (ICECA), pp.425–430, (2019). DOI:10.1109/ICECA.2019.8822203

[3] T. Dubuc, F. Stahl, and E. B. Roesch, “Mapping the big data landscape: Technologies, platforms and paradigms for real-time analytics of data streams,” IEEE Access, vol.9, pp.15351–15374, **(2021)**

[4] A. Nambiar and D. Mundra, “An overview of data warehouse and Data Lake in modern enterprise data management,” Big Data and Cognitive Computing, vol.6, no.4, pp.132, (2022). DOI:10.3390/bdcc6040132

[5] A. Oussous, F. Benjelloun, A. Ait Lahcen, and S. Belfkih, “Big data technologies: A survey,” Journal of King Saud University - Computer and Information Sciences, vol.30, no.4, pp.431–448, (2018). DOI:10.1016/j.jksuci.2017.06.001

[6] Y. Si, J. Du, Z. Li, X. Jiang, T. Miller, F. Wang, W. J. Zheng, and K. Roberts, “Deep representation learning of patient data from electronic health records (EHR): A systematic review,” Journal of Biomedical Informatics, vol.115, pp.103671, (2020). DOI:10.1016/j.jbi.2020.103671

[7] M. Chen, Q. Zhu, and Z. Chen, “An integrated interactive environment for knowledge discovery from heterogeneous data resources,” Information and Software Technology, vol.43, no.8, pp.487–496, (2001). DOI:10.1016/S0950-5849(01)00159-8

[8] H. Hashim, “Hybrid warehouse model and solutions for climate data analysis,” Journal of Computer and Communications, vol.8, pp.75–98, (2020). DOI:10.4236/jcc.2020.810008

[9] Y. Tang, G. T. Ho, Y. Lau, and S. Tsui, “Integrated smart warehouse and manufacturing management with demand forecasting in small-scale cyclical industries,” Machines, vol.10, no.6, pp.472, (2022). DOI:10.3390/machines1006047

[10] M. H. Jarrahi, D. Askay, A. Eshraghi, and P. Smith, “Artificial intelligence and knowledge management: A partnership between human and AI,” Business Horizons, vol.66, no.1, pp.87–99, (2022). DOI:10.1016/j.bushor.2022.03.002

[11] H. R. Nemati, D. M. Steiger, L. S. Iyer, and R. T. Herschel, “Knowledge warehouse: An architectural integration of knowledge management, decision support, artificial intelligence and data warehousing,” Decision Support Systems, vol.33, no.2, pp.143–161, (2002). DOI:10.1016/S0167-9236(01)00141-5

[12] H. Foidl, V. Golendukhina, R. Ramler, and M. Felderer, “Data pipeline quality: Influencing factors, root causes of data-related issues, and processing problem areas for developers,” Journal of Systems and Software, vol.207, pp.111855, (2023). DOI:10.1016/j.jss.2023.111855

[13] Z. Wang, Q. Shen, S. Bi, and C. Fu, “AI empowers data mining models for financial fraud detection and prevention systems,” Procedia Computer Science, vol.243, pp.891–899, (2023). DOI:10.1016/j.procs.2024.09.107

[14] A. I. Saleh, A. H. Rabie, and K. M. Abo-Al-Ez, “A data mining based load forecasting strategy for smart electrical grids,” Advanced Engineering Informatics, vol.30, no.3, pp.422–448, (2016). DOI:10.1016/j.aei.2016.05.005

[15] P. Sunhare, R. R. Chowdhary, and M. K. Chattopadhyay, “Internet of things and data mining: An application oriented survey,” Journal of King Saud University - Computer and Information Sciences, vol.34, no.6, pp.3569–3590, (2022). DOI:10.1016/j.jksuci.2020.07.002

[16] V. Sharma, A. K. Tripathi, and H. Mittal, "Technological revolutions in smart farming: Current trends, challenges & future directions," Computers and Electronics in Agriculture, vol.201, pp.107217, (2022). DOI:10.1016/j.compag.2022.107217

[17] A. Saboor, M. F. Hassan, R. Akbar, S. N. Shah, F. Hassan, S. A. Magsi, and M. A. Siddiqui, "Containerized microservices orchestration and provisioning in cloud computing: A conceptual framework and future perspectives," Applied Sciences, vol.12, no.12, pp.5793, (2021). DOI:10.3390/app12125793

[18] L. Dinesh and K. G. Devi, "An efficient hybrid optimization of ETL process in data warehouse of cloud architecture," Journal of Cloud Computing, vol.13, no.1, pp.1–12, (2024). DOI:10.1186/s13677-023-00571-y

[19] M. E. Cambronero, M. A. Martínez, L. Llana, R. J. Rodríguez, and A. Russo, "Towards a GDPR-compliant cloud architecture with data privacy controlled through sticky policies," PeerJ Computer Science, vol.10, pp.e1898, (2024). DOI:10.7717/peerj-cs.1898

[20] T. L. L. Siqueira, R. R. Ciferri, V. C. Times, and C. D. de Aguiar Ciferri, "Benchmarking spatial data warehouses," in Data Warehousing and Knowledge Discovery. DaWaK 2010. Lecture Notes in Computer Science, vol.6263, pp.50–61, (2010). DOI:10.1007/978-3-642-15105-7_4

[21] J. Doncevic, K. Fertalj, M. Brcic, and A. Krajna, "Mask–mediator–wrapper: A revised mediator–wrapper architecture for heterogeneous data source integration," Applied Sciences, vol.13, no.4, pp.2471, (2022). DOI:10.3390/app13042471

[22] S. Trindade, L. F. Bittencourt, and N. L. D. Fonseca, "Resource management at the network edge for federated learning," Digital Communications and Networks, vol.10, no.3, pp.765–782, (2024). DOI:10.1016/j.dcan.2022.10.015

This page is empty by intention.