International Journal of Technology, Innovation and Management (IJTIM) 5(2) -2025

Contents available at the publisher website: G A F T I M . C O M

International Journal of Technology, Innovation, and Management (IJTIM)

Journal homepage: https://journals.gaftim.com/index.php/ijtim/index

A Systematic Literature Review of Distributed Data Warehouse Architectures

Lama AlOud[1*], *Ohoud Alharbi*[2]

[1,2]*Department of Software Engineering, King Saud University, Sudia Arabia*

**Corresponding Author*

A R T I C L E I N F O

Keywords:
Distributed Data Warehouse, Data Warehouse Architecture, Federated Data Warehousing, Data Lakehouse Architecture,

Received: Jun, 23, 2025
Accepted: Aug, 10, 2025
Published: Dec, 25, 2025

A B S T R A C T

This research presents a systematic literature review (SLR) of distributed data warehouse (DDW) architectures, addressing challenges in governance, security, scalability, and real-time analytics. Conducted in accordance with PRISMA 2020 guidelines, the review synthesizes 29 peer-reviewed studies from 2020 to 2025. It identifies four major architectural themes: security-oriented, federated and data mesh–oriented, data lakehouse-based, and real-time/streaming-enabled architectures. These themes address recurring challenges such as data privacy, organizational autonomy, governance of diverse data types, and low-latency analytics. The review highlights the trend towards multi-paradigm designs that integrate multiple principles to balance autonomy, governance, performance, and security. Additionally, it outlines future research directions in autonomous architectures, AI-driven metadata management, and empirical evaluation of hybrid DDW models.

1. INTRODUCTION

The rapid growth of data volumes, the variety of data sources, and the increasing need for timely analytics are putting pressure on traditional centralized data warehouse systems. Classical data warehouses and data lakes are typically managed as centralized platforms, which can create bottlenecks in ownership, governance, and responsiveness when organizations become more dynamic and data becomes harder to control centrally.

Handling these constraints, distributed or decentralized approaches have gained attention. In particular, recent work highlights a shift toward organizing data around domains and distributing responsibility to reduce the overload and limitations created by centralized data teams. In parallel, research in data warehouse architectures also emphasizes that distribution is increasingly relevant when data is naturally spread across multiple sites and systems, and when scalability and traceability become important design requirements.

However, the literature on distributed data warehouse architectures is still scattered. Many studies address specific aspects, such as security and privacy controls in data warehouse environments or organizational and agile data management practices, without offering a consolidated architectural view of how distributed data warehouse systems are structured and how different designs compare (Pörtner et al., 2023).

To address this gap, this study presents a systematic literature review (SLR) of distributed data warehouse architectures. Following a rigorous selection process, 29 core studies were selected for in-depth synthesis in the Findings and Discussion section, as they provide explicit architectural descriptions and clear discussions of motivating challenges. The objectives of this review are to (i) identify and classify architectural models proposed for distributed data warehouses, and (ii) analyze the key challenges that motivate

their design (Thantilage et al., 2023).
By organizing the literature into coherent architectural themes and linking these themes to their motivating challenges, this review provides a structured overview of the current research landscape and highlights areas that remain underexplored.
The remainder of this paper is organized as follows. Section 2 describes the SLR methodology. Section 3 presents the findings and discussion. Section 4 concludes the paper and outlines directions for future research.

2. RESEARCH METHODOLOGY

This study employs a Systematic Literature Review (SLR) to examine architectural approaches to distributed data warehouses (DDWs). The SLR methodology was selected to enable a rigorous, transparent, and reproducible synthesis of existing research, in accordance with established guidelines for evidence-based software engineering research. The review process follows the PRISMA 2020 framework, ensuring systematic identification, screening, and selection of relevant studies.
The scope of the review is explicitly architectural. The analysis focuses on architectural models, design structures, and system-level approaches for distributed data warehouses. Studies that address implementation tools, ETL pipelines, query optimizations, or platform-specific deployments without contributing architectural insight are considered out of scope.

3.1 Search Strategy and Sources

The review is guided by the following research questions:
RQ1: What architectural models are proposed or implemented for distributed data warehouses?
RQ2: What challenges motivate the design of these architectures?
RQ1 targets the identification and classification of architectural patterns and design approaches, while RQ2 captures the technical and organizational challenges that drive architectural decisions.
The literature search was conducted using Web of Science (WoS). Selecting WoS due to its comprehensive coverage of high-quality, peer-reviewed publications across major publishers (e.g., IEEE, Elsevier, Springer, Wiley). Restricting the search to a single, authoritative database ensured consistency in indexing, reduced duplication, and supported a transparent screening process.
The search strategy was designed to capture studies addressing distributed data warehouse architecture and design motivations. Boolean search strings combined terms related to distribution paradigms, data warehousing, and architectural design.
Architectural Models (RQ1): TS=((distributed OR decentralized OR parallel OR federated OR hybrid OR cloud) AND ("data warehouse*" OR "data warehousing" OR "warehouse system*" OR "data warehouse architecture") AND (architecture OR "architectural model" OR design OR framework OR "system model"))
Motivating Challenges (RQ2): TS=((distributed OR decentralized OR parallel OR federated OR hybrid OR cloud) AND ("data warehouse*" OR "data warehousing" OR "warehouse system*" OR "data warehouse architecture") AND (architecture OR design OR framework) AND (challenge* OR limitation* OR issue* OR "research gap*" OR "future work"))
The search was limited to peer-reviewed studies published in English between 2020 and 2025, reflecting contemporary research in distributed data warehouse architectures.

3.2 Screening and Quality Appraisal

Clear inclusion and exclusion criteria were defined prior to screening.
Inclusion Criteria, studies were included if they:

- Propose or analyze architectural models or design approaches for distributed data warehouses
- Address analytical or OLAP-oriented data warehouse systems
- Discuss architectural challenges such as scalability, governance, security, or distribution

Exclusion Criteria, studies were excluded if they:

- Focus primarily on tools, platforms, or ETL implementations without architectural contribution
- Address BI, visualization, or query-level optimization in isolation
- Lack explicit discussion of architectural structure or design rationale
- Target non-analytical or operational data management systems

The study selection followed the PRISMA

workflow. After duplicate removal, titles and abstracts were screened to eliminate clearly irrelevant studies. The remaining papers were subjected to full-text review using the predefined criteria. Figure 1 presents the PRISMA flow diagram summarizing the study identification and selection process.

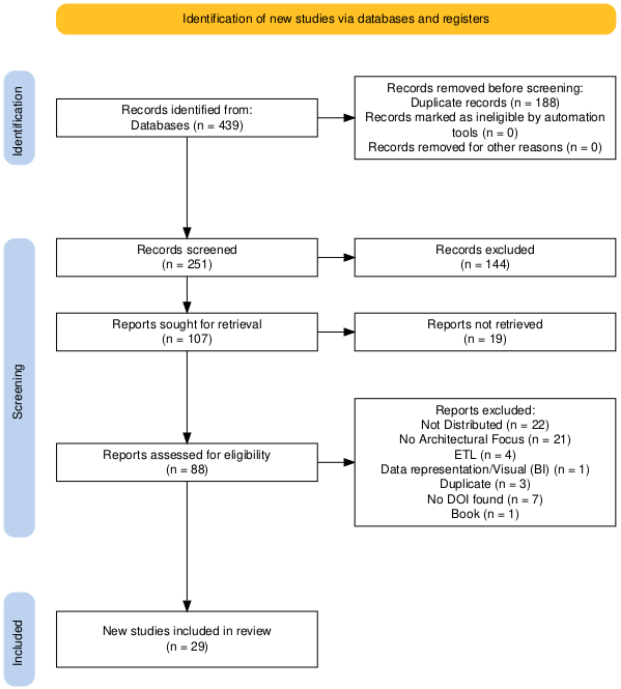

Figure 1. PRISMA 2020 Flow Diagram illustrating screening, and inclusion process for distributed data warehouse architecture studies (2020–2025).

This process resulted in 62 included studies, representing the full architectural literature base for the review. From this set, 29 core studies were selected for in-depth synthesis in the Findings and Discussion section. These studies were chosen because they provide explicit architectural descriptions and clearly articulate the challenges motivating their design.

3.3 Data Extraction and Synthesis Approach

A structured extraction protocol was applied to all included studies. The following data were extracted:

- Architectural model or design approach (RQ1)
- Motivating challenges addressed (RQ2)
- Supporting sections, figures, or architectural diagrams

Extraction was performed consistently to minimize subjectivity and support cross-study comparison.

Data synthesis was conducted using thematic analysis supported by narrative synthesis. Architectural approaches were grouped into recurring themes based on shared structural and design characteristics (RQ1). Motivating challenges were analyzed across studies to identify dominant drivers influencing architectural decisions (RQ2).

The synthesis prioritizes architectural structure and design rationale over implementation detail, providing a consolidated and comparative view of distributed data warehouse architectures.

3. findings and Discussion

This study systematically analyzed the literature on distributed data warehouse (DDW) architectures and identified four dominant architectural themes: security oriented, federated and data mesh oriented, data lakehouse based, realtime and streaming enabled, and autonomous and self adaptive of distributed data warehouse architectures. These themes reflect complementary architectural responses to distinct and represent overlapping challenges arising from scale, heterogeneity, governance, and data velocity in modern analytical environments.

3.1 Security Challenges and Architectural Responses in Distributed Data Warehouses

Security, privacy, and access control have long been critical concerns in data warehouse systems, particularly as analytical OLAP workloads increasingly demand scalability, high data velocity, and support for heterogeneous data sources. While these concerns exist in both centralized and distributed warehouses, they become more pronounced in distributed and cloud-based architectures, especially in sensitive domains such as healthcare, finance, and government. Addressing these challenges, several studies propose hybrid cloud deployments, in which private clouds manage sensitive data and access policies, while public clouds provide scalable analytical processing (Vestues et al., 2022). This architectural separation enables privacy-preserving analytical queries and fine-grained access control while balancing security and performance requirements. Collectively, these approaches indicate a shift toward treating security as a deployment-level architectural concern, rather than an isolated protection mechanism (Fugkeaw & Hak, 2024).

Beyond hybrid cloud solutions, blockchain technology has emerged as a promising approach

for enhancing security in distributed data warehouses. Recent studies leverage blockchain to prevent unauthorized access while maintaining analytical performance and scalability. In particular, blockchain-based architectures address the challenge of distributed integrity verification by providing tamper-resistant audit trails and mechanisms to confirm the authenticity of stored data and analytical results. This trend reflects an increasing emphasis on decentralized trust and auditability as fundamental requirements in distributed analytical systems (Bergers et al., 2021).

Other research focuses on strengthening authentication and authorization mechanisms within distributed warehouse environments. For example, the key-group distribution model proposed introduces a novel authentication framework, referred to as a data warehouse signature, to improve secure data sharing between users, managers, and execution entities (AlMeghari et al., 2021). Addtionaly, the ontology-based security models support security incident analysis and detection by enabling reasoning over security policies and events, as incidents or threaten detections, in the distributed nodes. Together, these approaches highlight a broader movement toward semantic, role-aware, and policy-driven security enforcement embedded directly within distributed data warehouse architectures (Butakova et al., 2020).

3.2 Federated and Data Mesh Oriented Distributed Data Warehouse Architectures

Federated and data mesh oriented architectures address limitations of centralized data warehouses in large and heterogeneous organizations. Instead of enforcing global schemas or physically consolidating data, these approaches enable analytical integration across distributed repositories while preserving local autonomy and governance (Vadim et al., 2020).

Federated distributed data warehouse architectures support analytics across independently managed and heterogeneous data sources without centralizing data. This is typically achieved through semantic abstraction layers, such as metadata and ontologies, which allow different data models to be queried in a unified manner. Analytical queries are mediated or translated at runtime, enabling each repository to participate in global analytics using its local schema. Ontology-based federated warehouses demonstrate that cross-institutional analytics can be performed using shared semantic models and metadata registries, particularly in multi-organization environments where data consolidation is impractical . Service-API based federation further supports this model by enabling loosely coupled coordination across distributed warehouse nodes

While federation enables interoperability, it does not fully address challenges related to data ownership and organizational coordination in large and evolving systems. Building on earlier federated ideas, data mesh inspired architectures emphasize domain-level data ownership and treat analytical data as a product managed by the domains that generate it. This perspective aligns responsibility for data quality and evolution with business boundaries rather than centralized integration teams. Studies indicate that such architectures can reduce coordination bottlenecks in centralized warehouses and better support scalability and adaptability in environments with frequent change, including agile development contexts (Loukiala et al., 2021).

Across both federated and data mesh approaches, metadata and semantic management play a central role. Federated query mediation allows analytical queries to be executed locally at each node, with results aggregated by a coordinating mediator, enabling cross-platform analytics while preserving autonomy. Other approaches propose centralized or hybrid metadata services enhanced with automation and machine learning to support schema discovery and semantic alignment across diverse data types. These mechanisms enable interoperability without enforcing tight coupling Barnes et al., 2022).

Recent studies also explore the use of blockchain technologies within federated and data mesh architectures to support governance and trust across decentralized data products. In these approaches, blockchain mechanisms are typically applied at the governance or metadata layer to provide auditability, policy enforcement, and verifiable data-sharing agreements without reintroducing centralized control (Rosenau & Ingenerf, 2024).

The practical value of federated and data mesh oriented architectures is evident in IoT Edge Cloud

environments, such as smart cities, where centralized warehouses struggle with latency, scalability, and data heterogeneity. Data mesh–based distributed warehouses enable local analytics and scalable expansion across distributed nodes, while federation supports cross-domain analytics without moving data to a central repository. In privacy-sensitive domains such as healthcare, federated analytics enable cross-institutional analysis without direct access to raw data, supporting data sovereignty and regulatory compliance when combined with privacy-preserving techniques such as differential privacy.

3.3 Data Lakehouse Based Distributed Architectures

Managing heterogeneous structured, semi-structured, and unstructured data at scale while maintaining governance, analytical usability, and performance is a fundamental challenge in distributed data warehouse systems. Traditional data warehouses provide strong schema enforcement and optimized OLAP performance but lack flexibility in handling diverse data types, whereas data lakes offer scalable and flexible storage with limited governance and analytical guarantees.

The reviewed literature collectively contributes the data lakehouse as an architectural unification model that resolves this trade-off by adopting a hybrid or layered design integrating both paradigms within a single distributed architecture. Rather than treating data lakes and data warehouses as separate systems, lakehouse architectures enable analytical workloads to operate directly over heterogeneous data using shared storage, open formats, and unified governance mechanisms (Silva et al., 2024). This integration combines warehouse-level schema management and query optimization with scalable data lake storage, allowing OLAP analytics to scale across distributed environments without duplicating data or sacrificing control.

Across multi-source and distributed deployments, the literature consistently identifies centralized metadata management as the key architectural enabler of this unification. Centralized metadata catalogs provide a single point for schema definition, access control, lineage tracking, and query coordination, ensuring consistent governance and efficient analytical access across heterogeneous data sources. This metadata-centric design further extends the contribution of lakehouse architectures beyond traditional analytics, enabling machine learning–ready data pipelines by ensuring data consistency, reproducibility, and data readiness in distributed data warehouse settings (Ghane, 2020).

3.4 Realtime and Streaming Enabled Distributed Data Warehouses

The continuity of OLAP analytics over streaming data in distributed operational systems highlights the need to achieve low-latency analytics while sustaining high data velocity. To address this requirement, real-time distributed data warehouse architectures increasingly process streaming data in memory and maintain continuously updated analytical summaries, enabling both snapshot OLAP (pre-aggregated and computed views) and continuous OLAP, where analytical results are incrementally updated as new data arrives. These capabilities overcome the limitations of traditional batch-oriented warehouses, which are unable to support timely analytics under continuous data ingestion (Rossini et al., 2024).

A recurring architectural solution in the literature is the separation of static and dynamic processing layers. The static layer preserves historical data optimized with indexes and materialized views, while the dynamic layer manages recent or streaming data using lightweight structures to support real-time loading and analysis. Queries are distributed across these layers and their results are merged to provide unified analytical views, enabling real-time responsiveness without degrading performance on historical data (Ryffel et al., 2025). This separation directly mitigates performance bottlenecks caused by simultaneous querying and data loading in centralized architectures.

The demand for real-time distributed analytics is further intensified by IoT-generated data, where high-frequency streams require immediate reflection in analytical results. To support this, distributed data warehouse architectures integrate hybrid batch, stream processing pipelines, enabling real-time ingestion alongside periodic batch consolidation (Harby & Zulkernine, 2025). Such designs ensure low-latency analytics for operational decision making while preserving analytical consistency across distributed storage and processing components (Levandoski et al., 2024).

Additionally, several studies identify the passive,

pull-based nature of traditional distributed data warehouses, where analytics are produced only in response to explicit queries, as a fundamental limitation for real-time continuity. To overcome this limitation, architectures incorporate active or continuous query mechanisms, allowing analytical results to be proactively updated and delivered as data changes. By integrating continuous queries across distributed components, these approaches enable uninterrupted analytical flows and transform distributed data warehouses from reactive systems into continuous analytical platforms (Kalmuk et al., 2024).

3.5 Autonomous and self adaptive distributed data warehouse architectures

Most distributed data warehouse architectures assume relatively stable workloads and rely on static design decisions or reactive optimization techniques. However, several studies argue that such assumptions no longer hold in modern distributed analytical environments, where workloads are highly dynamic and continuously evolving. To address this limitation, a distinct line of research proposes autonomous and self adaptive distributed data warehouse architectures. These architectures extend traditional parallel and distributed data warehouses with proactive adaptation capabilitiesinspired by autonomic computing principles. Instead of redesigning data placement only after performance degradation occurs, autonomous architectures continuously monitor workload behavior and adjust data partitioning and fragment allocation accordingly. For example, workload-aware designs employ query clustering and learning-based mechanisms to guide partitioning decisions, enabling the system to adapt incrementally as query patterns evolve, without requiring manual intervention (Tripathi et al., 2024).

Other studies emphasize that self adaptive data distribution strategies are essential for supporting ad-hoc and batch analytical workloads at scale. By learning from historical query behavior and shared query structures, these architectures dynamically reconfigure data placement to improve global query efficiency while maintaining scalability.

In addition, large-scale distributed data warehouse implementations demonstrate that adaptive partitioning and shard balancing mechanisms are critical for sustaining performance as data volumes and access patterns change over time. These systems show that autonomy is not only a conceptual design goal but also a practical architectural requirement in distributed analytical deployments. Overall, autonomous and self-adaptive distributed data warehouse architectures represent a distinct architectural direction within the distributed data warehouse landscape. Rather than introducing new storage paradigms, they enhance existing architectures with learning-driven and self-managing capabilities, directly addressing recurring challenges such as workload volatility, scalability limitations of static designs, and the operational cost of manual optimization.

Summaries findings through motivation challenges of the desgin Architecture for each theme shown in table 1.

Table 1. Key Findings: Distributed Data Warehouse Architecture Themes

Theme	Primary Challenge(s)	Core Architectural Response
Security Oriented Architectures	Privacy, access control, trust, regulatory constraints	Security-driven architectures using hybrid clouds, decentralized integrity, and policy-based access
Federated / Data Mesh Oriented Architectures	Organizational scale, data heterogeneity, domain autonomy	Logical integration via federation, domain ownership, and decentralized governance
Data Lakehouse Based Architectures	Heterogeneous data management with governance and performance	Unified lake–warehouse architectures with shared storage and centralized metadata
Realtime and Streaming Enabled Architectures	Low-latency analytics over continuous data streams	Continuous OLAP via stream processing and separation of static and dynamic layers
Autonomous and Self Adaptive Architectures	Workload variability and scalability of static designs	Self-managing architectures with adaptive partitioning and execution

4. CONCLUSION

This systematic literature review examined architectural models for distributed data warehouses with the objective of identifying dominant architectural paradigms and the underlying challenges that motivate their design. By synthesizing evidence from 29 primary studies, the review provides a consolidated architectural perspective on how distributed data warehouse systems are evolving in response to increasing scale, heterogeneity, autonomy, security, and real-time analytical demands.

The findings reveal four recurring architectural themes that characterize research in distributed data warehousing: security oriented, federated and data mesh oriented, data lakehouse based, and realtime and streaming enabled architectures. These themes represent distinct architectural responses to persistent and emerging challenges in distributed analytical environments.

Security oriented architectures are primarily motivated by concerns related to data privacy, trust, regulatory compliance, and controlled access in distributed and cloud-based settings. Rather than treating security as an auxiliary layer, these architectures embed security and governance mechanisms directly into architectural design, leveraging deployment strategies, decentralized trust models, and fine-grained access control to reconcile analytical scalability with strict security requirements.

Federated and data mesh oriented architectures address challenges associated with organizational scale, semantic heterogeneity, and autonomous data ownership. Federation enables analytical interoperability across independently managed data repositories without enforcing physical centralization, while data mesh architectures extend this principle by emphasizing domain-level ownership and distributed governance. Together, these approaches reflect a shift from centralized control toward coordination-based architectures aligned with organizational autonomy and distributed responsibility.

Data lakehouse based architectures respond to the challenge of managing heterogeneous data types at scale while preserving governance, analytical usability, and performance. By unifying data warehouse and data lake paradigms through shared storage layers and centralized metadata governance, lakehouse architectures support scalable OLAP analytics and machine learning ready data pipelines in distributed environments, reducing data duplication while maintaining consistency and control.

Realtime and streaming enabled architectures are motivated by increasing data velocity and timeliness requirements, particularly in IoT-driven and operational analytics contexts. These architectures extend traditional batch-oriented data warehouses by introducing continuous OLAP capabilities, in-memory stream processing, and the separation of static and dynamic data layers. As a result, distributed data warehouses evolve from passive, query-driven systems into continuous analytical platforms capable of supporting low-latency decision making.

Collectively, the findings indicate that no single architectural paradigm fully addresses all distributed data warehouse challenges. Instead, contemporary systems increasingly adopt multi-paradigm architectures, combining elements from multiple themes to balance governance, autonomy, heterogeneity, security, and latency requirements. This architectural convergence highlights the need to view distributed data warehouse design as a composition of complementary architectural principles rather than a choice of a single dominant model.

The findings also point to several directions for future research. While existing architectures address security, governance, and performance, most rely on static configurations; therefore, further work is needed on autonomous and self-adaptive distributed data warehouse architectures that can adjust policies and processing strategies at runtime. In addition, metadata and semantic management play a central role across federated, data mesh, and lakehouse designs, yet current solutions are largely manual, highlighting the need for more automated and intelligent, AI-driven, metadata management to support large-scale and heterogeneous environments. Finally, although many studies propose hybrid architectures that combine multiple architectural paradigms, empirical evaluations in real-world deployments remain limited, indicating the need for systematic studies that assess architectural trade-offs in practice.

REFERENCES

Pörtner, L., Möske, R., & Riel, A. (2023). Data Ecosystem for Industrial Product-Service Systems (IPS2) Based on a Decentralized Data Architecture. Procedia CIRP, 119, 1228–1233. https://doi.org/10.1016/j.procir.2023.02.190

Thantilage, R. D., Le-Khac, N.-A., & Kechadi, M.-T. (2023). Healthcare data security and privacy in Data Warehouse architectures. Information Medical Unlocked, 39, 101270. https://doi.org/10.1016/j.imu.2023.101270

Vestues, K., Hanssen, G. K., Mikalsen, M., Buan, T. A., & Conboy, K. (2022). Agile Data Management in NAV: A Case Study. In V. Stray, K.-J. Stol, M. Paasivaara, & P. Kruchten (Eds.), Agile Processes in Software Engineering and Extreme Programming (Vol. 445, pp. 220–235). Springer International Publishing. https://doi.org/10.1007/978-3-031-08169-9_14

Fugkeaw, S., & Hak, L. (2024). PPAC-CDW: A Privacy-Preserving Access Control Scheme With Fast OLAP Query and Efficient Revocation for Cloud Data Warehouse. IEEE Access, 12, 78743–78758. https://doi.org/10.1109/ACCESS.2024.3408221

Bergers, J., Shi, Z., Korsmit, K., & Zhao, Z. (2021). DWH-DIM: A Blockchain Based Decentralized Integrity Verification Model for Data Warehouses. In 2021 IEEE International Conference on Blockchain (Blockchain) (pp. 221–228). IEEE. https://doi.org/10.1109/Blockchain53845.2021.00037

AlMeghari, M., Taha, S., Elmahdy, H., & Shen, X. (2021). A proposed authentication and group-key distribution model for data warehouse signature, DWS framework. Egyptian Informatics Journal, 22(3), 245–255. https://doi.org/10.1016/j.eij.2020.09.002

Butakova, M. A., Chernov, A. V., Savvas, I. K., & Garani, G. (2020). Data Warehouse Design for Security Applications Using Distributed Ontology-Based Knowledge Representation. In I. Kotenko, C. Badica, V. Desnitsky, D. El Baz, & M. Ivanovic (Eds.), Intelligent Distributed Computing XIII (Vol. 868, pp. 140–145). Springer International Publishing. https://doi.org/10.1007/978-3-030-32258-8_16

Vadim, B., Dmitry, K., & Alexander, M. (2020). Intelligent Information Search Method Based on a Compositional Ontological Approach. In S. O. Kuznetsov, A. I. Panov, & K. S. Yakovlev (Eds.), Artificial Intelligence (Vol. 12412, pp. 371–381). Springer International Publishing. https://doi.org/10.1007/978-3-030-59535-7_27

Barnes, C., et al. (2022). The Biomedical Research Hub: A Federated Platform for Patient Research Data. Journal of the American Medical Informatics Association, 29(4), 619–625. https://doi.org/10.1093/jamia/ocab247

Loukiala, A., Joutsenlahti, J.-P., Raatikainen, M., Mikkonen, T., & Lehtonen, T. (2021). Migrating from a Centralized Data Warehouse to a Decentralized Data Platform Architecture. In L. Ardito, A. Jedlitschka, M. Morisio, & M. Torchiano (Eds.), Product-Focused Software Process Improvement (Vol. 13126, pp. 36–48). Springer International Publishing. https://doi.org/10.1007/978-3-030-91452-3_3

Rosenau, L., & Ingenerf, J. (2024). Structured Queries to AQL: Querying OpenEHR Data Leveraging a FHIR-Based Infrastructure for Federated Feasibility Queries. Studies in Health Technology and Informatics. https://doi.org/10.33/SHTI230922

Ghane, K. (2020). Big Data Pipeline with ML-Based and Crowd Sourced Dynamically Created and Maintained Columnar Data Warehouse for Structured and Unstructured Big Data. In 2020 3rd International Conference on Information and Computer Technologies (ICICT) (pp. 60–67). IEEE. https://doi.org/10.1109/ICICT50521.2020.00018

Rossini, E., Bicocchi, N., Hadjidimitriou, N. S., Pietri, M., Picone, M., & Mamei, M. (2024). Towards a Distributed Data Mesh Model for the IoT-Edge-Cloud Continuum in Smart Cities. In 2024 IEEE/ACM Symposium on Edge Computing (SEC) (pp. 383–388). IEEE. https://doi.org/10.1109/SEC62691.2024.00041

Ryffel, T., et al. (2025). Federated Analysis With Differential Privacy in Oncology Research: Longitudinal Observational Study Across Hospital Data Warehouses. JMIR Medical Informatics, 13, e59685–e59685. https://doi.org/10.2196/59685

Silva, D., et al. (2024). Review of open-source software for developing heterogeneous data management systems for bioinformatics applications. Bioinformatics Advances, 5(1), vbaf168. https://doi.org/10.1093/bioadv/vbaf168

Harby, A. A., & Zulkernine, F. (2025). Data Lakehouse: A Survey and Experimental Study. Information Systems, 127, 102460. https://doi.org/10.1016/j.is.2024.102460

Levandoski, J., et al. (2024). BigLake: BigQuery's Evolution toward a Multi-Cloud Lakehouse. In Companion of the 2024 International Conference on Management of Data (pp. 334–346). ACM. https://doi.org/10.1145/3626246.3653388

Kalmuk, D., et al. (2024). Native Cloud Object Storage in Db2 Warehouse: Implementing a Fast and Cost-Efficient Cloud Storage Architecture. In Companion of the 2024 International Conference on Management of Data (pp. 188–200). ACM. https://doi.org/10.1145/3626246.3653393

Tripathi, A., Waqas, A., Venkatesan, K., Yilmaz, Y., & Rasool, G. (2024). Building Flexible, Scalable, and Machine Learning-Ready Multimodal Oncology Datasets. Sensors, 24(5), 1634. https://doi.org/10.3390/s24051634