Seminars in Medical Writing and Education. 2024; 3:67
doi: 10.56294/mw202467

ORIGINAL

Data Lakehouse: Next Generation Information System

Data Lakehouse: Sistema de información de próxima generación

Mohamed Cherradi[1] ✉, Anass El Haddadi[1]

[1]Data Science and Competetive Intelligence Team (DSCI), ENSAH. Abdelmalek Essaâdi University (UAE). Tetouan, Morocco.

Cite as: Cherradi M, El Haddadi A. Data Lakehouse: Next Generation Information System. Seminars in Medical Writing and Education. 2024; 3:67. https://doi.org/10.56294/mw202467

Submitted: 30-10-2023 **Revised:** 20-01-2024 **Accepted:** 31-03-2024 **Published:** 01-04-2024

Editor: Dr. José Alejandro Rodríguez-Pérez

ABSTRACT

This paper introduces the Data Lakehouse Architecture, a transformative model in data architecture that seamlessly integrates the analytical strengths of traditional data warehouses with the schema flexibility inherent in data lakes. Departing from current frameworks, this comprehensive approach establishes a unified platform, overcoming limitations of conventional data management. Addressing the critical need for an integrated solution, our primary objective is to set a new standard for sophisticated data management. The distinctiveness of our proposal lies in the seamless fusion of data warehouse analytics and data lake schema flexibility, underscoring its originality. The full article delves into the research methodology, providing a comprehensive understanding of the study's framework proposal. The foundational outcomes showcase the successful implementation of our Data Lakehouse Architecture, revealing enhanced processing capabilities for structured data analysis, complex querying, and high-performance reporting. The conclusion emphasizes the paradigm shift and transformative impact on data management practices, reinforcing the significance of our innovative solution. This research not only contributes a novel technological framework but also highlights the importance of adaptability and performance in the face of evolving data landscapes.

Keywords: Big Data Management; Business Intelligence; Data Warehouse; Data Lake; Data Lakehouse.

RESUMEN

Este artículo presenta la arquitectura Data Lakehouse, un modelo transformador en la arquitectura de datos que integra a la perfección los puntos fuertes analíticos de los almacenes de datos tradicionales con la flexibilidad de esquemas inherente a los lagos de datos. Partiendo de los marcos actuales, este enfoque global establece una plataforma unificada que supera las limitaciones de la gestión de datos convencional. Atendiendo a la necesidad crítica de una solución integrada, nuestro principal objetivo es establecer un nuevo estándar para la gestión sofisticada de datos. El carácter distintivo de nuestra propuesta radica en la fusión sin fisuras de la analítica del almacén de datos y la flexibilidad del esquema del lago de datos, lo que subraya su originalidad. El artículo completo profundiza en la metodología de la investigación, proporcionando una comprensión exhaustiva de la propuesta marco del estudio. Los resultados fundamentales muestran el éxito de la implantación de nuestra arquitectura de lago de datos, revelando una mayor capacidad de procesamiento para el análisis de datos estructurados, consultas complejas e informes de alto rendimiento. La conclusión subraya el cambio de paradigma y el impacto transformador en las prácticas de gestión de datos, reforzando la importancia de nuestra innovadora solución. Esta investigación no sólo aporta un novedoso marco tecnológico, sino que también pone de relieve la importancia de la adaptabilidad y el rendimiento ante la evolución del panorama de los datos.

Palabras clave: Gestión de Big Data; Business Intelligence; Data Warehouse; Data Lake; Data Lakehouse.

INTRODUCTION

In recent years, the dynamic evolution of business intelligence systems has been apparent to their continuous development, increasing democratization, and adaptability. This progressive trajectory in decision-making systems gains momentum from the exponential surge in data volumes, a phenomenon that serves as a catalyst for transformative change. The collaborative contributions of both software and hardware infrastructure markets have played pivotal roles in reshaping business intelligence systems.[1] This transformative evolution not only facilitates the accelerated access, processing, transformation, storage, utilization, visualization, and analysis of data but does so in a manner that is notably faster, simpler, and more technologically open. This overarching technological revolution has rendered business intelligence systems accessible to a larger user base, transcending their traditional constraints limited to expert users. This democratization unfolds across the entire organizational workforce, marking a paradigm shift in how these systems empower diverse stakeholders.[2] Within this landscape, data lakehouses emerge as a trendy and highly sought-after tool for big data analytics, garnering widespread attention in both academic and industrial realms.[3] Despite their popularity, a notable gap exists—there is currently no formal and universally accepted definition for "data lakehouse." Diverse interpretations and definitions abound, each placing varying emphasis on key concepts, contributing to a nuanced yet undefined landscape within the realm of data analytics. This absence of a standardized definition underscores the need for a comprehensive exploration and conceptualization of data lakehouses, a gap that our research seeks to address by providing a clear and nuanced understanding of this evolving paradigm.

Thus, traditional data lakes play a crucial role in supporting teleportation, executing SQL queries, employing ACID transactions for data ingestion, and visualizing petabyte-scale datasets on cloud storage.[4] They serve as indispensable infrastructure for analytical applications, breaking down data silos, empowering data-driven decision-making, enhancing operational efficiency, and reducing costs. Functioning as a central repository, a data lake enables businesses to consolidate diverse data sources into a single location, offering improved control, governance, and analysis capabilities.[5] However, these traditional data lakes encounter limitations, particularly in applications such as natural language processing, audio processing, computer vision, and those utilizing non-tabular datasets, especially when employing deep learning approaches. In the contemporary landscape of digital information systems generating vast data volumes, the imperative for effective big data management solutions becomes evident.[6] To address challenges arising from high volume, velocity, diversity, and veracity, data management systems have evolved from structured databases to encompass big data storage systems, graph databases, data warehouses, and data lakes. Each solution brings distinct benefits and drawbacks. Recognizing the need to extract meaningful knowledge from unstructured data across diverse sources, a synthesis of data warehouses and data lakes, known as the "data lakehouse," becomes essential.[7] This approach seeks to combine the rapid unstructured data ingestion and storage capabilities of a data lake with post-storage transformation and analytics functionalities parrallel to a data warehouse. Drawing insights from a literature review comparing recent data warehouse and data lake solutions, we propose essential elements for the Lakehouse (LH) architecture. This architecture has garnered substantial interest in the big data management research community, offering a preferred and comprehensive solution that addresses the strengths and weaknesses of existing systems.

Nevertheless, the emergence of the term "data lakehouse" within the technology sphere is a direct response to the challenges encountered with traditional data management approaches. The fundamental concept driving the data lakehouse is to integrate the benefits of data lakes, characterized by economical storage and flexibility for various data types, with those of data warehouses, renowned for their prowess in structured data processing and optimized performance.[8] Despite the considerable attention garnered by the term "lakehouse" for its potential to streamline enterprise analytics architectures through the integration of data warehouse and data lake features, existing proposals often lack clarity. This lack of precision is primarily influenced by marketing strategies and tool-centric perspectives. This paper addresses the identified gap in the literature by conducting a thorough evaluation and proposing a novel architecture for lakehouses. The data management landscape assessment extends to popular data management tools such as Delta Lake, Snowflake, and Dremio, aiming to determine their alignment with the data lakehouse proposed characterization. This research contributes to a clearer understanding of lakehouses, providing a comprehensive framework for assessing their implementation across a spectrum of diverse data management tools. By delving into this exploration, our objective is to shed light on the intricacies of lakehouse definitions, ensuring a nuanced comprehension of this emerging concept.

The remainder of this paper is organized as follows. Section 2 systematically delves into the existing body of knowledge, offering a thorough review of related works to contextualize the research landscape. Moving forward, Section 3 establishes the typology of Data Warehouses (DW), Data Lakes (DL), and Data Lakehouse (LH). The essence of our contribution unfolds in Section 4, where we introduce and expound upon the intricacies of our innovative Data Lakehouse architecture. Finally, Section 5 concludes by some remarks and future perspectives.

https://doi.org/10.56294/mw202467

Related Works

The concept of data lakehouses has emerged as a focal point in contemporary discussions on data architecture evolution. As organizations grapple with escalating data volumes and diverse data types, the need for a comprehensive solution that seamlessly integrates the strengths of data warehouses and data lakes becomes increasingly apparent. In this section, we delve into the related works that have paved the way for understanding and exploring the dynamics of data lakehouses. Our objective is to provide a coherent narrative that builds upon and extends the collective contributions of scholars and practitioners in unraveling the intricacies of this emerging paradigm. In fact, Researchers are looking at a novel solution called Data Lakehouse (LH) in order to combine the desired features of the Data Lake (DL) and Data Warehouse (DW) and fix their weaknesses in order to accelerate efficient knowledge extraction.[9] To identify the key factors propelling the transformation from traditional DW and DL to the integrated LH. Hence, the concept of data lakehouses has been a subject of academic exploration since 2021, and scholars are increasingly drawn to this emerging option for big data analytics. Debauche[10] undertook an exhaustive analysis, offering a meticulous examination of the Lakehouse (LH) concept. The investigation defines LH as an emerging paradigm, incorporating flexible big data storage components derived from Data Lakes, along with Data Warehouse elements designed to enable Online Analytical Processing (OLAP) queries essential for business intelligence.

Yet, a more recent approach is to combine a data lake with a warehouse into a new system called a lakehouse.[11] Thus, the Lakehouse seeks to provide features unique to data warehouses (ACID characteristics, SQL queries) and data lakes (data versioning, lineage, indexing, polymorphism, inexpensive storage, semantic enrichment, etc.). Subsequently, the seminal academic papers on data lakehouse were pioneered by Armbrust et al.[12], addressing pivotal challenges linked to data warehouses. As the academic community delves into this area, a growing body of research reflects the industry's dynamic evolution towards lakehouse architectures. The authors examine this ongoing shift, shedding light on its potential implications for data management practices. Thus, data landscapes have changed over time from data DWs to data DLs and, more recently, data LHs. According to Armbrust et al.[12], big data presents DWs with several significant issues, including data staleness, reliability, total cost of ownership, data lock-in, and limited use-case support. It is more difficult to create efficient data linkages and processing pipelines when different DLs and DWs are used. In addition to supporting various data types, schemas, and SQL dialects, DLs and DWs can result in more ETL and ELT processes spanning several systems, which can raise the likelihood of fault and failures. Likewise, images, sensor data, and paper make up a large amount of unstructured data in many firms. Due to their complexity, SQL DWs and their APIs are unable to manage this data.

In addition to Armbrust's foundational contributions, subsequent research has expanded and deepened the understanding of data lakehouses. Noteworthy studies have explored the practical implementation, scalability, and performance aspects of these architectures. For instance, according to Schneider[13] conducted a comprehensive analysis of data lakehouse scalability, revealing key considerations for managing large-scale datasets efficiently. Furthermore, recent works by Orescanin[8] and Hambardzumyan et al.[14] have delved into the integration of machine learning and data science capabilities within the context of data lakehouses, providing valuable insights into their potential for advanced analytics. Nevertheless, Begoli et al.[15] presented a data-driven LH architecture for applications in biological research and health data analytics. The authors incorporate several elements to support the Findability, Accessibility, Interoperability, and Reuse (FAIR) standard, all of which are essential for the area of biomedical data. By automating manual procedures and integrating pre-processing tools into data ingestion, it is possible to improve data access.

Moreover, the state of the art in data lakehouses extends beyond individual case studies, with broader discussions encompassing the implications of these architectures for diverse industry sectors. Harby[3] conducted a comparative study evaluating the impact of data lakehouses on total cost of ownership, when compared to traditional data warehouses, shedding light on the economic considerations associated with this paradigm shift. These multifaceted explorations contribute to a nuanced understanding of data lakehouses, making evident their growing significance in contemporary big data analytics landscapes. The synthesis of these works underscores the dynamic and evolving nature of the data lakehouse landscape, emphasizing not only its theoretical underpinnings but also its practical implementations and potential applications in cutting-edge areas like machine learning and advanced analytics.

Typology of DW, DL, & LH

The speed of technological development has completely changed how we produce, gather, and process large datasets.[16] Big data is generated from a variety of sources at various times and in various formats. Traditional databases are unable to handle such complex, mixed-modality, unstructured data arriving at various speeds and requiring various transform techniques.[17] Additionally, in order to store and manage data effectively, the standard Extract, Transform, and Load (ETL) method cannot support high-speed data ingestion or handle variations in the structure and modality of the incoming data. Further, data warehouse gained more popularity

by using overnight ETL operations to extract, transform, and load data from numerous data sources, and merging the same into a data cube with additional analytical tools to provide business intelligence. With the use of DWs, sophisticated, routine analytical queries might be quickly executed to produce insights for business decisions. Moreover, the big data era presented additional challenges to DWs. Indeed, the ETL pipeline in a structured DW could no longer be used to quickly extract, process, integrate, and store complex mixed modality unstructured data. In order to provide value for just-in-time decision support, high-speed and high-volume data from social media that includes photos, text, and audio as well as data from numerous connected devices or the Internet of Things (IoT) must now be ingested, analyzed, and linked in close to real-time. Traditional DWs, which store structured, filtered data, cannot be effectively transformed to manage big data despite advanced information systems such as lakehouse.

Accordingly, DL have developed during the past two decades to address the storage needs of high-speed hybrid unstructured data. DLs use effective data ingestion techniques to gather information from many sources and enable quick storage in big data storage systems. Following data analytics, knowledge can be extracted from the data and managed using DWs and other structured storage to provide business intelligence (BI). Due to the latency in processing raw data, the DL system immediately triggered a new issue of "data swamps".[18] The requirement to constantly extract metadata to update the data transformation procedures to function effectively, changing data content and layout, inconsistencies and errors in the data, and other reasons all contributed to the delay. The resultant stale data in data swamps results in resource waste and value loss. Therefore, DLs lack a way that would allow them to compare or evaluate their performance objectively, particularly in terms of their metadata management systems.

Furthermore, LH data management is developing quickly to become a norm in the industry[19,20,21,22] because of the adaptability, scalability, analytical potential, and analytic capabilities of both DWs and DLs. For expanding businesses that are cost-conscious, a Data LH offers an appealing storage alternative.

Table 1 serves as a valuable resource for evaluating the effectiveness of DW, DL, and LH, offering a systematic and thorough comparison across multiple criteria. To ensure a focused and meaningful comparison, we carefully selected the most crucial elements, with cost and quality standing out as paramount industrial metrics. Additionally, we delve into critical aspects such as atomicity, consistency, isolation, and durability (ACID) to guarantee timely and robust database transactions. Acknowledging the significance of database structure, our comparison includes criteria for structural adaptability to empower developers with comprehensive insights. Beyond structural considerations, we collect both qualitative and quantitative usability data, presenting a succinct summary of currently compatible products and storage systems for the benefit of users. This multifaceted approach not only establishes the relevance of our study but also positions it as a valuable resource for industry professionals and researchers alike.

Table 1. Comparison between DW, DL and LH.

Criteria	DW	DL	LH
Importance	Data analytics and business intelligence	Machine learning and artificial intelligence	Both data analytics and machine learning
Data type	Structured	All data types	All data types
Cost	Expensive and time consuming	Inexpensive, quick, and adaptable	Inexpensive, quick, and adaptable
Structure	Unconfigurable	Customizable	Customizable
Schema	Defined before the data is stored	Developed after the data is saved	Developed after the data is saved
Usability	Users can easily access and report data	Analyzing vast amounts of raw data without tools that classify and catalog the data can be arduous	Combining the structure and simplicity of a DW with the wider use cases of a DL
ACID Conformity	Guarantees the greatest level of integrity, data is recorded in an ACID compliant way	Updates and deletes are difficult procedures that need no-ACID compliance	ACID-compliant to ensure consistency when several parties read or write at the same time
Quality	High	Low	High

Henceforth, the Data LH paradigm approach aims to get the most out of each of the warehouses and lakes. Data lakes make it easier and cheaper to store large amounts of raw data. Then, the data warehouse comes downstream from the data lake. Through the conventional ETL process, it receives the data that has

already been cleaned and refined in the data lake. Following this approach, the data warehouse is dedicated to industrialized analyses implemented with the help of classic multidimensional analysis languages, i.e., SQL and MDX. The data lake is used for more advanced analyses, but punctually, via scripts. Ultimately, the data lakehouse paradigm addresses the issues of the entreprise DW and data lake paradigms. However, it does present a unique set of difficulties that must be overcome. The following are a few of such difficulties, such as: Architectural challenge, Integral data governance is necessary and maintaining control and flexibility in check. Furthermore, the LH architecture, should accommodate large data storage, analytics, and decision query procedures while preserving existing technology if possible. By conducting an in-depth literature review on Data Warehouses (DWs), Data Lakes (DLs), and Data Lakehouses (LH), this study aims to analyze their respective strengths and weaknesses. It involves a comprehensive evaluation of design and architectural choices, along with an examination of metadata storage and processing functionalities. The goal is to facilitate the evolution of data structures and enhance analytic capabilities, ensuring robust decision-support mechanisms for future requirements. Through this investigation, the study seeks to optimize the application of current technologies, fostering their improvement and contributing to the development and validation of a dependable Data Lakehouse (LH).

PROPOSED DATA LAKEHOUSE ARCHITECTURE

In this section, we unveil our innovative Data Lakehouse Architecture, a noteworthy stride beyond current frameworks. Our architectural proposal embodies a comprehensive strategy, skillfully intertwining the robust attributes of data warehouses and the inherent flexibility of data lakes. What sets our contribution apart is the precision in design and concept, crafting a unified platform that surmounts the confines of conventional data management systems. By combining the analytical strength found in data warehouses with the schema flexibility unique to data lakes, we achieve a harmonious fusion. This integration gives rise to an architecture of exceptional power and adaptability, laying a resilient groundwork for the streamlined execution of big data analytics. Figure 1 represents our proposed Data Lakehouse Architecture, offering a clear illustration of its structural elements and integrative features.

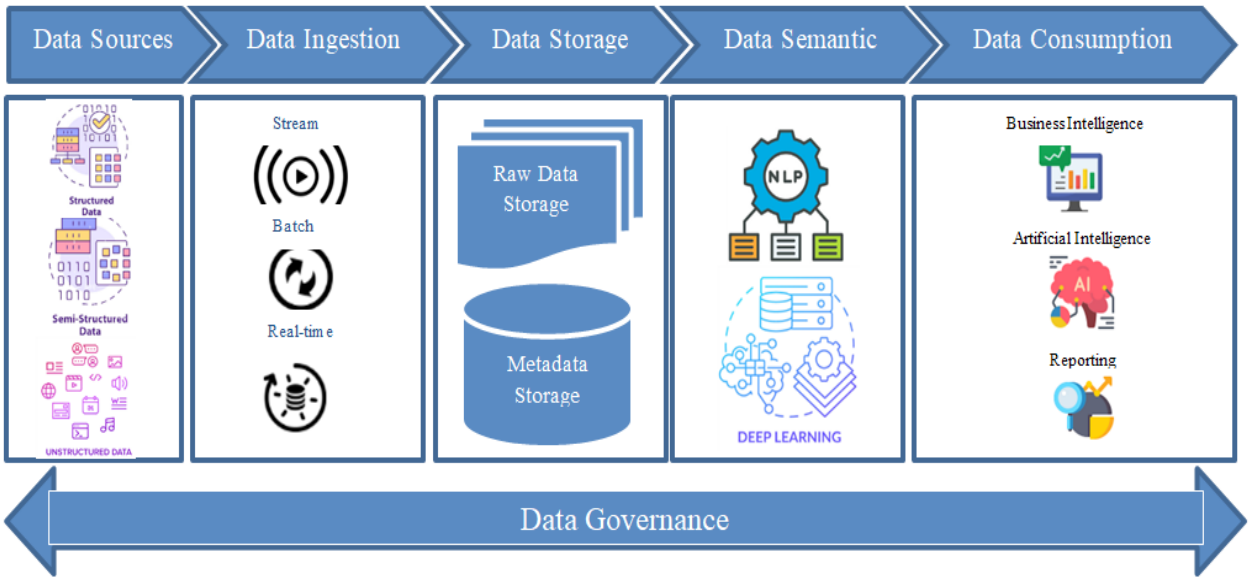

Figure 1. Proposed Innovative Data Lakehouse Architecture Design

Figure 1 visually illustrates our innovative Data Lakehouse Architecture, comprising five distinct layers meticulously crafted to ensure optimal functionality and adaptability. The architecture begins with a robust data collection layer that seamlessly integrates heterogeneous data from various sources, encompassing structured, semi-structured, and unstructured data formats. Subsequently, the data ingestion layer facilitates the seamless processing of data through stream, batch, and real-time methods, ensuring the continuous flow of information into the architecture. The third layer focuses on data storage, providing dedicated spaces for both metadata and raw data in their native formats. This structured storage approach enhances accessibility and retrieval efficiency. Moving forward, the fourth layer employs sophisticated artificial intelligence techniques, including Natural Language Processing (NLP) and cutting-edge Deep Learning models such as Recurrent Neural Networks (RNN), Long Short-Term Memory (LSTM), Bidirectional LSTM (Bi-LSTM), Gated Recurrent Unit (GRU), Seq2Seq, Auto-Encoders, and Transformers. These techniques empower the architecture with advanced semantics, enabling nuanced data interpretation and analysis. The final layer, data exploration, is designed to facilitate

https://doi.org/10.56294/mw202467

comprehensive insights by incorporating robust data governance measures. This layer ensures that users can navigate and explore the data seamlessly, promoting a user-friendly experience while adhering to stringent governance standards. The interplay of these layers results in a well-integrated and adaptive data lakehouse architecture. Importantly, the proposal architecture places a significant emphasis on ensuring robust data governance, aligning with industry best practices and standards.

As can be seen, data lakehouse architecture is made up of numerous components. The design of a data lakehouse must adhere to a set of architectural principles in order to achieve its objectives of being an adaptable platform for BI and AI as well as being agile enough to meet changing needs. Architecture principles set the underlying fundamental norms and usage guidelines for all architectural constructions. In fact, the goal of developing a new architecture paradigm is to make it agile and inventive, but it also needs to be constructively governed. It requires talent to maintain this balance. Being structured at the core implies that the layers where data is kept need to be controlled in how they handle data. These layers require comprehensive governance policies with no space for ambiguity. However, the layers where data is changed, forged, and made conducive to insights must be flexible at the borders of the data lakehouse. Flexibility does not imply adopting an irrational strategy. These layers continue to be governed by the data lakehouse's rules. However, they demonstrate some flexibility in adding new features as needed to meet requirements. Mixing raw data from the data lake layer and data warehouse from the data serving layer to generate an ML model is an example of being flexible at the edge. These datasets have various quality scores and features. Such adaptability is acceptable, yet, as it speeds up the generation of knowledge. In essence, the entreprise DW and DW paradigms were appropriate for the times. They had advantages and disadvantages. It was necessary for a fresh perspective to emerge, one that was rigid at the center yet flexible at the edges. Therefore, the data lakehouse remains the new data architectural paradigm. It aims to balance the benefits of the DL and DW paradigms while reducing their drawbacks.

CONCLUSION

In conclusion, our exploration into the proposed Data Lakehouse Architecture represents a significant leap forward in the realm of data management and analytics. By skillfully blending the analytical strengths of traditional data warehouses with the schema flexibility inherent in data lakes, our architecture offers a transformative solution that addresses key challenges in contemporary data ecosystems. The meticulous design and conceptualization of this unified platform provide a robust foundation for streamlined big data analytics, surpassing the limitations of conventional data management systems. This research contributes not only a novel technological framework but also emphasizes the importance of adaptability and performance in the face of evolving data landscapes.

Looking ahead, the future of data management holds exciting possibilities and avenues for further exploration. One avenue for future research involves extensive empirical validation and implementation of the proposed Data Lakehouse Architecture across diverse industry use cases. This will not only validate the theoretical underpinnings but also provide valuable insights into the real-world applicability and performance of our model. Additionally, advancements in machine learning and artificial intelligence continue to reshape the data analytics landscape. Future research could focus on enhancing the integration of these technologies within the Data Lakehouse Architecture, further expanding its capabilities for advanced analytics and predictive modeling. Moreover, ongoing developments in cloud computing and distributed computing infrastructures present opportunities for optimizing scalability and performance. As the data management field evolves, our proposed architecture lays the groundwork for continued innovation, adaptation, and refinement in the pursuit of more efficient and versatile solutions.

BIBLIOGRAPHIC REFERENCES

1. Hurbean, L., Miliaru, F., Muntean, M., & Danaiata, D. (2023). The Impact of Business Intelligence and Analytics Adoption on Decision Making Effectiveness and Managerial Work Performance. Scientific Annals of Economics and Business, 70, 43-54. https://doi.org/10.47743/saeb-2023-0012

2. Abu-AlSondos, I. (2023). The impact of business intelligence system (BIS) on quality of strategic decision-making. International Journal of Data and Network Science, 7, 1901-1912. https://doi.org/10.5267/j.ijdns.2023.7.003

3. Harby, A., & Zulkernine, F. (2022). From Data Warehouse to Lakehouse : A Comparative Review. https://doi.org/10.1109/BigData55660.2022.10020719

4. Cherradi, M., Bouhafer, F., & EL Haddadi, A. (2023). Data lake governance using IBM-Watson knowledge catalog. Scientific African, 21, e01854. https://doi.org/10.1016/j.sciaf.2023.e01854

5. Sawadogo, P. N., & Darmont, J. (2023). DLBench+ : A benchmark for quantitative and qualitative data lake assessment. Data & Knowledge Engineering, 145(C). https://doi.org/10.1016/j.datak.2023.102154

6. Dang, D., & Vartiainen, T. (2022). Digital Strategy in Information Systems : A Literature Review and an Educational Solution Based on Problem-Based Learning. Journal of Information Systems Education, 33, 261-282.

7. Errami, S., Hajji, H., Kenza, A. E. K., & Badir, H. (2022). Managing Spatial Big Data on the Data LakeHouse. In In book : Emerging Trends in Intelligent Systems & Network Security (p. 323-331). https://doi.org/10.1007/978-3-031-15191-0_31

8. Orescanin, D., & Hlupic, T. (2021). Data Lakehouse—A Novel Step in Analytics Architecture. International Convention on Information, Communication and Electronic Technology (MIPRO), 1242-1246. https://doi.org/10.23919/MIPRO52101.2021.9597091

9. Shiyal, B. (2021). Modern Data Warehouses and Data Lakehouses. In In book : Beginning Azure Synapse Analytics (p. 21-48). https://doi.org/10.1007/978-1-4842-7061-5_2

10. Debauche, O., Mahmoudi, S., Manneback, P., & Lebeau, F. (2022). Cloud and distributed architectures for data management in agriculture 4.0 : Review and future trends. Journal of King Saud University - Computer and Information Sciences, 34(9), 7494-7514. https://doi.org/10.1016/j.jksuci.2021.09.015

11. Zaharia, M., Ghodsi, A., Xin, R., & Armbrust, M. (2021). Lakehouse : A New Generation of Open Platforms that Unify Data Warehousing and Advanced Analytics. Conference on Innovative Data Systems Research. https://www.semanticscholar.org/paper/Lakehouse%3A-A-New-Generation-of-Open-Platforms-that-Zaharia-Ghodsi/451cf5fc9786ed4f7e1d9877f08d00f8b1262121

12. Armbrust, M., Das, T., Sun, L., Yavuz, B., Zhu, S., Murthy, M., Torres, J., Hovell, H., Ionescu, A., Łuszczak, A., Świtakowski, M., Szafrański, M., Li, X., Ueshin, T., Mokhtar, M., Boncz, P., Ghodsi, A., Paranjpye, S., Senster, P., & Zaharia, M. (2020). Delta lake : High-performance ACID table storage over cloud object stores. Proceedings of the VLDB Endowment, 13, 3411-3424. https://doi.org/10.14778/3415478.3415560

13. Schneider, J., Gröger, C., Lutsch, A., Schwarz, H., & Mitschang, B. (2023). Assessing the Lakehouse : Analysis, Requirements and Definition. https://doi.org/10.5220/0011840500003467

14. Hambardzumyan, S., Tuli, A., Ghukasyan, L., Rahman, F., Topchyan, H., Isayan, D., Harutyunyan, M., Hakobyan, T., Stranic, I., & Buniatyan, D. (2022). Deep Lake : A Lakehouse for Deep Learning. https://doi.org/10.48550/arXiv.2209.10785

15. Begoli, E., Goethert, I., & Knight, K. (2021). A Lakehouse Architecture for the Management and Analysis of Heterogeneous Data for Biomedical Research and Mega-biobanks. IEEE International Conference on Big Data (Big Data), 4643-4651. https://doi.org/10.1109/BigData52589.2021.9671534

16. Mahmoudian, M., Zanjani, s. M., Shahinzadeh, H., Kabalci, Y., Kabalcı, E., & Ebrahimi, F. (2023). An Overview of Big Data Concepts, Methods, and Analytics : Challenges, Issues, and Opportunities. https://doi.org/10.1109/GPECOM58364.2023.10175760

17. Al-Sai, Z., Abdullah, R., & Husin, H. (2019). Big Data Impacts and Challenges : A Review. https://doi.org/10.1109/JEEIT.2019.8717484

18. Cherradi, M., & El Haddadi, A. (2022). Data Lakes : A Survey Paper. In In book : Innovations in Smart Cities Applications Volume 5 (p. 823-835). https://doi.org/10.1007/978-3-030-94191-8_66

19. Park, S., Yang, C.-S., & Kim, J. (2023). Design of Vessel Data Lakehouse with Big Data and AI Analysis Technology for Vessel Monitoring System. Electronics, 12, 1943. https://doi.org/10.3390/electronics12081943

20. wong, B. (2023). Navigating the Data Architecture Landscape : A Comparative Analysis of Data Warehouse, Data Lake, Data Lakehouse, and Data Mesh. https://doi.org/10.20944/preprints202309.2113.v1

21. Kienzler, R., Blumenstiel, B., Nagy, Z., Mukkavilli, s. K., Schmude, J., Freitag, M., Behrendt, M., Salles Civitarese, D., & Hamann, H. (2023). TensorBank:Tensor Lakehouse for Foundation Model Training.

22. Ma, C., & Hu, X. (2023). A Data Analysis Privacy Regulation Compliance Scheme for Lakehouse. Proceedings of the 2023 2nd International Conference on Algorithms, Data Mining, and Information Technology, 1-5. https://doi.org/10.1145/3625403.3625405

FUNDING

None

CONFLICTS OF INTEREST

The authors declare no conflict of interest.

AUTHORSHIP CONTRIBUTION

Conceptualization: Mohamed Cherradi, Anass El Haddadi.
Data curation: Mohamed Cherradi, Anass El Haddadi.
Formal analysis: Mohamed Cherradi, Anass El Haddadi.
Fund acquisition: Mohamed Cherradi, Anass El Haddadi.
Research: Mohamed Cherradi, Anass El Haddadi.
Methodology: Mohamed Cherradi, Anass El Haddadi.
Project administration: Mohamed Cherradi, Anass El Haddadi.
Resources: Mohamed Cherradi, Anass El Haddadi.
Software: Mohamed Cherradi, Anass El Haddadi.
Supervision: Mohamed Cherradi, Anass El Haddadi.
Validation: Mohamed Cherradi, Anass El Haddadi.
Display: Mohamed Cherradi, Anass El Haddadi.
Writing – original draft: Mohamed Cherradi, Anass El Haddadi.
Writing – review and editing: Mohamed Cherradi, Anass El Haddadi.