EVENTO ON-LINE
III CIGELUBRA
Congresso Internacional de Gestão e Engenharia Luso-Brasileiro
02 a 04 de dezembro de 2026
Inovação e Transformação Digital como Estratégia Competitiva nas Engenharias

The Evolution of Data Management Architectures: From Data Warehouses to Implementing Data Lakehouses – A Managerial Perspective

Vitor Hugo Pagani
University of Huddersfield – United Kingdom
Beatryz de Almeida Alcantara
Departamento de Engenharia de Produção – Universidade Tecnológica Federal do Paraná, Ponta Grossa, Brasil
Regina Negri Pagani
Departamento de Engenharia de Produção – Universidade Tecnológica Federal do Paraná, Ponta Grossa, Brasil
Carla Alexandra Soares Geraldes
CeDRI, SusTEC, Bragança Polytechnic University, Campus de Santa Apolónia, Bragança, Portugal

Abstract: Data management architectures have evolved from traditional Data Warehouses to flexible Data Lakes, and recently to the hybrid Data Lakehouse model. While Data Warehouses excel at structured business intelligence, Data Lakes handle vast amounts of unstructured data. The Data Lakehouse combines the strengths of both, delivering superior performance, governance, and scalability for complex data environments. Despite these advancements, existing literature lacks clear guidance on these technologies from a managerial perspective—specifically regarding comparative benefits, organizational applicability, and strategic implementation. Furthermore, no structured implementation roadmap currently exists. To bridge this gap, this study presents a didactic, managerial-oriented roadmap for implementing a Data Lakehouse. It details core characteristics, costs, technical requirements, and cross-sector applications. Built upon an exploratory review of legacy architectures and a systematic literature review, this research establishes a solid theoretical foundation for actionable managerial guidelines. Ultimately, these findings empower decision-makers to evaluate data architecture investments, align technology choices with organizational needs, and manage both operational and strategic trade-offs. By integrating essential managerial considerations into technical decisions, organizations can better optimize data assets, manage costs, strengthen governance, and ensure long-term scalability.

Keywords: Data Management Architecture, Data Warehouse, Data Lake, Data Lakehouse, Managerial Guidelines.

1. Introduction

The evolution of data management architectures has been marked by the progression from Data Warehouses to Data Lakes and, more recently, to the innovative paradigm of Data Lakehouses (Solodovnikova; Niedrite, 2020). Each of these architectures addresses distinct challenges in data storage, processing, and analytics, reflecting the shifting demands of businesses in the age of big data and artificial intelligence (Boussaid, 2019). While Data Warehouses excel in handling structured data and business intelligence applications, Data Lakes have emerged as a flexible solution for managing large volumes of unstructured and semi-structured data. The Data Lakehouse, combining the strengths of both, represents a

transformative approach to modern data management, offering enhanced performance, governance, and scalability (Giebler et al., 2019; Nambiar; Mundra, 2022).

Despite the growing interest in Data Lakehouse s, much of the literature remains technical and targeted toward IT professionals and data architects. This creates a significant gap in addressing the needs of company managers, who are increasingly required to make strategic decisions about data infrastructure investments. Managers often lack a comprehensive understanding of the comparative benefits, costs, and applicability of Data Warehouses, Data Lakes, and Data Lakehouse s to their specific organizational contexts.

This paper seeks to bridge that gap by offering a structured roadmap to implement a Data Lakehouse, highlighting its key characteristics, costs, and potential use cases across various sectors. Additionally, the paper provides a comparative analysis of Data Warehouses, Data Lakes, and Data Lakehouse, elucidating the advantages and limitations of each. To reach this purpose, an exploratory search on the literature was done. Then, in order to propose the roadmap, a systematic analysis on the extant literature was performed. The roadmap is tailored to provide actionable insights for managers, enabling them to make informed decisions about the best data architecture for their companies based on size, budget, and data needs. The results show a roadmap that companies and managers might use to structure a Data Lakehouse, as well as the best features for each company's size and sector. Despite using technical language, the paper was aimed to clarify to Managers, more than IT professionals, the characteristics and benefits of employing these technologies, and how to follow a path for practical strategies.

Thus, this paper presents a double contribution. The first one lies in its focus on the managerial perspective, which is often overlooked in technical discussions. By addressing this gap, the paper aims to empower organizations to adopt Data Lakehouse s strategically, enhancing their ability to harness the full potential of their data assets while optimizing costs and ensuring scalability for future growth. The second one is to bridge the gap on the literature, offering a roadmap for structuring a Data Lakehouse, which was not identified in the extant literature.

This paper is structured as follows: The introduction outlines the evolution of data management concepts and technologies. Section 2 provides the theoretical background, discussing Data Warehouses, Data Lakes, and the hybrid Data Lakehouse, highlighting their characteristics, purposes, and limitations. The Material and Methods section details the data collection and analysis process. The Results and Discussion section presents findings on key sectors using these technologies, explaining their specific applications. It also includes a roadmap for implementing a Data Lakehouse. The conclusion offers actionable insights for managers, bridging the gap between IT professionals and organizational leaders in leveraging modern data management solutions.

2. Theoretical background

2.1 Data Management

In the ever-evolving field of data management, two central concepts have emerged in the last decades: the Data Warehouse and the Data Lake. Both serve as essential repositories for storing vast amounts of data, yet they cater to different needs and use cases. Data Warehouses have been around since the late 1980s and early 1990s, designed to support business intelligence activities by providing a centralized repository for structured data, optimized for querying and reporting (Herden, 2020; El Aissi et al., 2022).

A Data Warehouse is a centralized repository designed to store large volumes of structured and semi-structured data from various sources (El Aissi et al., 2022). A Data Warehouse aggregates data from multiple sources, including transactional systems, log files, and other databases (Vaisman; Zimányi, 2014). This data is cleaned, transformed, and loaded into the

warehouse through processes known as Extract, Transform, and Load (ETL) or Extract, Load, and Transform (ELT) (Venkateswarlu; Vasista, 2023). The primary purpose of a Data Warehouse is to enable organizations to perform complex queries and analyses on historical data, providing valuable insights for decision-making.

Key features of Data Warehouse (Herden, 2020; Harby; Zulkernine, 2022) included:

- Structured Data: Data Warehouses primarily handled structured data, which is organized in a predefined schema.
- ETL Processes: Extract, Transform, Load (ETL) processes were used to integrate data from different sources into the warehouse.
- Analytical Capabilities: They provided tools for querying and analyzing data, making them essential for decision-making.

As the volume and variety of data grew, traditional Data Warehouses faced several challenges (Batwada et al., 2020; Cuzzocrea et al., 2013):

- Scalability: Handling large volumes of data became difficult and expensive.
- Flexibility: They struggled to manage unstructured data like text, images, and videos.
- Real-Time Processing: Data Warehouses were not designed for real-time data processing, which became increasingly important.

To address these challenges, a more recent development of the concept concerning data management, particularly in big data projects (Giebler et al., 2019), was introduced, the Data Lakes. They emerged in the early 2010s as a response to the growing need to handle large volumes of unstructured and semi-structured data. A Data Lake is a storage repository that holds a vast amount of raw data in its native format until it is needed (Harby; Zulkernine, 2022). It has emerged as a transformative solution for managing vast, diverse data (Boussaid, 2019). Unlike traditional Data Warehouses, which require data to be transformed and structured before storage (schema-on-write), Data Lakes use a schema-on-read approach, allowing data to remain in its raw form and be transformed during analysis (Khine; Wang, 2018; Sawadogo; Darmont, 2021). This flexibility enables organizations to store unprocessed data from multiple sources, making it easier to explore and analyze data without fixed schema limitations, especially in contexts requiring real-time analysis (Khine; Wang, 2018; Sawadogo; Darmont, 2021).

Typically built on scalable, cost-effective technologies, Data Lakes support large-scale data storage and processing, with frameworks for real-time processing (Macedo, 2023). The primary advantage of Data Lakes is their ability to eliminate data silos by integrating diverse data sources, offering a holistic view of organizational data (Sawadogo; Darmont, 2021). Chart 1 offers a comparison between Data Lake and Data Warehouse (Salamkar; Allam, 2019; Errami et al., 2023).

Chart 1. Data Lake vs. Data Warehouse

Characteristic	Data Lake	Data Warehouse	References
Type of Data	Raw, unstructured, semi-structured, and structured data	Structured and processed data	Azzabi et al. (2024); Lamer et al. (2024); Nambiar & Mundra (2022)
Purpose	Data science, machine learning, big data analytics	Reporting and business analytics	El Aissi et al. (2022); Wieder & Nolte (2022)
Data Storage	Flat, distributed storage (e.g., Amazon S3, Azure Data Lake)	Optimized relational databases (e.g., Snowflake, Redshift)	Bimonte et al. (2022); Lamer et al. (2024)
Schema	Schema-on-read (flexible; defined during data usage)	Schema-on-write (rigid; defined before loading data)	Nambiar & Mundra (2022); El Aissi et al. (2022)

Scalability	Highly scalable for massive data volumes	Limited scalability depending on the platform	Wieder & Nolte (2022); Azzabi et al. (2024)
Cost	Cost-efficient for storage, expensive for processing large-scale queries	Higher storage costs but optimized for query performance	Lamer et al. (2024); Nambiar & Mundra (2022)
Users	Data scientists, advanced analysts	Business analysts, executives	El Aissi et al. (2022); Wieder & Nolte (2022)
Governance	Challenging governance and risk of "data swamp"	Well-established governance and data quality	Bimonte et al. (2022); Nambiar & Mundra (2022)
Performance	May require tuning for queries on large datasets	High performance for structured query processing	Wieder & Nolte (2022); Lamer et al. (2024)
Security & Compliance	Requires additional tools for robust security and regulatory compliance	Built-in security and compliance features	Azzabi et al. (2024); Nambiar & Mundra (2022)
Integration	Flexible with modern tools for IoT, streaming, and unstructured data	Integrated with traditional enterprise systems	Bimonte et al. (2022); El Aissi et al. (2022)
Use Cases	IoT analytics, machine learning model development, customer behavior analysis	Financial reporting, sales dashboards, executive decision support	Wieder & Nolte (2022); Nambiar & Mundra (2022)

Source: Authors (2026)

From Chart 1, it is possible to view a comparison of the key characteristics of Data Lakes and Data Warehouses, highlighting differences in data types, purposes, storage methods, scalability, costs, user demographics, governance challenges, performance expectations, security measures, integration capabilities, and typical use cases. As observed, Data Lakes complement Data Warehouses by offering a flexible platform for storing raw data and enabling transformations as needed, whereas Data Warehouses are optimized for structured, complex analytical queries. Overall, Data Lakes provide a scalable, flexible solution for integrating and analyzing large data volumes, but their successful implementation requires strong metadata management and governance to maintain data quality.

The latest evolution in data storage is the Data Lakehouse, which is approached in the next subsection.

2.2 Data Lakehouse

The concept of the Data Lakehouse emerged around 2017. It was introduced as a solution to combine the best aspects of Data Warehouses and Data Lakes, addressing the limitations of both. Data Lakehouse s provide the structured data management and analytical capabilities of Data Warehouses while maintaining the flexibility and scalability of Data Lakes. This hybrid architecture aims to deliver the performance and management capabilities of Data Warehouses alongside the adaptability and scalability of Data Lakes. Based on the extant literature (Errami et al., 2023; Harby; Zulkernine, 2022, 2025; Hassan, 2024; Janssen et al., 2024; Park et al., 2023; Schneider et al., 2024), Chart 2 details a generic structure of a Data Lakehouse.

Chart 2. Generic structure of a Data Lakehouse

Layer	Purpose	Technologies	References
Storage Layer	Stores raw data in its native format, supporting structured, semi-structured, and unstructured data.	Amazon S3, Azure Data Lake Storage, Google Cloud Storage, Hadoop Distributed File System (HDFS).	Harby & Zulkernine (2022); Errami et al. (2023); Schneider et al. (2024).
Metadata Management	Maintains metadata for schema, governance, and	Apache Hive, Apache Iceberg, Delta Lake, AWS Glue.	Harby & Zulkernine (2025); Hassan (2024); Janssen et al. (2024).

	indexing for efficient querying and organization.		
Data Processing Layer	Processes and transforms data for analytics, machine learning, and reporting.	Apache Spark, Apache Flink, Databricks, Presto, Apache Beam.	Errami et al. (2023); Park et al. (2023); Schneider et al. (2024).
Query and Analytics	Enables SQL-like querying and business analytics for structured data.	Trino, Databricks SQL, Snowflake, Google BigQuery.	Harby & Zulkernine (2022); Schneider et al. (2024).
Governance & Security	Ensures data access control, compliance, and auditing.	Apache Ranger, AWS Lake Formation, Azure Purview, Databricks Unity Catalog.	Hassan (2024); Janssen et al. (2024).
Machine Learning Layer	Supports AI/ML model development and deployment.	MLflow, TensorFlow, PyTorch, Databricks Machine Learning.	Errami et al. (2023); Harby & Zulkernine (2025).
Medallion Architecture	Organizes data into Bronze (raw), Silver (enriched), and Gold (curated) layers for progressive refinement.	Delta Lake, Apache Hudi, Iceberg for table formats; ETL/ELT tools for data ingestion processes.	Databricks (2020); Schneider et al. (2024); Harby & Zulkernine (2025).

Source: Authors (2026)

The Storage Layer emphasizes the importance of cloud object storage for scalability and cost-effectiveness. The Metadata Management layer is vital for effective governance and efficient querying. The Data Processing Layer utilizes various processing engines to support diverse analytical workloads. The Query and Analytics layer focuses on enabling business intelligence capabilities through SQL-like querying. The Governance & Security layer ensures compliance and robust access control measures. The Machine Learning Layer supports advanced analytics and model deployment. The Medallion Architecture provides a structured approach to data refinement through its tiered layers.

The idea behind the Data Lakehouse is to create a unified platform that can handle structured, semi-structured, and unstructured data, making it easier for organizations to perform advanced analytics and machine learning tasks without needing to maintain separate systems for different types of data. This hybrid approach supports a wide range of data types and use cases, including machine learning and AI. This evolution reflects the growing need for more efficient, flexible, and scalable data storage solutions in the era of big data and advanced analytics.

3. Material and methods

This research employes an exploratory review of literature (Pagani, 2023) in order to uncover the concepts of Data Warehouse, Data Lake and Data Lakehouse. The review was performed Scopus database, covering journals, books and conference papers based on the title, without time delimitation. The systematic review followed the Methodi Ordinatio 2.0 protocol, where papers were ranked according to their scientific relevance (InOrdinatio index), considering the impact factor of the journals, the number of citations, and the year of publication. Only documents with high relevance and alignment with data architecture roadmaps were selected for the final analysis of 100 papers. This preliminary search served to build the literature review.

Then, once it was noticed missing on the preliminary search works with roadmaps proposition, a systematic review was performed, using the principles of the methodology Methodi Ordinatio (Pagani et al., 2023). The search parameters for the systematic review, which results were used to propose the roadmap, it detailed on Table 1.

Table 1. Parameters for the material search on Scopus

Combinations	Parameters and procedures Applied on Scopus Database	Papers found
("Data Warehouse" AND "Data Lake")	Seach on TAK, with no temporal delimitation	216
("Data Warehouse" AND "Data Lake") AND "Data Lakehouse "	Seach on TAK, with no temporal delimitation	18
("Data Warehouse" OR "Data Lake" OR "Data	Seach on Title, with no temporal	0

Lakehouse ") AND roadmap	delimitation	
Eliminated by filtering out procedures	Duplicates, books, conference papers	(-134)
Total papers left for reading and analysis		100

Source: Authors (2026)

The content analysis of the papers provided material to propose the roadmap and other suggestions for practical applications.

4. Results

This section provides a roadmap and for structuring of a Data Lakehouse. From the content analysis it was possible to identify the main sectors that make use of Data Warehouse, Data Lake and Data Lakehouse. They are described on Chart 3.

Chart 3. Main sectors and use of Data Warehouse, Data Lake and Data Lakehouse

Sector	Technology	How They Use It	Why They Use It	References
Healthcare	Data Lake	Storing and analyzing large-scale patient data, genomics, and real-time monitoring.	To manage diverse datasets for improved patient outcomes, personalized medicine, and operational efficiency.	Azzabi et al. (2024); Lamer et al. (2024)
Retail and E-commerce	Data Lakehouse	Analyzing customer behavior, real-time pricing, and inventory optimization.	To enhance customer experience through targeted marketing and optimized inventory management.	El Aissi et al. (2022); Wieder & Nolte (2022)
Financial Services	Data Warehouse	Fraud detection, risk analysis, and portfolio management.	To enable complex queries on structured data for regulatory compliance and financial reporting.	Nambiar & Mundra (2022); Bimonte et al. (2022)
Logistics and Supply Chain	Data Lakehouse	Real-time tracking, route optimization, and demand forecasting.	To improve operational efficiency and responsiveness to market changes through real-time data insights.	Wieder & Nolte (2022); Azzabi et al. (2024)
Media and Entertainment	Data Lake	Streaming analytics, recommendation systems, and content personalization.	To provide personalized content experiences and optimize streaming services based on user preferences.	Lamer et al. (2024); Nambiar & Mundra (2022)
Manufacturing	Data Warehouse	Quality control analytics, supply chain management, and production optimization.	To enhance operational efficiency and reduce costs through data-driven decision-making processes.	El Aissi et al. (2022); Wieder & Nolte (2022)

Source: Authors (2026)

Building a Data Lakehouse involves several key steps to ensure a robust, scalable, and efficient data management system (Azzabi et al., 2024; Errami et al., 2023; Gade, 2022; Harby; Zulkernine, 2022, 2025; Hassan, 2024; Janssen et al., 2024; Park et al., 2023; Schneider et al., 2024).

4.1 Roadmap for Building a Lakehouse

The comprehensive generic roadmap serves as a guide through the process. Despite it is directed to structure a Lakehouse, it can be used as a guide for any data managing technology approached in this paper, for any sector desired:

1. Define Objectives and Use Cases
    - Identify Business Needs: Determine the primary objectives for implementing a Data Lakehouse, such as improving analytics capabilities, enhancing data accessibility, or supporting machine learning initiatives.
    - Select Use Cases: Prioritize specific use cases that can deliver quick wins, such as customer 360 views, real-time analytics, or operational reporting. This helps in gaining executive support and securing initial funding.
2. Assemble the Right Team

- Stakeholder Engagement: Form a cross-functional team that includes an executive sponsor, data analysts, data engineers, architects, and governance managers. This diverse team will ensure that all perspectives are considered during implementation.
- Training and Skill Development: Ensure team members are trained in the technologies and methodologies relevant to the Lakehouse architecture.

3. Design the Architecture

Layered Architecture: Plan for a layered architecture consisting of:

- Storage Layer: Utilize cloud storage solutions (e.g., Amazon S3, Azure Data Lake Storage) to store raw data in its native format.
- Metadata Management Layer: Implement tools for metadata management (e.g., Apache Hive, Delta Lake) to facilitate data governance and organization.
- Data Processing Layer: Choose processing engines (e.g., Apache Spark, Presto) to handle batch and real-time data processing.
- Query and Analytics Layer: Integrate SQL-like querying capabilities using tools like Trino or Databricks SQL for business analytics.
- Governance & Security Layer: Establish governance frameworks (e.g., AWS Lake Formation) to ensure compliance and security.

A critical aspect of this design is interoperability. By utilizing open table formats such as Delta Lake or Apache Iceberg, the Lakehouse architecture prevents 'vendor lock-in,' ensuring that data remains accessible across different cloud providers and processing engines. This technological flexibility is a key managerial consideration, as it protects long-term investments and ensures the infrastructure can adapt to future technological shifts.

4. Data Ingestion Strategy

- ETL/ELT Processes: Develop robust ETL/ELT processes to ingest data from various sources into the Lakehouse. This includes both historical and real-time data ingestion.
- Schema-on-Read Approach: Implement a schema-on-read strategy to allow flexibility in how data is structured upon querying while maintaining raw data integrity.

5. Implementation Phases

- Phase 1: Initial Deployment: Start with a pilot project that focuses on one high-priority use case (e.g., migrating semi-structured customer data).
- Phase 2: Expand Functionality: After demonstrating success with the initial project, expand the scope to include additional datasets (e.g., finance or supply chain records).
- Phase 3: Full Scale-Up: Gradually scale up the Lakehouse by integrating more complex datasets and use cases, ensuring continuous stakeholder engagement throughout.

6. Performance Optimization

- Indexing and Caching: Implement indexing and caching strategies to enhance query performance on large datasets.
- Monitoring and Tuning: Regularly monitor performance metrics and tune the architecture as necessary to meet evolving business needs.

7. Governance and Compliance

- Data Quality Measures: Establish processes for ensuring data quality through validation checks and automated monitoring.
- Access Control Policies: Implement role-based access control (RBAC) to secure sensitive data and comply with regulatory requirements.

8. Machine Learning Integration

- Support for AI/ML Workflows: Integrate machine learning frameworks (e.g., MLflow, TensorFlow) into the Lakehouse architecture to facilitate model development and deployment.

- Real-Time Analytics Capability: Ensure the architecture supports real-time analytics for immediate insights into operational metrics.

9. Continuous Improvement
    - Feedback Loop: Create mechanisms for gathering feedback from users to continuously improve the Lakehouse functionality.
    - Iterative Development: Adopt an agile approach to incrementally enhance features based on user needs and technological advancements.

By following this roadmap, organizations can effectively transition from traditional data management architectures to a Data Lakehouse that combines the strengths of both Data Lakes and warehouses. This structured approach ensures alignment with business objectives while fostering innovation through enhanced data accessibility and analytical capabilities. Chart 4 summarizes the steps, tasks, key tools/technologies for the roadmap.

Chart 4. Steps, tasks, key tools/technologies for the roadmap

Step	Task Description	Key Tools/Technologies	References
1. Define Objectives	Identify business needs, data sources, volume, and specific use cases (e.g., real-time analytics).	Business workshops, data profiling tools.	Harby & Zulkernine (2022); Schneider et al. (2024).
2. Assemble the Team	Form a cross-functional team with stakeholders including data analysts, engineers, and governance managers.	Team collaboration tools (e.g., Slack, Microsoft Teams).	Errami et al. (2023); Janssen et al. (2024).
3. Design the Architecture	Plan a layered architecture including storage, metadata management, processing, and governance layers.	Cloud storage (Amazon S3, Azure Data Lake), Apache Hive, Apache Spark.	Harby & Zulkernine (2025); Schneider et al. (2024).
4. Develop Ingestion Strategy	Create ETL/ELT processes for data ingestion from various sources into the Lakehouse.	Apache NiFi, Talend, Airflow for orchestration; Delta Lake for data formats.	Park et al. (2023); Azzabi et al. (2024).
5. Implement Metadata	Establish metadata management to handle schemas and governance policies effectively.	Delta Lake, AWS Glue, Apache Iceberg.	Ghodsi et al. (2020); Hassan (2024).
6. Deploy Processing Layer	Set up processing engines for batch and real-time analytics and transformations.	Apache Spark, Databricks, Apache Flink.	Errami et al. (2023); Schneider et al. (2024).
7. Enable Querying	Integrate analytics platforms to support SQL-like querying and business intelligence needs.	Trino, Databricks SQL, Snowflake, Google BigQuery.	RisingWave (2024); Harby & Zulkernine (2025).
8. Add Governance & Security	Implement data access control measures and compliance protocols to ensure data security and quality.	Apache Ranger, Azure Purview, AWS Lake Formation.	Hassan (2024); Janssen et al. (2024).
9. Integrate AI/ML	Introduce tools for AI/ML workflows and model deployment within the Lakehouse environment.	MLflow, TensorFlow, PyTorch; Databricks Machine Learning capabilities.	Park et al. (2023); Schneider et al. (2024).
10. Optimize & Scale	Continuously monitor performance and optimize architecture for scalability based on business needs.	Auto-scaling features in cloud platforms; performance monitoring tools like Grafana.	Gandomi & Haider (2015); Schneider et al. (2024).

Source: Authors (2026)

Each step is designed to build upon the previous one to create a cohesive strategy for implementing a Data Lakehouse. The selected tools and technologies are aligned with best practices identified in the referenced papers.

Chart 5 that integrates the advisory information with the comparison based on company type and budget.

Chart 5. Recommendations on company type, budget and technologies

Company Type	Recommended Architecture	Budget	Technologies/Key Tools	Key Use Cases	References
Startups	Data Lake	Low to Medium	Amazon S3, Apache Spark, AWS Glue	Innovation, exploratory analytics, AI/ML research	Harby & Zulkernine (2022); Schneider et al. (2024).

Medium-Sized Companies	Data Lake or Lakehouse	Medium	Azure Data Lake, Delta Lake, Databricks	Real-time analytics, AI/ML, scalability	Errami et al. (2023); Park et al. (2023).
Large Enterprises	Data Warehouse or Lakehouse	High	Snowflake, Google BigQuery, AWS Lake Formation	BI, compliance, hybrid analytics, real-time decision-making	Harby & Zulkernine (2025); Janssen et al. (2024).
General Advisory	Data Lakehouse	SMEs: $50K - $200K; Large: $500K+	Amazon S3, Azure Data Lake Storage, Apache Hive, Delta Lake	Customer analytics, operational reporting	Gandomi & Haider (2015); Hassan (2024).
		Ongoing Costs: 20-30% of initial setup annually	Apache Spark, Databricks	Advanced analytics and machine learning	Errami et al. (2023); Schneider et al. (2024).
			Trino, Snowflake	SQL-like querying and business intelligence	RisingWave (2024); Harby & Zulkernine (2025).
			Apache Ranger, AWS Lake Formation	Data access control and compliance	Hassan (2024); Janssen et al. (2024).
			MLflow, TensorFlow	AI/ML workflows and model deployment	Park et al. (2023); Schneider et al. (2024).

Source: Authors (2026)

This table provides a clear overview of recommended architectures, budgets, technologies, key use cases, and references for implementing a Data Lakehouse.

4.2 Discussions

The analysis of how various sectors utilize Data Warehouse, Data Lake, and Data Lakehouse technologies underscores significant differences in implementation and application strategies. These differences are often shaped by the size of the company, its budget, and its absorptive capacity. Understanding these dimensions is crucial for tailoring data management solutions that align with organizational needs and capabilities.

Takin the company size and budget into consideration, startups generally adopt Data Lake architectures due to their lower initial investment and flexibility in handling unstructured data. These companies often prioritize innovation and exploratory analytics, where the cost-effective storage and processing capabilities of technologies like Amazon S3 and Apache Spark are advantageous. Harby and Zulkernine (2022) and Schneider et al. (2024) highlight how these technologies allow startups to engage in AI/ML research without the burden of substantial upfront costs.

Medium-sized companies tend to favor Data Lakehouses, combining the best of Data Lakes and warehouses to support both real-time analytics and AI/ML applications. With a medium budget, these companies can afford technologies like Azure Data Lake and Databricks, which provide scalable solutions suitable for growing data demands (Errami et al., 2023; Park et al., 2023). The ability to handle diverse datasets and offer rapid analytics capabilities makes the Lakehouse model appealing to this segment. The alignment between company size and technology choice is deeply rooted in the concept of Absorptive Capacity. A Data Lakehouse architecture enhances an organization's ability to recognize the value of new, diverse data, assimilate it through integrated metadata layers, and apply it for commercial ends. By reducing architectural complexity, it allows managers to focus on transforming raw information into strategic knowledge, effectively bridging the gap between data acquisition and value creation.

Large enterprises often implement Data Warehouses or Lakehouses, given their higher budgets and more complex data management needs. Tools such as Snowflake and Google BigQuery cater to their requirements for business intelligence, compliance, and real-time decision-making (Harby; Zulkernine, 2025; Janssen et al., 2024). The robust infrastructure

and advanced features of these technologies enable large companies to efficiently manage structured data and support extensive analytical processes.

In healthcare, the adoption of Data Lakes is driven by the need to manage large-scale and diverse datasets, such as patient records and genomics data (Azzabi et al., 2024). The sector's high absorptive capacity, fostered by ongoing research and development, allows for the integration of these complex datasets to improve patient outcomes and operational efficiency.

The retail and e-commerce sector exemplifies the use of Data Lakehouse s for analyzing customer behavior and optimizing inventory. Companies in this sector benefit from the real-time analytics capabilities of Data Lakehouse s, which enhance customer experiences and operational efficiency (El Aissi et al., 2022). Furthermore, the implementation of a Lakehouse model significantly reduces 'Decision Latency'—the time elapsed between a business event and the subsequent action. By unifying Business Intelligence (historical analysis) and Machine Learning (predictive analysis) into a single source of truth, the architecture eliminates the need for time-consuming data movements between separate systems. This provides managers with the agility required for real-time responsiveness in dynamic markets, such as logistics tracking and inventory optimization.

In financial services, Data Warehouses are essential for handling structured data necessary for regulatory compliance and financial reporting. The sector’s absorptive capacity enables it to utilize complex queries and risk analysis tools effectively (Nambiar; Mundra, 2022).

4.2.1 Strategic Implications

For sectors like logistics and supply chain, which rely on real-time tracking and demand forecasting, Data Lakehouse s offer a strategic advantage by providing timely insights into market changes (Azzabi et al., 2024). The integration of real-time data processing tools such as Apache Spark and Trino facilitates agile decision-making and enhances operational responsiveness.

Media and entertainment companies leverage Data Lakes to offer personalized content experiences and optimize streaming services (Nambiar; Mundra, 2022). The absorptive capacity in this sector supports the continuous adaptation to user preferences, ensuring a competitive edge in a dynamic market.

Manufacturing entities benefit from Data Warehouses for quality control analytics and production optimization. The structured approach to data management aligns with their need for precision and efficiency in operations (El Aissi et al., 2022).

In conclusion, the choice of data management technology is closely linked to a company's size, budget, and absorptive capacity. By aligning these factors with sector-specific needs, companies can adopt data solutions that enhance their operational efficiency, innovation potential, and competitive advantage.

5. Conclusions

The data management landscape has evolved significantly, driven by the distinct roles and limitations of Data Warehouses and Data Lakes. Since their emergence in the late 1980s, Data Warehouses have served as the foundation for business intelligence by enabling complex analysis on structured historical data. However, growing data volumes and rigid schema-on-write requirements make traditional warehouses costly to scale and ill-equipped for unstructured media. In response, Data Lakes emerged in the early 2010s, utilizing a flexible schema-on-read model to store raw, diverse datasets for real-time analysis. Despite these advantages, Data Lakes require rigorous governance and metadata management to avoid becoming disorganized "data swamps."

To bridge this gap, the Data Lakehouse has gained traction as a hybrid architecture. It combines the structured management of traditional warehouses with the scalability of lakes, allowing organizations to run advanced analytics and machine learning on a unified platform without deploying separate systems. Ultimately, Data Warehouses remain ideal for structured reporting and compliance, Data Lakes excel at raw data exploration, and Data Lakehouses offer a versatile alternative, provided an organization has the necessary resources and expertise to implement them.

Because existing literature lacks practical frameworks to assist managers with these choices, this paper introduced a decision-making roadmap tailored to organizational size, budget, and technological maturity. Under this framework, startups benefit most from cost-effective Data Lakes to foster innovation. Medium-sized companies are advised to leverage Data Lakes or Lakehouses for real-time scalability, while large enterprises should utilize Data Warehouses or Lakehouses to satisfy demanding analytical and compliance requirements. A limitation of this work is that the proposed roadmap was not tested in a real-world setting; therefore, future research should focus on applying and validating this roadmap within a practical corporate environment.

The main limitation of this work it that it does not present a practical application with the roadmap. Therefore, future works could present the application of this roadmap on a company.

Acknowledgments

The authors thank the National Council for Scientific and Technological Development (CNPq) and the Coordination for the Improvement of Higher Education Personnel – Brazil (CAPES) for the support received. This work was financially supported by the Research Unit on Governance, Competitiveness and Public Policies (UIDB/04058/2020) + (UIDP/04058/2020), funded by national funds through FCT (Fundação para a Ciência e a Tecnologia), Portugal.

References

AZZABI, S.; ALFUGHI, Z.; OUDA, A. Data Lakes: a survey of concepts and architectures. **Computers**, v. 13, n. 7, p. 183, 2024.

BATWADA, R. K.; MITTAL, N.; PILLI, E. S. Uncovering Data Warehouse issues and challenges in Big Data management. In: **BIG DATA, MACHINE LEARNING, AND APPLICATIONS: FIRST INTERNATIONAL CONFERENCE, BigDML 2019**, Silchar, India, 16-19 Dec. 2019. **Proceedings...** [S. l.]: Springer International Publishing, 2020. p. 48-59.

BOUSSAID, O. Data Lakes: new generation of information systems? In: **ADVANCES IN COMPUTING SYSTEMS AND APPLICATIONS: PROCEEDINGS OF THE 3RD CONFERENCE ON COMPUTING SYSTEMS AND APPLICATIONS**, 2019. [S. l.]: Springer International Publishing, 2019. p. 91.

CUZZOCREA, A.; BELLATRECHE, L.; SONG, I. Y. Data warehousing and OLAP over big data: current challenges and future research directions. In: **INTERNATIONAL WORKSHOP ON DATA WAREHOUSING AND OLAP**, 16., 2013. **Proceedings...** [S. l.: s. n.], 2013. p. 67-70.

EL AISSI, M. E. M.; BENJELLOUN, S.; LOUKILI, Y.; LAKHRISSI, Y.; BOUSHAKI, A. E.; CHOUGRAD, H.; ELHAJ BEN ALI, S. Data Lake versus Data Warehouse architecture: a comparative study. In: **INTERNATIONAL CONFERENCE ON WIRELESS TECHNOLOGIES, EMBEDDED, AND INTELLIGENT SYSTEMS**, 6., 2020. **WITS 2020: proceedings of the 6th International Conference on Wireless Technologies, Embedded, and Intelligent Systems**. Singapore: Springer, 2022. p. 201-210.

ERRAMI, S. A.; HAJJI, H.; EL KADI, K. A.; BADIR, H. Spatial big data architecture: from Data Warehouses and Data Lakes to the Lakehouse. **Journal of Parallel and Distributed Computing**, v. 176, p. 70-79, 2023.

GADE, K. R. Data Lakehouses: combining the best of Data Lakes and Data Warehouses. **Journal of Computational Innovation**, v. 2, n. 1, 2022.

GIEBLER, C.; GRÖGER, C.; HOOS, E.; SCHWARZ, H.; MITSCHANG, B. Leveraging the Data Lake: current state and challenges. In: **INTERNATIONAL CONFERENCE ON BIG DATA ANALYTICS AND KNOWLEDGE DISCOVERY – DaWaK**, 21., 2019, Linz, Austria. **Big Data Analytics and Knowledge Discovery: proceedings...** [S. l.]: Springer International Publishing, 2019. p. 179-188.

HARBY, A. A.; ZULKERNINE, F. From Data Warehouse to Lakehouse: a comparative review. In: **IEEE INTERNATIONAL CONFERENCE ON BIG DATA**, 2022. **Proceedings...** [S. l.]: IEEE, 2022. p. 389-395.

HARBY, A. A.; ZULKERNINE, F. Data Lakehouse: a survey and experimental study. **Information Systems**, v. 127, p. 102460, 2025.

HASSAN, I. Storage structures in the era of big data: from Data Warehouse to Lakehouse. **Journal of Theoretical and Applied Information Technology**, v. 102, n. 6, 2024.

HERDEN, O. Architectural patterns for integrating Data Lakes into Data Warehouse architectures. In: **INTERNATIONAL CONFERENCE ON BIG DATA ANALYTICS – BDA**, 8., 2020, Sonepat, India. **Big Data Analytics: proceedings...** [S. l.]: Springer International Publishing, 2020. p. 12-27.

JANSSEN, N.; ILAYPERUMA, T.; JAYASINGHE, J.; BUKHSH, F.; DANEVA, M. The evolution of data storage architectures: examining the secure value of the Data Lakehouse. **Journal of Data, Information and Management**, p. 1-26, 2024.

KHINE, P. P.; WANG, Z. S. Data Lake: a new ideology in big data era. **ITM Web of Conferences**, 2018. p. 03025.

MACEDO, R. P. C. C. **Implementation of a Data Lake in a Microservices Architecture**. 2023. Dissertação (Mestrado em Engenharia Informática) – Universidade de Lisboa, Lisboa, 2023.

NAMBIAR, A.; MUNDRA, D. An overview of Data Warehouse and Data Lake in modern enterprise data management. **Big Data and Cognitive Computing**, v. 6, n. 4, p. 132, 2022.

PAGANI, R. N. **Revisão Sistemática de Literatura: Methodi Ordinatio na era da IA generativa**. Ponta Grossa: Texto e Contexto, 2023.

PAGANI, R. N.; PEDROSO, B.; DOS SANTOS, C. B.; PICININ, C. T.; KOVALESKI, J. L. Methodi Ordinatio 2.0: revisited under statistical estimation, and presenting FInder and RankIn. **Quality & Quantity**, v. 57, n. 5, p. 4563-4602, 2023.

PARK, S.; YANG, C. S.; KIM, J. Design of vessel Data Lakehouse with big data and AI analysis technology for vessel monitoring system. **Electronics**, v. 12, n. 8, p. 1943, 2023.

SALAMKAR, M. A.; ALLAM, K. Data Lakes vs. Data Warehouses: comparative analysis on when to use each, with case studies illustrating successful implementations. **Distributed Learning and Broad Applications in Scientific Research**, v. 5, 2019.

SAWADOGO, P.; DARMONT, J. On Data Lake architectures and metadata management. **Journal of Intelligent Information Systems**, v. 56, n. 1, p. 97-120, 2021.

SCHNEIDER, J.; GRÖGER, C.; LUTSCH, A.; SCHWARZ, H.; MITSCHANG, B. The Lakehouse: state of the art on concepts and technologies. **SN Computer Science**, v. 5, n. 5, p. 1-39, 2024.

SCHNEIDER, J., LUTSCH, A., GRÖGER, C., SCHWARZ, H., & MITSCHANG, B. First Experiences on the Application of Lakehouses in Industrial Practice. **In: GvDB**. p. 3-8. 2024.

SOLODNIKOVA, D.; NIEDRITE, L. Handling evolution in big data architectures. **Baltic Journal of Modern Computing**, v. 8, n. 1, p. 21-47, 2020.

VAISMAN, A.; ZIMÁNYI, E. **Data Warehouse systems**. Data-Centric Systems and Applications, v. 9, 2014.

VENKATESWARLU, M.; VASISTA, T. G. Extraction, transformation and loading process in the cloud computing scenario. **International Journal of Engineering Applied Sciences and Technology**, v. 8, p. 232-236, 2023.