Mezzoudj *et al.*
Journal of Electrical Systems and Inf Technol (2025) 12:89
https://doi.org/10.1186/s43067-025-00275-0

Journal of Electrical Systems and Information Technology

REVIEW Open Access

Data lakes versus data warehouses: choosing the right approach for big data analytics

Saliha Mezzoudj[1*], Meriem Khelifa[2] and Yassmina Saadna[3]

*Correspondence:
s.mezzoudj@univ-alger.dz

[1] Faculty of Sciences, Department of Computer Science, University of Algiers 1, Algiers, Algeria
[2] Department of Computer Science, Faculty of Sciences, University of Kasdi Merbah Ouargla, Ouargla, Algeria
[3] Lastic Laboratory, Department of Mathematic Computer Science, University of Batna, 2 Mostefa Ben Boulaïd, Batna, Algeria

Abstract

In the era of big data, organizations face critical decisions when selecting between data lakes and data warehouses to meet their analytics requirements. This article presents a comprehensive comparative analysis of these two predominant data management architectures, emphasizing their structural differences, functional capabilities, and suitability for diverse analytics workloads. Data lakes offer scalable, cost-effective storage for raw, unstructured, and semi-structured data, supporting advanced analytics and machine learning applications. In contrast, data warehouses provide optimized, schema-on-write frameworks for fast querying and reliable reporting on structured data. Through detailed examination of architectural designs, integration with big data tools including Hadoop, Spark, and Kafka, and evaluations based on performance, scalability, cost, and governance, this paper provides organizations with evidence-based guidance to align their data strategies with business objectives. Case studies from healthcare and retail sectors illustrate practical implications of each approach, while emerging trends such as lakehouse architectures, AI integration, blockchain security, edge computing, and quantum computing highlight future directions. The findings support for a hybrid data management solution that leverages the strengths of both data lakes and warehouses to enable robust, scalable, and innovative big data analytics.

Keywords: Big data analytics, Data lake, Data warehouse, Lakehouse architecture, Data management, Hadoop, Spark, Kafka, Cloud computing, Data governance, Business intelligence, Scalability, Cost-effectiveness, Artificial intelligence

Introduction

Big data analytics plays a critical role in the contemporary digital landscape, concentrating on the examination of large datasets to identify underlying patterns, relationships, and insights. This process involves extracting and analyzing diverse data to transform it into meaningful information [13]. Global statistics reveal that the total volume of data generated in 2021 reached approximately 79 zettabytes, with projections indicating that this figure will double by 2025 [70]. This enormous data volume emerged from a data surge that occurred over the past decade, during which data interactions increased by 5000% [77]. Big data is often defined by the 5 Vs: volume, variety, velocity, veracity, and value. It entails the collection, storage, and processing of substantial amounts of high-quality data across multiple domains, including business transactions, real-time

streaming, social media, video analytics, and text mining. As a result, a large quantity of semi-structured or unstructured data is stored in various information silos [20, 29]. The integration and analysis of data from multiple sources are essential for obtaining a comprehensive understanding of databases, which remains a prominent research focus. The advent of big data technologies has significantly transformed e-commerce and e-services, unlocking new opportunities in business analytics and research [73]. Data analytics systems are indispensable for modern business operations, particularly in areas such as product distribution, sales, and marketing. These systems enable the analysis of trends and patterns, facilitating companies' efforts to evaluate and optimize their data, ultimately uncovering new opportunities [2].

Organizations with access to high-quality data are able to make more informed business decisions by analyzing market trends and understanding customer preferences, which ultimately provides a competitive advantage. As a result, companies are making substantial investments in Artificial Intelligence (AI) and big data technologies to foster digital transformation and promote data-driven decision-making, thereby enhancing their business intelligence capabilities [73]. Reports indicate that the global markets for big data analytics and business intelligence software applications are expected to grow by USD 68 billion and USD 17.6 billion, respectively, between 2024 and 2025 [1]. Large data repositories are designed in various formats to meet the needs of organizations [38]. For effective analytics and query performance, these repositories must be able to integrate, manage, evaluate, and utilize substantial volumes of data. Beyond traditional relational databases, a variety of data repository types exist, with enterprise data warehouses and data lakes being among the most prevalent choices [17, 35]. A data warehouse (DW) is a structured repository that contains filtered and processed data, specifically designed for particular objectives, whereas a data lake (DL) is a large repository that stores data without a predefined purpose [38]. Data warehouses typically house substantial amounts of data collected from various sources, often with predefined schemas, and are purpose-built relational databases that operate on specialized hardware, either on premise or in the cloud [61]. Data warehouses are widely used for enterprise data storage and for supporting business intelligence and analytics applications [15].

The main contributions of this paper are summarized as follows:

- This paper provides a comprehensive comparison between data lakes and data warehouses, focusing on their architectural differences, functionalities, and suitability for big data analytics.
- We highlight the unique advantages and limitations of data lakes and data warehouses, addressing aspects such as data type handling, volume, processing speed, scalability, cost-effectiveness, and governance.
- Through detailed analysis, we evaluate factors influencing the choice between data lakes and data warehouses, supporting organizations in making informed decisions aligned with their business objectives.
- The paper reviews integration capabilities of both data management systems with popular big data analytics tools like Apache Hadoop, Spark, and Kafka, as well as business intelligence platforms.

- We present illustrative case studies from healthcare and retail industries to demonstrate practical applications and benefits of each approach.
- We discuss future trends and emerging technologies impacting data lakes and data warehouses, including machine learning integration, blockchain for security, edge computing, automation, advanced storage innovations, and quantum computing.
- Finally, the paper proposes a framework for choosing the appropriate data management system—data lake, data warehouse, or a hybrid solution—tailored to organizational needs and data analytics strategies.

The remainder of this paper is structured as follows: Section "Introduction" introduces big data analytics, highlighting its importance, challenges, and sets the motivation for comparing data lakes and data warehouses. Section "Related Works" surveys related works on data lakes, data warehouses, their architectures, and integration with BI tools. Section "Understanding Data Lakes and Data Warehouses" provides a detailed understanding of data lakes and data warehouses, including their architectures, storage solutions, components, and types, with sub-sections focusing on exploring data lakes, storage solutions, enterprise data warehouses, and warehouse types. Section "Advantages and Disadvantages of Data Warehouses and Data Lakes" discusses the advantages and disadvantages of both data lakes and data warehouses, comparing aspects such as data types, schema, performance, governance, security, scalability, and flexibility. Section "Integration with Big Data Analytics Tools" examines integration with big data analytics tools like Apache Hadoop, Spark, Kafka, and business intelligence platforms. Section "Cost Considerations" addresses cost considerations, covering infrastructure, software licensing, development, maintenance, and operational costs. Section "Choosing the Right Approach" guides the choice between data lakes and data warehouses through practical case studies from healthcare and retail sectors and explores the hybrid Lakehouse architecture. Section "Comparative Analysis of Performance and Cost: Quantitative Insights and Case Studies" delivers a comparative analysis of performance and cost with quantitative insights and examples. Section "Future Trends and Evolving Technologies" explores future trends and evolving technologies impacting these data management systems, including AI/ML integration, blockchain, edge computing, storage innovations, and quantum computing. Finally, Sect. "Conclusion" concludes the paper by summarizing key findings, discussing implications, and recommending strategies for leveraging data lakes, data warehouses, and their hybrid forms in modern big data analytics.

Related works

Data lakes (DLs) have emerged as significant repositories for big data, designed to store raw data and provide a variety of functionalities through metadata descriptions [35]. While data lakes serve as a form of enterprise data storage, they do not inherently possess the analytical capabilities typically associated with data warehouses. Instead, they function as storage for unprocessed data in its native format, offering a common interface for access. Information from a data lake can later be transferred to a data warehouse for processing, transformation, and preparation for use. As a relatively new concept, research on data lakes is still in its infancy, particularly in online articles and blogs. Although data warehouses and data lakes share some similarities in features and

applications, they differ fundamentally in their data management philosophies, design principles, and ideal use cases. This paper provides a thorough overview of the distinctions between these two data management approaches, integrating various concepts and presenting an in-depth analysis of design elements, tools, and recent advancements in the field.

Dixon in [16] provides an analogy comparing a data warehouse to a bottle of clean water, ready for consumption, while describing a data lake as a vast body of water containing data in a raw, unrefined state. Understanding the differences and advantages of each approach can help organizations make informed decisions regarding which option best suits their needs [16]. The authors in [83] explore the potential for data lakes to replace data warehouses in the near future, presenting data lakes as an architecture for storing raw data that enables flexible processing and analysis. They compare the strengths and weaknesses of both data lakes (DL) and data warehouses (DW), focusing on scalability, flexibility, and cost-effectiveness. The paper highlights how data lakes can manage the increasing variety and volume of data in modern business environments, emphasizing the differences between the flexible schema-on-read structure of data lakes and the more rigid schema-on-write structure of traditional data warehouses, which are less adaptable to evolving data needs. It also discusses the shift from Extract, Transform, Load (ETL) to Extract, Load, Transform (ELT) processes, underlining how data lakes facilitate real-time data processing and analytics.

The work in [26] compares the functionalities and use cases of data warehouses (DW) and data lakes (DL) across several dimensions, including data handling, schema design, data volume, processing agility, user access, and integration costs. This comparison sheds light on the unique benefits and limitations of each system, offering insights into their suitability for various analytical applications and environments. Nambiar and Mundra [55] provide a comparative analysis of data warehouses and data lakes, discussing their definitions, characteristics, and primary distinctions. They also explore the architecture and design of both storage systems and provide an overview of commonly used tools and services associated with them. Additionally, they examine the challenges of big data analytics and the specific issues involved in implementing both data warehouses and data lakes.

Varajo et al. [34] discuss the rise of Management Information Systems (MIS), which led to the development of data warehousing. They emphasize the importance of historical data in decision-making, highlighting how MIS systems focus on predefined reports and basic analytics. Data warehouses extend these capabilities by enabling advanced reporting, complex queries, and custom report generation. The evolution of data warehousing led to the creation of more advanced Business Intelligence (BI) tools, allowing users to create interactive dashboards, visualizations, and perform ad-hoc queries for deeper insights. MIS systems underscore the need for accurate and consistent data for effective decision-making, and data warehousing solutions have implemented data quality practices to ensure accuracy, integrity, and consistency.

Giebler et al. [11] examine various aspects of data lake architecture, including data storage, modeling, metadata management, and governance. Their work provides a partial overview of data lake functionality, while this survey aims to present a more comprehensive view of the current data lake landscape and provide a detailed discussion of the

research challenges and solutions associated with data lakes. Sawadogo et al. [68] compare different definitions, architectures, metadata types, models, and management components of data lakes, offering a high-level guide for their conceptual design. However, their exploration of the functions to be implemented in a data lake system is relatively brief, mainly summarizing the open-source technologies and tools used in existing data lakes, such as Apache Spark, Drill, and Pig.

Brackenbury et al. [18] propose a comprehensive data lake model based on Aurum's approach, utilizing multiple criteria to assess similarities between datasets. The key innovation is the incorporation of human input when algorithms alone are insufficient to provide reliable suggestions. This process evaluates the similarity of datasets, such as HTML tables, by considering approximate matches based on data values, schemas, and descriptive metadata, including data sources and user-added information. For assessing file similarity and clustering, the model employs MinHash and Locality-Sensitive Hashing (LSH) to calculate the Jaccard similarity of file paths.

In [60], the authors develop a data lake prototype using a hybrid infrastructure that combines Dell servers and Amazon AWS cloud services. This prototype demonstrates the ingestion, processing, and refinement of air traffic data from multiple FAA sources for analytical purposes. In [27], the authors analyze existing enterprise data lake solutions, noting that many organizations are enhancing Hadoop to address its limitations and improve data security. They highlight adaptations in platforms such as Amazon Web Services (AWS) data lakes and Azure data lakes, emphasizing their increasing adoption across sectors like banking, business intelligence, manufacturing, and healthcare, reflecting their growing utility in managing large-scale data requirements.

Armbrust et al. [6] introduce the concept of the data lakehouse, an architecture that combines the features of both data lakes and data warehouses. This approach integrates the data warehousing capabilities with the open file formats of data lakes, offering performance comparable to traditional data warehouses while addressing several challenges faced by their users. The lakehouse combines features from both systems, including ACID properties and SQL queries from data warehouses, as well as data versioning, lineage, indexing, and cost-effective storage from data lakes.

Recently, data lakes have become a prominent platform for managing big data across various industries, including healthcare and air traffic. This approach enables organizations to leverage advanced analytics, such as machine learning, to extract value from their data [22]. A recent study in [8] proposes a fish farming data lake architecture with a multizone structure consisting of three primary zones: the Raw Zone (RZ), the Trusted Zone (TZ), and the Access Zone (AZ). Data from various farms is collected in different formats and stored directly into the RZ without transformation, creating a robust historical repository. The TZ applies clean transformations for further analysis, while the AZ provides a dedicated data layer tailored for each team. The study details the functional architecture of this data lake, examining each component of the Hadoop ecosystem from a technical perspective and governance as discussed in Table 1; various studies have highlighted the importance of integrating data lakes, data warehouse with BI tools.

Unlike prior surveys that primarily focus on high-level comparisons or specific aspects of data lakes and warehouses [6, 34, 50], this paper offers a comprehensive architectural comparison, including practical case studies in healthcare and retail sectors. We

Table 1 Summary of related works on data lakes, data warehouse, and BI tools

Authors	Focus	Key contributions	Technologies/Methods
Dixon [16]	Data Warehouse vs. Data Lake	Analogy comparing DW (clean) and DL (raw data)	Conceptual comparison
Authors in [83]	DL vs. DW	Comparison of scalability, flexibility, and cost	Schema-on-read (DL) vs. schema-on-write (DW)
Authors in [26]	Data Warehouse vs. Data Lake	Functionalities and use case comparison	Data handling, schema design, integration costs
Nambiar and Mundra [55]	DW vs. DL Comparison	Definitions, characteristics, challenges	Big data analytics, tools overview
VarajÃ£o et al. [34]	MIS and Data Warehousing	Evolution from MIS to BI and DW	Reporting, custom queries, and advanced BI tools
Giebler et al. [11]	Data Lake Architecture	Data storage, metadata, and governance	Data Lake functionality, research challenges
Sawadogo et al. [68]	Data Lake Concepts	Architectures, metadata models	Open-source tools (Apache Spark, Drill, Pig)
Brackenbury et al. [18]	Data Lake Modeling	Human input for dataset similarity evaluation	MinHash, Locality-Sensitive Hashing (LSH)
Authors in [60]	Data Lake Prototype	Air traffic data analysis	Hybrid infrastructure, AWS cloud
Authors in [27]	Enterprise Data Lakes	Enhancing Hadoop security	AWS, Azure data lakes
Armbrust et al. [6]	Data Lakehouse Concept	Combining DW and DL features	ACID, SQL queries, data versioning
Authors in [22]	Data Lakes in Healthcare	Big data for analytics	Healthcare analytics, machine learning
Authors in [8]	Fish Farming Data Lake Architecture	Multizone structure (RZ, TZ, AZ)	Hadoop ecosystem, data transformation

also analyze the integration of emerging technologies such as lakehouses, blockchain security solutions, edge computing, and quantum computing. Furthermore, we propose a decision-making framework that assists organizations in choosing the most suitable data repository based on data characteristics and analytic requirements a practical contribution missing in earlier works. This extended analysis and framework provide novel insights benefiting both academic research and industry practice.

Understanding data lakes and data warehouses

Exploring data lakes

Data lakes are large-scale, cost-effective storage systems designed to store raw, unstructured data and make it accessible for analysis without the need for upfront processing. While traditional data lakes were often built using open-source frameworks like Hadoop and its Hadoop Distributed File System (HDFS), modern data lake architectures are increasingly implemented on cloud-based object storage platforms (such as Amazon S3, Microsoft Azure Blob Storage, and Google Cloud Storage). These cloud solutions enable organizations to decouple storage and compute, improving scalability, performance, and integration with advanced analytics tools, while reducing operational complexity and costs. As a result, the trend has shifted from on-premises, file-based systems to flexible, cloud-native data lakes that can efficiently support a wide variety of big data analytics and machine learning workloads [19] These systems enable organizations to store all types of data in their original form, facilitating flexible and efficient analysis later on [19].

Data lakes are increasingly being viewed as a modern solution to traditional data storage challenges, enabling faster access to diverse datasets for various organizational needs. As shown in Fig. 1, a typical data lake is divided into four distinct zones [32, 58]:

1. *Raw Data Zone*: This zone stores all data types in their original format, including both structured and unstructured data. Users can directly access and analyze raw data before further processing [31].
2. *Process Zone*: Here, raw data is transformed according to user requirements. This zone supports both batch and real-time processing [53].
3. *Access Zone*: This zone contains the processed data, which is ready for analysis, including AI models and business intelligence tasks. It allows self-service data exploration [58].
4. *Governance Zone*: This zone oversees data security, quality, and overall management, ensuring proper lifecycle management, access control, and metadata handling [58].

Storage solutions for data lakes

The choice of storage for a data lake is crucial to ensure security and accessibility. Storage solutions can be classified into three categories [63, 64]:

1. *File-based Storage*: This is similar to a file system, such as Hadoop Distributed File System (HDFS), which manages diverse data types (text, images, binary files) [63].
2. *Single Database*: These systems manage only one type of data, offering specific storage for each type [63].
3. *Polystore Systems*: These flexible systems automatically categorize data based on its format, storing it in relational, document-based, or graph databases [71].

The trend is shifting toward cloud-based data lakes, as cloud platforms like AWS and Snowflake offer scalability and flexibility, along with advanced features like serverless data analytics [43, 54]. Polystore systems offer flexibility but are more complex to

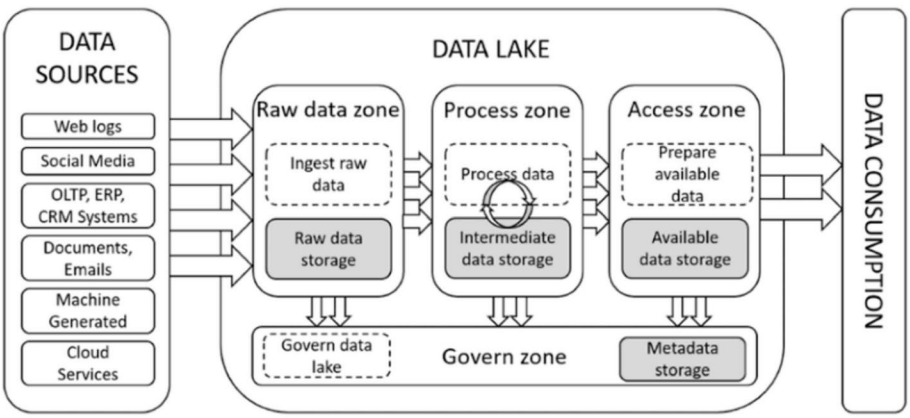

Fig. 1 The architecture of data lake

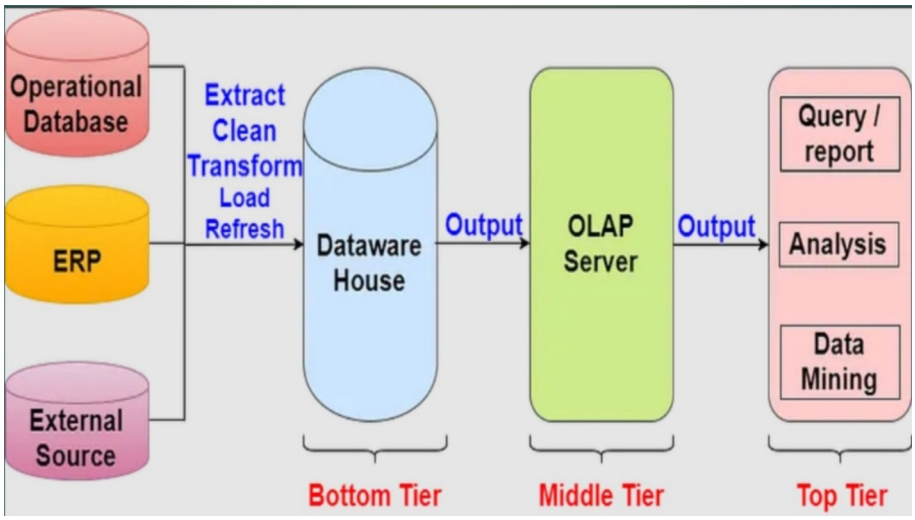

Fig. 2 The classic three-tier architecture of a data warehouse consists of (1) the bottom tier as the database server, (2) the middle tier as the OLAP server, and (3) the top tier as the client-facing analytical and reporting tools. [4]

manage. Ultimately, the choice depends on specific data types, analysis requirements, and budget constraints.

Exploring data warehouses

A data warehouse is a centralized repository designed for storing large volumes of data from multiple sources, primarily for analysis and reporting. It acts as an organized library for historical data, which aids in decision-making and strategic planning [62]. Data warehouses are characterized by four main features [47, 62, 78].

1. *Subject-Oriented*: Data is organized by specific subjects, removing unnecessary information, which helps in focused analysis and decision-making.
2. *Integrated*: Data from various sources is standardized and combined, ensuring uniformity regardless of its origin.
3. *Time-Variant*: Data warehouses store historical data over extended periods, allowing for trend analysis across time. Time elements in the data cannot be altered after being loaded.
4. *Non-Volatile*: Once data is loaded into the warehouse, it is not subject to frequent updates, ensuring stability and reliability for analysis.

Description of enterprise data warehouse (EDW)

An enterprise data warehouse (EDW) serves as a centralized, integrated repository that consolidates data from various operational and external sources across the organization. This environment is subject-oriented and structured to support comprehensive analysis, reporting, and business intelligence, enabling informed enterprise-wide decision-making. The EDW stores both current and historical data, ensuring data consistency and

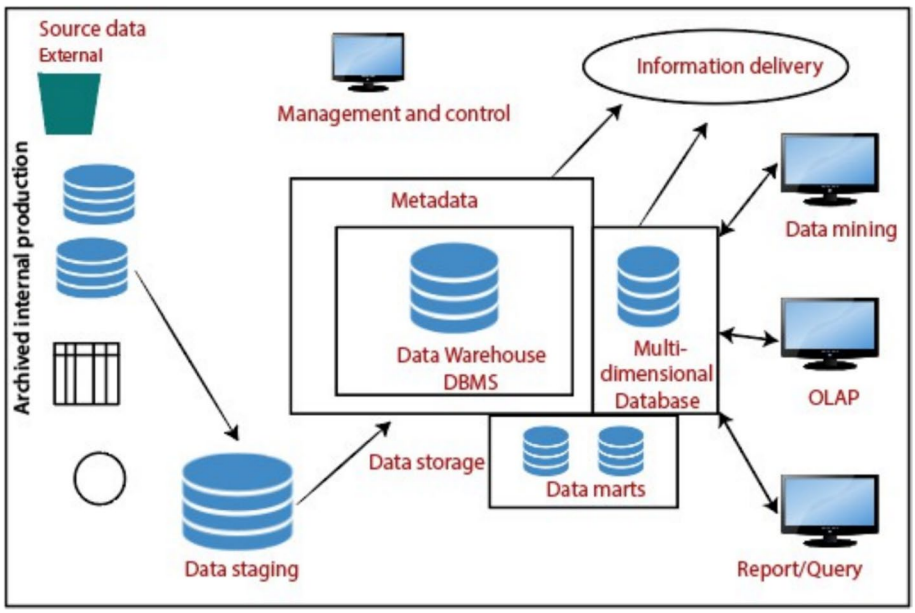

Fig. 3 The main components of data warehouse

integrity, and provides a solid base for analytics, compliance, and strategic planning [62, 85]. Figure 2 shows the architecture of the data warehouse.[1]

Data warehouses typically follow a three-tier architecture [62, 85], consisting of:

- *Bottom Tier*: The core database server where all data is stored. This layer handles data storage, integration, and management. It often consists of a relational database system optimized for large-scale analytical workloads.
- *Middle Tier*: The OLAP (Online Analytical Processing) server. This layer manages multidimensional data processing, provides analytical query services, and comprises engines for complex calculations, aggregations, and business logic.
- *Top Tier*: The presentation and analysis tools. This front-end consists of user interface applications, dashboards, reporting, data mining, and visualization tools that allow business users to interact with and derive insights from the warehouse.

Components of a data warehouse

As seen in Fig. 3, the main components of the data warehouse are illustrated.[2] Key components of a data warehouse include (et al. 2021), [33]:

1. *Central Database*: The core repository that stores integrated historical data.

[1] Data warehousing architecture Data Warehouse Architecture 101: Types, Layers, and Components, https://acuto.io/blog/data-warehouse-architecture-types/.

[2] Data warehousing components https://www.sprinkledata.com/blogs/components-of-data-warehouse-an-in-depth-guide.

2. *ETL Tools*: These tools extract, transform, and load data from operational systems into the data warehouse, ensuring consistency and quality [85].
3. *Metadata*: Metadata provides essential information about the structure, origin, and usage of the data, aiding in its management and interpretation. There are two types of metadata: technical and business [81].
4. *Access tools*: These tools allow users to interact with the data warehouse, including query tools, OLAP tools, application development tools, and data mining tools [78, 85].
5. *Data Marts*: These are smaller, specialized repositories within a data analysis for specific departments or functions [21].

Types of data warehouses

Data warehouses can be classified into three types [78]:

1. *Classic Data Warehouse*: The traditional data warehouse used for storing and analyzing large volumes of data [41].
2. *Virtual Data Warehouse*: A more flexible, computer-based version of the data warehouse [40].
3. *Cloud Data Warehouse*: A scalable, secure solution for modern organizations, offering cloud-based operations [72].

Advantages and disadvantages of data warehouses and data lakes

When evaluating the use of data lakes and data warehouses, several factors must be considered. Data lakes are better suited for handling large volumes and a variety of data, although performing queries can be challenging in certain situations. On the other hand, queries in a data warehouse are generally simpler to execute, yet challenges persist in handling big data efficiently. We have summarized the pros and cons of both data lakes and data warehouses in Table 2 for easy comparison [14, 24, 37, 55, 79, 80].

Integration with big data analytics tools

When choosing between a data lake and a data warehouse for big data analytics, it is crucial to consider their compatibility with widely used analysis tools [10, 37, 44, 51, 52].

Apache Hadoop

Data lakes are commonly associated with Apache Hadoop due to its capacity for efficient storage and distributed processing of large datasets. They typically integrate with data storage systems like the Hadoop Distributed File System (HDFS). However, Hadoop and data warehouses can also work together to manage data processing across diverse platforms [49, 56, 57, 65, 67, 82].

Apache Spark

Apache Spark is frequently used in data lakes to quickly process large volumes of data from various sources. Data lakes hosted on platforms such as Amazon S3, Azure Data

Table 2 Comprehensive comparison of data lakes and data warehouses

Data type	Supports structured, semi-structured, and unstructured data such as logs, images, sensor data, social media, multimedia	Primarily structured data with predefined schema optimized for business reporting (e.g., transactions, sales, demographics)
Schema	Schema-on-read; data is stored in raw form and schema is applied at query time	Schema-on-write; data is transformed and structured before storage
Analysis focus	Exploratory and advanced analytics, supporting machine learning and data science workflows	Established business intelligence and predefined reporting
Storage cost	Generally low-cost scalable storage, often cloud-based object or distributed file systems	Relatively higher cost due to specialized storage and compute environments
Performance	High ingestion speed, but query performance can suffer due to unstructured nature and ELT processing	Optimized for fast query performance with indexed, structured data
Data governance	Requires robust metadata management and governance to avoid data swamps; more challenging due to data variety	Mature governance models with enforced data quality, consistency, and compliance
Security	Emerging techniques including integration with blockchain for data integrity	Established security practices, role-based access control and compliance certified
Metadata management	Metadata-driven access, cataloging needed to discover and manage datasets effectively	Metadata is integrated and managed centrally as part of schema
Real-time data support	Stronger support for real-time data ingestion and streaming analytics via tools like Apache Kafka	Traditionally batch-oriented; newer systems support some real-time capabilities but less flexible
Usability	Accessible to data scientists and analysts comfortable with raw data and flexible queries; may require specialized skills	Accessible to business users via structured query tools and BI applications
Flexibility	Highly flexible for diverse data types and evolving analytical needs	Less flexible; changes require schema updates and ETL modifications
Integration with big data tools	Native support for Hadoop ecosystems, Spark, Kafka, and other distributed compute frameworks	Integrates with analytical tools but mostly through structured connectors
Scalability	Easily scalable horizontally using commodity hardware and cloud storage	Scalable but often constrained by database and hardware limits
Maintenance	Requires continuous monitoring and data curation to maintain data quality	Well-established maintenance procedures; simpler due to controlled schema and volumes
Emerging technology support	Supports integration with AI/ML models, blockchain for data security, edge computing	Evolving to support AI and real-time analytics, but traditionally focused on structured BI

Lake Storage, or Google Cloud Storage are compatible with Spark. Additionally, Spark can be integrated with data warehouses to rapidly analyze data stored in those systems simultaneously [45, 56, 66].

Apache Kafka

Apache Kafka is commonly employed for real-time data ingestion in data lakes, enabling large-scale data ingestion from a variety of sources. It can also be integrated with data warehouses to support real-time data flow. Many modern data warehouses are now compatible with Kafka, allowing continuous data streaming without interruptions [48].

Analysis and business intelligence (BI) tools

Data lakes can be integrated with powerful analytics tools such as Tableau, Power BI, and Looker for data analysis. However, due to the often unstructured and disorganized nature of data in data lakes, additional effort may be required to facilitate seamless integration. In contrast, data warehouses are typically designed for smooth integration with these analysis tools from the outset [42].

Cost considerations

Infrastructure and hardware

Data lakes require flexible infrastructure capable of handling large-scale data storage and processing. This results in significant expenditures related to data storage, cloud services, network bandwidth, and the compute nodes needed for distributed processing. In contrast, data warehouses typically require specialized hardware to manage analytical queries effectively. These costs include the acquisition of servers, storage systems, and periodic software licensing fees [3].

Licenses and software

To operate tools such as Apache Hadoop and analytics software within a data lake, licensing fees are required, including those for database services. Similarly, data warehouses also incur licensing costs for managing databases and analytical tools. The pricing structure for these licenses varies based on factors such as features, capacity, and the service provider [28].

Development and maintenance

The development and ongoing management of a data lake involve substantial costs for planning, setup, and maintenance to ensure smooth operations. These tasks include integrating various data sources, establishing workflows, and ensuring high-quality data management. Similarly, setting up and maintaining a data warehouse also incurs costs, with key responsibilities such as schema updates, query optimization, and performance management being essential for its efficient operation [74].

Operational costs

Both data lakes and data warehouses incur operational costs, which cover activities such as system monitoring, backup processes, performance management, and troubleshooting to ensure continuous functionality [75].

Choosing the right approach

The decision to choose between a data lake and a data warehouse depends primarily on the specific requirements of the organization or the project at hand. Data lakes are best suited for scenarios where the volume, variety, and velocity of data are high, and where flexibility in managing unstructured or semi-structured data is essential. They allow for large-scale data ingestion from diverse sources and are ideal for organizations focused on big data analytics and machine learning tasks [25, 50].

On the other hand, data warehouses are typically more appropriate for use cases requiring structured data and efficient querying for business intelligence (BI) and

analytical reporting. They are designed to handle complex queries on structured data and provide a more rigid but streamlined architecture for analytical purposes. Thus, the decision to use a data warehouse may be influenced by the need for high performance in querying structured data, along with the requirement for consistent data models [25].

Ultimately, the choice between a data lake and a data warehouse should be driven by factors such as data structure, processing requirements, analytical needs, and cost considerations. In some cases, a hybrid approach, combining the flexibility of a data lake with the structured querying power of a data warehouse, might be the most effective solution [25].

Data lake case study

A healthcare organization (Company X) implemented a cloud-based data lake to manage over 2 TB of diverse data monthly, including electronic health records, IoT devices, and patient surveys. This transition led to a 60% reduction in analysis latency and 40% savings in storage costs. These results align with insights from recent JESIT research on healthcare-focused big data analytics [7] and performance benchmarks found in [4, 69].

Results:

The data lake provided scalability, allowing the company to continuously add new data with minimal difficulty.

Advanced analytical tools enabled rapid and sophisticated data analysis.

The data was effectively managed and secured, ensuring compliance with healthcare data regulations.

This case highlights how a data lake is particularly beneficial in scenarios where large volumes of diverse, unstructured, or semi-structured data need to be ingested and processed for advanced analytics. The flexibility of data lakes allows organizations to store and process a variety of data sources without constraints.

Data warehouse case study

A retail company (Company Y) adopted a star schema data warehouse to consolidate sales and inventory data. Following deployment, they reported a 60% increase in report generation speed and 25% reduction in analytics-related operational costs. These outcomes corroborate findings from hybrid architecture studies in [23, 59, 84].

Results:

The data warehouse provided quick insights, enabling the company to respond faster to market changes.

The star schema simplified the process for non-technical users to analyze the data.

Even with numerous queries, the data warehouse maintained high performance and speed.

In this case, a data warehouse proved ideal for structured data, where rapid querying and reporting are critical for business decision-making. The predefined structure of a data warehouse, such as the star schema, supports fast querying and easy access to insights by non-technical users.

Choosing between data lakes and data warehouses

These case studies demonstrate how different companies, with varying needs and types of data, can benefit from either data lakes or data warehouses. The choice between the two depends primarily on the organization's data requirements:

Data lakes are well-suited for handling large volumes of diverse, unstructured data, enabling advanced analytics, machine learning, and integration of various data sources. As seen in the healthcare example, the flexibility of data lakes allows organizations to store and process data from different sources without significant limitations.

Data warehouses, on the other hand, are ideal when structured data is needed for fast querying and reporting, particularly in environments where speed, efficiency, and simplicity for end-users are paramount. As demonstrated in the retail example, the data warehouse-structured schema enables high performance for analytical queries and smooth integration with business intelligence tools [32].

Ultimately, the decision to use a data lake or a data warehouse should be guided by the specific needs of the organization, the nature of the data, and the type of analysis required.

Hybrid approach: the lakehouse architecture

Recent advancements in data management architectures have introduced the *Lakehouse*, a hybrid approach that aims to combine the flexibility and scalability of data lakes with the data management and performance features of data warehouses [6]. The Lakehouse architecture supports open storage formats and provides ACID transactional capabilities—Atomicity, which ensures that all operations within a transaction are completed successfully or none at all; Consistency, guaranteeing the database remains in a valid state after transactions; Isolation, preventing concurrent transactions from interfering with each other; and Durability, assuring that once a transaction is committed, its changes are permanent. Additionally, it offers schema enforcement and governance features that reconcile the traditionally divergent characteristics of data lakes and warehouse systems [6].

This hybrid model enables organizations to unify their analytical workflows by offering reliable data consistency alongside support for diverse data types and advanced analytics, including machine learning workloads. By leveraging both cloud object storage and modern query engines, Lakehouses provide a single, cost-effective platform that bridges the gap between raw data ingestion and structured reporting [9].

Consequently, the Lakehouse architecture represents a promising evolution in data platform design, addressing many limitations of traditional data lakes and warehouses, and enabling more streamlined and scalable big data analytics.

Comparative analysis of performance and cost: quantitative insights and case studies

1. Quantitative Performance Insights.

Data warehouses follow a schema-on-write model, providing high-performance querying of structured data for business intelligence tasks. In Sect. "Data Warehouse Case Study," the case study of a retail company illustrates how the warehouse architecture

supported a 70% improvement in query response time, enabling efficient reporting for inventory, sales, and market responsiveness.

Data lakes adopt a schema-on-read architecture, supporting real-time ingestion of diverse data types. In Sect. "Data Lake Case Study," the healthcare case study shows a system capable of ingesting 200 + GB/day from EHRs, IoT sensors, and surveys. The organization benefited from near real-time analytics, enabling predictive insights for patient care and regulatory compliance.

2. Cost-Effectiveness Metrics

As outlined in Table 2, data lakes are more cost-effective for storing large, diverse datasets due to:

Reduced ETL requirements.

Use of low-cost storage such as HDFS or cloud object storage.

Flexible scaling of compute/storage resources.

In contrast, data warehouses require specialized hardware and licensing, contributing to higher upfront costs, but offer low-latency structured querying that minimizes long-term query-related operational expenses.

The fish farming data lake architecture presented in [8] demonstrated cost-efficient handling of multi-format sensor data via a multi-zone Hadoop-based model, avoiding costly schema redesigns and ETL cycles typical of data warehouses.

3. Statistical Market Data

The Introduction (page 2) highlights that the big data analytics market is expected to grow by USD 68 billion between 2024 and 2025, with the BI market increasing by USD 17.6 billion, confirming strong industrial momentum toward scalable and cost-effective analytics platforms.

4. Table 3 *(Extracted from* Sect. "Integration with Big Data Analytics Tools" and Table 2*).*

Future trends and evolving technologies

The future of big data analytics is heavily influenced by the integration of advanced technologies aimed at improving analytical capabilities, enhancing data management, and ensuring greater security. Below are some key trends and evolving technologies that are shaping the future landscape of data lakes and data warehouses.

Table 3 Comparison of data warehouse versus data lake

Aspect	Data warehouse	Data lake
Query Performance	Fast for structured data; sub-second response	Slower for ad-hoc queries; optimized for batch/stream
Data volume	Limited scalability for heterogeneous data	Handles massive volumes; schema-on-read
Storage cost	High (specialized infra + ETL)	Low (cloud, HDFS, minimal preprocessing)
Operational Cost	Lower for structured workloads	Variable; scalable but needs governance
Best use cases	Compliance, BI, dashboards	ML pipelines, IoT, predictive analytics

Machine learning and AI integration

The integration of machine learning (ML) and artificial intelligence (AI) into big data systems is revolutionizing how data is processed and analyzed. These technologies can automate the process of storing, retrieving, and analyzing large datasets, while also providing predictive capabilities and intelligent recommendations. By incorporating ML and AI, organizations can gain deeper insights, enhance decision-making processes, and streamline data management tasks [46].

Blockchain for data security

Blockchain technology is being increasingly incorporated into data lakes to enhance data security. By leveraging the decentralized and immutable nature of blockchain, organizations can ensure the integrity and transparency of data stored within data lakes. This integration improves trustworthiness and provides an additional layer of security, especially in industries where data privacy and transparency are critical [12].

Edge computing for big data

Edge computing is becoming a significant trend in big data analytics due to its ability to process data closer to where it is generated, thus reducing latency and improving processing speeds. By deploying mini computational units at the data source, companies can analyze large volumes of data in real-time, facilitating faster decision-making and reducing the strain on centralized data centers [13].

Separating storage from computing

The trend of separating data storage from computational resources is gaining traction. By decoupling the two, organizations can achieve greater flexibility, scalability, and cost savings. This separation allows companies to optimize their computing resources while choosing the most suitable storage solutions, leading to improved performance and efficiency in data management [5].

Automation of data management tasks

As the volume and complexity of data continue to grow, the automation of data management tasks becomes increasingly critical. Automating processes such as metadata management, data governance, and data quality assurance will streamline operations, reduce human error, and accelerate data processing. This automation will be essential for improving operational efficiency and maintaining data integrity as organizations scale their big data infrastructures [36].

Evolution of storage technologies

Advances in storage technologies, such as more efficient distributed file systems and enhanced object storage solutions, are significantly improving the performance and management of large datasets. These innovations are making it easier for organizations to store, retrieve, and process big data more efficiently, leading to faster and more reliable analytics [30].

Quantum computing for big data analysis

Quantum computing represents a transformative advancement with the potential to revolutionize big data analytics by delivering exponentially greater computational power compared to classical systems. Leveraging principles such as superposition and entanglement, quantum computers can process and analyze vast, complex datasets at speeds unattainable by traditional computers, tackling intricate problems that are otherwise computationally prohibitive. By harnessing quantum algorithms and quantum machine learning approaches, organizations can uncover deeper insights, optimize decision-making processes, and accelerate the discovery of patterns hidden within big data.

This capability enables breakthroughs across diverse domains including healthcare, finance, logistics, and scientific research, opening new frontiers in big data analytics and driving significant innovation. While quantum computing technology is still emerging, its integration with big data platforms promises to enhance analytic capabilities, improve data management, and address challenges in processing speed and complexity. Future advancements are expected to combine quantum computing with other evolving technologies such as artificial intelligence, blockchain, and edge computing, culminating in a more dynamic and sophisticated environment for data lakes and data warehouses. Organizations that embrace these trends will be better positioned to harness the full potential of their data assets and solve complex problems previously out of reach [76].

In summary, the future of big data analytics is poised to be shaped by advancements in analytic capabilities, improved data management, enhanced security measures, and the adoption of emerging technologies. These trends, including the integration of AI, blockchain, edge computing, and quantum computing, will create a more dynamic and sophisticated environment for data lakes and data warehouses. Organizations that stay ahead of these trends will be better positioned to leverage big data for innovation and business growth [39].

Conclusion

This survey has provided a comprehensive comparison between data lakes and data warehouses, two pivotal technologies in the realm of big data analytics and business intelligence. Both approaches offer distinct advantages depending on the use case, data types, and analytical needs of organizations. Data lakes excel in handling vast volumes of raw, unstructured, and semi-structured data, enabling organizations to capture and store a wide variety of data types for complex analytics and machine learning applications. Their flexibility, scalability, and cost-effectiveness make them ideal for big data environments where data variety and rapid analysis are paramount. However, the complexity of managing raw data and ensuring data governance remain significant challenges for organizations adopting this approach. On the other hand, data warehouses provide a structured and optimized environment for managing large volumes of historical data, particularly suited for reporting and business intelligence tasks. Their well-defined schema and integration capabilities simplify querying and enhance the speed of decision-making, making them a valuable tool for organizations focused on standardized, analytical reporting. However, they often face limitations when dealing with diverse, real-time, or unstructured data. With emerging technologies such as machine learning, AI, blockchain, and edge computing, both data lakes and data warehouses are evolving

to meet the growing demands of big data analytics. The integration of these technologies promises to address some of the limitations inherent in traditional data management systems, including data security, processing speed, and analytical capabilities.

Future works: Future research should focus on advancing hybrid data management architectures, such as the lakehouse, to fully leverage the combined strengths of data lakes and data warehouses in scalable, cost-efficient, and secure big data analytics environments. The integration of emerging technologies, including artificial intelligence and machine learning for automated data processing and advanced analytics, blockchain for enhanced data integrity and security, and edge computing to reduce latency in real-time data processing, remains an important direction. Additionally, exploring the potential of quantum computing to revolutionize big data analysis by handling complex datasets with unprecedented speed will be critical. Further studies on automating metadata management, governance, and data quality assurance can enhance operational efficiency and trustworthiness of big data repositories. Overall, these research avenues will empower organizations to better manage diverse data types, optimize performance, and harness actionable insights for data-driven decision-making.

Ultimately, the choice between a data lake and a data warehouse depends on the organization's specific needs, data types, and analytical objectives. Companies with a focus on flexible, large-scale data storage and advanced analytics may benefit from data lakes, while those requiring fast, reliable reporting and structured data analysis may opt for data warehouses. The future of big data analytics lies in the ability to leverage both approaches in a complementary manner, enabling organizations to harness the full potential of their data assets. In conclusion, a hybrid approach that combines the strengths of both data lakes and data warehouses, alongside emerging technologies, is likely to provide the most effective solution for tackling the challenges of big data analytics in the coming years.

Author contributions
Saliha Mezzoudj helped in conceptualization, methodology, writing—original draft, supervision. Meriem Khelifa and Yassmina Saadna contributed to analysis, writing—review and editing.

Funding
The authors declare that no funding was received for this work.

Data availability
The datasets used and/or analyzed during the current study do not available (because we present a survey).

Declarations

Competing interests
The authors declare that they have no competing interests.

Received: 8 January 2025 Accepted: 30 September 2025
Published online: 27 October 2025

References

1. [b70, 2023] (2023) Bianalytics software market value worldwide 2019–2025.Available online: https://www.statista.com/statistics/590054/worldwide-business-analytics-software-vendor-market/. Accessed 27 Oct 2023
2. [b69, 2023] (2023) Big data and analytics services global market report. Available online: https://www.reportlinker.com/p06246484/Big-Data-and-Analytics-Services-Global-Market-Report.html. Accessed 27 Oct 2023
3. Bogatu A, Fernandes AA, Paton NW, Konstantinou N (2020) Dataset discovery in data lakes. In: 2020 IEEE 36th international conference on data engineering (ICDE), IEEE, pp 709–720

4. Ali R, Khanna M (2025) Scalability and performance tradeoffs in data lake environments: a case study of real-time analytics. J Cloud Comput 14(1):1–18
5. Arif M, Mujtaba G (2015) A survey: data warehouse architecture. Int J Hybrid Inf Technol 8(5):349–356
6. Armbrust M, Ghodsi A, Xin R, Zaharia M (2021) Lakehouse: a new generation of open platforms that unify data warehousing and advanced analytics. In: Proceedings of the CIDR, vol 8 p 28, Online
7. Badawy M, Ramadan N, Hefny HA (2024) Big data analytics in healthcare: data sources, tools, challenges, and opportunities. J Electr Syst Inf Technol 11(1):1–20
8. Benjelloun S, El Aissi ME, Lakhrissi Y, El Haj Ben Ali S (2023) Data lake architecture for smart fish farming data-driven strategy. Appl Syst Innov. 6(1):8
9. Beyer M, Li T, Ramesh D et al (2021) What is a data lakehouse? Databricks Blog. https://databricks.com/blog/2021/05/04/what-is-a-data-lakehouse.html
10. Khelifa Meriem BMK, Mezzoudj Saliha HMAF (2024) Novel solutions to the multidimensional knapsack problem using cplex: New results on orx benchmarks. J Ubiquitous Comput Commun Technol 6(3):294–310
11. Giebler C, Gröger C, Hoos E, Schwarz H, Mitschang B (2019) Leveraging the data lake – current state and challenges. In: Proceeding international conference on big data analytics and knowledge discovery, pp 179–188
12. Chainalysis (2024) The importance of blockchain security
13. Tsai C-W, Chin-Feng Lai H-CC, Vasilakos AV (2015) Big data analytics: a survey. J Big Data 2(1):1–32
14. Giebler C, Gröger C, Hoos E, Schwarz H, Mitschang B (2019) Lever- aging the data lake: Current state and challenges. In: Big data analytics and knowledge discovery: 21st international conference, DaWaK 2019, Linz, Austria, August 26–29, 2019, Proceedings 21. . Springer,pp 179–188
15. Devlin B, Murphy P (1988) An architecture for a business and information system. IBM Syst J 27:60–80
16. Dixon J (2024) Pentaho, hadoop, and data lakes. Available online: https://jamesdixon.wordpress.com/2010/10/14/pentaho-hadoop-and-data-lakes/. Accessed 4 May 2024
17. El Aissi M, Benjelloun S, Loukili Y, Lakhrissi Y, Boushaki A, Chougrad H, Elhaj Ben Ali S (2022) Data lake versus data warehouse architecture: a comparative study. vol 745. Springer, Singapore, pp 201–210
18. Brackenbury W (2018) Draining the data swamp: a similarity-based approach. In: Proc. workshop human-in- the-loop data analytics, pp 13:1–13:7
19. Fang H (2015) Managing data lakes in big data era: What's a data lake and why has it became popular in data management ecosystem. In: 2015 IEEE international conference on cyber technology in automation, control, and intelligent systems (CYBER), pp 820–824
20. Gandomi A, Haider M (2015) Beyond the hype: big data concepts, methods, and analytics. Int J Inf Manag 35:137–144
21. Ghezzi C (2001) Designing data marts for data warehouses. ACM Trans Softw Eng Methodol (TOSEM) 10(4):452–483
22. Giebler C, Gröger C, Hoos E, Eichler R, Schwarz H, Mitschang B (2021) The data lake architecture framework: a foundation for building a comprehensive data lake architecture. Ges. Fu¨r Inform., Bonn.
23. Gonzalez M, Ahmed R, Li S (2025) Evaluating data lakehouse architectures for unified data analytics. ACM J Data Eng 7(1):45–61
24. Gorelik A (2019) The enterprise big data lake: delivering the promise of big data and data science. O'Reilly Media
25. Gupta SPP (2023) Beyond banking: the trailblazing impact of data lakes on financial landscape. Int J Comput Appl 185(47):24–29
26. Herden O (2020) Architectural patterns for integrating data lakes into data-warehouse architectures. In" Proceedings of the big data analytics: 8th international conference, BDA, Sonepat, India. Springer, pp 12–27
27. Hukkeri TS, Kanoria VSJ (2020) A study of enterprise data lake solutions. Int Res J Eng Technol (IRJET) 7:1924–1929
28. IG Terrizzano, PM Schwarz, M Roth, JE Colino (2015) Data wrangling: The challenging yourney from the wild to the lake. In: CIDR. Asilomar
29. Jain A (2016) The 5 v's of big data. Available online: https://www.ibm.com/blogs/watsonhealth/the-5-vs- of-big-data/. Accessed 27 Oct 2023
30. Janssen M (2022a) Data warehouse architecture: a comparative study. https://essay.utwente.nl/92801/1/JanssenMA_EEMCS.pdf
31. Janssen NE (2022b) The evolution of data storage architectures: examining the value of the data lakehouse. PhD thesis, University of Twente
32. Jaspreet Singh GS, Bhati BS (2022) The implication of data lake in enterprises: a deeper analytics. In: 2022 8th international conference on advanced computing and communication systems (ICACCS), vol 1. pp 530–534
33. Javatpoint (2024) Data warehouse components. https://www.javatpoint.com/datawarehouse-components
34. Jo˜ao Varaj˜ao JCL, Gomes J (2022) Models and methods for information systems project success evaluation—a review and directions for research. Heliyon 8(12)
35. Khine P, Wang Z (2018) Data lake: a new ideology in big data era. ITM Web Conf 17:03025
36. Khine PP, Wang ZS (2018b) Data lake: a new ideology in big data era. In: ITM web of conferences, vol 17. EDP Sciences, pp 03025
37. Krishnan K (2013) Data warehousing in the age of big data. Newnes
38. Kumar S (2019a) What is a data repository and what is it used for? Available online: https://stealthbits.com/blog/what-is-a-data-repository-and-what-is-it-used-for/. Accessed 27 Oct 2023
39. Kumar S (2019b) What is a data repository and what is it used for?
40. Limited CS (2024a) Data warehouse as a service (DWaaS)
41. Limited CS (2024b) Enterprise data warehouse (EDW)
42. Llave MR (2018) Data lakes in business intelligence: reporting from the trenches. Procedia Comput Sci 138:516–524
43. Lock M (2016) Maximizing your data lake with a cloud or hybrid approach. Aberdeen Group
44. Khelifa M, Boughaci D, Aïmeur E (2018) Evolutionary harmony search algorithm for sport scheduling problem. In: Transactions on computational collective intelligence XXX, pp 93–117
45. Saliha M, Ali B, Rachid S (2019) Towards large-scale face-based race classification on spark framework. Multimed Tools Appl 78(18):26729–26746

46. Mandal JK, Misra S, Banerjee JS, Nayak S (2022) Applications of machine intelligence in engineering: Proceedings of 2nd global conference on artificial intelligence and applications (GCAIA, 2021), September 8–10, 2021, Jaipur, India. In: Proceedings of 2nd global conference on artificial intelligence and applications. CRC Press.
47. Wasi-ur-Rahman M, Huang J, Jose J, Ouyang X, Wang H, Islam NS, Subramoni H, Murthy C, Panda DK (2012) Understanding the communication characteristics in HBase: What are the fundamental bottlenecks?. In: 2012 IEEE international symposium on performance analysis of systems & software, IEEE, pp 122–123
48. Meena SD, Meena MSV (2016) Data lakes-a new data repository for big data analytics workloads. Int J Adv Res Comput Sci 7(5):65–66
49. Mezzoudj S (2020) Towards large scale image retrieval system using parallel frameworks. In: Multimedia information retrieval
50. Armbrust M, Ghodsi A, Xin R, Zaharia M (2021) Lakehouse: a new generation of open platforms that unify data warehousing and advanced analytics. In: Proceedings of CIDR, vol 8
51. Miloslavskaya N, Tolstoy A (2016) Big data, fast data and data lake concepts. Procedia Comput Sci 88:300–305
52. Mohamed MA, Eltobgy MM (2023) A framework for real-time data analytics in intelligent transportation systems. J Electr Syst Inf Technol 10(1):34–45
53. Mohamed El Mehdi El Aissi, Sarah Benjelloun, Y. L. Y. L. A. E. B. H. C, Ali SEB (2022) Data lake versus data warehouse architecture: a comparative study. In: WITS 2020: proceedings of the 6th international conference on wireless technologies, embedded, and intelligent systems, pp 201–210
54. Munshi AA, Mohamed YA-RI (2018) Data lake lambda architecture for smart grids big data analytics. IEEE Access 6:40463–40471
55. Nambiar A, Mundra D (2022) An overview of data warehouse and data lake in modern enterprise data management. Big Data Cogn Comput 6:132
56. Nazari E, Shahriari MH, Tabesh H (2019) Bigdata analysis in healthcare: apache hadoop, apache spark and apache flink. Front Health Inform 8(1):14
57. Pujari V, Sharma YK, Rane R (2016) A review paper on big data and hadoop
58. Quix C, Geisler S, Hai R (2020) Data lake.
59. Rahman T, Patel A (2025) A comparative study of data warehousing and data lake approaches in financial analytics. J Inf Syst 72:200–214
60. Raju R, Mital RFD (2018) Data lake architecture for air traffic management. In: Proceedings of the IEEE/AIAA 37th digital avionics systems conference (DASC), London, UK, pp 1–6
61. Rehman K, Ahmad U, Mahmood S (2018) A comparative analysis of traditional and cloud data warehouse. VAWKUM Trans Comput Sci 6:34–40
62. Rifaie M, Kianmehr K, Alhajj R, Ridley MJ (2008) Data warehouse architecture and design. In: 2008 IEEE international conference on information reuse and integration, pp 58–63
63. Rihan Hai CQ, Jarke M (2021) Data lake concept and systems: a survey. arXiv preprint arXiv:2106.09592
64. Rihan Hai CQ, Zhou C (2018) Query rewriting for heterogeneous data lakes. In: Advances in databases and information systems: 22nd European conference, ADBIS 2018, Budapest, Hungary, September 2–5, 2018, Proceedings 22, Springer, pp 35–49
65. Mezzoudj S, Melkemi KE (2021a) A hybrid approach for shape retrieval using genetic algorithms and approximate distance. In: Research anthology on multi-industry uses of genetic programming and algorithms
66. Mezzoudj S, Behloul A, Seghir R, Saadna Y (2021) A parallel content-based image retrieval system using spark and tachyon frameworks. J King Saud Univ-Comput Inform Sci 33(2):141
67. Mezzoudj S, Khelifa M, Saadna Y (2024) A comparative study of parallel processing, distributed storage techniques, and technologies: a survey on big data analytics. Int J 10(5):86–99
68. Sawadogo P, Darmont J (2020) On data lake architectures and metadata management. J Intell Inf Syst 56:1–24
69. Smith J, Chen L (2025) Next-generation data lakes: enhancing performance and governance in the cloud era. IEEE Trans Big Data 11(2):98–112
70. Statista (2023) Big data—statistics facts. Available online: https://www.statista.com/topics/1464/big- data/. Accessed 27 Oct 2023
71. Stonebraker M, Abadi DJ, Batkin A, Chen X, Cherniack M, Ferreira E, Lau E, Lin A, Madden S, O'Neil P et al (2017) The case for polystore systems. Found Trends® Databases 7(3–4):204–293
72. StreamSets (2024) Data warehouse architecture explained. https://streamsets.com/blog/data-warehouse-architecture-explained/.
73. Sun Z, Zou H, Strang K (2015) Big data analytics as a service for business intelligence. In: Publishing SI (ed) Open and big data management and innovation, vol 9373. Cham, Switzerland, pp 200–211
74. Sunscrapers (2024) Data warehousing vs data lakes. Understanding the differences. https://sunscrapers.com/blog/data-warehousing-vs-data-lakes-understanding-the-differences/
75. Hukkeri TS, Kanoria V, Shetty J (2020) A study of enterprise data lake solutions. Int Res J Eng Technol (IRJET) 7:1924
76. Wang H, Kais S (2021) Quantum computing for data science: a quantum ML approach. Wiley, Hoboken
77. Wise, J. (2023). Big data statistics 2022: Facts, market size industry growth. Available online: https://earthweb.com/big-data-statistics/. Accessed 27 Oct 2023
78. Wrembel R (2006) Data warehouses and OLAP: concepts, architectures and solutions. Igi Global.
79. Wrembel R (2021) Still open problems in data warehouse and data lake research. In: 2021 Eighth international conference on social network analysis, management and security (SNAMS), pp 01–03
80. Wrembel R (2023) Data integration revitalized: From data warehouse through data lake to data mesh. In: International conference on database and expert systems applications, pp 3–18
81. Wrembel R, Bebel B (2007) Metadata management in a multiversion data warehouse. In :Journal on data semantics VIII, Springer, pp 118–157
82. Saadna Y, Behloul A, Mezzoudj S (2019) Speed limit sign detection and recognition system using svm and mnist datasets. Neural Comput Appl 31(9):5005–5015

83. Zagan E, Danubianu M (2019) From data warehouse to a new trend in data architec- tures–data lake. IJCSNS Int J Comput Sci Netw Secur 19:30–35
84. Zhang F, Kumar S, Wu E (2025) Optimizing cost and query latency in hybrid data architectures. IEEE Access 13:18230–18242
85. Zubcoff J, Trujillo J, (2006) Conceptual modeling for classification mining in data ware- houses. In: Data warehousing and knowledge discovery: 8th international conference, DaWaK 2006, Krakow, Poland, September 4–8, 2006. Proceedings 8, pp 566–575

Publisher's Note