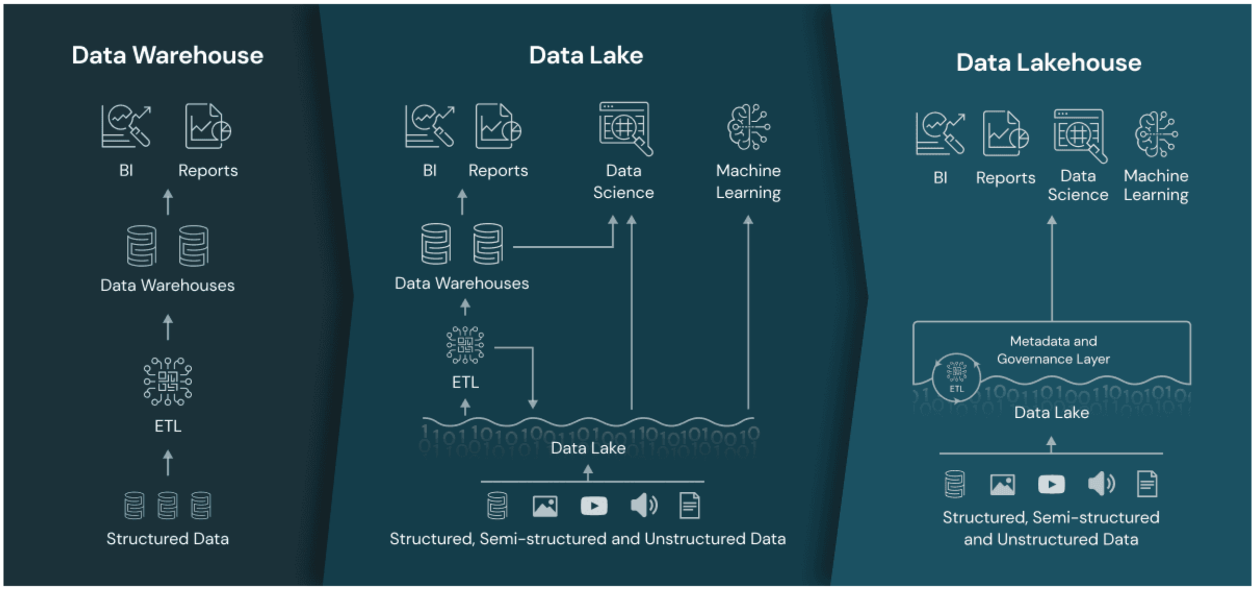
Explain in couple of paragraphs the difference between Data Lakehouse and Data Warehouse
o Focus on the benefits of the Data Lakehouse
o Explain the main enablers of the Data Lakehouse
o Try to suggest a reference architecture for a Data Lakehouse in cloud (e. g. AWS)



Data Warehouses were purpose-built for BI and reporting, however...

- No support for video, audio, text
- No support for data science,
- ML Limited support for streaming Closed & proprietary formats

Therefore, most data is stored in data lakes & blob stores

The benefits of a data warehouse

- Data warehouses, when implemented, offer tremendous advantages to an organization. Some of the benefits include:
- Improving data standardization, quality, and consistency: Organizations generate data from various sources, including sales, users, and transactional data. Data warehousing consolidates corporate data into a consistent, standardized format that can serve as a single source of data truth, giving the organization the confidence to rely on the data for business needs.
- Delivering enhanced business intelligence: Data warehousing bridges the gap between voluminous raw data, often collected automatically as a matter of practice, and the curated data that offers insights. They serve as the data storage backbone for organizations, allowing them to answer complex questions about their data and use the answers to make informed business decisions.

- Increasing the power and speed of data analytics and business intelligence workloads: Data warehouses speed up the time required to prepare and analyze data. Since the data warehouse’s data is consistent and accurate, they can effortlessly connect to data analytics and business intelligence tools. Data warehouses also cut down the time required to gather data and give teams the power to leverage data for reports, dashboards, and other analytics needs.
- Improving the overall decision-making process: Data warehousing improves decision-making by providing a single repository of current and historical data. Decision-makers can evaluate risks, understand customers’ needs, and improve products and services by transforming data in data warehouses for accurate insights.

The disadvantages of a data warehouse

Data warehouses empower businesses with highly performant and scalable analytics. However, they present specific challenges, some of which include:

- Lack of data flexibility: Although data warehouses perform well with structured data, they can struggle with semi-structured and unstructured data formats such as log analytics, streaming, and social media data. This makes it hard to recommend data warehouses for machine learning and artificial intelligence use cases.
- High implementation and maintenance costs: Data warehouses can be expensive to implement and maintain. This article by Cooladata estimates the annual cost of an in-house data warehouse with one terabyte of storage and 100,000 queries per month to be $468,000. Additionally, the data warehouse is typically not static; it becomes outdated and requires regular maintenance, which can be costly.

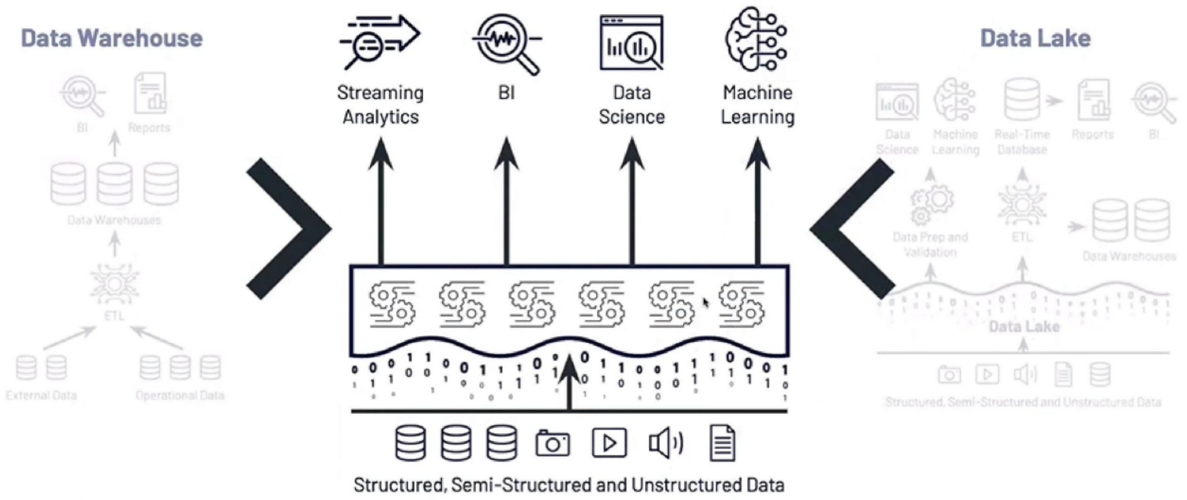
Data Lakes could handle all your data for data science and ML, however…

- Poor BI support
- Complex to Setup
- Poor performance
- Unreliable data swamps

Coexistence is not a desirable strategy, that means maintaining and usage of both at the same time.

Here comes the **Lakehouse**

A data lakehouse is a new, open data management architecture that combines the flexibility, cost-efficiency, and scale of data lakes with the data management and ACID transactions of data warehouses, enabling business intelligence (BI) and machine learning (ML) on all data.



Data lakehouses are enabled by a new, open system design: implementing similar data structures and data management features to those in a data warehouse, directly on the kind of low-cost storage used for data lakes. Merging them together into a single system means that data teams can move faster as they are able to use data without needing to access multiple systems. Data lakehouses also ensure that teams have the most complete and up-to-date data available for data science, machine learning, and business analytics projects.

A lakehouse has the following key features:

- Transaction support: In an enterprise lakehouse many data pipelines will often be reading and writing data concurrently. Support for ACID transactions ensures consistency as multiple parties concurrently read or write data, typically using SQL.
- Schema enforcement and governance: The Lakehouse should have a way to support schema enforcement and evolution, supporting DW schema architectures such as star/snowflake-schemas. The system should be able to reason about data integrity, and it should have robust governance and auditing mechanisms.
- BI support: Lakehouses enable using BI tools directly on the source data. This reduces staleness and improves recency, reduces latency, and lowers the cost of having to operationalize two copies of the data in both a data lake and a warehouse.
- Storage is decoupled from compute: In practice this means storage and compute use separate clusters, thus these systems are able to scale to many more concurrent users and larger data sizes. Some modern data warehouses also have this property.
- Openness: The storage formats they use are open and standardized, such as Parquet, and they provide an API so a variety of tools and engines, including machine learning and Python/R libraries, can efficiently access the data directly.

- Support for diverse data types ranging from unstructured to structured data: The lakehouse can be used to store, refine, analyze, and access data types needed for many new data applications, including images, video, audio, semi-structured data, and text.
- Support for diverse workloads: including data science, machine learning, and SQL and analytics. Multiple tools might be needed to support all these workloads but they all rely on the same data repository.
- End-to-end streaming: Real-time reports are the norm in many enterprises. Support for streaming eliminates the need for separate systems dedicated to serving real-time data applications.
- The benefits of a data lakehouse
- Data lakehouse architecture combines a data warehouse's data structure and management features with a data lake's low-cost storage and flexibility. The benefits of this implementation are enormous and include:
- Reduced data redundancy: Data lakehouses reduce data duplication by providing a single all-purpose data storage platform to cater to all business data demands. Because of the advantages of the data warehouse and the data lake, most companies opt for a hybrid solution. However, this approach could lead to data duplication, which can be costly.
- Cost-effectiveness: Data lakehouses implement the cost-effective storage features of data lakes by utilizing low-cost object storage options. Additionally, data lakehouses eliminate the costs and time of maintaining multiple data storage systems by providing a single solution.
- Support for a wider variety of workloads: Data lakehouses provide direct access to some of the most widely used business intelligence tools (Tableau, PowerBI) to enable advanced analytics. Additionally, data lakehouses use open-data formats (such as Parquet) with APIs and machine learning libraries, including Python/R, making it straightforward for data scientists and machine learning engineers to utilize the data.
- Ease of data versioning, governance, and security: Data lakehouse architecture enforces schema and data integrity making it easier to implement robust data security and governance mechanisms.
- The disadvantages of a data lakehouse
- The main disadvantage of a data lakehouse is it's still a relatively new and immature technology. As such, it's unclear whether it will live up to its promises. It may be years before data lakehouses can compete with mature big-data storage solutions. But with the current speed of modern innovation, it's difficult to predict whether a new data storage solution could eventually usurp it.
- This table summarizes the differences between the data warehouse vs. data lake vs. data lakehouse.

	Data Warehouse	Data Lake	Data Lakehouse
Storage Data Type	Works well with structured data	Works well with semi-structured and unstructured data	Can handle structured, semi-structured, and unstructured data
Purpose	Optimal for data analytics and business intelligence (BI) use-cases	Suitable for machine learning (ML) and artificial intelligence (AI) workloads	Suitable for both data analytics and machine learning workloads
Cost	Storage is costly and time-consuming	Storage is cost-effective, fast, and flexible	Storage is cost-effective, fast, and flexible
ACID Compliance	Records data in an ACID-compliant manner to ensure the highest levels of integrity	Non-ACID compliance: updates and deletes are complex operations	ACID-compliant to ensure consistency as multiple parties concurrently read or write data

Data Lakehouse vs. Data Warehouse vs. Data Lake: Which One Is Right for Your Needs?

- Data lakehouses can be complex to build from scratch. And you'll most likely use a platform built to support open data lakehouse architecture. So, ensure you research each platform's different capabilities and implementations before making a purchase.
- A data warehouse is a good choice for companies seeking a mature, structured data solution that focuses on business intelligence and data analytics use cases. However, data lakes are suitable for organizations seeking a flexible, low-cost,

big-data solution to drive machine learning and data science workloads on unstructured data.

- Suppose the data warehouse and data lake approaches aren’t meeting your company’s data demands, or you’re looking for ways to implement both advanced analytics and machine learning workloads on your data. In that case, a data lakehouse is a reasonable choice.