Data Lakes compared to Data Warehouses – Two Different Approaches

Depending on the requirements, a typical organization will require both a data warehouse and a data lake as they serve different needs, and use cases.

A data warehouse is a database optimized to analyze relational data coming from transactional systems and line of business applications. The data structure, and schema are defined in advance to optimize for fast SQL queries, where the results are typically used for operational reporting and analysis. Data is cleaned, enriched, and transformed so it can act as the “single source of truth” that users can trust.

A data lake is different, because it stores relational data from line of business applications, and non-relational data from mobile apps, IoT devices, and social media. The structure of the data or schema is not defined when data is captured. This means you can store all of your data without careful design or the need to know what questions you might need answers for in the future. Different types of analytics on your data like SQL queries, big data analytics, full text search, real-time analytics, and machine learning can be used to uncover insights.

As organizations with data warehouses see the benefits of data lakes, they are evolving their warehouse to include data lakes, and enable diverse query capabilities, data science use-cases, and advanced capabilities for discovering new information models. Gartner names this evolution the “Data Management Solution for Analytics” or “DMSA.”

Data warehouse vs. Data lake

Data warehouse vs data lake

Characteristics	Data Warehouse	Data Lake
Data	Relational data from transactional systems, operational databases, and line of business applications	All data, including structured, semi-structured, and unstructured
Schema	Often designed prior to the data warehouse implementation but also can be written at the time of analysis (schema-on-write or schema-on-read)	Written at the time of analysis (schema-on-read)
Price/Performance	Fastest query results using local storage	Query results getting faster using low-cost storage and decoupling of compute and storage
Data quality	Highly curated data that serves as the central version of the truth	Any data that may or may not be curated (i.e. raw data)
Users	Business analysts, data scientists, and data developers	Business analysts (using curated data), data scientists, data developers, data engineers, and data architects
Analytics	Batch reporting, BI, and visualizations	Machine learning, exploratory analytics, data discovery, streaming, operational analytics, big data, and profiling

Data Warehouse vs. Database

Data warehouse vs database

Characteristics	Data Warehouse	Transactional Database
Suitable workloads	Analytics, reporting, big data	Transaction processing
Data source	Data collected and normalized from many sources	Data captured as-is from a single source, such as a transactional system
Data capture	Bulk write operations typically on a predetermined batch schedule	Optimized for continuous write operations as new data is available to maximize transaction throughput
Data normalization	Denormalized schemas, such as the Star schema or Snowflake schema	Highly normalized, static schemas
Data storage	Optimized for simplicity of access and high-speed query performance using columnar storage	Optimized for high throughout write operations to a single row-oriented physical block
Data access	Optimized to minimize I/O and maximize data throughput	High volumes of small read operations

More information on Data Lake and Data Warehouses can be found at:

Data Lake vs. Data Warehouse - Snowflake - https://www.snowflake.com/trending/data-lake-vs-data-warehouse