Tejashree G
tejashreeg.ai2021@dce.edu.in

Database vs Data Warehouse vs Data Lake vs Delta Lake Assignment

1. Database

A database is a collection of data that is stored and accessed electronically. It is mainly used for day-to-day operations and supports Online Transaction Processing (OLTP). Databases can store structured and semi-structured data.

Key features:

- Can be relational (tables) or non-relational (JSON, key-value, graphs).
- Follows ACID rules to keep data correct and reliable.
- Works in real-time for storing and retrieving data quickly.

Examples: MySQL, PostgreSQL, MongoDB, Oracle.
When to use:

- For applications that need to store and update live data like online shopping carts, banking transactions, etc

2. Data Warehouse

A data warehouse stores structured and semi-structured data from different sources. It is designed for analytics and supports Online Analytical Processing (OLAP). It has a fixed schema and data is loaded after it is cleaned and transformed (ETL process).

Key features:

- Stores large amounts of current and historical data.
- Optimized for reporting and BI tools.
- Data may not be real-time (depends on ETL schedule).

Examples: Snowflake, Google BigQuery, Amazon Redshift.
When to use:

- For creating reports, dashboards, and business insights from multiple sources of data.

3. Data Lake

A data lake is a large storage system that keeps raw data in its original format. It can store structured, semi-structured, and unstructured data like videos, PDFs, images, and text files.

Key features:

- Schema-on-read (structure is applied when data is read).
- Stores data in a variety of formats such as JSON, BSON, CSV, TSV, Avro, ORC, and Parquet.
- Cheaper storage than a data warehouse.
- Can handle huge amounts of data for big data analytics and machine learning.

Examples: AWS S3, Azure Data Lake Storage, Google Cloud Storage.
When to use:

- When we want to store all kinds of data for future analysis without needing to transform it right away.

4. Delta Lake

Delta Lake is built on top of a data lake. It adds ACID transactions, versioning, and schema enforcement to make data lakes more reliable. It combines the flexibility of a data lake with the reliability of a data warehouse.

Key features:

- Supports both batch and streaming data.
- Handles updates and deletes easily.
- Works well for analytics and machine learning.

Example: Databricks Delta Lake.

Comparison Table

Feature	Database	Data Warehouse	Data Lake	Delta Lake
Purpose	Real-time transactions	Business analytics	Raw data storage	Reliable big data analytics
Data Type	Structured/Semi-structured	Structured	Structured/Semi/Unstructured	Structured/Semi/Unstructured
Schema	Fixed or Flexible	Fixed (pre-defined)	No schema (schema-on-read)	Schema support
Example	MySQL, MongoDB	Snowflake, Redshift	AWS S3, Hadoop	Databricks Delta Lake