11.8.25 – Assignments

Database vs Data Warehouse vs Data Lake vs Delta Lake

1. Database

- Purpose: Store and manage transactional data (OLTP - Online Transaction Processing).
- Structure: Highly structured with predefined schema (tables, rows, columns).
- Data Type: Primarily current, clean, and processed data.
- Use Cases: Applications, websites, ERP, CRM systems for real-time operations.
- Examples: MySQL, PostgreSQL, Oracle DB, SQL Server.

2. Data Warehouse

- Purpose: Store historical and aggregated data for reporting and analytics (OLAP - Online Analytical Processing).
- Structure: Structured data with well-defined schemas optimized for fast query performance.
- Data Type: Cleaned, processed, and transformed data from multiple sources.
- Use Cases: Business intelligence, dashboards, complex queries, trend analysis.
- Examples: Amazon Redshift, Google BigQuery, Snowflake, Azure Synapse.

3. Data Lake

- Purpose: Store vast amounts of raw, unprocessed data from various sources in native formats.
- Structure: Schema-on-read (flexible), stores structured, semi-structured, and unstructured data.
- Data Type: Raw data like logs, images, videos, JSON, CSV, etc.
- Use Cases: Data science, machine learning, big data analytics requiring diverse data types.
- Examples: Amazon S3, Azure Data Lake Storage, Google Cloud Storage.

4. Delta Lake

- Purpose: Enhance data lakes with ACID transactions, scalable metadata handling, and data reliability.
- Structure: Built on top of data lakes; provides structured, transactional storage with schema enforcement and versioning.
- Data Type: Raw and processed data with data reliability guarantees.
- Use Cases: Reliable big data pipelines, streaming and batch processing, upserts, time travel queries.
- Examples: Delta Lake on Databricks, open-source Delta Lake project.

Comparison Table

Feature	Database	Data Warehouse	Data Lake	Delta Lake
Data Type	Structured	Structured	Structured & unstructured	Structured + transactional
Schema	Schema-on-write	Schema-on-write	Schema-on-read	Schema-on-write + enforcement
Use Case	Transactions	BI & reporting	Data science, ML	Reliable lakehouse workloads
Data Freshness	Real-time	Periodic batch	Raw, variable freshness	Batch & streaming with ACID
Examples	MySQL, Oracle	Redshift, Snowflake	S3, ADLS, GCS	Delta Lake on Databricks