Data Modeling Concepts:

  - Database vs Data Warehouse vs Data Lake vs Data Lakehouse

  - Delta Lake

  - Types of Data Models: [Conceptual Data Model, Logical Data Model, Physical Data Model]

  - Star Schema vs Snowflake Schema

  - Data Vault Schema

  - Galaxy Schema (Fact Constellation Schema)

  - Fact Table vs Dimension Table

  - Measures and Attributes

  - Types of Dimensions

    -> Slowly Changing Dimensions (SCDs)

    -> Rapidly Changing Dimensions

    -> Conformed Dimensions

    -> Time Dimensions

    -> Role-Playing Dimensions

    -> Degenerate Dimensions

    -> Junk Dimensions

    -> Hierarchical Dimensions

    -> Inferred Dimensions

    -> Shrunken Dimensions

    -> Static Dimensions

  - Types of Facts

    -> Additive Facts

    -> Semi-Additive Facts

    -> Non-Additive Facts

    -> Snapshot Facts

    -> Cumulative Facts

    -> Factless Facts

    -> Aggregated Facts

    -> Derived Facts

    -> Measure Facts

    -> Event-Based Facts

  - Grain (Granularity) in Fact Tables

  - Primary Key vs Surrogate Key

  - OLTP vs OLAP

  - Normalization vs Denormalization

  - Schema on Read vs Schema on Write

Data Loading Methods:

  - Full Load

  - Delta Load

  - Incremental Loading

  - Change Data Capture (CDC)

  - Truncate and Load

  - Different types of Slowly Changing Dimensions (SCDs)

**✅ Data Modeling Concepts**

**Database vs Data Warehouse vs Data Lake vs Data Lakehouse**

* **Database**: Stores structured data for OLTP operations.
* **Data Warehouse**: Stores structured data optimized for analytics (OLAP).
* **Data Lake**: Stores raw structured, semi-structured, and unstructured data.
* **Data Lakehouse**: Combines the reliability of data warehouses with the flexibility of data lakes.

**Delta Lake**  
An open-source storage layer on top of data lakes that brings ACID transactions, schema enforcement, and time travel.

**Types of Data Models**

* **Conceptual**: High-level view of entities and relationships.
* **Logical**: Adds attributes, keys, and detailed structure.
* **Physical**: Implementation-specific with tables, indexes, partitions.

**Star Schema vs Snowflake Schema**

* **Star**: Denormalized dimensions for performance.
* **Snowflake**: Normalized dimensions for storage efficiency.

**Data Vault Schema**  
Hybrid modeling technique using **Hubs (keys)**, **Links (relationships)**, and **Satellites (descriptive data)** for agility and auditability.

**Galaxy Schema (Fact Constellation)**  
Multiple fact tables sharing dimension tables; supports complex reporting and subject areas.

**Fact Table vs Dimension Table**

* **Fact Table**: Contains measurable, quantitative data.
* **Dimension Table**: Provides descriptive context for facts.

**Measures and Attributes**

* **Measures**: Numeric values for analysis (e.g., sales).
* **Attributes**: Descriptive properties (e.g., product name).

**📌 Types of Dimensions**

* **SCDs**: Track historical changes in dimension data (Type 1, 2, 3).
* **Rapidly Changing**: Frequently updated; may be stored in separate structures.
* **Conformed**: Shared across fact tables for consistency.
* **Time**: Prebuilt calendar attributes for time-based analysis.
* **Role-Playing**: Same dimension used in different roles (e.g., Order Date, Ship Date).
* **Degenerate**: Fact table key with no related dimension (e.g., invoice number).
* **Junk**: Combines low-cardinality flags or indicators into one dimension.
* **Hierarchical**: Contain levels (e.g., Country → State → City).
* **Inferred**: Placeholder dimensions when actual data is delayed.
* **Shrunken**: Subset of a dimension for specific use cases.
* **Static**: Never change after initial load.

**📌 Types of Facts**

* **Additive**: Summable across all dimensions.
* **Semi-Additive**: Summable on some dimensions (e.g., balance over time).
* **Non-Additive**: Not summable (e.g., percentages).
* **Snapshot**: Point-in-time facts (e.g., daily inventory).
* **Cumulative**: Running total facts.
* **Factless**: No measures, only relationships (e.g., attendance).
* **Aggregated**: Pre-summarized for faster reporting.
* **Derived**: Calculated from other facts.
* **Measure**: Raw numeric values in facts.
* **Event-Based**: Records events (e.g., transactions, logins).

**💡 Other Key Concepts**

**Grain in Fact Tables**  
Defines the lowest level of detail stored in a fact table (e.g., daily sales per product).

**Primary Key vs Surrogate Key**

* **Primary Key**: Natural, business-defined.
* **Surrogate Key**: Artificial, system-generated for uniqueness.

**OLTP vs OLAP**

* **OLTP**: Real-time transactional systems.
* **OLAP**: Analytical systems optimized for read-heavy queries.

**Normalization vs Denormalization**

* **Normalization**: Reduces redundancy, improves integrity.
* **Denormalization**: Improves read performance by introducing redundancy.

**Schema on Read vs Schema on Write**

* **Schema on Read**: Applied during data access (flexible).
* **Schema on Write**: Enforced at ingestion (structured).

**🔁 Data Loading Methods**

**Full Load**  
Replaces all data by truncating and loading fresh.

**Delta Load**  
Loads only new or changed records.

**Incremental Loading**  
Identifies and loads records added or modified since last load.

**Change Data Capture (CDC)**  
Tracks changes in source data in real time or batch (insert, update, delete).

**Truncate and Load**  
Deletes existing data and reloads completely — simple but not efficient for large datasets.

**Slowly Changing Dimensions (SCDs)**

* **Type 1**: Overwrite old data.
* **Type 2**: Add new row with versioning.
* **Type 3**: Add new column for recent change.
* **Type 4**: Store changes in a separate history table.
* **Type 6**: Hybrid of Types 1, 2, and 3.