COURSE GLOSSARY

# Database Design

Data lake: A large, cost-effective storage repository that holds raw or lightly processed data of all types (structured, semi-structured, unstructured) using object storage and schema-on-read, intended for scalable storage and diverse downstream analytics.

Data warehouse: A centralized storage system optimized for OLAP that consolidates data from multiple sources, uses denormalized or dimensional schemas for fast analytical queries, and often employs massively parallel processing for performance.

Database model: A high-level conceptual framework that defines how data is logically organized and related (for example, relational, document, key-value, columnar, or graph models), guiding schema design and data access patterns.

Database view: A virtual table defined by a stored query whose result set can be queried like a table but is not physically stored as data

DBMS (Database Management System): System software that creates, manages, and provides access to databases by handling data storage, schema enforcement, query processing, transaction management, and access control for users and applications.

Dimension table: In dimensional modeling, a table that stores descriptive attributes or context about facts (e.g., product, time, store, customer) and is used to filter, group, and label measures from the fact table.

Dimensional modeling: A design technique for data warehouses that organizes data into fact and dimension tables to make analytical queries intuitive and performant, often using star or snowflake schemas.

Data type: A classification that defines the kind of values a field can hold (such as Boolean, Numeric, String, DateTime, or Spatial) and determines how operations and profiling behave on that field

Discrete (time): A way of treating date/time values as separate categories (e.g., “January”, “Monday”) so each distinct value is shown as its own mark or bin

Fact table: In dimensional modeling, a central table that stores measurable, often numeric metrics or events (e.g., sales amount, quantity) and typically contains foreign keys referencing related dimension tables.

Materialized view: A view whose query results are physically stored on disk so reads are fast and precomputed, requiring explicit refreshes to update underlying data and making them suitable for expensive analytical queries where slightly stale results are acceptable.

Normal forms (1NF, 2NF, 3NF): Standardized levels of normalization where 1NF requires atomic values and unique records, 2NF removes partial dependencies on composite keys, and 3NF removes transitive dependencies among non-key attributes to minimize redundancy and anomalies.

Normalization: A set of design techniques that decompose tables into smaller related tables and enforce rules to reduce redundancy and update anomalies, improving data integrity and consistency.

OLAP (Online Analytical Processing): A class of systems and databases optimized for complex, read-heavy analytical queries over large volumes of historical data to support reporting and decision-making

OLAP systems prioritize fast reads and aggregated analysis.

OLTP (Online Transaction Processing): A class of systems and databases optimized for handling a large number of short, atomic transactions such as inserts, updates, and deletes for day-to-day operations

OLTP systems prioritize fast writes and transaction integrity.

Operational database: A database designed to support OLTP use cases that stores current transactional data used by applications and business processes, typically optimized for quick inserts, updates, and concurrency control.

Relational model: A database model that represents data in tables (relations) of rows and columns with primary keys to enforce uniqueness and foreign keys to define relationships, typically accessed via SQL.

Schema-on-read vs Schema-on-write: Schema-on-read defers schema application until data is read (common in data lakes), allowing flexible ingestion of raw data, whereas schema-on-write enforces a schema at write time (common in relational databases and warehouses), ensuring structured, consistent data at storage.

Schema: The concrete blueprint or implementation of a database model that defines specific tables/collections, fields/columns, data types, relationships, indexes, and constraints for a database.

Snowflake schema: A dimensional model variant where dimension tables are normalized into multiple related tables (creating a deeper hierarchy), reducing redundancy but increasing the number of joins needed for queries.

Star schema: A dimensional model layout where a central fact table (containing measures) is directly connected to multiple denormalized dimension tables, producing a simple, easy-to-query structure for analytics.

Structured data: Data that conforms to a predefined schema of tables, columns, and data types (e.g., relational databases), making it easy to validate and analyze but less flexible to change.

Unstructured data: Data that lacks a predefined tabular schema and is stored in raw formats such as text, images, audio, and video, which offers flexibility but requires additional processing to analyze.

views simplify complex queries, provide abstraction, and can restrict user access to underlying data.