datacamp

COURSE GLOSSARY

Data Warehousing Concepts

Column store: A storage format that stores column values together, improving performance and compression for analytical queries that scan or aggregate a subset of columns

Data cube (OLAP cube): A multidimensional data structure that organizes measures across multiple dimensions (e.g., product, time, region) to enable quick aggregation and drill-down analysis

Data governance: The organizational policies, standards, and processes that define data quality, security, definitions, and ownership to ensure trusted and compliant use of data

Data lake: A large centralized repository that stores raw structured, semi-structured, and unstructured data (e.g., logs, audio, video) for future analysis and flexible use

Data mart: A department-focused, smaller relational data store designed to support analysis for a specific business area, typically containing a subset of a data warehouse

Data modeling: The practice of designing how data is organized into tables, defining relationships, grain, dimensions, and metrics to meet business requirements for reporting and analysis

Data pipeline: An automated end-to-end workflow that extracts data from source systems, applies transformations or validations, and delivers data to downstream systems like a warehouse or reports

Hallucination: When a model produces confident but incorrect or fabricated information, often due to gaps or biases in its training data or reasoning process

Hallucination: When a model produces confident but incorrect or fabricated information, often due to gaps or biases in its training data or reasoning process

Denormalization: The deliberate consolidation of related data into fewer tables to reduce joins and speed up analytical queries, often used in data marts and star schemas

Dimension table: A reference table that stores descriptive attributes (dimensions) about facts, such as customer, product, time, or location, used to filter and group measures

ELT (Extract, Load, Transform): A data integration process that extracts data, loads raw copies into the warehouse, and then performs transformations inside the warehouse environment

ETL (Extract, Transform, Load): A data integration process that extracts data from sources, transforms and cleans it externally, and then loads the transformed data into the data warehouse

Fact table: A central table in a data model that stores measurable, numeric metrics or events (facts) at a defined grain, with foreign keys linking to dimensions

Normalization: A design principle that organizes data to minimize redundancy by splitting information into multiple related tables, commonly used in transactional systems

OLAP (Online Analytical Processing): Systems and technologies optimized for fast, multidimensional analysis of large volumes of data, often implemented via data cubes to enable slicing and dicing

OLTP (Online Transaction Processing): Systems optimized for processing many small, fast transactional operations on relational tables, used for day-to-day business activities rather than analysis

Presentation layer: The layer where end users and applications interact with warehouse data via BI tools, dashboards, data mining tools, or direct queries

Row store: A physical storage format that stores complete rows together in blocks, which is efficient for transactional workloads that access full records

Slowly Changing Dimension (SCD): A set of techniques for handling updates to dimensional attributes over time to preserve appropriate historical context, commonly implemented as Type I (overwrite), Type II (new row with new key), or Type III (additional column for prior value)

Snowflake schema: A variation of the star schema where one or more dimension tables are normalized into related sub-dimensions, requiring additional joins for some queries

Staging layer: A temporary storage area in the ETL/ELT process where extracted data is placed and cleaned or transformed before being loaded into the main storage layer

Star schema: A denormalized data model with a single central fact table directly connected to multiple dimension tables, optimized for fast and simple queries

Transactional database (OLTP source): A database optimized for recording and processing high volumes of simple transactional operations (inserts, updates, deletes) used as source systems feeding a warehouse