

COURSE GLOSSARY

Understanding Data Engineering

- **Data Engineering:** The practice of designing, building, and maintaining systems for collecting, storing, and processing data.
- **ETL (Extract, Transform, Load):** A process in data engineering that extracts data from sources, transforms it into a usable format, and loads it into storage systems.
- **Data Pipeline:** A series of processes that move data from sources to destinations, often including cleaning and transformation.
- **Batch Processing:** Handling large volumes of data at scheduled intervals.
- **Stream Processing:** Real-time processing of continuously generated data.
- **Data Warehouse:** A centralized repository optimized for storing and analyzing structured data.
- **Data Lake:** A storage system that holds raw, unstructured, and structured data for future processing.
- **Data Mart:** A subset of a data warehouse focused on a specific business area or department.
- **Relational Database:** A structured data storage system that organizes data into tables with rows and columns.
- **SQL (Structured Query Language):** A language used to query and manage relational databases.
- **NoSQL Database:** A non-relational database designed to store unstructured or semi-structured data, such as documents or key-value pairs.
- **Schema:** The structure that defines how data is organized in a database, including tables, fields, and relationships.
- **Data Modeling:** The process of designing how data is structured and related within databases.
- **OLAP (Online Analytical Processing):** A system optimized for analytical queries, such as trend analysis and reporting.
- **OLTP (Online Transaction Processing):** A system optimized for transaction-oriented tasks, such as order processing.
- **Data Governance:** Policies and practices ensuring data quality, security, and compliance.
- **Data Quality:** A measure of the accuracy, consistency, and reliability of data.
- **Data Integration:** Combining data from multiple sources into a unified view.
- **Data Transformation:** The process of cleaning, aggregating, or reformatting data to make it usable.
- **Airflow:** An open-source tool for orchestrating and scheduling data workflows.
- **Apache Spark:** A distributed data processing framework for large-scale batch and stream data.
- **Hadoop:** A framework for distributed storage and processing of large datasets.
- **Scalability:** The ability of a system to handle increasing amounts of data or workload.
- **Latency:** The delay between data input and the system's response or processing.
- **Data Engineer:** A professional responsible for building and managing data infrastructure.
- **DataOps:** A practice that combines DevOps principles with data engineering to improve collaboration, automation, and data delivery.