COURSE GLOSSARY

ETL and ELT in Python

API (Application Programming Interface): A set of rules and endpoints that allow software systems to request or send data or services without direct access to the underlying database

Connection URI: A formatted string that encodes database connection details (driver/schema, username, password, host, port, database) used by tools like SQLAlchemy to create a connection engine

Data pipeline: A sequence of processes that move data from one or more source systems to a destination while optionally transforming the data along the way

DataFrame: A two-dimensional, tabular data structure provided by pandas that organizes data as labeled rows and columns and is used for in-memory data manipulation

Destination (landing zone): The storage or system where pipeline output is persisted for downstream use, such as files, databases, or a data warehouse

ELT (Extract, Load, Transform): A pipeline design pattern that extracts data, loads it into a destination (often a data warehouse), and performs transformations there rather than before loading

ETL (Extract, Transform, Load): A pipeline design pattern that first extracts data from sources, then transforms it into the desired shape, and finally loads the transformed data to a destination

loc: A pandas DataFrame indexer that selects data by label-based indexing and boolean masks for rows and columns

Logging: The practice of emitting timestamped messages during program execution to record events, successes, warnings, and errors for monitoring and debugging

Orchestration tool (e.g., Apache Airflow): Software that schedules, manages, and monitors execution of pipelines or workflows, providing features such as scheduling, retries, resource management, and observability

pandas: A Python library for data manipulation and analysis that provides DataFrame objects and functions for reading, transforming, and writing tabular data

Parquet (Apache Parquet): An open-source, columnar file format optimized for efficient storage and retrieval of tabular data, often faster than CSV for large datasets

read_csv: A pandas function that reads a CSV file from a file path or buffer and returns a DataFrame, with optional parameters for delimiter, header, and parsing behavior

read_sql: A pandas function that executes a SQL query against a database connection and returns the result as a DataFrame

Source system: Any origin of data such as CSV, Parquet, JSON files, APIs, databases, data lakes, or web sources that a pipeline reads from

to_csv: A pandas DataFrame method that writes the DataFrame to a CSV file, with configurable arguments such as header, index, and separator

to_datetime: A pandas function that converts strings or numeric timestamp representations into pandas datetime objects for time-based analysis

to_sql: A pandas DataFrame method that writes the DataFrame to a SQL table, with options for table name, connection object, append/replace behavior, and index handling

try-except: A Python control structure that attempts to run code in the try block and, if an exception occurs, executes fallback or error-handling code in the except block instead of crashing

Unit test: A small automated test that verifies the behavior of a single component or function in code, typically written to assert expected outputs or side effects