Data Engineering Fundamentals with Pipelines, Analytics, and Agentic AI

Duration: 5 days / 40 hours
Prerequisites: Basic SQL knowledge and basic programming knowledge in Python is helpful.

Day 1: Data Engineering Fundamentals and SQL

Topics

- Introduction to Data Engineering
- Role of a Data Engineer
- Data lifecycle
- Structured, semi-structured, and unstructured data
- Databases vs data warehouses vs data lakes
- SQL fundamentals
- SELECT, WHERE, GROUP BY, ORDER BY
- Joins
- Subqueries
- Views
- Basic data modeling concepts

Labs

- Write SQL queries on sample business data
- Create tables and relationships
- Perform joins and aggregations
- Analyze sales or employee data using SQL

Outcome

Participants will understand core data engineering concepts and write SQL queries for data analysis.

Day 2: Python for Data Engineering and File-Based Data Processing

Topics

- Python essentials for data engineering
- Reading and writing files
- CSV, JSON, Excel data processing
- Introduction to pandas
- Data cleaning
- Handling missing values
- Data transformation
- Data validation
- Logging and error handling in pipelines

Labs

- Read CSV and JSON datasets using Python
- Clean and transform raw data using pandas
- Generate a processed output file
- Add logging and error handling to a data processing script

Outcome

Participants will be able to process and transform data using Python.

Day 3: ETL Pipelines and Database Integration

Topics

- ETL vs ELT
- Batch processing concepts
- Data extraction from files and databases
- Data transformation rules
- Loading data into databases
- Incremental load concepts
- Data quality checks
- Pipeline failure handling
- Introduction to workflow orchestration
- Overview of tools such as Airflow / Azure Data Factory / cloud data pipelines

Labs

- Build a simple ETL pipeline using Python and SQL
- Extract data from CSV
- Transform and clean data
- Load processed data into a database
- Add validation checks and error logs

Outcome

Participants will be able to design and implement a basic ETL pipeline.

Day 4: Data Warehousing, Cloud Data Platforms, and Analytics Readiness

Topics

- Data warehouse concepts
- Fact and dimension tables
- Star schema and snowflake schema
- Slowly Changing Dimensions overview
- Data lake and lakehouse concepts
- Introduction to cloud data platforms
- Data storage formats: CSV, JSON, Parquet
- Partitioning basics
- Data quality and governance basics

- Preparing data for BI and analytics

Labs

- Design a simple star schema
- Create fact and dimension tables
- Load transformed data into warehouse-style tables
- Prepare an analytics-ready dataset
- Create simple reporting queries

Outcome

Participants will understand warehouse design and prepare data for reporting and analytics use cases.

Day 5: Agentic AI for Data Engineering and Capstone Project

Topics

- Introduction to AI in Data Engineering
- What is Agentic AI?
- Agentic AI use cases in Data Engineering:
    - Data quality monitoring
    - Pipeline troubleshooting
    - SQL query generation
    - Metadata understanding
    - Data catalog assistance
    - Automated documentation
- Prompt engineering for data tasks
- Using AI for SQL generation and optimization suggestions
- AI-assisted data validation
- AI-assisted pipeline documentation
- Responsible AI and data privacy
- Human-in-the-loop validation for AI-generated outputs

Labs

- Use AI prompts to generate SQL queries from business requirements
- Use AI to document an ETL pipeline
- Create a basic AI-assisted data quality checker
- Capstone: Build a mini data pipeline that extracts, transforms, validates, loads data, and generates AI-assisted documentation

Outcome

Participants will understand how Agentic AI can improve data engineering productivity and complete a mini data pipeline project.
