DATA ENGINEERING FUNDAMENTALS – 2025 EDITION

" Build the pipelines that power data-driven decisions."

Market Demand Note

Data engineering is the foundation of every AI, analytics, and machine learning project. Organizations need skilled data engineers to collect, clean, store, and process data efficiently. This program focuses on essential data engineering skills to handle real-world data pipelines.

Duration: 6 weeks | **Mode**: Online/Offline

Data Engineering fundamentals @ https://udaan.x-fuzion.com/

📘 Data Engineering Fundamentals — 2025 Edition

Table of Contents

Detailed Content

Week 1: Introduction to Data Engineering

- **Understand the data engineer's role:**
  Design, build, and maintain scalable data pipelines and infrastructure for analytics, reporting, and ML.
- **Key responsibilities:**
  - **Collect, process, and store vast volumes of structured/unstructured data.**
  - **Ensure data quality, reliability, and accessibility.**
  - **Automate workflows and optimize performance.**
- **Data pipeline overview:**
  - **ETL (Extract, Transform, Load): Data is cleaned and transformed before loading into warehouse.**
  - **ELT (Extract, Load, Transform): Raw data is loaded into warehouse, then transformed as needed.**
  - **Comparison: When to use ETL vs. ELT, pros/cons, trends in modern data stacks.**
- **Hands-on:**
  - **Research real-world data pipelines (e.g., from e-commerce, IoT, fintech).**
  - **Diagram a simple ETL/ELT pipeline for a business case.**

Week 2: Data Storage & Databases

- **Relational databases:**
  - **Tables, rows, columns, schemas, SQL.**
  - **Examples: PostgreSQL, MySQL, SQL Server.**
- **NoSQL databases:**
  - **Key-value, document, columnar, graph.**
  - **Examples: MongoDB, Cassandra, DynamoDB.**
  - **When to use each? Consistency, scalability, flexibility needs.**
- **SQL basics:**
  - **SELECT: Query data, filter with WHERE.**

    - **JOIN: Combine data from multiple tables (INNER, LEFT, RIGHT, FULL).**
    - **GROUP BY: Aggregate data (COUNT, SUM, AVG, MIN, MAX).**
- **Hands-on:**
    - **Set up a local database (e.g., PostgreSQL).**
    - **Import sample data (e.g., sales, users, products).**
    - **Write and execute SELECT, JOIN, and GROUP BY queries.**

Week 3: Data Processing with Python & Pandas

- **Reading data:**
    - **CSV, JSON (Pandas: read_csv, read_json).**
    - **SQL databases (Pandas: read_sql).**
- **Writing data:**
    - **CSV, JSON (to_csv, to_json).**
    - **SQL databases (to_sql).**
- **Cleaning data:**
    - **Handle missing values, duplicates, outliers, incorrect types.**
- **Transforming data:**
    - **Create new columns, merge datasets, reshape (pivot/melt).**
- **Hands-on:**
    - **Write Python scripts to clean and transform a real dataset (e.g., missing sales records, messy customer data).**
    - **Export cleaned data back to CSV/JSON/database.**

Week 4: Workflow Automation

- **Apache Airflow overview:**
    - **Open-source platform to programmatically author, schedule, and monitor workflows.**
    - **Core concepts: DAGs (Directed Acyclic Graphs), Tasks, Operators.**
- **Scheduling:**

    - **Define dependencies between tasks.**
    - **Set up hourly/daily/weekly runs.**
    - **Handle failures, retries, alerts.**
- **Hands-on:**
    - **Install Airflow locally or in the cloud.**
    - **Create a DAG to automate your Week 3 data cleaning/transformation pipeline.**
    - **Monitor job status, logs, and troubleshoot.**

Week 5: Big Data Processing

- **Apache Spark overview:**
    - **Distributed computing framework for processing large datasets.**
    - **Core concepts: RDDs, DataFrames, transformations, actions.**
- **Use cases:**
    - **Batch processing, streaming, ML.**
- **Hands-on:**
    - **Set up Spark locally or in the cloud (e.g., Databricks Community Edition).**
    - **Read a large dataset (millions of rows), clean, and aggregate using Spark.**
    - **Compare performance to Python/Pandas.**

Week 6: Final Project – Sales Data Pipeline for Analytics

- **Project scope:**
    - **Ingest sales data (CSV/JSON).**
    - **Clean, validate, and transform the data.**
    - **Load into a cloud data warehouse (e.g., Google BigQuery, AWS Redshift, Azure Synapse).**
    - **Schedule the entire pipeline with Airflow.**

    - **(Optional) Add basic analytics or dashboards.**
- **Hands-on:**
    - **Build and document the pipeline end-to-end.**
    - **Deploy to a cloud environment.**
    - **Present your architecture, code, and results.**
- **Output:**
    - **Fully automated, production-ready ETL pipeline.**
    - **GitHub repo with code, README, and visuals.**
    - **Portfolio-ready project for job interviews.**

Tools Covered

- **Python, Pandas, Jupyter Notebook (core data processing)**
- **SQL (relational databases)**
- **Apache Airflow (workflow automation)**
- **Apache Spark (big data processing)**
- **Google BigQuery, AWS Redshift, Azure Synapse (cloud data warehouses)**

Final Project: "Sales Data Pipeline for Analytics"

- **Goal: Automate the ingestion, transformation, and storage of sales data for analytics.**
- **Tasks:**
    - **Ingest: Pull data from CSV/JSON sources.**
    - **Transform: Clean, validate, and enrich the data (Python/Pandas/Spark).**
    - **Load: Store in a cloud data warehouse.**
    - **Automate: Schedule the pipeline with Airflow.**
- **Output:**
    - **Code: Python scripts, Airflow DAGs, SQL.**
    - **Documentation: README, architecture diagram.**

- **Presentation: Demo your pipeline and explain design choices.**