Designing and Implementing Data Onboarding Pipelines

Course Description

This course provides a comprehensive exploration of data onboarding pipelines, focusing on the end-to-end process of ingesting, transforming, validating, and preparing data for enterprise analytics and AI platforms. Participants will gain hands-on experience with modern data engineering tools and frameworks, learning how to design scalable, automated, and resilient onboarding workflows that align with enterprise data strategies.

Duration : 4 Days

Pre-requisites

- Basic understanding of databases (SQL) and data formats (CSV, JSON, Parquet).
- Familiarity with at least one programming language (Python, R, or Scala).
- Awareness of cloud platforms (Azure, AWS, or GCP) is helpful but not mandatory.
- Prior exposure to ETL/ELT concepts is beneficial.

Learning Objectives

By the end of this course, participants will be able to:

- Understand the architecture and lifecycle of data onboarding pipelines.
- Design ingestion workflows for batch and streaming data sources.
- Apply data validation, cleansing, and transformation techniques.
- Implement scalable pipelines using modern frameworks (Spark, Databricks, Kafka).
- Automate pipeline orchestration and monitoring.
- Align onboarding pipelines with enterprise data governance and compliance needs.

Content Coverage

Day 1 – Foundations of Data Onboarding Pipelines

- **Introduction to Data Onboarding**
    - Definition and importance in enterprise data strategy
    - Batch vs. streaming onboarding
    - Key challenges and best practices

- **Pipeline Architecture**
    - Components: ingestion, staging, transformation, validation, loading
    - Data lake vs. data warehouse onboarding
    - Role of metadata and schema management
- **Hands-on Lab**: Designing a simple ingestion pipeline with SQL and Python

Day 2 – Data Ingestion and Staging

- **Data Source Connectivity**
    - APIs, databases, flat files, cloud storage
    - Streaming sources (Kafka, Event Hubs)
- **Batch Ingestion Techniques**
    - Bulk load strategies
    - Incremental loads and change data capture (CDC)
- **Streaming Ingestion Techniques**
    - Real-time event processing
    - Handling late-arriving and out-of-order data
- **Staging Layer Design**
    - Raw zone vs. curated zone
    - Schema evolution and versioning
- **Hands-on Lab**: Building ingestion pipelines with Spark and Kafka

Day 3 – Data Transformation and Validation

- **Data Cleansing & Standardization**
    - Handling nulls, duplicates, and anomalies
    - Normalization and enrichment
- **Transformation Frameworks**
    - ELT vs. ETL approaches
    - Using Spark SQL, Databricks notebooks, and Python scripts
- **Data Validation & Quality Checks**
    - Rule-based validation
    - Profiling and anomaly detection

    - Automating data quality checks
- **Hands-on Lab**: Implementing transformation and validation workflows

Day 4 – Pipeline Orchestration, Monitoring, and Governance

- **Pipeline Orchestration**
    - Workflow automation with Airflow, Azure Data Factory, or Prefect
    - Scheduling and dependency management
- **Monitoring & Logging**
    - Metrics collection and alerting
    - Error handling and retries
- **Governance & Compliance**
    - Data lineage and cataloging
    - Security, privacy, and regulatory compliance (GDPR, HIPAA)
- **Enterprise Integration**
    - Connecting onboarding pipelines to downstream analytics and AI platforms
- **Capstone Lab**: End-to-end pipeline design, orchestration, and monitoring