

Forging Data Excellence: From Spark to Kafka (11-Day Program)

Day 1: Foundations of Data Engineering

- Introduction to Data Engineering
 - Roles and responsibilities
 - Real-world examples
- Data Engineering Ecosystem
 - Databases, Pipelines, Warehousing, Orchestration
- Hands-on Setup
 - PostgreSQL installation
 - Apache Hop configuration
- Case Studies
 - Modern data engineering stacks
- Guided Hands-on Labs
 - Data loading techniques
 - Basic SQL queries in PostgreSQL

Day 2: Data Modeling and Lakehouse Architecture

- OLTP vs OLAP
- Data Modeling: Star Schema, Snowflake, Data Vault
- Data Lakehouse & Medallion Architecture
- Hands-on Schema Design in PostgreSQL
- Medallion & Vault Model Implementation

Day 3: Data Ingestion & Batch ETL

- Batch vs Stream Ingestion Strategies
- ETL vs ELT Architecture & Design
- Introduction to Apache Hop
- Hands-on ETL Pipeline: CSV ingestion, cleansing, PostgreSQL loading, logging

Forging Data Excellence: From Spark to Kafka (11-Day Program)

Day 4: Change Data Capture (CDC) & Slowly Changing Dimensions (SCD)

- CDC Overview & Methods (Log-based, Trigger-based, Query-based)
- Hands-on CDC in PostgreSQL
- SCD Types (Type 1, 2, 3) & Hands-on Versioning Implementation

Day 5: Orchestration with Apache Airflow

- Importance of Orchestration in Data Pipelines
- Apache Airflow Concepts (DAGs, Operators, Sensors)
- Hands-on Airflow Setup
- Build Basic DAG & ETL Pipeline with Monitoring

Day 6: Data Lakes, Lakehouses & Mini Project 1

- Data Storage Concepts: Data Lake, Warehouse, Lakehouse
- Lakehouse Technologies: Delta Lake, Iceberg
- Mini Project 1:
 - Bronze/Silver/Gold layers
 - Ingestion, transformation, validation

Day 7: Governance, Metadata & Security

- Data Lineage, Cataloging, Metadata (Purview)
- RBAC, Encryption, Row/Column-Level Security
- Data Quality Fundamentals
- Hands-on: RBAC in PostgreSQL & Metadata Management

Day 8: Performance Optimization & Materialization

Forging Data Excellence: From Spark to Kafka (11-Day Program)

- Query Optimization, PostgreSQL Analysis (EXPLAIN/ANALYZE)
- MPP Concepts, Caching, Materialized Views
- Hands-on SQL Tuning, Performance Improvements
- Case Study Review

Day 9: DataOps & CI/CD

- DataOps Lifecycle & Principles
- Git for ETL, DVC Basics
- CI/CD for Pipelines: GitHub Actions, Azure Pipelines
- Intro to IaC (Terraform)
- Hands-on: Git-based Project Setup, DAG Deployment

Day 10: Big Data, HDFS & Spark Introduction

- Big Data Architectures: Lambda & Kappa
- HDFS Basics: Blocks, Replication, Namenode
- MapReduce & Spark Ecosystem Overview
- Hands-on: Spark Setup, PySpark Shell, RDD Operations

Day 11: Spark SQL & Streaming

- Spark SQL & DataFrames
- Structured Streaming Overview
- Hands-on Batch & Streaming Pipelines
 - CSV to DataFrame -> Transformations -> Parquet
 - Directory/Socket ingestion -> Aggregation -> Console Output