PySpark for Data Engineers

Course Description: Unlock the power of big data with PySpark! This comprehensive course is tailored for data engineers aiming to master distributed computing, scalable data processing, and real-time analytics. Participants will dive deep into PySpark's architecture, APIs, and advanced features while gaining hands-on experience with real-world use cases and practical labs.

Duration: 5 days (40 hours)

Prerequisites:

- Familiarity with Python programming
- Basic knowledge of data engineering concepts
- Exposure to big data fundamentals (e.g., Hadoop, Spark)
- Awareness of cloud platforms (e.g., Azure, AWS, or GCP)

Course Contents

Day 1: Foundations of PySpark

- Introduction to PySpark
    - Spark Ecosystem Overview
    - Role of PySpark in Data Engineering
    - Advantages of PySpark for Distributed Computing
- PySpark Architecture
    - Driver & Executor
    - Cluster Manager Options
    - DAG Scheduler
- Spark Basics
    - RDDs: Overview, Creation, Transformations, Actions
    - SparkContext
- Hands-on: Setting up PySpark locally and creating basic RDD operations

Day 2: Working with Spark DataFrames

- Introduction to DataFrames

    - Differences between RDDs and DataFrames
    - Optimizations with Catalyst and Tungsten
- DataFrames APIs
    - Reading Data (CSV, JSON, Parquet)
    - Transformations (filter, groupBy, join)
    - Actions (count, show, write)
- Spark SQL
    - Running SQL Queries on DataFrames
    - Query Optimization Techniques
- Hands-on: Build a DataFrame pipeline for data cleaning and aggregation

Day 3: Advanced PySpark Concepts

- Spark Streaming
    - Structured Streaming Basics
    - Event Time vs Processing Time
    - Window Operations
- Advanced Transformations and Actions
    - Aggregate Functions
    - User-Defined Functions (UDFs)
- Performance Optimization in PySpark
    - Partitioning and Bucketing
    - Caching and Persistence
- Hands-on: Streaming data pipeline with transformations and real-time outputs

Day 4: Machine Learning with PySpark

- Introduction to MLlib
    - Overview of MLlib APIs
    - Supervised and Unsupervised Learning Techniques
- Data Preprocessing in PySpark
    - Feature Engineering
    - Handling Missing Data

- Building Machine Learning Pipelines
    - Train-Test Split
    - Model Selection and Evaluation
- Hands-on: Build a Spark ML pipeline for predictive analytics

Day 5: Integration and Real-World Applications

- Integrating PySpark with Other Tools
    - Apache Kafka and Spark Streaming
    - Cloud Integration: Deploying PySpark on Azure/AWS/GCP
- Real-World Use Cases
    - IoT Data Processing
    - Social Media Analytics
    - Financial Transaction Monitoring
- Mini Project:
    - Design, implement, and optimize a PySpark solution for a real-world dataset
- Visualization with PySpark
    - PySpark Integration with Apache Superset
    - Hands-on: Build dashboards to visualize processed data