Practical Apache Spark– 40 Hours

Course description

This hands-on course provides a comprehensive introduction to Apache Spark and its role in building modern, large-scale data processing pipelines. Participants will learn Spark's architecture, execution model, and parallel processing concepts, and gain practical experience working with RDDs, DataFrames, Spark SQL, and Spark Streaming. The course covers batch and streaming processing, structured streaming with Kafka, partitioning strategies, performance tuning, and monitoring using Spark UI. By the end of the course, participants will be able to design, build, and optimize scalable data pipelines using Spark in real-world data engineering environments.

What Participants Gain?

- Deep understanding of Oracle architecture and how it impacts performance
- Ability to analyze execution plans and identify performance bottlenecks
- Practical skills in SQL tuning and query optimization
- Using tracing and monitoring tools (AWR, ASH, ADDM, SQL Trace, TKPROF) to diagnose issues
- Designing and using indexes efficiently, including advanced index types
- Applying partitioning strategies to improve performance and data management
- Managing optimizer statistics and ensuring plan stability in production
- Rewriting SQL and applying hints to improve execution plans in real-world scenarios

Prerequisites

- Experience with Python or Scala
- Basic knowledge of SQL
- Familiarity with data processing concepts and distributed systems – an advantage
- No prior experience with Spark is required

Target Audience

- Data Engineers working with large-scale data processing systems
- Backend and software engineers building data pipelines
- Big Data and analytics engineers
- BI professionals moving into large-scale data processing
- Developers integrating batch and streaming data platforms

Course Syllabus

- **Introduction to Spark**
  - Spark architecture
  - Running modes
  - Parallelism concept
  - SparkContent & SQLContent & SparkSession
  - Common file formats
- **RDD – Low Level API**
  - What is it RDD
  - Lazy execution
  - Transformations & Actions
  - Collecting data to driver
- **Broadcast & Accumulators**
- **DataFrame Operations**
  - Introduction to DataFrame
  - DataFrame optimizers
  - Working with Data Sources
  - DataFrame API
  - Spark functions
  - UDF
  - Analytic function
- **Spark SQL API**
- **Spark Partitioning**
- **Spark Streaming**
  - Spark Streaming introduction
  - Spark Streaming concepts
- **Structured Streaming**
  - Structured Streaming introduction
  - Structured Streaming with Kafka
  - Stateful Streaming Application
- **Management & Monitoring**
  - Spark UI Introduction
  - Read Execution Plan
  - Find bottlenecks
  - Access logs
  - Performance tuning