CONCEPT

Big Data Analytics with Apache Spark: PySpark, Spark SQL and Performance Tuning

Big Data Analytics with Apache Spark: PySpark, Spark SQL and Performance Tuning

Introduction:

Big data analytics becomes necessary when clickstream, transaction, sensor and log volumes outgrow a single machine, so reports time out, samples replace full history and analysts wait on overnight extracts. This Core Concept course builds hands-on Apache Spark capability: distributed storage and partitioning, Spark architecture, DataFrames and Spark SQL, PySpark transformations, columnar file formats, performance tuning, streaming and machine learning at scale. Every session runs as a lab on multi-gigabyte datasets, and participants leave with an End-to-End PySpark Analytics Pipeline and Tuning Report built on a realistic business question.

Course Objectives:

- Judge when a dataset needs distributed processing by assessing volume, velocity, variety, veracity and value against single-machine limits
- Explain how the Spark driver, executors, cluster manager and DAG scheduler execute a job, and read the Spark UI to trace stages and tasks
- Write PySpark DataFrame and Spark SQL code that filters, joins, aggregates and windows large datasets stored in Parquet, ORC and Delta Lake tables
- Diagnose slow Spark jobs and apply partitioning, caching, broadcast joins and adaptive query execution to reduce shuffle and run time
- Build a Structured Streaming query and an MLlib pipeline to extend batch analysis to incoming events and predictive scoring at scale
- Deliver an End-to-End PySpark Analytics Pipeline and Tuning Report that colleagues can rerun, review and schedule

Target Audience:

- Data analysts whose SQL and spreadsheet workloads have outgrown a single database or laptop
- Data engineers responsible for ingesting and transforming log, event and transaction data
- Business intelligence developers who prepare large aggregated datasets for reporting layers
- Analytics engineers maintaining batch jobs on shared cluster or cloud data platforms
- Technical analysts in telecoms, retail, finance, energy and public service handling high-volume records

Course Outline:

Day 1: Big Data Foundations and Distributed Computing Concepts

- Five Vs Assessment: Volume, Velocity, Variety, Veracity and Value
- Single-Machine Limits Versus Horizontal Scaling Decision Checklist
- Hadoop HDFS Blocks, Replication and Object Storage Buckets Compared
- MapReduce Model and the Move to In-Memory Processing
- Workload Inventory: Current Data Sizes, Latency Needs and Pain Points

Day 2: Apache Spark Architecture and the PySpark Environment

- Driver, Executors and Cluster Managers: Standalone, YARN and Kubernetes
- RDD Lineage, Lazy Evaluation and the DAG Scheduler
- Jobs, Stages, Tasks and Partitions Traced in the Spark UI
- SparkSession Setup, Notebook Environment and spark-submit Basics
- DataFrame and Dataset APIs Versus the RDD API

Day 3: DataFrames, Spark SQL and PySpark Transformations

- Reading and Writing CSV, JSON, Parquet and ORC With Explicit Schemas
- Narrow and Wide Transformations: select, filter, withColumn and join
- groupBy Aggregations, Window Functions and User-Defined Functions
- Spark SQL Temporary Views and Catalyst Optimiser Query Plans with explain
- Delta Lake Tables: ACID Writes, Schema Enforcement and Version History

Day 4: Performance Tuning, Streaming and Machine Learning at Scale

- Partition Sizing, repartition and coalesce and Shuffle Partition Settings
- Caching and Persistence Levels and When to unpersist
- Broadcast Joins, Data Skew and Adaptive Query Execution
- Structured Streaming Micro-Batches, Sources, Sinks and Checkpoints
- MLlib Pipelines: Transformers, Estimators and Distributed Model Training

Day 5: Lab Project and the Analytics Pipeline Report

- Managed Spark Services and Cloud Object Storage Options Overview
- E-Commerce Clickstream Lab: Ingest, Cleanse and Partition Raw Events
- Utility Meter Lab: Sessionisation and Time-Window Aggregation
- End-to-End PySpark Analytics Pipeline and Tuning Report Build
- Peer Code Review and Pipeline Walkthrough

Skills You Will Gain:

- Distributed Data Processing
- PySpark Programming
- Spark SQL Querying
- Columnar Storage Design
- Spark Job Performance Tuning
- Stream Processing
- Scalable Machine Learning
- Spark UI Diagnostics

Why Attend This Course:

- Return with an End-to-End PySpark Analytics Pipeline and Tuning Report built and reviewed during the labs
- Analyse full history instead of samples by moving heavy queries from single-machine tools to Spark
- Cut job run time and cluster cost by recognising shuffle, skew and small-file problems before they reach production
- Practise on clickstream and meter datasets alongside analysts and engineers from several sectors

Conclusion:

Spark turns data that once needed samples and overnight extracts into something an analyst can query in minutes, provided the code and storage layout respect how distributed processing works. The course moves from the five Vs and distributed storage, through Spark architecture, DataFrames, Spark SQL and PySpark transformations, to tuning, streaming and MLlib. The final day brings these together in clickstream and meter labs that produce an End-to-End PySpark Analytics Pipeline and Tuning Report for the workplace.