

# Apache Spark for Big Data Processing

**Duration:** 3 Days

## Description

---

Apache Spark for Big Data Processing is a technical course covering the core components of Apache Spark and application development methods for fast and effective processing of large-scale data on distributed systems. It helps technical professionals understand the principles and practices required to build scalable data processing solutions. Through hands-on exercises, delegates explore Apache Spark architecture, distributed data processing, PySpark, RDD and DataFrame APIs, Spark SQL, data ingestion and output, transformations, and fundamental performance optimization. Participants gain the ability to read, clean, transform, analyse, and process data from different sources while developing foundational ETL workflows.

### Delegates will learn how to

- Understand Apache Spark architecture, components, and distributed processing principles
- Develop data processing applications using PySpark, RDDs, and DataFrames
- Query and analyse large-scale data using Spark SQL
- Read and write data in common file formats and through JDBC connections
- Build data cleansing, transformation, aggregation, and ETL workflows
- Apply fundamental Spark performance optimization techniques
- Develop basic Structured Streaming applications

### Audience

- Data Engineers
- Big Data Developers
- Data Analysts
- ETL Developers
- Software Developers and technical teams working with data processing and big data technologies

## Outline

---

### 1. Introduction to Big Data and Apache Spark

- Big Data concepts and core characteristics
- Overview of big data processing technologies
- Apache Hadoop and Apache Spark comparison

- Apache Spark use cases and ecosystem
- Overview of Spark Core, Spark SQL, Structured Streaming, and MLlib
- Open Data Table formats including Delta Lake and Iceberg

## **2. Apache Spark Architecture**

- Apache Spark architecture and application lifecycle
- Driver, Executor, and Cluster Manager concepts
- Jobs, Stages, and Tasks
- Distributed computing and lazy evaluation
- Transformations and Actions
- SparkSession and SparkContext

## **3. Apache Spark Development Environment**

- Apache Spark installation and configuration
- Python and PySpark development environment
- Creating a SparkSession
- Running Spark applications
- Local Mode and Cluster Mode
- Using Spark Shell and PySpark

## **4. Resilient Distributed Datasets**

- RDD concepts and creation methods
- RDD Transformations and Actions
- Map, Filter, and Reduce operations
- Key-Value RDDs
- RDD Persistence and Caching
- RDD and DataFrame comparison

## **5. PySpark and DataFrame API**

- DataFrame concepts, architecture, and creation
- Schemas and data types
- Data selection, filtering, and column operations
- Data transformations, GroupBy, and aggregations
- Join operations and missing value management
- Duplicate record management and DataFrame practice

## **6. Spark SQL**

- Spark SQL architecture and its relationship with the DataFrame API
- Creating Temporary Views
- Spark SQL queries with SELECT, WHERE, GROUP BY, and ORDER BY

- JOIN operations and aggregate functions
- Window Functions
- Data analysis and querying practice with Spark SQL

## **7. Data Sources and File Formats**

- Reading and writing data with Spark
- CSV, JSON, Parquet, and ORC file formats
- Schema Inference and explicit schema definition
- Database connectivity through JDBC
- Partitioned Data and data read and write options
- Working with different data sources

## **8. Data Processing and ETL with Spark**

- ETL concepts and data processing workflows
- Batch Data Processing
- Data cleansing, transformation, and consolidation
- Data validation and aggregation operations
- Building ETL Pipelines
- Developing a basic ETL process with PySpark

## **9. Spark Performance Fundamentals**

- Performance concepts in Spark applications
- Partitioning, Repartition, and Coalesce
- Caching and Persistence
- Shuffle Operations and Broadcast Variables
- Basic join optimization
- Using Spark UI and reviewing Execution Plans

## **10. Introduction to Structured Streaming**

- Batch and streaming data processing
- Structured Streaming concepts and Streaming DataFrames
- Streaming Sources and Sinks
- Micro-Batch Processing
- Running streaming queries and Checkpointing
- Developing a simple Structured Streaming application

## **11. Hands-on Workshop: Apache Spark Data Processing**

- Preparing a sample data set and reading data with PySpark
- Creating DataFrames and performing data cleansing and transformation
- Analysing data with Spark SQL

- Combining data from different sources
- Developing a basic ETL Pipeline
- Saving results in Parquet format and reviewing the application in Spark UI

## Prerequisites

---

- Basic Python programming knowledge
- SQL querying knowledge
- Basic knowledge of databases and data processing workflows
- Prior familiarity with big data technologies is recommended