Training Course on Apache Spark for Advanced Data Processing

Professional Development Training Course Outline

Category	Data Science	**Duration**	10 Days
Base Fee	$3,000 USD	**Delivery Mode**	Online / On-site Hybrid

Course Introduction

Training Course on Apache Spark for Advanced Data Processing offers an in-depth exploration of **Apache Spark**, the leading **unified analytics engine for large-scale data processing**. Participants will gain mastery over **distributed data processing**, leveraging Spark's powerful capabilities for batch and real-time analytics. The curriculum focuses on practical application, enabling professionals to build robust, scalable, and high-performance **big data solutions** that drive impactful business insights. Through hands-on exercises and real-world **case studies**, attendees will unlock the full potential of Spark's core components, including **Spark SQL** for structured data manipulation and **Spark MLlib** for **machine learning** at scale.

Designed for data professionals seeking to advance their skills, this course dives deep into Spark's architecture and optimization techniques. From understanding **Resilient Distributed Datasets (RDDs)** and **DataFrames** to implementing advanced **ETL pipelines** and **real-time data streaming**, learners will acquire the expertise to tackle complex **data engineering** challenges. Emphasizing **performance tuning** and **fault tolerance**, the training equips participants with the knowledge to deploy and manage efficient Spark applications across diverse environments, from on-premise clusters to cloud-based platforms like **Databricks** and **AWS EMR**.

Programme Curriculum

Training Course on Apache Spark for Advanced Data Processing

Introduction

Training Course on Apache Spark for Advanced Data Processing offers an in-depth exploration of **Apache Spark**, the leading **unified analytics engine for large-scale data processing**. Participants will gain mastery over **distributed data processing**, leveraging Spark's powerful capabilities for batch and real-time analytics. The curriculum focuses on practical application, enabling professionals to build robust, scalable, and high-performance **big data solutions** that drive

impactful business insights. Through hands-on exercises and real-world **case studies**, attendees will unlock the full potential of Spark's core components, including **Spark SQL** for structured data manipulation and **Spark MLlib** for **machine learning** at scale.

Designed for data professionals seeking to advance their skills, this course dives deep into Spark's architecture and optimization techniques. From understanding **Resilient Distributed Datasets (RDDs)** and **DataFrames** to implementing advanced **ETL pipelines** and **real-time data streaming** , learners will acquire the expertise to tackle complex **data engineering** challenges. Emphasizing **performance tuning** and **fault tolerance**, the training equips participants with the knowledge to deploy and manage efficient Spark applications across diverse environments, from on-premise clusters to cloud-based platforms like **Databricks** and **AWS EMR**.

Course Duration

10 days

Course Objectives

1. Master **distributed computing** principles with Apache Spark.
2. Develop high-performance **Spark applications** for **big data analytics**.
3. Utilize **Spark SQL** for advanced **data manipulation** and **query optimization**.
4. Implement **real-time data processing** using **Spark Structured Streaming**.
5. Build and evaluate **machine learning models** with **Spark MLlib**.
6. Understand **Spark architecture** and **cluster management**.
7. Perform **performance tuning** and **troubleshooting** of Spark jobs.
8. Integrate Spark with various **data sources** and **data lakes**.
9. Apply **data warehousing** concepts using Spark.
10. Explore **GraphX** for **graph processing** and analysis.
11. Design **scalable ETL pipelines** using Spark.
12. Leverage Spark for **predictive analytics** and **business intelligence**.
13. Implement **data governance** and **security** best practices in Spark.

Organizational Benefits

- Faster analysis of massive datasets, leading to quicker insights and informed decision-making.
- Ability to handle growing data volumes efficiently, reducing infrastructure costs through optimized resource utilization.
- Rapid development and deployment of data-driven applications, enabling organizations to respond swiftly to market changes.
- Unlocking predictive power through machine learning and real-time analytics, driving innovation and competitive advantage.
- Streamlined data pipelines and improved troubleshooting capabilities, leading to less downtime and more reliable operations.
- Equipping data teams with cutting-edge skills in a highly demanded technology, fostering internal expertise and reducing reliance on external consultants.

Target Audience

- Data Engineers
- Data Scientists
- Big Data Developers
- BI Developers
- Machine Learning Engineers

- Solutions Architects
- ETL Developers
- Anyone working with large-scale datasets and seeking to leverage Apache Spark**.**

Course Outline

Module 1: Apache Spark Fundamentals & Architecture

- Introduction to Big Data and Distributed Computing
- Apache Spark Ecosystem: Spark Core, Spark SQL, Spark Streaming, MLlib, GraphX
- Spark Architecture: Driver, Executors, Cluster Manager (YARN, Mesos, Standalone)
- Resilient Distributed Datasets (RDDs): Transformations and Actions
- Setting up a Spark Development Environment (Local, Databricks, Cloud)
- **Case Study:** Analyzing web server logs for common errors and traffic patterns using RDDs.

Module 2: Spark Core: Advanced Concepts

- Spark Internals: DAGScheduler, TaskScheduler, Memory Management
- DataFrames and Datasets: Unified API for Structured Data
- Optimizing DataFrames: Catalyst Optimizer, Tungsten
- Wide vs. Narrow Transformations and Shuffles
- Partitioning and Caching Strategies for Performance
- **Case Study:** Optimizing a large-scale ETL job for a retail company to improve data loading times by 50%.

Module 3: Spark SQL: Structured Data Analysis

- Working with Spark SQL and DataFrames
- Loading and Saving Data: Parquet, ORC, JSON, CSV
- Advanced SQL Queries: Joins, Aggregations, Window Functions
- User-Defined Functions (UDFs) and User-Defined Aggregate Functions (UDAFs)
- Integrating Spark SQL with External Data Sources (Hive, JDBC, Cloud Storage)
- **Case Study:** Building a data warehouse layer for a financial institution, enabling complex SQL queries for risk analysis.

Module 4: Spark Structured Streaming: Real-time Data Processing

- Introduction to Stream Processing and Micro-batching
- Structured Streaming Architecture and Concepts
- Sources and Sinks: Kafka, File Stream, Socket Stream
- Stateful Operations: Aggregations, Watermarking, Joins
- Handling Late Data and Fault Tolerance in Streaming
- **Case Study:** Developing a real-time fraud detection system for an e-commerce platform by analyzing transaction streams.

Module 5: Spark MLlib: Machine Learning at Scale

- Machine Learning Fundamentals and Workflow
- ML Pipelines: Transformers, Estimators, Models
- Feature Engineering and Selection with Spark MLlib
- Supervised Learning: Regression, Classification (Linear Regression, Logistic Regression, Decision Trees)
- Unsupervised Learning: Clustering (K-Means)

- **Case Study:** Building a customer churn prediction model for a telecommunications company using historical customer data.

Module 6: Advanced Spark MLlib Techniques

- Collaborative Filtering for Recommendation Systems
- Model Evaluation and Hyperparameter Tuning
- Model Persistence and Deployment
- Integrating MLlib with other Data Science tools
- Scalability challenges in ML and Spark's solutions
- **Case Study:** Implementing a personalized recommendation engine for a streaming service to suggest movies and shows to users.

Module 7: Performance Tuning & Optimization

- Understanding Spark UI and Metrics
- Debugging Spark Applications: Common Errors and Solutions
- Shuffle Optimization Techniques
- Memory Management and Garbage Collection in Spark
- Resource Allocation and Cluster Configuration Best Practices
- **Case Study:** Diagnosing and resolving performance bottlenecks in a slow-running Spark batch job for a large dataset.

Module 8: Integrating Spark with Data Lakes & Cloud

- Working with Apache HDFS and Cloud Storage (S3, ADLS, GCS)
- Delta Lake: ACID Transactions on Data Lakes
- Integrating Spark with Kafka for message queuing
- Using Spark with NoSQL Databases (Cassandra, MongoDB)
- Deploying Spark on Cloud Platforms (AWS EMR, Azure Databricks, Google Cloud Dataproc)
- **Case Study:** Migrating an on-premise data processing pipeline to a cloud-based Spark environment using Delta Lake for data reliability.

Module 9: Advanced Data Engineering with Spark

- Building Robust ETL Pipelines: Data Ingestion, Transformation, Loading
- Schema Evolution and Data Governance
- Data Quality and Validation Techniques
- Orchestrating Spark Jobs with Airflow or Oozie
- Designing and Implementing Data Warehousing Solutions
- **Case Study:** Developing a robust ETL process to consolidate data from various sources into a unified data lake for a healthcare provider.

Module 10: Spark GraphX for Graph Processing

- Introduction to Graph Theory and Graph Processing
- GraphX API: Vertices, Edges, Property Graphs
- Graph Algorithms: PageRank, Connected Components
- Use Cases for Graph Analysis
- Building Graph-Based Applications with Spark
- **Case Study:** Analyzing social network connections to identify influential users or communities.

Module 11: Security and Monitoring in Spark

- Authentication and Authorization in Spark Clusters
- Data Encryption at Rest and In Transit
- Auditing and Logging for Compliance
- Monitoring Spark Applications with Prometheus, Grafana
- Best Practices for Securing Big Data Environments
- **Case Study:** Implementing security measures for a Spark cluster processing sensitive customer data to meet regulatory requirements.

Module 12: Advanced Deployment and Operations

- Spark on Kubernetes: Containerized Deployments
- Spark on YARN: Advanced Configurations
- Autoscaling Spark Clusters
- Backup and Disaster Recovery Strategies
- CI/CD for Spark Applications
- **Case Study:** Automating the deployment and scaling of a Spark application using Kubernetes for a continuous data pipeline.

Module 13: Best Practices & Future Trends

- Code Optimization Best Practices for Spark
- Performance Benchmarking and Capacity Planning
- Emerging Trends: Spark 3.x Features, Photon, Data Lakehouse
- Real-world Spark Use Cases and Architectures
- Career Paths and Certification Opportunities
- **Case Study:** Reviewing the architecture of a successful large-scale Spark deployment by a leading tech company.

Module 14: Capstone Project

- Participants will work on a comprehensive project applying learned concepts.
- Project options include building an end-to-end data pipeline, developing a machine learning application, or optimizing an existing Spark job.
- Mentored guidance and peer review.
- Presentation of final solutions.
- **Case Study:** Building a complete data pipeline from data ingestion to advanced analytics and machine learning model deployment for a chosen industry.

Module 15: Workshop & Q&A

- Hands-on workshop to reinforce key concepts.
- Troubleshooting clinic: bring your Spark challenges.
- Open discussion and Q&A with instructors.
- Advanced tips and tricks for Spark development.
- Resource sharing for continued learning.
- **Case Study:** Collaborative problem-solving session on a complex Spark performance issue encountered in a real-world scenario.

Training Methodology

This course adopts an **interactive and hands-on learning approach** to maximize knowledge retention and practical skill development. The methodology includes:

- **Instructor-Led Sessions:** Expert-led lectures with clear explanations of complex concepts.
- **Live Coding Demonstrations:** Practical examples showcasing real-world Spark implementations.
- **Hands-on Labs & Exercises:** Extensive practical assignments to solidify understanding and build proficiency.
- **Case Studies:** Analysis and discussion of industry-specific use cases to illustrate practical applications.
- **Group Discussions & Collaborative Problem-Solving:** Fostering peer learning and diverse perspectives.
- **Q&A Sessions:** Dedicated time for addressing participant queries and clarifying doubts.
- **Capstone Project:** A comprehensive project to apply all learned skills in a real-world scenario.
- **Resource Materials:** Access to comprehensive slides, code repositories, and supplementary readings.

Register as a group from 3 participants for a Discount

Send us an email: info@fineskilltrainingcenter.org or call *+254769199797*

Certification

Upon successful completion of this training, participants will be issued with a globally- recognized certificate.

Tailor-Made Course

We also offer tailor-made courses based on your needs.

Key Notes

a. Upcoming Scheduled Sessions

Start Date	End Date	Location	Price
23 Nov 2026	04 Dec 2026	Online	$2,000
23 Nov 2026	04 Dec 2026	Physical	$3,000
23 Nov 2026	04 Dec 2026	Physical	$6,000
23 Nov 2026	04 Dec 2026	Physical	$6,000
23 Nov 2026	04 Dec 2026	Physical	$12,000
23 Nov 2026	04 Dec 2026	Physical	$9,000

Start Date	End Date	Location	Price
23 Nov 2026	04 Dec 2026	Physical	$12,000
23 Nov 2026	04 Dec 2026	Physical	$12,000

Ready to enroll or request a customized program? Book online or contact us:

Register Online Now

Email: info@fineskilltrainingcenter.com | Website: www.fineskilltrainingcenter.com