Certified Spark and Scala Course – Curriculum

The Certified Spark and Scala course by DataFlair is a perfect blend of in-depth theoretical knowledge and strong practical skills via implementation of real life projects to give you a headstart and enable you to bag top Big Data jobs in the industry.

Course duration: 24+10 Hours

Module 1: Diving into Scala

This module introduces you to the rudiments of Scala, how to get it up and running, its functions and procedures, and different operations with APIs for those.

- What is Scala
- Setup and configuration of Scala
- Developing and running basic Scala Programs
- Scala operations
- Functions and procedures in Scala
- Different Scala APIs for common operations
- Loops and collections- Array, Map, Lists, Tuples
- Pattern matching for advanced operations
- Eclipse with Scala

Module 2: Object-Oriented and Functional Programming

We then introduce you to the concepts of object-oriented programming and their constructs- nested classes, constructors, and more. Finally, we will take a good look at call-by-name and call-by-value.

- Introduction to object-oriented programming
- Different OOPS concepts
- Constructor, getter, setter, singleton, overloading, and overriding
- Nested Classes and visibility rules
- Functional structures
- Functional programming constructs
- Call by Name, Call by Value

Module 3: Big Data and the need for Spark

This module deals with the challenges to older Big Data solutions and introduces you to the alternatives. We discuss the limitations of each of those.

- Introduction to Big Data
- Challenges to old Big Data solutions
- Batch vs Real-time vs in-Memory processing

- MapReduce and its limitations
- Apache Storm and its limitations
- Need for a general purpose solution - Apache Spark

Module 4: Diving deep into Apache Spark

Next, we discuss Spark and its components. This is much about its features and design principles.

- What is Apache Spark?
- Components of Spark architecture
- Apache Spark design principles
- Spark features and characteristics
- Apache Spark ecosystem components and their insights

Module 5: Deploying Spark in local mode

After finishing this module, you will be comfortable with Spark and its structures. This module deals with setting it up on your machine in different modes. This will also guide you with issues you will likely encounter.

- Setting up the Spark Environment
- Installing and configuring prerequisites
- Installing Apache Spark in local mode
- Working with Spark in local mode
- Troubleshooting encountered problems in Spark

Module 6: Deploying Spark in different modes

Time to dig deeper into Spark! This module tells you about many more modes to install Spark in.

- Installing Spark in standalone mode
- Installing Spark in YARN mode
- Installing & configuring Spark on a real multi-node cluster
- Playing with Spark in cluster mode
- Best practices for Spark deployment

Module 7: Demystifying Apache Spark

More than halfway through the course now, we begin to demystify Spark. We take you right to the Spark shell so you can expect a full hands-on experience.

- Playing with the Spark shell
- Executing Scala and Java statements in the shell
- Understanding the Spark context and driver
- Reading data from the local filesystem

- Integrating Spark with HDFS
- Caching the data in memory for further use
- Distributed persistence
- Testing and troubleshooting

Module 8: Basic abstraction RDDs

This module teaches you all about RDDs in Spark. You will learn about the operations, transformations, and fault tolerance.

- What is an RDD in Spark
- How do RDDs make Spark a feature-rich framework
- Transformations in Apache Spark RDDs
- Spark RDD action and persistence
- Spark Lazy Operations - Transformation and Caching
- Fault tolerance in Spark
- Loading data and creating RDD in Spark
- Persist RDD in memory or disk
- Pair operations and key-value in Spark
- Spark integration with Hadoop
- Apache Spark practicals and workshops

Module 9: Spark Streaming

We move on to Spark Streaming. In this module, we talk of its need, operations, and execution flow. Finally, we discuss ways to optimize performance.

- The need for stream analytics
- Comparison with Storm and S4
- Real-time data processing using Spark streaming
- Fault tolerance and check-pointing
- Stateful stream processing
- DStream and window operations
- Spark Stream execution flow
- Connection to various source systems
- Performance optimizations in Spark

Module 10: Spark SQL

This module familiarizes you with Spark SQL and explains its features, components, and techniques. We also talk about Data-frames and Hive queries.

- What is Spark SQL
- Apache Spark SQL features and data flow
- Spark SQL architecture and components
- Hive and Spark SQL together

- Play with Data-frames and data states
- Data loading techniques in Spark
- Hive queries through Spark
- Various Spark SQL DDL and DML operations
- Performance tuning in Spark

Module 11: Spark MLlib and Spark GraphX

Before we move on to the final project of this course, let's learn about machine learning and its libraries with Spark. Algorithms like clustering and classification form a perfect fit for this purpose.

- Why Machine Learning is needed
- What is Spark Machine Learning
- Various Spark ML libraries
- Algorithms for clustering, statistical analytics, classification etc.
- What is GraphX
- The need for different graph processing engines
- Graph handling using Apache Spark

Module 12: Real Life Spark Project

We conclude this course with a live Spark project to prepare you for the industry. Here, we make use of various constructs of Scala and Spark to solve real-world problems in Big Data Analytics.

- *Set Top Box Data Analysis* - Learn to analyze Set-Top-Box data and generate insights about smart tv usage patterns. Analyze set top box media data and generate patterns of channel navigation and VOD. This Spark Project includes details about users' activities tuning a channel or duration, browsing for videos, or purchasing videos using VOD.
- *Twitter Trends Analysis* - Collect Twitter data in real-time and find out current trends in various categories. In this Apache Spark project, you will collect live Twitter streams and analyze them using Spark Streaming to generate insights like finding current trends in Politics, Finance, Entertainment, and such.
- *Titanic Data Analysis* - Titanic was one of the most colossal disasters in the history of mankind, and it happened because of both natural events and human mistakes. The objective of this Spark project is to analyze multiple Titanic data sets to generate essential insights pertaining to age, gender, survived, class, and embarked.
- *E-commerce Reviews Analysis* - Learn to analyze e-commerce review data and generate various insights of products. Companies use these reports and patterns to understand the sentiments of users about their products. E-commerce reviews are made of fields like product-id, star-rating, reviews, timestamp, and reviewer-id.
- *YouTube Data Analysis* - Yearn to analyze YouTube Data and generate insights like the 10 topmost videos in various categories, user demographics, no. of views, ratings and such. The data holds fields like id, age, category, length, views, ratings, and comments.
- And so many more projects of retail, telecom, media, etc..