Course Curriculum: Your 10 Module Learning Plan

Apache Spark and Scala

About Edureka

Edureka is a leading e-learning platform providing live instructor-led interactive online training. We cater to professionals and students across the globe in categories like *Big Data & Hadoop, Business Analytics, NoSQL Databases, Java & Mobile Technologies, System Engineering, Project Management and Programming.*

We have an easy and affordable learning solution that is accessible to millions of learners. With our students spread across countries like the US, India, UK, Canada, Singapore, Australia, Middle East, Brazil and many others, we have built a community of over 1 million learners across the globe.

About The Course

Apache Spark Certification Training Course is designed to provide knowledge and skills to become a successful Big Data Developer.

You will understand basics of Big Data and Hadoop. You will learn how Spark enables in-memory data processing and runs much faster than Hadoop MapReduce. You will also learn about RDDs, different APIs, which Spark offers such as Spark Streaming, MLlib, Clustering, and Spark SQL. This Edureka course is an integral part of a Big Data Developer's Career path. It will also encompass the fundamental concepts like data capturing using Flume, data loading using Sqoop, Kafka cluster, Kafka API.

This course is designed to provide knowledge and skills to become a successful Spark and Hadoop Developer and would help to clear the CCA Spark and Hadoop Developer (CCA175) Examination.

Module 1
Introduction to Scala for Apache Spark

Learning Objectives - In this module, you will understand the basics of Scala that are required for programming Spark applications. You can learn about the basic constructs of Scala such as variable types, control structures, collections, and more.

Topics

- ✓ What is Scala?
- ✓ Why Scala for Spark?
- ✓ Scala in other frameworks
- ✓ Introduction to Scala REPL
- ✓ Basic Scala operations
- ✓ Variable Types in Scala
- ✓ Control Structures in Scala
- ✓ Foreach loop, Functions and Procedures
- ✓ Collections in Scala- Array
- ✓ ArrayBuffer, Map, Tuples, Lists, and more

Hands On:

- ✓ Scala REPL Detailed Demo

Module 2
OOPS and Functional Programming in Scala

Learning Objectives - In this module, you will learn about object oriented programming and functional programming techniques in Scala.

Topics

- ✓ Class in Scala
- ✓ Getters and Setters
- ✓ Custom Getters and Setters
- ✓ Properties with only Getters
- ✓ Auxiliary Constructor
- ✓ Primary Constructor
- ✓ Singletons
- ✓ Extending a Class
- ✓ Overriding Methods
- ✓ Traits as Interfaces
- ✓ Layered Traits
- ✓ Functional Programming
- ✓ Higher Order Functions
- ✓ Anonymous Functions and more.

Hands On:

- ✓ Case Class Demo
- ✓ Layered Traits

Module 3
Introduction to Big Data

Learning Objectives - In this module, you will understand Big Data, the limitations of the existing solutions for Big Data problem, how Hadoop solves the Big Data problem, Hadoop ecosystem components, Hadoop Architecture, HDFS, Rack Awareness and Replication. You will learn about the Hadoop Cluster Architecture, important configuration files in a Hadoop Cluster. You will get an overview of Apache Sqoop and how it is used in importing and exporting tables from RDBMS to HDFS & vice versa.

Topics

- ✓ What is Big Data?
- ✓ Big Data Customer Scenarios
- ✓ Limitations and Solutions of Existing Data Analytics Architecture with Uber Use Case
- ✓ How Hadoop Solves the Big Data Problem
- ✓ What is Hadoop?
- ✓ Hadoop's Key Characteristics
- ✓ Hadoop Ecosystem and HDFS
- ✓ Hadoop Core Components
- ✓ Rack Awareness and Block Replication
- ✓ Edureka's VM Tour
- ✓ YARN and Its Advantage
- ✓ Hadoop Cluster and Its Architecture
- ✓ Hadoop: Different Cluster Modes
- ✓ Data Loading using Sqoop

Hands On:

- ✓ A Tour of Edureka's Hadoop & Spark VM
- ✓ Basic Hadoop Commands
- ✓ Importing and Exporting Data Using Sqoop

Module 4
Apache Spark Framework

Learning Objectives - In this module, you will understand different frameworks available for Big Data Analytics and the module also includes a first-hand introduction to Spark, demo on Building and Running a Spark Application and Web UI.

Topics

- ✓ Big Data Analytics with Batch & Real-Time Processing
- ✓ Why Spark is Needed?
- ✓ What is Spark?
- ✓ How Spark Differs from Its Competitors?
- ✓ Spark at eBay
- ✓ Spark's Place in Hadoop Ecosystem
- ✓ Spark Components & It's Architecture
- ✓ Running Programs on Scala IDE & Spark Shell
- ✓ Spark Web UI
- ✓ Configuring Spark Properties

Hands On:

- ✓ Building and Running Spark Application
- ✓ Spark Application Web UI
- ✓ Configuring Spark Properties

Module 5
Playing with RDDs

Learning Objectives - In this module, you will learn one of the fundamental building blocks of Spark - RDDs and related manipulations for implementing business logics (Transformations, Actions and Functions performed on RDD). You will learn about Spark applications, how it is developed and configuring Spark properties.

Topics:

- ✓ Challenges in Existing Computing Methods
- ✓ Probable Solution & How RDD Solves the Problem
- ✓ What is RDD, It's Functions, Transformations & Actions?
- ✓ Data Loading and Saving Through RDDs
- ✓ Key-Value Pair RDDs and Other Pair RDDs
- ✓ RDD Lineage
- ✓ RDD Persistence
- ✓ WordCount Program Using RDD Concepts
- ✓ RDD Partitioning & How It Helps Achieve Parallelization

Hands On:

- ✓ Loading data in RDDs
- ✓ Saving data through RDDs
- ✓ RDD Transformations
- ✓ RDD Actions and Functions

Module 6
DataFrames and Spark SQL

Learning Objectives - In this module, you will learn about Spark SQL which is used to process structured data with SQL queries. You will learn about data-frames and datasets in Spark SQL and perform SQL operations on data-frames.

Topics

- ✓ Need for Spark SQL
- ✓ What is Spark SQL?
- ✓ Spark SQL Architecture
- ✓ SQL Context in Spark SQL
- ✓ Data Frames & Datasets
- ✓ Interoperating with RDDs
- ✓ JSON and Parquet File Formats
- ✓ Loading Data through Different Sources

Hands On:

- ✓ Spark SQL – Creating data frames
- ✓ Loading and transforming data through different sources
- ✓ Stock Market Analysis

Module 7
Machine Learning using Spark MLlib

Learning Objectives – In this module you will learn about what is the need for machine learning, types of ML concepts, clustering and MLlib (i.e. Spark's machine learning library), various algorithms supported by MLlib and implement K-Means Clustering.

Topics:

- ✓ What is Machine Learning?
- ✓ Where is Machine Learning Used?
- ✓ Different Types of Machine Learning Techniques
- ✓ Face Detection: USE CASE
- ✓ Understanding MLlib
- ✓ Features of MLlib and MLlib Tools
- ✓ Various ML algorithms supported by MLlib
- ✓ K-Means Clustering & How It Works with MLlib
- ✓ Analysis on US Election Data: K-Means MLlib USE CASE

Hands On:

- ✓ Machine Learning MLlib
- ✓ K- Means Clustering

Module 8
Understanding Apache Kafka and Kafka Cluster

Learning Objectives - In this module, you will understand Kafka and Kafka Architecture. Afterwards you will go through the details of Kafka Cluster and you will also learn how to configure different types of Kafka Cluster.

Topics:

- ✓ Need for Kafka
- ✓ What is Kafka?
- ✓ Core Concepts of Kafka
- ✓ Kafka Architecture
- ✓ Where is Kafka Used?
- ✓ Understanding the Components of Kafka Cluster
- ✓ Configuring Kafka Cluster
- ✓ Producer and Consumer

Hands On:

- ✓ Configuring Single Node Single Broker Cluster
- ✓ Configuring Single Node Multi Broker Cluster

Module 9
Capturing Data with Apache Flume and Integration with Kafka

Learning Objectives – In this module you will get an introduction to Apache Flume and its basic architecture and how it is integrated with Apache Kafka for event processing.

Topics:

- ✓ Need of Apache Flume
- ✓ What is Apache Flume?
- ✓ Basic Flume Architecture
- ✓ Flume Sources
- ✓ Flume Sinks
- ✓ Flume Channels
- ✓ Flume Configuration
- ✓ Integrating Apache Flume and Apache Kafka

Hands On:

- ✓ Flume Commands
- ✓ Setting up Flume Agent
- ✓ Streaming Twitter Data into HDFS

Module 10
Apache Spark Streaming

Learning Objectives – In this module you will get an opportunity to work on Spark streaming which is used to build scalable fault-tolerant streaming applications. You will learn about DStreams and various Transformations performed on it. You will get to know about main streaming operators, Sliding Window Operators and Stateful Operators.

Topics:

- ✓ Drawbacks in Existing Computing Methods
- ✓ Why Streaming is Necessary?
- ✓ What is Spark Streaming?
- ✓ Spark Streaming Features
- ✓ Spark Streaming Workflow
- ✓ How Uber Uses Streaming Data
- ✓ Streaming Context & DStreams
- ✓ Transformations on DStreams
- ✓ WordCount Program using Spark Streaming
- ✓ Describe Windowed Operators and Why it is Useful
- ✓ Important Windowed Operators
- ✓ Slice, Window and ReduceByWindow Operators
- ✓ Stateful Operators
- ✓ Perform Twitter Sentimental Analysis Using Spark Streaming

Hands On:

- ✓ Creating DStreams
- ✓ Transactions and Actions performed on DStreams.
- ✓ Output Operations in DStreams
- ✓ Sliding Window Operations
- ✓ Stateful Operations
- ✓ Twitter Sentimental Analysis

Certification Project

Project #1: US Election
Industry: Government

Technologies Used:
1. HDFS (for storage)
2. Spark SQL (for transformation)
3. Spark MLlib (for machine learning)
4. Zeppelin (for visualization)

PROBLEM STATEMENT:
In the US Primary Election 2016, Hillary Clinton was nominated over Bernie Sanders from Democrats and on the other hand, Donald Trump was nominated from Republican Party to contest for the presidential position. As an analyst, you have been tasked to understand different factors that led to the winning of Hillary Clinton and Donald Trump in the primary elections based on demographic features to plan their next initiatives and campaigns.

Project #2: Design a system to replay the real time replay of transactions in HDFS using Spark.

Technologies Used:
1. Spark Streaming
2. Kafka (for messaging)
3. HDFS (for storage)
4. Core Spark API (for aggregation)

Project #3: Instant Cabs
Industry: Transportation

Technologies Used:
1. HDFS (for storage)
2. Spark SQL (for transformation)
3. Spark MLlib (for machine learning)
4. Zeppelin (for visualization)

PROBLEM STATEMENT:
A US cab service start-up (i.e. Instant cabs) wants to meet the demands in an optimum manner and maximize the profit. Thus, they hired you as a data analyst to interpret the available Uber's data set and find out the beehive customer pick-up points & peak hours for meeting the demand in a profitable manner.

[Apache Spark and Scala]