Course Outline

Hortonworks HDP Developer: Apache Spark 2.3

Course Summary

Description

This course introduces the Apache Spark distributed computing engine, and is suitable for developers, data analysts, architects, technical managers, and anyone who needs to use Spark in a hands-on manner. It is based on the Spark 2.x release. The course provides a solid technical introduction to the Spark architecture and how Spark works. It covers the basic building blocks of Spark (e.g. RDDs and the distributed compute engine), as well as higher-level constructs that provide a simpler and more capable interface. It includes in-depth coverage of Spark SQL, DataFrames, and DataSets, which are now the preferred programming API. This includes exploring possible performance issues and strategies for optimization. The course also covers more advanced capabilities such as the use of Spark Streaming to process streaming data, and integrating with the Kafka server.

Topics

- Scala Ramp Up, Introduction to Spark
- RDDs and Spark Architecture, Spark SQL, DataFrames and DataSets
- Shuffling, Transformations and Performance, Performance Tuning
- Creating Standalone Applications and Spark Streaming

Audience

This course is designed for software engineers that are looking to develop in-memory applications for time sensitive and highly iterative applications in an Enterprise HDP environment.

Prerequisites

Before taking this course, students should be familiar with programming principles and have previous experience in software development using Scala. Previous experience with data streaming, SQL, and HDP is also helpful, but not required.

Duration

Four days

Course Outline

Hortonworks HDP Developer: Apache Spark 2.3

Course Outline

I. Scala Ramp Up, Introduction to Spark

A. Scala Introduction
B. Working with: Variables, Data Types, and Control Flow
C. The Scala Interpreter
D. Collections and their Standard Methods (e.g. map())
E. Working with: Functions, Methods, and Function Literals
F. Define the Following as they Relate to Scale: Class, Object, and Case Class
G. Overview, Motivations, Spark Systems
H. Spark Ecosystem
I. Spark vs. Hadoop
J. Acquiring and Installing Spark
K. The Spark Shell, SparkContext

Labs

- Setting Up the Lab Environment
- Starting the Scala Interpreter
- A First Look at Spark
- A First Look at the Spark Shell

II. RDDs and Spark Architecture, Spark SQL, DataFrames and DataSets

A. RDD Concepts, Lifecycle, Lazy Evaluation
B. RDD Partitioning and Transformations
C. Working with RDDs Including: Creating and Transforming
D. An Overview of RDDs
E. SparkSession, Loading/Saving Data, Data Formats
F. Introducing DataFrames and DataSets
G. Identify Supported Data Formats
H. Working with the DataFrame (untyped) Query DSL
I. SQL-based Queries
J. Working with the DataSet (typed) API
K. Mapping and Splitting
L. DataSets vs. DataFrames vs. RDDs

Labs

- RDD Basics
- Operations on Multiple RDDs
- Data Formats
- Spark SQL Basics
- DataFrame Transformations
- The DataSet Typed API
- Splitting Up Data

III. Shuffling, Transformations and Performance, Performance Tuning

A. Working with: Grouping, Reducing, Joining
B. Shuffling, Narrow vs. Wide Dependencies, and Performance Implications
C. Exploring the Catalyst Query Optimizer
D. The Tungsten Optimizer
E. Discuss Caching, Including: Concepts, Storage Type, Guidelines
F. Minimizing Shuffling for Increased Performance
G. Using Broadcast Variables and Accumulators
H. General Performance Guidelines

Labs

- Exploring Group Shuffling
- Seeing Catalyst at Work
- Seeing Tungsten at Work
- Working with Caching, Joins, Shuffles, Broadcasts, Accumulators
- Broadcast General Guidelines

IV. Creating Standalone Applications and Spark Streaming

A. Core API, SparkSession.Builder
B. Configuring and Creating a SparkSession
C. Building and Running Applications
D. Application Lifecycle (Driver, Executors, and Tasks)
E. Cluster Managers (Standalone, YARN, Mesos)
F. Logging and Debugging
G. Introduction and Streaming Basics
H. Spark Streaming (Spark 1.0+)
I. Structured Streaming (Spark 2+)
J. Consuming Kafka Data

Labs

- Spark Job Submission
- Additional Spark Capabilities
- Spark Streaming
- Spark Structured Streaming
- Spark Structured Streaming with Kafka