Developer

HDP Developer: Apache Spark Using Python

Overview

This course is designed for developers who need to create applications to analyze Big Data stored in Apache Hadoop using Spark. Topics include: Hadoop, YARN, HDFS, using Spark for interactive data exploration, building and deploying Spark applications, optimization of applications, creating Spark pipelines with multiple libraries, working with different filetypes, building data frames, exploring the Spark SQL API, using Spark Streaming and an introduction to Spark MLlib.

Duration

3 days

Target Audience

Software engineers that are looking to develop time sensitive applications for Hadoop.

Course Objectives

- Describe Hadoop, HDFS, YARN, and uses cases for Hadoop
- Describe Spark and Spark specific use cases
- Understand the HDFS architecture
- Use the HDFS commands to insert and retrieve data
- Explain the differences between Spark and MapReduce
- Explore data interactively through the spark shell utility
- Explain the RDD concept
- Understand concepts of functional programming
- Use the Python or Scala Spark APIs
- Create all types of RDDs: Pair, Double, and Generic
- Use RDD type-specific functions
- Explain interaction of components of a Spark Application
- Explain the creation of the DAG schedule
- Build and package Spark applications
- Use application configuration items
- Deploy applications to the cluster using YARN
- Use data caching to increase performance of applications
- Implement advanced features of spark
- Learn general application optimization guidelines/tips
- Create applications using the Spark SQL library
- Create/transform data using dataframes
- Read, use, and save to different Hadoop file formats
- Understand the concepts of Spark Streaming
- Create a streaming application
- Use Spark MLlib to gain insights from data

Hands-On Labs

- Create a Spark “Hello World” word count application
- Use HDFS commands to add and remove files and folders
- Use advanced RDD programming to perform sort, join, pattern matching and regex tasks
- Explore partitioning and the Spark UI
- Increase performance using data caching
- Checkpoint iterative applications
- Build/package a Spark application using Maven
- Use a broadcast variable to efficiently join a small dataset to a massive dataset
- Use an accumulator for reporting data quality issues
- Create a data frame and perform analysis
- Load/transform/store data using Spark with Hive tables
- Create a point-in-time spark stream application
- Create a spark stream application using window functions
- Create a Spark MLlib application using K-Means

Prerequisites

Students should be familiar with programming principles and have previous experience in software development. SQL knowledge is helpful. No prior Hadoop experience required, but is very helpful.

Format

50% Lecture/Discussion

50% Hands-on Labs

Certification

Hortonworks offers a comprehensive certification program that identifies you as an expert in Apache Hadoop. Visit *hortonworks.com/training/certification* for more information.

Hortonworks University

Hortonworks University is your expert source for Apache Hadoop training and certification. Public and private on-site courses are available for developers, administrators, data analysts and other IT professionals involved in implementing big data solutions. Classes combine presentation material with industry-leading hands-on labs that fully prepare students for real-world Hadoop scenarios.

About Hortonworks
Hortonworks develops, distributes and supports the only 100 percent open source distribution of Apache Hadoop explicitly architected, built and tested for enterprise-grade deployments.

US: 1.855.846.7866
International: 1.408.916.4121
www.hortonworks.com

5470 Great America Parkway
Santa Clara, CA 95054 USA