

HDP Developer: Apache Spark Using Python H0MC2X

HPE course number	H0MC2X
Course length	3 days
Delivery mode	ILT
Contact us	View now
View related courses	View now

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 file types, building data frames, exploring the Spark SQL API, using Spark Streaming, and an introduction to Spark MLlib.

Why HPE Education Services?

- IDC MarketScape leader 4 years running for IT education and training*
- Recognized by IDC for leading with global coverage, unmatched technical expertise, and targeted education consulting services*
- Key partnerships with industry leaders OpenStack®, VMware®, Linux®, Microsoft®, ITIL, PMI, CSA, and (ISC)²
- Complete continuum of training delivery options—self-paced eLearning, custom education consulting, traditional classroom, video on-demand instruction, live virtual instructor-led with hands-on lab, dedicated onsite training
- Simplified purchase option with HPE Training Credits

Audience

Software engineers that are already familiar with Python looking to develop time sensitive applications for Hadoop using Python.

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.

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 data frames
- 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

Learn more at
hpe.com/ww/learnbigdata

Follow us:



© Copyright 2016 Hewlett Packard Enterprise Development LP. The information contained herein is subject to change without notice. The only warranties for Hewlett Packard Enterprise products and services are set forth in the express warranty statements accompanying such products and services. Nothing herein should be construed as constituting an additional warranty. Hewlett Packard Enterprise shall not be liable for technical or editorial errors or omissions contained herein.

Microsoft is either a registered trademark or trademark of Microsoft Corporation in the United States and/or other countries. The OpenStack Word Mark is either a registered trademark/service mark or trademark/service mark of the OpenStack Foundation, in the United States and other countries and is used with the OpenStack Foundation's permission. We are not affiliated with, endorsed or sponsored by the OpenStack Foundation or the OpenStack community. Pivotal and Cloud Foundry are trademarks and/or registered trademarks of Pivotal Software, Inc. in the United States and/or other countries. Linux is the registered trademark of Linus Torvalds in the U.S. and other countries. VMware is a registered trademark or trademark of VMware, Inc. in the United States and/or other jurisdictions. All other third-party trademark(s) is/are property of their respective owner(s).

c05105581, October 2016, Rev. 2