Course Data Sheet

Data Science for the Hortonworks Data Platform

Overview

This 3-day, hands-on training course introduces the fundamentals of Data Science and how to apply these concepts in Hadoop using machine learning, Mahout, Pig, Python and various machine learning libraries like SciPy and Scikit-Learn.

Description

Data Science for the Hortonworks Data Platform covers data science principles and techniques through lecture and hands-on experience. During this 3-day class, students will learn the processes and practice of data science, including machine learning and natural language processing. Students will also learn the tools and programming languages used by data scientists, including Python, IPython, Mahout, Pig, NumPy, pandas, SciPy, Scikit-Learn and Spark MLlib.

Duration

3 days

Price

For information about pricing, email us at sales-training@hortonworks.com

Prerequisites

Students must have:

- Experience with at least one programming or scripting language
- Knowledge of statistics and/or mathematics
- A basic understanding of big data and Hadoop principles

Students new to Hadoop are encouraged to attend the Hadoop Essentials course.

Target Audience

Architects, software developers, analysts and data scientists who need to understand how to apply data science and machine learning on Hadoop will get the most from this course.

Format

50% Instructor-led lecture/discussion
50% hands-on labs.

Course Objectives

At the completion of the course students should be able to:

- Recognize use cases for data science
- Describe the architecture of Hadoop and YARN
- Explain the differences between supervised and unsupervised learning
- List the six machine learning tasks
- Recognize use cases for clustering, outlier detection, affinity analysis, classification, regression, and recommendation
- Use Mahout to run a machine learning algorithm on Hadoop
- Write Pig scripts to transform data on Hadoop
- Use Pig to prepare data for a machine learning algorithm
- Write a Python script
- Use NumPy to analyze big data
- Use the data structure classes in the pandas library
- Write a Python script that invokes a SciPy machine learning algorithm
- Explain the options for running Python code on a Hadoop cluster
- Write a Pig User Defined Function in Python
- Use Pig streaming on Hadoop with a Python script
- Write a Python script that invokes a scikit-learn machine learning algorithm
- Use the k-nearest neighbor algorithm to predict values based on a data set
- Run the k-means clustering algorithm on a distributed data set on Hadoop
- Describe use cases for Natural Language Processing (NLP)
- Run an NLP algorithm on a Hadoop cluster
- Run machine learning algorithms on Hadoop using Spark MLlib

Agenda

Day 1

- Unit 1: Using Hadoop for Data Science
- Unit 2: Hadoop Architecture
- Unit 3: Machine Learning

- Unit 4: Introduction to Pig

Day 2
- Unit 5: Python Programming
- Unit 6: Analyzing Data with Python
- Unit 7: Running Python on Hadoop

Day 3
- Unit 8: Implementing Machine Learning
- Unit 9: Natural Language Processing
- Unit 10: Using Spark MLlib

Hands-on Labs
Students will complete the following hands-on labs using their own 7-node Hadoop cluster (HDP 2.1) and IPython Notebook:

- Setting Up a Development Environment
- Using HDFS Commands
- Using Mahout for Machine Learning
- Getting Started with Pig
- Exploring Data with Pig
- Using the IPython Notebook
- Data Analysis with Python
- Interpolating Data Points
- Define a Pig UDF in Python
- Streaming Python with Pig
- K-Nearest Neighbor
- K-Means Clustering
- Natural Language Processing
- Running Data Science Algorithms using Spark MLlib

Additional Information

- For availability of individual seats in our open enrollment classes please visit us at www.hortonworks.com/training
- Please contact sales-training@hortonworks.com to discuss your specific training needs