Hortonworks UNIVERSITY

Data Analyst

HDP Analyst: Data Science

Overview

This course Provides instruction on the processes and practice of data science, including machine learning and natural language processing. Included are: tools and programming languages (Python, IPython, Mahout, Pig, NumPy, pandas, SciPy, Scikit-learn), the Natural Language Toolkit (NLTK), and Spark MLlib.

Duration

3 days

Target Audience

Architects, software developers, analysts and data scientists who need to apply data science and machine learning on Hadoop.

Course Objectives

- Recognize use cases for data science
- Describe the architecture of Hadoop and YARN
- Describe supervised and unsupervised learning differences
- List the six machine learning tasks
- Use Mahout to run a machine learning algorithm on Hadoop
- Describe the data science life cycle
- Use Pig to transform and prepare data on Hadoop
- 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 SciPy machine learning
- Describe 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 scikit-learn
- Use the k-nearest neighbor algorithm to predict values
- Run a machine learning algorithm on a distributed data set
- Describe use cases for Natural Language Processing (NLP)
- Perform sentence segmentation on a large body of text
- Perform part-of-speech tagging
- Use the Natural Language Toolkit (NLTK)
- Describe the components of a Spark application
- Write a Spark application in Python
- Run machine learning algorithms using Spark MLlib
- Take data science into production

Hands-On Labs

- 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 and K-Means Clustering
- Using NLTK for Natural Language Processing
- Classifying Text using Naive Bayes
- Spark Programming and Spark MLlib
- Spam Classification with MLlib

Prerequisites

Students must have experience with at least one programming or scripting language, knowledge in statistics and/or mathematics, and a basic understanding of big data and Hadoop principles. Students new to Hadoop are encouraged to attend the *HDP Overview: Apache Hadoop Essentials* course.

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.

Hortonworks

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