Hewlett Packard Enterprise

Hortonworks Data Platform Analyst: Data Science H7G69S

Data Science for the Hortonworks Data Platform covers data science principles and techniques through lecture and handson experience. During this three day course, 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, the Natural Language Toolkit (NLTK), and Spark MLlib.

Hortonworks Data Platform Analyst: Data Science	
Price	USD $2,095
Links to local schedules, pricing and registration	**US/Canada** **Mexico/Latin America** **Brazil**
HP course #	H7G69S
Category	Big Data
Duration	3 days

Audience

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

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

Course objectives

At the completion of the course students will 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 training data set
- Run a machine learning algorithm on a distributed data set on Hadoop
- 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) for implement NLP tasks and machine learning algorithms
- Explain the components of a Spark application
- Write a Spark application in Python
- Run machine learning algorithms on Hadoop using Spark MLlib

Benefits to you

- This course will provide in depth explanation on how to perform Hadoop 2.0 application development

Course outline

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: 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
- Using NLTK for Natural Language Processing
- Classifying Text using Naive Bayes
- Spark Programming
- Running Data Science Algorithms using Spark MLlib

Learn more at

hpe.com/us/training/bigdata

H7G69S Ver A.00

c04680809 April 2016

Hewlett Packard Enterprise