SUMMIT TRAINING OFFERING

HORTONWORKS DATA PLATFORM (HDP®) DATA SCIENCE

2 DAYS 

Learn Data Science techniques and best practices leveraging the Hadoop ecosystem and tools in this **2 day** course.

PREREQUISITES

Students must have experience with at least one programming language such as Python, or scripting language, knowledge in statistics and/or mathematics, and a basic understanding of big data and Hadoop principles.

TARGET AUDIENCE

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

FORMAT

50% Lecture/Discussion

50% Demos/Hands-0n Labs

COURSE OBJECTIVES

- 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

Contact

For further information visit www.hortonworks.com

+1 408 675-0983
+1 855 8-HORTON
INTL: +44 (0) 20 3826 1405

LABS

- Describe the architecture of Hadoop and YARN
- Explain the differences between supervised and unsupervised learning
- Recognize use cases for clustering, outlier detection, affinity analysis, classification, regression, and recommendation
- Write Pig scripts to transform data on Hadoop
- Use Pig to prepare data for a machine learning algorithm
- Write a Python script using NumPy, Scipy, Matplotlib, Pandas, and Scikit-learn to analyze big data
- Exercise 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 scriptRun a Hadoop Streaming job
- Understand some key tasks in Natural Language Processing (NLP)
- Run an NLP algorithms on IPython
- Run machine learning algorithms on Hadoop using Spark MLlib

Contact

For further information visit www.hortonworks.com

+1 408 675-0983
+1 855 8-HORTON
INTL: +44 (0) 20 3826 1405