[SCI-221]: HDP Analyst: Data Science Foundations

Length : 3 Days
Delivery Method : Instructor-led (Classroom)

Course 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, Jupyter, Mahout, Hive, NumPy, Pandas, SciPy, Scikit-learn), and Spark MLlib.

Audience Profile

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

Pre-Requisites

Participants 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.

Formats

Lecture/Discussion	50%
Hands-on Labs	50%

Course Outline

Module 1: Introduction to Data Science, Python, Hadoop, and Machine Learning
Lessons

- Define Data Science and Explain What a Data Scientist Does
- Differentiate Between Different Types of Data Roles
- List a Number of Data Science Use Cases
- Present an Overview of Python
- List and Describe Python Programming Components
- Import Python Modules
- Develop Python Code
- List the Python Packages that Comprise the Scientific Python Ecosystem and Explain their Use Cases
- Utilize the Jupyter Notebook
- Demonstrate How to Use NumPy Core Functionality
- Explain the Data Structures in the Ecosystem
- Describe the Components of the Big Data Scientific Stack
- Explain the Benefits of Big Data for Machine Learning

- Describe How Files are Stored in a Distributed Manner
- Give a High-Level Overview of Zeppelin and Spark
- Explain the Different Task Types of Machine Learning
- Explain How Supervised Learning Differs from Unsupervised Learning
- Describe the Modeling Process
- Explain What a Feature and a Target or Label is in Machine Learning

Labs

- Using IPython
- Data Analysis with Python
- Using HDFS Commands
- Introduction to Spark REPLs and Zeppelin
- Using Apache Mahout

Module 2: Working with Spark RDDs, DataFrames and SparkSQL, Visualization in Zeppelin
Lessons

- Explain what an RDD is
- Explain How RDDs are Partitioned
- Create, Manipulate, and Restore RDDs
- Describe what Lazy Evaluation Means and the Two Types of Spark Operations
- Define, Create and Perform Common Functions with Pair RDDs
- Explain what a DataFrame is and How it Differs from an RDD and a Dataset
- Demonstration How to Create and Manipulate DataFrames
- Explain the Benefit of Catalyst Optimizer
- Use SparkSQL to Create Tables
- Demonstrate How to Use Visualization and Collaboration in Zeppelin
- Use Dynamic Forms
- Create an Application to Submit to the Cluster
- Describe Client vs Cluster Submission with YARN
- Submit an Application to the Cluster
- List and Set Important Configurations Items

LABS

- Create and Manipulate RDDs
- Create and Save DataFrames
- Build and Submit Spark Applications

Module 3: Machine Learning Algorithms, Natural Language Processing, and Spark MLlib
Lessons

- Describe common machine learning applications
- List the pros and cons of various algorithms
- Explain what Natural Language Processing is
- Describe common tasks in the field of NLP
- Utilize NLTK

- Explain the Difference Between spark.mllib and spark.ml
- Explain what Pipelines do
- Explain the Feature Engineering Capabilities of Spark MLlib
- Build a Classifier with MLlib

LABS

- Use the Python Natural Language Toolkit (NLTK)
- Classify text using Naïve Bayes
- Compute K-nearest neighbors
- Creating a Spam Classifier with MLlib
- Sentiment Analysis with Spark MLlib