Cloudera Data Scientist Training (DATA-SCI-TRAIN)

COURSE OVERVIEW:

This four-day workshop covers data science and machine learning workflows at scale using Apache Spark 2 and other key components of the Hadoop ecosystem. The workshop emphasizes the use of data science and machine learning methods to address real-world business challenges. Using scenarios and datasets from a fictional technology company, students discover insights to support critical business decisions and develop data products to transform the business. The material is presented through a sequence of brief lectures, interactive demonstrations, extensive hands-on exercises, and discussions. The Apache Spark demonstrations and exercises are conducted in Python (with PySpark) and R (with sparklyr) using the Cloudera Data Science Workbench (CDSW) environment. The workshop is designed for data scientists who currently use Python or R to work with smaller datasets on a single machine and who need to scale up their analyses and machine learning models to large datasets on distributed clusters. Data engineers and developers with some knowledge of data science and machine learning may also find this workshop useful.

WHO WILL BENEFIT FROM THIS COURSE?

The workshop is designed for data scientists who currently use Python or R to work with smaller datasets on a single machine and who need to scale up their analyses and machine learning models to large datasets on distributed clusters. Data engineers and developers with some knowledge of data science and machine learning may also find this workshop useful.

PREREQUISITES:

Workshop participants should have a basic understanding of Python or R and some experience exploring and analyzing data and developing statistical or machine learning models. Knowledge of Hadoop or Spark is not required.

COURSE OBJECTIVES:

- Overview of data science and machine learning at scale
- Overview of the Hadoop ecosystem
- Working with HDFS data and Hive tables using Hue
- Introduction to Cloudera Data Science Workbench
- Overview of Apache Spark 2
- Reading and writing data
- Inspecting data quality
- Cleansing and transforming data
- Summarizing and grouping data
- Combining, splitting, and reshaping data
- Exploring data
- Configuring, monitoring, and troubleshooting Spark applications
- Overview of machine learning in Spark MLlib
- Extracting, transforming, and selecting features

- Building and evaluating regression models
- Building and evaluating classification models
- Building and evaluating clustering models
- Cross-validating models and tuning hyperparameters
- Building machine learning pipelines
- Deploying machine learning models
- Spark, Spark SQL, and Spark MLlib
- PySpark and sparklyr
- Cloudera Data Science Workbench (CDSW)
- Hue

COURSE OUTLINE:

Overview of CDSW

- Introduction to CDSW
- Who Can Use CDSW
- How to Access CDSW
- Navigating around CDSW
- User Settings
- Hadoop Authentication

Projects in CDSW

- Creating a New Project
- Navigating around a Project
- Project Settings

The CDSW Workbench Interface

- Using the Workbench
- Using the Sidebar
- Using the Code Editor
- Engines and Sessions
- Running Python and R Code in CDSW
- Running Code
- Using the Session Prompt
- Using the Terminal
- Installing Packages
- Using Markdown in Comments

Using Apache Spark 2 in CDSW

- Scenario and Dataset
- Copying Files to HDFS
- Interfaces to Apache Spark 2
- Connecting to Spark
- Reading Data
- Inspecting Data

Data Science and Machine Learning in CDSW

- Transforming Data
- Using SQL Queries
- Visualizing Data from Spark
- Machine Learning with MLlib
- Session History

Experiments and Models in CDSW

- Machine Learning Workflow
- Running Experiments
- Using Packages in Experiments
- Deploying Models
- Calling Models
- Using Packages in Models

Teams and Collaboration in CDSW

- Collaboration in CDSW
- Teams in CDSW
- Using Git for Collaboration
- Conclusion

SUNSET LEARNING INSTITUTE (SLI) DIFFERENTIATORS:

Sunset Learning Institute (SLI) has been an innovative leader in developing and delivering authorized technical training since 1996. Our goal is to help our customers optimize their cloud technology investments by providing convenient, high quality technical training that our customers can rely on. We empower students to master their desired technologies for their unique environments.

What sets SLI apart is not only our immense selection of trainings options, but our convenient and consistent delivery system. No matter how complex your environment is or where you are located, SLI is sure to have a training solution that you can count on!

Premiere World Class Instruction Team

- All SLI instructors have a four-year technical degree, instructor level certifications and field consulting work experience.
- Sunset Learning has won numerous Instructor Excellence and Instructor Quality Distinction awards since 2012

Enhanced Learning Experience

- The goal of our instructors during class is ensure students understand the material, guide them through our labs and encourage questions and interactive discussions.

Convenient and Reliable Training Experience

- You have the option to attend classes at any of our established training facilities or from the convenience of your home or office with the use of our HD-ILT network (High Definition Instructor Led Training)
- All Sunset Learning Institute classes are guaranteed to run – you can count on us to deliver the training you need when you need it!

Outstanding Customer Service

- Dedicated account manager to suggest the optimal learning path for you and your team
- Enthusiastic Student Services team available to answer any questions and ensure a quality training experience