CLOUDERA DATA SCIENCE WORKBENCH TRAINING

Accelerate data science in the enterprise

Cloudera Data Science Workbench enables fast, easy, and secure self-service data science for the enterprise.

Cloudera Data Science Workbench Training prepares learners to complete data science and machine learning projects using Cloudera Data Science Workbench (CDSW).

Get Hands-On Experience

Through narrated demonstrations and hands-on exercises, learners achieve proficiency in CDSW and develop the skills required to:

- Navigate CDSW's options and interfaces with confidence
- Create projects in CDSW and collaborate securely with other users and teams
- Develop and run reproducible Python and R code
- Customize projects by installing packages and setting environment variables
- Connect to a secure (Kerberized) Cloudera or Hortonworks cluster
- Work with large-scale data using Apache Spark 2 with PySpark and sparklyr
- Perform end-to-end machine learning workflows in CDSW using Python or R (read, inspect, transform, visualize, and model data)
- Measure, track, and compare machine learning models using CDSW's Experiments capability
- Deploy models as REST API endpoints serving predictions using CDSW's Models capability
- Work collaboratively using CDSW together with Git

What to Expect

This OnDemand course is designed for learners at organizations using CDSW under a trial license or a commercial license. The learner must have access to a CDSW environment on a Cloudera or Hortonworks cluster running Apache Spark 2. Some experience with data science using Python or R is helpful but not required. No prior knowledge of Spark or other Hadoop ecosystem tools is required.

Course Details:

Overview of CDSW
- Introduction to Cloudera Data Science Workbench
- 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
- The Workbench Interface
- 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
- Introducing PySpark (Python track)
- Introducing sparklyr (R track)
- Connecting to Spark
- Reading Data
- Inspecting Data

Data Science and Machine Learning in CDSW
- Transforming Data (Python track)
- Transforming Data Using dplyr (R track)
- Using SQL Queries
- Spark DataFrames Functions (R track)
- 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
- Cloning a Git Repository with SSH
- Using Git for Collaboration

What's New in CDSW 1.6
- Changes to the User Interface
- Using Third-Party Editors
- Desktop-Based Third-Party Editors

Cloudera_Datasheet_DataScienceWorkBench_105 : 191031