Training: Cloudera

DSCI-272 Predicting with Cloudera Machine Learning

CLOUDERA Training Partner

TRAINING GOALS:

Enterprise data science teams need collaborative access to business data, tools, and computing resources required to develop and deploy machine learning workflows. Cloudera Machine Learning (CML), part of the Cloudera Data Platform (CDP), provides the solution, giving data science teams the required resources.

This four-day course covers machine learning workflows and operations using CML. Participants explore, visualize, and analyze data. You will also train, evaluate, and deploy machine learning models.

The course walks through an end-to-end data science and machine learning workflow based on realistic scenarios and datasets from a fictitious technology company. The demonstrations and exercises are conducted in Python (with PySpark) using CML.

What you'll learn

Through lecture and hands-on exercises, you will learn how to:

- Utilize Cloudera SDX and other components of the Cloudera Data Platform to locate data for machine learning experiments
- Use an Applied ML Prototype (AMP)
- Manage machine learning experiments
- Connect to various data sources and explore data
- Utilize Apache Spark and Spark ML
- Deploy an ML model as a REST API
- Manage and monitor deployed ML models

What to expect

The course is designed for data scientists who need to understand how to utilize Cloudera Machine Learning and the Cloudera Data Platform to achieve faster model development and deliver production machine learning at scale. Data engineers, developers, and solution architects who collaborate with data scientists will also find this course valuable.

CONSPECT:

Compendium Centrum Edukacyjne Spółka z o.o.
ul. Tatarska 5, 30-103 Kraków, tel.: (12) 298 47 77
ul. Postępu 18B, 02-676 Warszawa, tel.: (22) 417 41 70
BNP Paribas Bank Polska S.A.
EUR: PL 79 1600 1462 1853 6451 3000 0006
USD: PL 52 1600 1462 1853 6451 3000 0007

- Introduction to CML
  - Overview
  - CML Versus CDSW
  - ML Workspaces
  - Workspace Roles
  - Projects and Teams
  - Settings
  - Runtimes/Legacy Engines
- Introduction to AMPs and the Workbench
  - Editors and IDE
  - Git
  - Embedded Web Applications
  - AMPs
- Data Access and Lineage
  - SDX Overview
  - Data Catalog
  - Authorization
  - Lineage
- Data Visualization in CML
  - Data Visualization Overview
  - CDP Data Visualization Concepts
  - Using Data Visualization in CML
- Experiments
  - Experiments in CML
- Introduction to the CML Native Workbench
  - Entering Code
  - Getting Help
  - Accessing the Linux Command Line
  - Working With Python Packages
  - Formatting Session Output
- Spark Overview
  - How Spark Works
  - The Spark Stack
  - File Formats in Spark
  - Spark Interface Languages
  - Introduction to PySpark

Compendium Centrum Edukacyjne Spółka z o.o.
ul. Tatarska 5, 30-103 Kraków, tel.: (12) 298 47 77
ul. Postępu 18B, 02-676 Warszawa, tel.: (22) 417 41 70
BNP Paribas Bank Polska S.A.
EUR: PL 79 1600 1462 1853 6451 3000 0006
USD: PL 52 1600 1462 1853 6451 3000 0007

- How DataFrame Operations Become Spark Jobs
- How Spark Executes a Job

- Running a Spark Application
  - Running a Spark Application
  - Reading data into a Spark SQL DataFrame
  - Examining the Schema of a DataFrame
  - Computing the Number of Rows and Columns of a DataFrame
  - Examining a Few Rows of a DataFrame
  - Stopping a Spark Application
- Inspecting a Spark DataFrame
  - Inspecting a DataFrame
  - Inspecting a DataFrame Column
- Transforming DataFrames
  - Spark SQL DataFrames
  - Working with Columns
  - Working with Rows
  - Working with Missing Values
- Transforming DataFrame Columns
  - Spark SQL Data Types
  - Working with Numerical Columns
  - Working with String Columns
  - Working with Date and Timestamp Columns
  - Working with Boolean Columns
- Complex Types
  - Complex Collection Data Types
  - Arrays
  - Maps
  - Structs
- User-Defined Functions
  - User-Defined Functions
  - Example 1: Hour of Day
  - Example 2: Great-Circle Distance
- Reading and Writing DataFrames
  - Working with Delimited Text Files
  - Working with Text Files
  - Working with Parquet Files

Compendium Centrum Edukacyjne Spółka z o.o.
ul. Tatarska 5, 30-103 Kraków, tel.: (12) 298 47 77
ul. Postępu 18B, 02-676 Warszawa, tel.: (22) 417 41 70
BNP Paribas Bank Polska S.A.
EUR: PL 79 1600 1462 1853 6451 3000 0006
USD: PL 52 1600 1462 1853 6451 3000 0007

- Working with Hive Tables
- Working with Object Stores
- Working with Pandas DataFrames

- Combining and Splitting DataFrames
  - Combining and Splitting DataFrames
  - Joining DataFrames
  - Splitting a DataFrame
- Summarizing and GroupingDataFrames
  - Summarizing Data with Aggregate Functions
  - Grouping Data
  - Pivoting Data
- Window Functions
  - Window Functions
  - Example: Cumulative Count and Sum
  - Example: Compute Average Days Between Rides for Each Rider
- Machine Learning Overview
  - Introduction to Machine Learning
  - Machine Learning Tools
- Apache Spark MLlib
  - Introduction to Apache Spark MLlib
- Exploring and Visualizing DataFrames
  - Possible Workflows for Big Data
  - Exploring a Single Variable
  - Exploring a Pair of Variables
- Monitoring, Tuning, and Configuring Spark Applications
  - Monitoring Spark Applications
  - Configuring the Spark Environment
- Fitting and Evaluating Regression Models
  - Assemble the Feature Vector
  - Fit the Linear Regression Model
- Fitting and Evaluating Classification Models
  - Generate Label
  - Fit the Logistic Regression Model
- Tuning Algorithm Hyperparameters Using Grid Search
  - Requirements for Hyperparameter Tuning
  - Tune the Hyperparameters Using Holdout Cross-Validation

Compendium Centrum Edukacyjne Spółka z o.o.
ul. Tatarska 5, 30-103 Kraków, tel.: (12) 298 47 77
ul. Postępu 18B, 02-676 Warszawa, tel.: (22) 417 41 70
BNP Paribas Bank Polska S.A.
EUR: PL 79 1600 1462 1853 6451 3000 0006
USD: PL 52 1600 1462 1853 6451 3000 0007

- Tune the Hyperparameters Using K-Fold Cross-Validation
- Fitting and Evaluating Clustering Models
  - Print and Plot the Home Coordinates
  - Fit a Gaussian Mixture Model
  - Explore the Cluster Profiles
- Processing Text: Fitting and Evaluating Topic Models
  - Fit a Topic Model Using Latent Dirichlet Allocation
- Fitting and Evaluating Recommender Models
  - Recommender Models
  - Generate Recommendations
- Working with Machine Learning Pipelines
  - Fit the Pipeline Model
  - Inspect the Pipeline Model
- Applying a Scikit-Learn Model to a Spark DataFrame
  - Build a Scikit-Learn Model
  - Apply the Model Using a Spark UDF
- Deploying a Machine Learning Model as a REST API in CML
  - Load the Serialized Model
  - Define a Wrapper Function to Generate a Prediction
  - Test the Function
- Autoscaling, Performance, and GPU Settings
  - Autoscaling Workloads
  - Working with GPUs
- Model Metrics and Monitoring
  - Why Monitor Models?
  - Common Models Metrics
  - Models Monitoring With Evidently
  - Continuous Model Monitoring
- Appendix: Workspace Provisioning
  - Workspace and Environment

REQUIREMENTS:

The course is designed for data scientists who need to understand how to utilize Cloudera Machine Learning and the Cloudera Data Platform to achieve faster model development and deliver production machine learning at scale. Data engineers, developers, and solution architects who collaborate with data scientists will also find this course valuable.

Compendium Centrum Edukacyjne Spółka z o.o.
ul. Tatarska 5, 30-103 Kraków, tel.: (12) 298 47 77
ul. Postępu 18B, 02-676 Warszawa, tel.: (22) 417 41 70
BNP Paribas Bank Polska S.A.
EUR: PL 79 1600 1462 1853 6451 3000 0006
USD: PL 52 1600 1462 1853 6451 3000 0007

Difficulty level



CERTIFICATE:

The participants will obtain certificates signed by Cloudera (course completion).

Upon completion of the course, attendees are encouraged to continue their study and register for the Cloudera Certified Administrator (CCA) exam
https://www.cloudera.com/about/training/certification/cdhhdp-certification/cca-admin.html

Certification is a great differentiator. It helps establish you as a leader in the field, providing employers and customers with tangible evidence of your skills and expertise.

TRAINER:

Certified Cloudera Instructor

Compendium Centrum Edukacyjne Spółka z o.o.
ul. Tatarska 5, 30-103 Kraków, tel.: (12) 298 47 77
ul. Postępu 18B, 02-676 Warszawa, tel.: (22) 417 41 70
BNP Paribas Bank Polska S.A.
EUR: PL 79 1600 1462 1853 6451 3000 0006
USD: PL 52 1600 1462 1853 6451 3000 0007