Cloudera Data Scientist Training (DST)

ID DST   **Preis** auf Anfrage   **Dauer** 4 Tage

Zielgruppe

The workshop is designed for data scientists who use Python or R to work with small datasets on a single machine and who need to scale up their data science and machine learning workflows to large datasets on distributed clusters. Data engineers, data analysts, developers, and solution architects who collaborate with data scientists will also find this workshop valuable. Workshop participants walk through an end-to-end data science and machine learning workflow based on realistic scenarios and datasets from a fictitious technology company. The material is presented through a sequence of brief lectures, interactive demonstrations, extensive hands-on exercises, and lively discussions. The demonstrations and exercises are conducted in Python (with PySpark) using Cloudera Data Science Workbench (CDSW). Supplemental examples using R (with sparklyr) are provided.

Kursziele

Through narrated lecture, recorded demonstrations, and hands-on exercises, you will learn how to:
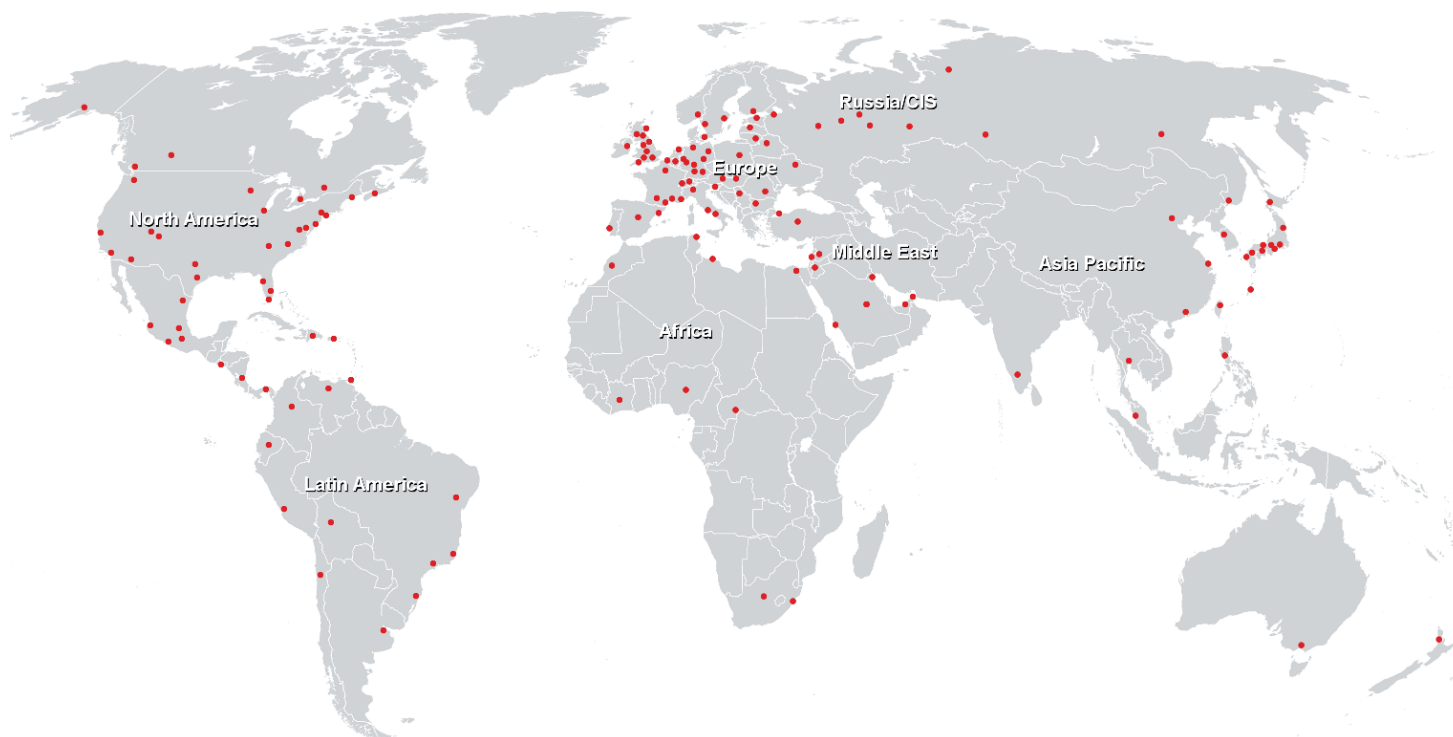
- How to use Apache Spark to run data science and machine learning workflows at scale
- How to use Spark SQL and DataFrames to work with structured data
- How to use MLlib, Spark's machine learning library
- How to use PySpark, Spark's Python API
- How to use sparklyr, a dplyr-compatible R interface to Spark
- How to use Cloudera Data Science Workbench (CDSW)
- How to use other Cloudera platform components including HDFS, Hive,
- Impala, and Hue

Kursinhalt

- Data Science Overview
- Cloudera Data Science Workbench (CDSW)
- Science Workbench
- Workbench Works
- Workbench

- Case Study
- Apache Spark
- Summarizing and Grouping DataFrames
- Window Functions
- Exploring DataFrames
- Apache Spark Job Execution
- Processing Text and Training and Evaluating Topic Models
- Training and Evaluating Recommender Models
- Running a Spark Application from (CDSW)
- Columns of a DataFrame
- Inspecting a Spark SQL DataFrame
- Transforming DataFrames
- Monitoring, Tuning, and Configuring Spark Applications
- Machine Learning Overview
- Training and Evaluating Regression Models
- Working with Machine Learning Pipelines
- Deploying Machine Learning Pipelines
- Transforming DataFrame Columns
- Complex Types
- User-Defined Functions
- Reading and Writing Data
- Combining and Splitting DataFrames
- Training and Evaluating Classification Models
- Tuning Algorithm Hyperparameters Using Grid Search
- Training and Evaluating Clustering Models
- Overview of sparklyr
- Introduction to Additional CDSW Features

Cloudera Data Scientist Training (DST)

Weltweite Trainingscenter

Fast Lane Institute for Knowledge Transfer GmbH

Husacherstrasse 3
CH-8304 Wallisellen
Tel. +41 44 832 50 80

info@flane.ch, https://www.flane.ch