cloudera

DATA SCIENTIST TRAINING

Kai Voigt, Senior Instructor, Cloudera University

HIGHLIGHTS

A course on Data Science and Machine Learning at Scale

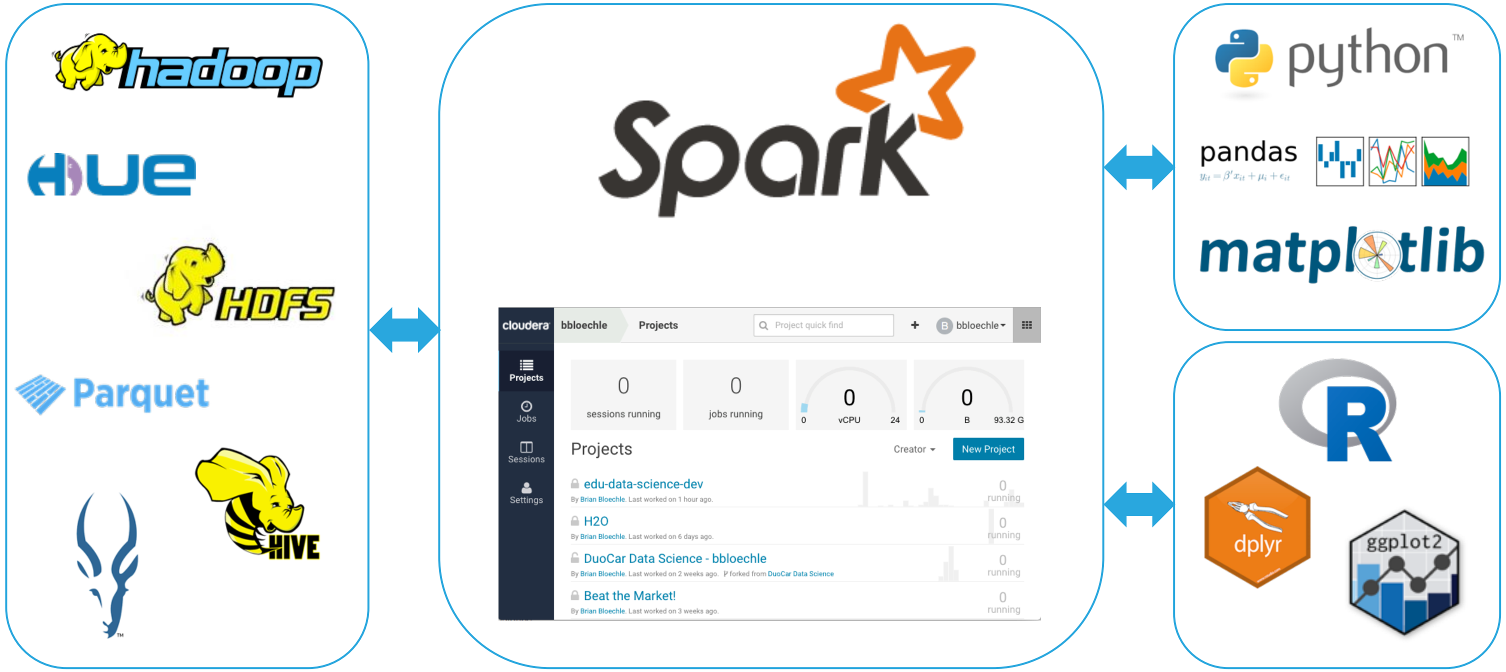
- Four-Day workshop using interactive code modules instead of slides
- Designed for both
  - Data Scientists with limited Hadoop experience
  - Hadoop Data Engineers with limited Data Science experience
- Focused on industry-leading tools
  - Hadoop, Spark, Python, R, SQL
  - Cloudera Data Science Workbench (CDSW)

OBJECTIVES

- Data Engineering
  - Reading and Writing Data
  - Inspecting Data Quality
  - Cleansing and Transforming Data
  - Combining and Splitting Data
  - Summarizing and Grouping Data
  - Exploring and Visualizing Data

- Data Science and Machine Learning
  - Extracting and Selecting Features
  - Building and Evaluation of
    - Regression Models
    - Classification Models
    - Clustering Models
  - Cross-Validating Models and Tuning Hyperparameters
  - Building Pipelines
  - Deploying Models

ARCHITECTURE AND TOOLS



0 sessions running

0 jobs running

0 vCPU 0 24

0 B 0 93.32 G

Projects

edu-data-science-dev
By Brian Bloechle. Last worked on 1 hour ago.
0 running

H2O
By Brian Bloechle. Last worked on 6 days ago.
0 running

DuoCar Data Science - bbloechle
By Brian Bloechle. Last worked on 2 weeks ago. forked from DuoCar Data Science
0 running

Beat the Market!
By Brian Bloechle. Last worked on 3 weeks ago.
0 running

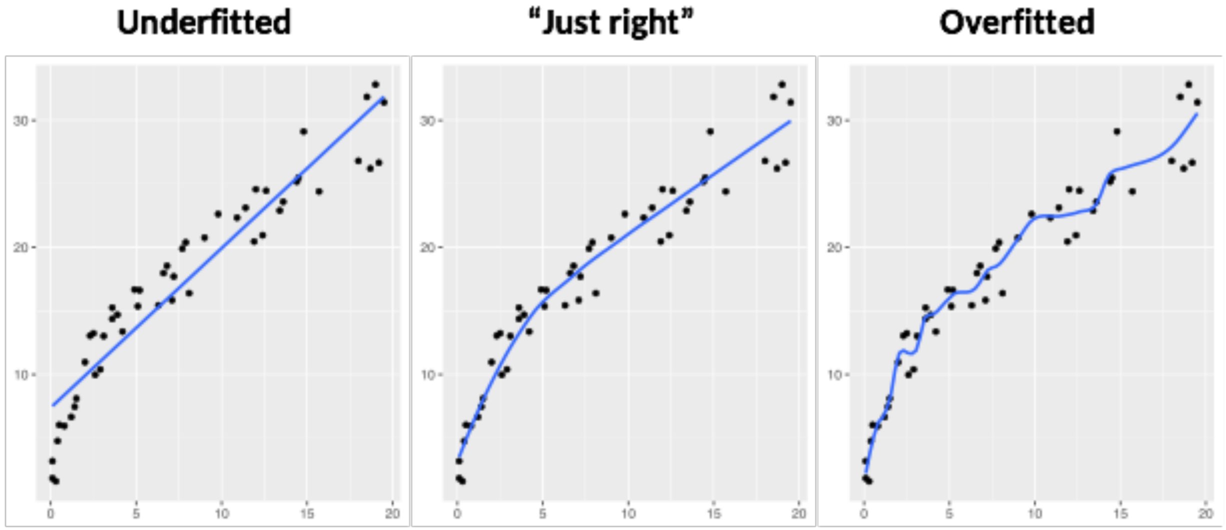
OUR STORY

Ficticious, but applicable to the real world

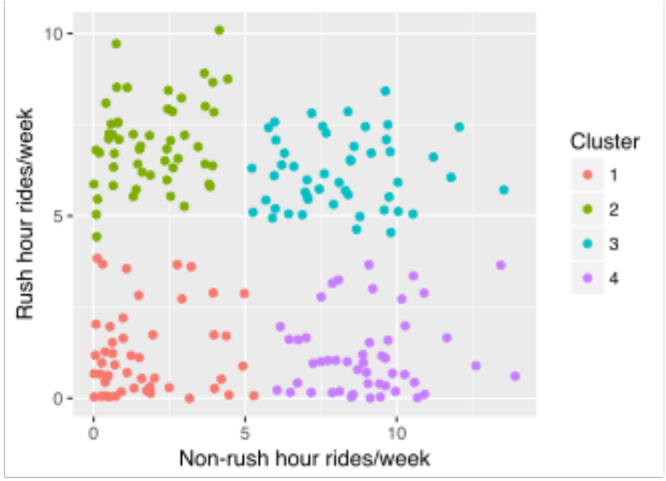
- Ridesharing company with lots of data
  - Drivers
  - Riders
  - Rides
  - Ride Routes
  - Reviews
  - Weather
  - Demographics

duocar

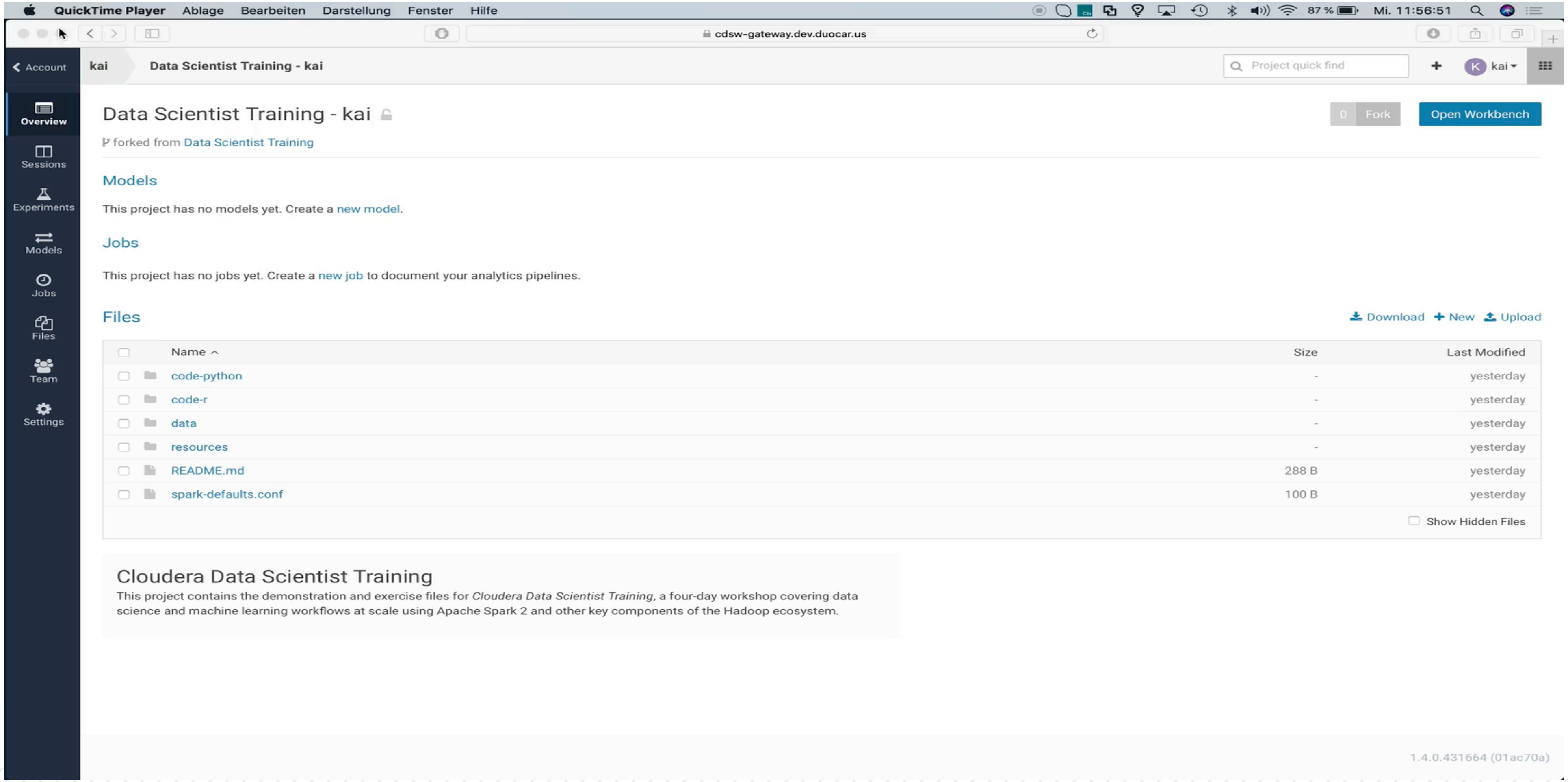
DISCUSSION EXAMPLE



ANOTHER EXAMPLE



DEMO TIME



Data Scientist Training - kai

forked from Data Scientist Training

Models

This project has no models yet. Create a new model.

Jobs

This project has no jobs yet. Create a new job to document your analytics pipelines.

Files

Name	Size	Last Modified
code-python	-	yesterday
code-r	-	yesterday
data	-	yesterday
resources	-	yesterday
README.md	288 B	yesterday
spark-defaults.conf	100 B	yesterday

Cloudera Data Scientist Training

This project contains the demonstration and exercise files for *Cloudera Data Scientist Training*, a four-day workshop covering data science and machine learning workflows at scale using Apache Spark 2 and other key components of the Hadoop ecosystem.

THANK YOU