Day 1

What you will learn

- What is Spark?
- The concept of Big Data modelling
- How to load differently formatted data
- Basic data manipulation
- How to analyse and visualise data
- How to transform data using Spark's inbuilt functions
- How to create a set of transformations to modify data in a single step
- How to use logistic regression as an example of modelling in Spark
- How to explore the model's outputs and choose the best one

Programme

Introductions; schedule

- A quick introduction of participants and familiarisation with the workshop schedule
- Access verification
- Distributing and installing the tools and data that will be used during the workshop

Cloudera Data Science Workbench (CDSW)

- A presentation of the CDSW analytical tool
- Principles of work at CDSW – project creation, team management, adjusting jobs and dependencies

Data analysis and visualisation

- Loading data from various sources
- Basic work with data
- Discovering descriptive statistics of individual variables
- Visualising data using packages in Python

An overview of Spark tools for data transformation

- A demonstration of algorithms for transforming variables
  - Continuous, categorical, text
- A demonstration of functions to select variables for the model and reduce dimensionality

Practice with the model; exploring the model's outputs

- Setting parameters for the logistic regression model
- Selecting evaluation criteria
- Practice with the model
- Choosing the best model
- Exploring the model's properties
- Applying the model to test data to determine its real predictive abilities

Individual work

- Each of the sessions mentioned above will be followed by a block of time designated for individual work, so you will be able to put the theoretical knowledge you have gained into practice.

Programme

Solutions for classification

- The specifics of classification
- An overview of the classification approaches offered by Spark
- Tasks focused on classification trees and random forests; Multilayer Perceptrons
- A demonstration of individual approaches

Solutions for regression

- The specifics of regression
- An overview of the regression approaches offered by Spark
- Tasks focused on regression trees and random forests
- Tasks focused on gradient boosting and other approaches
- A demonstration of individual approaches

Applying the model to new data

- Deploying the model on a new data file
- Criteria assessment and testing evaluation

Solutions for segmentation

- An overview of the segmentation approaches offered by Spark
- A demonstration of individual approaches

Deploying the model to streaming data

- The specifics of modelling on streaming data
- Deploying the selected model on new streaming production data and exporting the results

Individual work

- Each of the sessions mentioned above will be followed by a block of time designated for individual work, so you will be able to put the theoretical knowledge you have gained into practice.

Day 2

What you will learn

- Which analytics algorithms Spark includes
- How to use individual techniques for advanced analytics and machine learning
- How to deploy and apply the final model to the newly generated data
- How to create a complete distributed data science pipeline
- How to work with the Notebook tool and how to use it for teamwork