Big Data Analytics with RapidMiner

Course Overview

Big data is worthless without the capability to analyze and visualize. Big Data Analytics with RapidMiner Radoop is a two day course designed to help leverage huge data collection by converting raw data into valuable information using RapidMiner Radoop. RapidMiner Radoop provides ETL, analytics and visualization in a single package and integrates seamlessly into new and existing RapidMiner processes to bring analytics into your Hadoop cluster.

After successfully completing this course, participants will have a solid understanding of how RapidMiner Radoop integrates with Hadoop. Participants will be able to connect to a Hadoop cluster, explore, extract and load data, and integrate in-cluster analyses into RapidMiner processes.

Practical exercises during the course prepare students to take the knowledge gained and apply to their own big data challenges. Since the class labs are hands-on and performed on the participants' personal laptops, students will take actual classwork home with them, which will provide a jumpstart to the real world.

Target Audience

Advanced Analysts, Data Scientists

Prerequisites

Basic knowledge of computer programs and mathematics
RapidMiner & DataScience: Foundations
RapidMiner & DataScience: Advanced
RapidMiner Server: Deployment and Web Apps

Course Objectives

After the training, students will have the ability to:

- Understand Hadoop infrastructure
- Connect to a Hadoop cluster
- Explore large data stores
- Perform data extraction and loading tasks
- Integrate in-cluster analyses into RapidMiner processes

Course Outline

- What is Big Data?
- How does Big Data fit into modern analytics?
- Introduction to Hadoop
    - ◊ Distributions
    - ◊ General Infrastructure
- Hadoop Integration with RapidMiner: Radoop
- Introduction to the Radoop GUI
- Connecting to a Hadoop Cluster
- Data Exploration
    - ◊ Browsing Tables
    - ◊ Viewing Statistics and High Level Information
- Data extraction and Loading
    - ◊ Formulation of Queries
    - ◊ Pushing Data into Hadoop
- Integration of In-cluster Analyses into RapidMiner Processes
    - ◊ Modeling Algorithms
    - ◊ Natural Aggregation
    - ◊ In-memory Training, in-Hadoop Scoring
- Beyond natural Aggregation:
    - ◊ Chunking
    - ◊ Voting
    - ◊ In-Hadoop Modeling
    - ◊ Clustering