Trainings

Apache Spark für Data Scientists

Training rund um das Framework Apache Spark zur Echtzeit-Datenanalyse.

Duration: 3 Tage

Target group: Data Scientists

Egal ob Batch- oder Stream-Processing – Apache Spark hat sich dank seiner Performance als verteilte In-Memory-Technologie innerhalb von kurzer Zeit einen Stammplatz im Ökosystem der Big Data Tools erarbeitet.

Dieses Training richtet sich primär an Data Scientists und vermittelt den grundlegenden Aufbau und die Architektur von Spark, sowie den Umgang mit mächtigen Frontend-Tools aus dem Spark-Ökosystem zur Durchführung der Analysen.

Ein inhaltlicher Schwerpunkt des Trainings ist Machine Learning. Nach einer allgemeinen Einführung wird die Spark MLlib eingehend vorgestellt, eine Bibliothek, welche dem Anwender viele mächtige Machine-Learning-Algorithmen „out of the box“ zur Verfügung stellt.

In diesem Training steht immer die Praxis im Vordergrund: Grundlage des Trainings ist eine komplexe Datenbasis an welcher Methoden, Tools & Techniken von den Teilnehmern geübt werden. Dabei wird Python als Programmiersprache verwendet.

Agenda:

Tag 1 — Spark

- Einführung in Apache Spark
- Einführung in Apache Zeppelin
- Spark API und RDDs

- KeyValue-RDD und Joins
- Spark SQL und Dataframes/DataSets

Tag 2 — Machine Learning

- Einführung in Machine Learning
  - Supervised / Unsupervised Learning
  - Features Extraction
  - Validation

Tag 3 — Machine Learning in der Praxis

- Überblick über Modelle, Algorithmen und ihre Einsatzgebiete
- Vor- und Aufbereitung der Daten
- Machine Learning in der Praxis:
  - Anwendung von Spark ML auf einer großen Datenbasis

Hinweis:

- Die Kursgebühr beinhaltet Schulungsunterlagen, Mittagessen, Getränke und Snacks.
- Die Teilnehmer:innen müssen ein eigenes Notebook zum Training mitbringen.