*Trainings*

# Apache Spark für Data Scientists

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

Dauer: 2 Tage

Zielgruppe: 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 führt in den Umgang mit Spark zur Analyse großer Datenmengen ein. Dabei werden sowohl Batch- als auch Streamingverfahren diskutiert. Ein Schwerpunkt des Trainings ist die Formulierung von analytischen Anfragen und die Nutzung maschineller Lernverfahren. Ausgehend von konkreten Business Anforderungen lernen die Teilnehmer:innen geeignete Architekturen, Techniken und Tools kennen, um Lösungen zu implementieren, welche die Business Bedürfnisse befriedigen.

In diesem Training steht immer die Praxis im Vordergrund: Grundlage des Trainings ist eine komplexe Datenbasis an welcher Methoden, Tools & Techniken von den Teilnehmer:innenn geübt werden.

## Agenda:

- Spark Grundlagen und Architektur
- Spark APIs und die RDD Datenstruktur

- Abfragen formulieren mit Spark SQL
- Transformationen und Aktionen im Spark Kontext
- Machine Learning mittels der Spark MLlib
- Überblick über das Apache Spark Ökosystem
- Design von Spark-Architekturen zur Umsetzung konkreter Usecases