Our **trainings** introduce the foundations of data analysis and relevant tools for data scientists. During our sessions, we always keep data-driven use cases in sight. Our modular training concept consists of skill building blocks from introductory to expert level. We emphasize applications and a hands-on approach. All trainings take place in a virtual environment or on-site. Our trainers are senior data scientists with extensive experience in research and industry.

Our **TechTrainings** are geared towards technical users - data scientists, analysts, engineers - building data-driven innovation hands-on.

Our **BusinessTrainings** provide insights for decision makers - shedding light on how to put data science and artificial intelligence to work in the enterprise.

We use our **digital laboratory in the cloud**, thus each participant can work at his workspace independently while the trainer is presenting. We focus on interactivity, hands-on and individual support. This is what we aim with our developed and well tested trainings material in Jupyter Notebooks, small group sizes and enough time for practice.
Big Data Analysis with PySpark
BDAS

Scale your data analysis to really, really big data. Strengthen your data science skills and learn how to work with PySpark and distributed computing on a cluster.

Level: ♦♦ (Intermediate)

Duration: 1 day

Prerequisites: DAP (or similar know-how)

Language: english, german; materials are in english

1. Processing Big Data
   Which strategies are available to compute efficiently with increasing amounts of data? What is a cluster, and when do we need one?

2. Spark Fundamentals
   An overview of Spark - a framework for programming distributed computation, using PySpark, its Python API - core data structures and operations.

3. Submitting Spark Jobs
   How to submit jobs to a Spark cluster for batch processing.

4. Spark and Structured Data
   Working with structured data in Spark.

5. Exercises:
   A. Museums of France
      An exercise with a clear task, requiring you to apply the learnings from the course.
   B. Counting Bigrams
      Using Spark to count bigrams in big text data.