Contribution ID: **41**

Type: **not specified**

Apache Spark in Scientific Applications

Thursday, September 1, 2016 1:00 PM (5 hours)

The workshop Spark in Scientific Applications covers fundamentale development and data analysis techniques using Apache Hadoop and Apache Spark. Beside an introduction into the theoretical background about Map-Reduce- and Bulk-Synchronous-Parallel processing, also the machine learning library MLlib and the graph processing framework GraphX are used.

We work on sample data sets from Wikipedia, financial market data, and from a generic data generator. During the tutorial sessions we illustrate the Data Science Workflow and present the right tools for the right task. All practical exercises are well prepared in a pre-configured virtual machine. Participants get access to required data sets on a „one node pseudo-distributed“cluster with all tools inside. This VM is also a starting point for further experiments after the workshop.

Presenter: KÄMPF, Mirko (Cloudera)