



Contribution ID: 30

Type: **not specified**

Apache Spark in Scientific Applications [B]

Thursday, September 10, 2015 1:00 PM (5 hours)

This tutorial is limited to 12 participants. Another session of this tutorial is also available
</p>

The workshop Spark in Scientific Applications covers fundamental 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.

</p>

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.

</p>

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.

Author: Mr MIRKO, Kaempf (Cloudera)

Presenter: Mr MIRKO, Kaempf (Cloudera)

Session Classification: Apache Spark in Scientific Applications

Track Classification: Data Analysis