GridKa School 2015
Big Data Virtualization Modern Programming

Contribution ID: **29** Type: **not specified**

Apache Spark: The next Generation of Hadoop Processing

Thursday, September 10, 2015 9:00 AM (40 minutes)

Apache Spark is known as the "Next Generation Framework" of Hadoop based data processing. Why, and what Apache Spark offers to the scientific community is explained in this talk. The convergence of different analysis techniques into one flexible and highly efficient processing engine allows completely new interdisciplinary analysis methods beside cheap analysis prototypes. In this presentation I shown examples in Scala and Python. Beside fundamental techniques and the core features of Apache Spark we look into development practices and data analysis techniques. Therefore we recap the theoretical background about Map-Reduce- and Bulk-Synchronous-Parallel processing before I introduce the machine learning library MLlib and the graph processing framework GraphX. Apache Spark uses the concept of data frames, and allows SQL operations on data sets, after this presentation you know how this works and how you can save a lot of time. Finally, you can see how data can be collected and analyzed on the fly, using Spark Streaming.

Author: Mr KAEMPF, Mirko (Cloudera)

Presenter: Mr KAEMPF, Mirko (Cloudera)

Session Classification: Plenary talks

Track Classification: Plenary Talks