Hadoop for HPC Users: Overview and first steps

Elmar Kiesling

TU Wien

In this tutorial, we will cover key concepts of the Hadoop framework, including data locality, replication, and data-centric parallelization. Next, we will take a look at the architecture and core components of the Hadoop stack and discuss some of the challenges and changes in Hadoop 2.0.

In the hands-on part of the tutorial, we will first cover the basics of HDFS and MapReduce programming. Based on this foundation, we will then move up the stack and provide a glimpse at technologies such as HBase and Hive, two large-scale distributed data stores on top of Hadoop. The final part of the tutorial will focus on Apache Spark, which provides several more general APIs to parallelise data-intensive computing problems for applications such as machine learning, stream processing, and interactive analytics.

After completing this tutorial, participants will have a basic understanding of the kinds of problems that Hadoop-based clusters can tackle, as well as an overview of some of the key tools in the Hadoop ecosystem.