MapReduce and Hadoop Eco System, Prof. Zoran Đorđević (Harvard University)

sreda, 12.06.2013. od 12:00 do 14:00, sala 2, SANU, Knez Mihajlova 35

We will review basic principles of MapReduce technique and Apache Hadoop, its open source implementation. The Apache Hadoop is a framework that allows for the distributed processing of large data sets across clusters of computers using simple programming models. Hadoop is designed to scale up from single servers to thousands of machines, each offering local computation and storage. Rather than relying on hardware to deliver high-availability, the library itself is designed to detect and handle failures at the application layer, so delivering a highly-available service on top of a cluster of computers, each of which may be prone to failures. HDFS, Hadoop's file system, and Java MapReduce API will be introduced and discussed, as well as Hadoop Streaming API which allows us to write code and run MapReduce processes using any major programing language. Some of the most important frameworks in the Hadoop eco system will be introduced: PIG Latin, a flow language for batch processing, Hive a SQL like data warehousing framework atop of Hadoop, HBase a columnar database, Zookeeper, Hadoop cluster coordinator, and Avro, a serialization system.

Statistical and Real Time Processing of Big Data, Prof. Zoran Đorđević (Harvard University)

sreda, 12.06.2013. od 14:00 do 18:00, sala 2, SANU, Knez Mihajlova 35

Analysis of very large data sets, so called Big Data, requires use of statistical techniques and tools. R, a language and environment for statistical computing and graphic, has emerged as the statistical tool of choice for Big Data Analytics. R provides a wide variety of statistical (linear and nonlinear modeling, classical statistical tests, time-series analysis, classification, clustering) and graphical techniques, and is highly extensible. We will review R itself and also discuss the combination of R with Hadoop (MapReduce) which provides ability for massively scalable statistical analysis. We will introduce three popular techniques to couple R and MapReduce: Hadoop Streaming, Rhipe and RHadoop. We will introduce Mahout, a Java Framework for highly scalable machine learning processing on Hadoop. A few illustrative clustering and classification algorithms will be demonstrated. Finally we will touch on Impala which adds ad hoc query capability to Apache Hadoop, complementing traditional MapReduce batch processing. Impala enables real time queries against large volumes of streaming data collected by Storm or Flume or queries of data stored in HDFS or Apache HBase. Impala queries are written in familiar SQL syntax, including SELECT, JOIN, and aggregate functions and executed in real time.