SORS: Building a Unified Data Pipeline in Apache Spark

Speakers: Aaron Davidson (Apache Spark committer and Software Engineer at Databricks) and Paco Nathan (Community Evangelism Director at Databricks)

Title: Building a Unified Data Pipeline in Apache Spark

Date: 20 November 2014, 18.30

Venue: Sala d'Actes de la FIB (edifici B6)

Summary:
One of the promises of Apache Spark is to let users build unified data analytic pipelines that combine diverse processing types. In this talk, we'll demo this live by building a machine learning pipeline with 3 stages: ingesting JSON data from Hive; training a k-means clustering model; and applying the model to a live stream of tweets. Typically this pipeline might require a separate processing framework for each stage, but we can leverage the versatility of the Spark runtime to combine Shark, MLlib, and Spark Streaming and do all of the data processing in a single, short program. This allows us to reuse code and memory between the components, improving both development time and runtime efficiency. Spark as a platform integrates seamlessly with Hadoop components, running natively in YARN and supporting arbitrary Hadoop InputFormats, so it brings the power to build these types of unified pipelines to any existing Hadoop user.

This talk will be a fully live demo and code walkthrough where we'll build up the application throughout the session, explain the libraries used at each step, and finally classify raw tweets in real-time.

Places are limited, please confirm your attendance by sending email to jordi [dot] torres [at] bsc [dot] es

Barcelona Supercomputing Center - Centro Nacional de Supercomputación