Contribution ID: **231**

Type: **Oral**

Hadoop and friends - first experience at CERN with a new platform for high throughput analysis steps

Thursday 13 October 2016 14:45 (15 minutes)

The statistical analysis of infrastructure metrics comes with several specific challenges, including the fairly large volume of unstructured metrics from a large set of independent data sources. Hadoop and Spark provide an ideal environment in particular for the first steps of skimming rapidly through hundreds of TB of low relevance data to find and extract the much smaller data volume that is relevant for statistical analysis and modelling.

This presentation will describe the new Hadoop service at CERN and the use of several of its components for high throughput data aggregation and ad-hoc pattern searches. We will describe the hardware setup used, the service structure with a small set of decoupled clusters and the first experience with co-hosting different applications and performing software upgrades. We will further detail the common infrastructure used for data extraction and preparation from continuous monitoring and database input sources.

Secondary Keyword (Optional)

Monitoring

Primary Keyword (Mandatory)

Analysis tools and techniques

Tertiary Keyword (Optional)

Author: DUELLMANN, Dirk (CERN)

Co-authors: SURDY, Kacper (CERN); MENICHETTI, Luca (CERN); KOTHURI, Prasanth (CERN); TOEBBICKE, Rainer (CERN); MENON, Vineet (Bhabha Atomic Research Centre (IN))

Presenter: KOTHURI, Prasanth (CERN)

Session Classification: Track 5: Software Development

Track Classification: Track 5: Software Development