Contribution ID: **96** Type: **Oral**

# Exploiting Apache Spark platform for CMS computing analytics

*Tuesday, August 22, 2017 3:40 PM (20 minutes)*

The CERN IT provides a set of Hadoop clusters featuring more than 5 PB of raw storage. Different open-source user-level tools are installed for analytics purposes. For this reason, since early 2015, the CMS experiment has started to store a large set of computing metadata, including e.g. a massive number of dataset access log.. Several streamers have registered some billions traces from heterogeneous providers. These trace logs represent a valuable yet scarcely investigated set of information that needs to be cleansed, categorized and correlated; in the case of the CMS dataset access information, this work may lead to discover useful patterns to enhance the overall efficiency of the distributed infrastructure in terms of CPU utilization and task completion time. This work presents an evaluation of Apache Spark platform for CMS needs. We demonstrate a few use-cases how to efficiently process metadata information stored on CERN HDFS system in a scalable manner by harnessing a variety of languages of choice. Among them, Scala and Python offer the best approach to CMS use cases for executing extremely I/O intensive queries that leverage in-memory and persistence Spark API as well as assess streamlining predictive models that can learn dataset properties using machine learning approaches.

**Primary authors:** Prof. BONACORSI, Daniele (University of Bologna); KUZNETSOV, Valentin Y (Cornell University (US)); BOCCALI, Tommaso (INFN Sezione di Pisa, Universita' e Scuola Normale Superiore, P); MEONI, Marco (INFN Sezione di Pisa, Universita' e Scuola Normale Superiore, P); MENICHETTI, Luca (CERN)

**Presenter:** MEONI, Marco (INFN Sezione di Pisa, Universita' e Scuola Normale Superiore, P)

**Session Classification:** Track 1: Computing Technology for Physics Research

**Track Classification:** Track 1: Computing Technology for Physics Research