Contribution ID: **68** Type: **Oral Presentation**

Exploiting big data analytics for CMS computing operations cost-effectiveness

Thursday, 24 March 2022 16:10 (20 minutes)

Computing operations at the Large Hadron Collider (LHC) at CERN rely on the Worldwide LHC Computing Grid (WLCG) infrastructure, designed to efficiently allow storage, access, and processing of data at the pre-exascale level.
A close and detailed study of the exploited computing systems for the LHC physics mission represents an increasingly crucial aspect in the roadmap of High Energy Physics (HEP) towards the exascale regime. A deep knowledge of these systems will be essential to effectively express the discovery potential of LHC in the High Luminosity phase, which is estimated to collect up to 30 times more data than LHC in the next few years, and with novel part of the detectors which also adds large complexities to the overall picture.
In this context, the Compact Muon Solenoid (CMS) experiment has been collecting and storing over the last years a large set of heterogeneous non-collision-data (e.g. meta-data about data placement choices, transfer operations, and actual user access to physics datasets): all this data richness is currently residing on a distributed Hadoop cluster, and the data is organized so that running fast and arbitrary queries using the Spark analytics framework is a viable approach for focussed big data mining efforts.
CMS relies on its Monte Carlo Management (MCM) system, a tool to collect and monitor all Monte Carlo sample requests, namely a system that can be used to gain access to additional information about the simulated datasets produced for physics analysis.
Exploiting these sources, and using a data-driven approach oriented to the analysis of the aforementioned meta-data, we started to focus on data storage and data transfers over the WLCG infrastructure, and drafted an embrional software toolkit that can provide useful indicators about recurrent patterns and their correlations, for a deeper and more accurate exploration of the CMS computing beating heart in terms data movement and access. This aims – as a short-to-medium term goal – at exploring the effectiveness and adequateness of various choices in a data lake context, and – as a long term goal – at contributing to the overall design of a predictive/adaptive system that would eventually reduce cost and complexity of the CMS computing operations, while taking into account the stringent requests and the boundary conditions set by the physics analysts community.

Primary authors: BONACORSI, Daniele (University of Bologna); Dr GASPERINI, Simone (University of Bologna); ROSSI TISBENI, Simone (INFN-CNAF)

Presenters: Dr GASPERINI, Simone (University of Bologna); ROSSI TISBENI, Simone (INFN-CNAF)

Session Classification: Data Management & Big Data

Track Classification: Track 6: Data Management & Big Data