21st International Conference on Computing in High Energy and Nuclear Physics **CHEP2015** Okinawa Japan: April 13 - 17, 2015

Contribution ID: **335** Type: **oral presentation**

Exploiting CMS data popularity to model the evolution of data management for Run-2 and beyond

Monday, April 13, 2015 5:00 PM (15 minutes)

During the LHC Run-1 data taking, all experiments collected large data volumes from proton-proton and heavy-ion collisions. The collisions data, together with massive volumes of simulated data, were replicated in multiple copies, transferred among various Tier levels, transformed/slimmed in format/content. These data were then accessed (both locally and remotely) by large groups of distributed analysis communities exploiting the WorldWide LHC Computing Grid infrastructure and services. While efficient data placement strategies - together with optimal data redistribution and deletions on demand - have become the core of static versus dynamic data management projects, little effort has so far been invested in understanding the detailed data-access patterns which surfaced in Run-1. These patterns, if understood, can be used as input to simulation of computing models at the LHC, to optimise existing systems by tuning their behaviour, and to explore next-generation CPU/storage/network co-scheduling solutions. This is of great importance, given that the scale of the computing problem will increase far faster than the resources available to the experiments, for Run-2 and beyond.

Studying data-access patterns involves the validation of the quality of the monitoring data collected on the "popularity" of each dataset, the analysis of the frequency and pattern of accesses to different datasets by analysis end-users, the exploration of different views of the popularity data (by physics activity, by region, by data type), the study of the evolution of Run-1 data exploitation over time, the evaluation of the impact of different data placement and distribution choices on the available network and storage resources and their impact on the computing operations.

This work presents some insights from studies on the popularity data from the CMS experiment. We present the properties of a range of physics analysis activities as seen by the data popularity, and make recommendations for how to tune the initial distribution of data in anticipation of how it will be used in Run-2 and beyond.

Primary author: Prof. BONACORSI, Daniele (University of Bologna)

Co-authors: Dr GIORDANO, Domenico (CERN); Dr GIRONE, Maria (CERN); MATTEO, Neri (University of Bologna); MAGINI, Nicolo (CERN); BOCCALI, Tommaso (Sezione di Pisa (IT)); Dr WILDISH, Tony (Princeton University (US)); KUZNETSOV, Valentin Y (Cornell University (US))

Presenter: Prof. BONACORSI, Daniele (University of Bologna)

Session Classification: Track 5 Session

Track Classification: Track5: Computing activities and Computing models