Contribution ID: 13

Type: not specified

Access Pattern Analysis in the EOS Storage System at CERN

Tuesday, June 9, 2020 11:00 AM (15 minutes)

EOS is a CERN-developed storage system that serves several hundred petabytes of data to the scientific community of the Large Hadron Collider (LHC). In particular, it provides services to the four largest LHC particle detectors: LHCb, CMS, ATLAS and ALICE. Each of these collaborations uses different workflows to process and analyse its data. EOS has a monitoring system that collects detailed information on the file accesses and can give important insights about the specifics of the physics experiments' workflows. In our study, we analyse the monitoring information accumulated over a six months period and amounting to over 1.3 terabytes and have the goal to help the IT department and the experiments' operations teams to better understand the EOS data flows.

In this contribution, we describe a pipeline, mainly developed in R, for processing large volumes of access logs and perform a comparative analysis of the storage usage in scientific workflows. In particular, we calculate aggregated statistics over a six months period and provide a high-level overview of the experiments' data flows. Additionally, we study how the frequency of data accesses changes over time and estimate to what extent different experiments may benefit from an additional caching layer.

Author: CHUCHUK, Olha (CERN, Taras Shevchenko KNU)
Co-author: Dr DUELLMANN, Dirk (CERN)
Presenter: CHUCHUK, Olha (CERN, Taras Shevchenko KNU)
Session Classification: Session 4