Contribution ID: **176** Type: **Oral**

XRootD Popularity on Hadoop Clusters

Thursday 13 October 2016 14:30 (15 minutes)

The CMS experiment has implemented a computing model where distributed monitoring infrastructures are collecting any kind of data and metadata about the performance of the computing operations. This data can be probed further by harnessing Big Data analytics approaches and discovering patterns and correlations that can improve the throughput and the efficiency of the computing model.

CMS has already begun to store a large set of operational data - user activities, job submissions, resources, file transfers, site efficiencies, software releases, network traffic, machine logs - in a Hadoop cluster. This offers the ability to run fast arbitrary query on the data and test several computing MapReduce-based frameworks.

In this work we analyze the XrootD logs collected in Hadoop through Gled and Flume and we benchmark their aggregation at the level of dataset for monitoring purpose of popularity queries, thus proving how dashboard and monitoring systems can benefit from Hadoop parallelism. Processing time on existing Oracle DBMS of XrootD time-series logs does not scale linearly with data volume. Conversely, Big Data architectures do and make it very effective re-processing any user-defined time interval. The entire set of existing Oracle queries is replicated in the Hadoop data store and result validation is performed accordingly.

These results constitute the set of features on top of which a mining platform is designed to predict the popularity of a new dataset, the best location for replicas or the proper amount of CPU and storage in future timeframes. Learning techniques applied to Big Data architectures are extensively explored to study the correlations between aggregated data and seek for patterns in the CMS computing ecosystem. Examples of this kind are primarily represented by operational information like file access statistics or dataset attributes, which are organised in samples suitable for feeding several classifiers.

Tertiary Keyword (Optional)

Secondary Keyword (Optional)

Databases

Primary Keyword (Mandatory)

Analysis tools and techniques

Author: MEONI, Marco (Universita di Pisa & INFN (IT))

Co-authors: GIORDANO, Domenico (CERN); MENICHETTI, Luca (CERN); MAGINI, Nicolo (Fermi National Accelerator Lab. (US)); BOCCALI, Tommaso (Universita di Pisa & INFN (IT))

Presenters: MENICHETTI, Luca (CERN); MEONI, Marco (Universita di Pisa & INFN (IT)); MAGINI, Nicolo (Fermi National Accelerator Lab. (US))

Session Classification: Track 5: Software Development

Track Classification: Track 5: Software Development