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

Contribution ID: **303** Type: **oral presentation**

Disk storage management for LHCb based on Data Popularity estimator

Tuesday, 14 April 2015 18:15 (15 minutes)

The amount of data produced by the LHCb experiment every year consists of several petabytes. This data is kept on disk and tape storage systems. Disks are much faster than tapes, but are way more expensive and hence disk space is limited. It is impossible to fit the whole data taken during the experiment's lifetime on disk, but fortunately fast access to datasets are no longer needed after the analysis requiring them are over. So it is highly important to identify which datasets should be kept on disk and which ones should be kept as archives on tape. The metrics to be used for deprecating datasets'caching is based on the "popularity"of this dataset, i.e. whether it is likely to be used in the future or not. We discuss here the approach and the studies carried out for optimizing such a Data Popularity estimator.

Input information to the estimator are the dataset usage history and metadata (size, type, configuration etc). The system is designed to select the datasets which may be used in the future and thus should remain on disk. Studies have therefore been performed on how to optimize the usage of dataset information from the past for predicting its future popularity. In particular, we have carried out a detailed comparison of various time series analysis, machine learning classifier, clustering and regression algorithms. We demonstrate that our approach is capable of improving significantly the disk usage efficiency.

Primary author: HUSHCHYN, Mikhail (Moscow Institute of Physics and Technology, Moscow)

Co-authors: USTYUZHANIN, Andrey (ITEP Institute for Theoretical and Experimental Physics (RU)); Dr CATTANEO, Marco (CERN); CHARPENTIER, Philippe (CERN)

Presenter: HUSHCHYN, Mikhail (Moscow Institute of Physics and Technology, Moscow)

Session Classification: Track 3 Session

Track Classification: Track3: Data store and access