Contribution ID: **515** Type: **Oral**

Analyzing how we do Analysis and consume data, Results from the SciDAC-Data Project.

Monday, 10 October 2016 15:45 (15 minutes)

One of the principle goals of the Dept. of Energy funded SciDAC-Data project is to analyze the more than 410,000 high energy physics “datasets”that have been collected, generated and defined over the past two decades by experiments using the Fermilab storage facilities. These datasets have been used as the input to over 5.6 million recorded analysis projects, for which detailed analytics have been gathered. The analytics and meta information regarding these for these datasets and analysis projects are being combined with knowledge of their part of the HEP analysis chains for major experiments to understand how modern computing and data delivery is being used.

We present the first results of this project, which examine in detail how the CDF, DØ and NO⊠A experiments have organized, classified and consumed petascale datasets to produce their physics results. The results include the analysis of the correlations in dataset/file overlap, data usage patterns, data popularity, dataset dependency and temporary dataset consumption. The results provide critical insight into how workflows and data delivery schemes can be combined with different caching strategies to more efficiently perform the work required to mine these large HEP data volumes and to understand the physics analysis requirements for the next generation of HEP computing facilities.

In particular we present detailed analysis of the NO⊠A data organization and consumption model corresponding to their first and second oscillation results (2014-2016) and the first look at the analysis of the Tevatron Run II experiments. We present statistical distributions for the characterization of these data and data driven models describing their consumption.

Tertiary Keyword (Optional)

Data processing workflows and frameworks/pipelines

Secondary Keyword (Optional)

Computing facilities

Primary Keyword (Mandatory)

Data model

Primary author: Dr NORMAN, Andrew (Fermilab)

Co-authors: Dr LYON, Adam (Fermilab); Dr TSARIS, Aristeidis (Fermilab); Dr ALIAGA SOPLIN, Leonidas (Fermilab); Dr MUBARAK, Misbah (Argonne); Dr DING, Pengfei (Fermilab); Dr ROSS, Robert (Argonne)

Session Classification: Track 7: Middleware, Monitoring and Accounting

Track Classification: Track 7: Middleware, Monitoring and Accounting