24th International Conference on Computing in High Energy & Nuclear Physics

Contribution ID: **422** Type: **Poster**

Striped Data Analysis Framework

Thursday, 7 November 2019 16:15 (15 minutes)

Traditionally, High Energy data analysis is based on the model where data are stored in files and analyzed by running multiple analysis processes, each reading one or more of the data files. This process involves repeated data reduction step, which produces smaller files, which is time consuming and leads to data duplication. We propose an alternative approach to data storage and analysis, based on the Big Data technologies. The idea is to store each element of data once and only once in a distributed scalable database and analyze data by reading only "interesting" pieces of data from the database.

To make this approach possible, we developed columnar Striped Data Representation Format as the basis of the framework. Traditional columnar approach allows for efficient analysis of complex data structures. While keeping all the benefits of columnar data representation, striped mechanism goes further by enabling efficient parallelization of computations and flexible distribution of data analysis.

The framework includes scalable and elastic data storage, compute and user analysis backend components. The framework uses off-the shelf web services and data caching technologies as the compute/data co-location mechanism. Flexible architecture allows the framework run in the cloud using container technologies. The framework offers Python/Jupyter as the user analysis backend platform, but can also run in command line or batch mode.

In the article, we will present the results of the FNAL-LDRD-2016-032 FNAL LDRD project, the design, implementation, features and performance characteristics of the Striped Data Analysis Framework.

Consider for promotion

No

Authors: Mr MANDRICHENKO, Igor (Fermi National Accelerator Lab. (US)); GUTSCHE, Oliver (Fermi National Accelerator Lab. (US))

Presenter: GUTSCHE, Oliver (Fermi National Accelerator Lab. (US))

Session Classification: Posters

Track Classification: Track 6 –Physics Analysis