Contribution ID: **68** Type: **Poster**

Striped Data Server for Scalable Parallel Data Analysis

Thursday, August 24, 2017 4:30 PM (15 minutes)

Columnar data representation is known to be an efficient way to store and access data, specifically in cases when the analysis is often done based only on a small fragment of the available data structure. Data representations like Apache Parquet, on the other hand, split data horizontally to allow for easy parallelization of data analysis. Based on the general idea of columnar data storage, working on the LDRD Project FNAL-LDRD-2016-032, we have developed Striped data representation, which, we believe, is better suited to the needs of High Energy Physics data analysis.

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.

We present simple and efficient striped data representation model based on Numpy arrays and unified API, which have been implemented for a range of different types of physical storage from local file system to distributed no-SQL database. We further demonstrate a Python-based analysis application platform, which leverages the striped data representation.

We have also implemented Striped Data Server (SDS) as a web service, which hides storage implementation details from the end user and exposes data to WAN users via the web service. Such web service can be deployed as a part of the enterprise computing facility or as a cloud service.

We plan to explore SDS as an enterprise scale data analysis platform for High Energy Physics community and hope to expand it to the other areas that require similar high performance analysis with massive datasets. We have been testing this architecture with 2TB CMS dark matter search dataset and plan to expand it to full CMS public dataset, which is close to 10PB in size.

Primary authors: Mr MANDRICHENKO, Igor (FNAL); Mr PIVARSKI, James (Princeton University); Mr CHANG, Jin (FNAL); MANDRICHENKO, Igor Vasilyevich

Presenters: Mr MANDRICHENKO, Igor (FNAL); MANDRICHENKO, Igor Vasilyevich

Session Classification: Poster Session

Track Classification: Track 2: Data Analysis - Algorithms and Tools