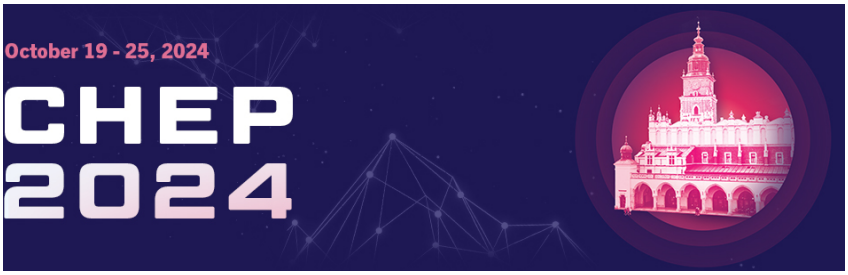

Contribution ID: **47** Type: **Talk**

On Demand Column Joining for End User Analysis

Monday 21 October 2024 14:06 (18 minutes)

The high luminosity LHC (HL-LHC) era will deliver unprecedented luminosity and new detector capabilities for LHC experiments, leading to significant computing challenges with storing, processing, and analyzing the data. The development of small, analysis-ready storage formats like CMS NanoAOD (4kB/event), suitable for up to half of physics searches and measurements, helps achieve necessary reductions in data processing and storage. However, a large fraction of analyses frequently require very computationally expensive machine learning output or data only stored in larger and less accessible formats, such as CMS MiniAOD (45kB/eevent) or AOD (450kB/event). This necessitates the non-volatile storage of derived data in custom formats. In this work, we present research on the development of workflows and integration of tools with ServiceX to efficiently fetch, cache, and join together data for use with columnar analysis tools.
We leverage scaleable, distributed SQL query engines like Trino to join disparate columns sourced from multiple files and without a restriction on relative row ordering. By replacing many customized datasets, containing largely overlapping contents, with smaller and unique sets of information that can be joined on demand with common central data, duplication can be reduced. Caching of these results keeps the cost of subsequent retrieval low, fitting well with modern physics analysis paradigms.

Primary authors: GALEWSKY, Benjamin (Univ. Illinois at Urbana Champaign (US)); Dr HOLZMAN, Burt (Fermi National Accelerator Lab. (US)); MANGANELLI, Nick (University of Colorado Boulder (US))

Co-authors: WATTS, Gordon (University of Washington (US)); ULMER, Keith (University of Colorado, Boulder (US)); GRAY, Lindsey (Fermi National Accelerator Lab. (US))

Presenter: MANGANELLI, Nick (University of Colorado Boulder (US))

Session Classification: Parallel (Track 5)

Track Classification: Track 5 - Simulation and analysis tools