Contribution ID: **127** Type: **Poster**

columnflow: Fully automated analysis through flow of columns over arbitrary, distributed resources

Thursday 14 March 2024 16:10 (30 minutes)

To study and search for increasingly rare physics processes at the LHC, a staggering amount of data needs to be analyzed with progressively complex methods. Analyses involving tens of billions of recorded and simulated events, multiple machine learning algorithms for different purposes, and an amount of 100 or more systematic variations are no longer uncommon. These conditions impose a complex data flow on an analysis workflow and render its steering and bookkeeping a serious challenge.
For this purpose, a toolkit for columnar HEP analysis, called *columnflow*, has been developed. It is written in Python, experiment agnostic in its core, and supports any flat file format, such as ROOT-based trees or Parquet files. Leveraging on the vast Python ecosystem, vectorization and convenient physics objects representation can be achieved through NumPy, awkward arrays and other libraries. Based upon the Luigi Analysis Workflow (*law*) package, *columnflow* provides full analysis automation over arbitrary, distributed computing resources. Despite the end-to-end nature, this approach allows for persistent, intermediate outputs for purposes of debugging, caching, and exchange with collaborators. Job submission to various batch systems, such as HTCondor, Slurm, or CMS-CRAB, is natively supported. Remote files can be seamlessly accessed via various protocols using either the Grid File Access Library (GFAL2) or the fsspec file system interface. In addition, a sandboxing mechanism can encapsulate the execution of parts of a workflow into dedicated environments, supporting subshells, virtual environments, and containers.
This contribution introduces the key components of *columnflow* and highlights the benefits of a fully automated workflow for complex and large-scale HEP analyses, showcasing an implementation of the Analysis Grand Challenge.

Significance

References

Experiment context, if any

CMS

Primary authors: WIEDERSPAN, Bogdan (Hamburg University (DE)); RIEGER, Marcel (Hamburg University (DE))

Presenter: WIEDERSPAN, Bogdan (Hamburg University (DE))

Session Classification: Poster session with coffee break

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