Contribution ID: **189** Type: **Oral**

RDataFrame: a flexible and scalable analysis experience

Wednesday 26 October 2022 14:15 (20 minutes)

The growing amount of data generated by the LHC requires a shift in how HEP analysis tasks are approached. Usually, the workflow involves opening a dataset, selecting events, and computing relevant physics quantities to aggregate into histograms and summary statistics. The required processing power is often so high that the work needs to be distributed over multiple cores and multiple nodes. This contribution establishes ROOT RDataFrame as the single entry point for virtually all HEP data analysis use cases. In fact, the typical steps of an analysis workflow can be easily and flexibly written with RDataFrame. Data ingestion from multiple sources is streamlined through a single interface. Relevant metadata can be made available to the dataframe and used during analysis execution. A declarative API offers the most common operations to the users, while transparently taking care of data processing optimisations. For example, it is possible to inject user-defined code to compute complex quantities, gather them into histograms or other relevant statistics, include large sets of systematic variations and use machine-learning inference kernels. A Pythonic layer allows dynamic injection of Python functions in the main C++ event loop. Finally, any RDataFrame application can seamlessly scale out to hundreds of cores on the same machine or multiple distributed nodes by changing a single line of code. The latest performance validation studies are also included in this contribution to demonstrate the efficiency of the tool on both the computation complexity and the scalability spectra.

Significance

This contribution demonstrates how a physics analysis can be written from begin to end with a single interface. All users can benefit from having a coherent interface that removes the burden of thinking about the programming implementation and just focus on the desired results. Virtually all analysis workflows can be written with RDataFrame, which is now more flexible than ever. Most relevant new features of the tool are: the possibility of writing a Python-only application that still exploits a fast C++ core; the inclusion of machine learning kernels in the event loop; handling metadata of large data samples directly within RDataFrame itself, thus enabling usage of such information directly within the event loop on a per-sample basis. The attendees will see a HEP analysis being written step by step with this tool, from easier tasks with the aim of exploring the dataset to complex operations that represent a full analysis in production.

References

Experiment context, if any

Authors: GUIRAUD, Enrico (EP-SFT, CERN); PADULANO, Vincenzo Eduardo (Valencia Polytechnic University

(ES)); TEJEDOR SAAVEDRA, Enric (CERN); KABADZHOV, Ivan (Albert Ludwig University of Freiburg); PAWAN, Pawan

Presenter: PADULANO, Vincenzo Eduardo (Valencia Polytechnic University (ES))

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

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