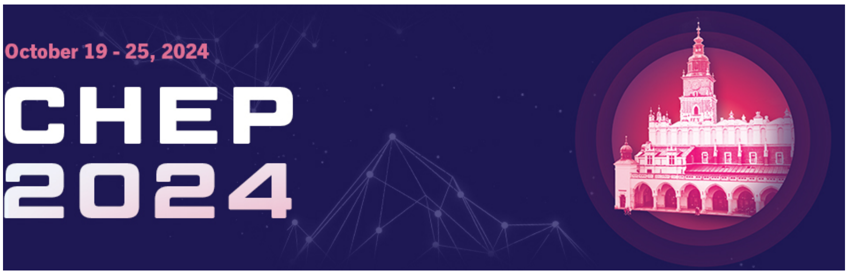
Conference on Computing in High Energy and Nuclear Physics



Contribution ID: **106** Type: **Talk**

Distributed analysis in production with RDataFrame

Thursday 24 October 2024 16:15 (18 minutes)

The ROOT software package provides the data format used in High Energy Physics by the LHC experiments. It offers a data analysis interface called RDataFrame, which has proven to adapt well to the requirements of modern physics analyses. However, with increasing data collected by the LHC experiments, the challenge to perform an efficient analysis expands. One of the solutions to ease this challenge, is the leverage of modern high performing distributed computing environments for which RDataFrame provides an easy-to-use interface layer - the distributed RDataFrame.

In this talk, we show that the Distributed RDataFrame is out of the experimental testing phase, and it is now ready for production thanks to a stabilized user interface. We delve into recent improvements of the distributed RDataFrame, including Pythonizations of the interface that allow running the workflows seamlessly (for example, with the XGBoost library). As the variety and geographical locations of distributed environments are available, we show the reproducibility and compare the performance across several of them.

Primary author: CZURYLO, Marta (CERN)

Co-authors: OLA MEJICANOS, Andrea Maria (University of Wisconsin Madison (US)); PIPARO, Danilo (CERN); Dr PADULANO, Vincenzo Eduardo (CERN)

Presenter: CZURYLO, Marta (CERN)

Session Classification: Parallel (Track 9)

Track Classification: Track 9 - Analysis facilities and interactive computing