Contribution ID: **22** Type: **Lightning talk**

RDataFrame: interactive analysis at scale by example

Tuesday 10 October 2023 16:30 (10 minutes)

With the increasing dataset sizes brought by the current LHC data analysis workflows and the future expectations that estimate even greater computational needs, data analysis software must strive to optimise the processing throughput on a single core and ensure an efficient distribution of tasks across multiple cores and computing nodes. RDataFrame, the high-level interface for data analysis offered by ROOT, natively supports distributing Python applications seamlessly across a plethora of computing cluster deployments, via the Spark and Dask execution engines. This contribution reports known use cases of the distributed RDataFrame tool at scale on different deployments, including resources at CERN and externally. In particular, a focus is devoted to the user experience aspect of distributed RDataFrame, showing how notable functionality for an interactive workflow is included natively in the tool. For example, the possibility to visualise in real time the events processed by the distributed tasks.

Authors: Dr PADULANO, Vincenzo Eduardo (CERN); TAIDER, Silia; TEJEDOR SAAVEDRA, Enric (CERN); CZURYLO, Marta (CERN); BOULIS, Joseph; GUIRAUD, Enrico (Princeton University, CERN); FALKO, Andrii

Presenter: Dr PADULANO, Vincenzo Eduardo (CERN)

Session Classification: Plenary Session Tuesday