Contribution ID: **18** Type: **Notebook talk**

A Python package for distributed ROOT RDataFrame analysis

Monday, July 5, 2021 3:40 PM (30 minutes)

The declarative approach to data analysis provides high-level abstractions for users to operate on their datasets in a much more ergonomic fashion compared to previous imperative interfaces. ROOT offers such a tool with RDataFrame, which has been tested in production environments and used in real-world analyses with optimal results. Its programming model acts by creating a computation graph with the operations issued by the user and executing it lazily only when the final results are queried. It has always been oriented towards parallelisation, with native support for multi-threading execution on a single machine.
Recently, RDataFrame has been extended with a Python layer that is capable of steering and executing the RDataFrame computation graph over a set of distributed resources. In addition, such layer requires minimal code changes for an RDataFrame application to run distributedly. The new tool features a modular design, such that it can support multiple backends in order to exploit the vast ecosystem of distributed computing frameworks with Python bindings.
This work presents Distributed RDataFrame, its programming model and design. It also demonstrates its current compatibility with two different distributed computing frameworks, namely Apache Spark and Dask, with more to come in the future.

Primary authors: TEJEDOR SAAVEDRA, Enric (CERN); PADULANO, Vincenzo Eduardo (Valencia Polytechnic University (ES))

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

Session Classification: Plenary Session Monday