Contribution ID: **58** Type: **Short Talk**

Fine-grained data caching approaches to speedup a distributed RDataFrame analysis

Wednesday, May 19, 2021 6:06 PM (13 minutes)

Thanks to its RDataFrame interface, ROOT now supports the execution of the same physics analysis code both on a single machine and on a cluster of distributed resources. In the latter scenario, it is common to read the input ROOT datasets over the network from remote storage systems, which often increases the time it takes for physicists to obtain their results. Storing the remote files much closer to where the computations will run can bring latency and execution time down. Such a solution can be improved further by caching only the actual portion of the dataset that will be processed on each machine in the cluster, reusing it in subsequent executions on the same input data. This paper shows the benefits of applying different means of caching input data in a distributed ROOT RDataFrame analysis. Two such mechanisms will be applied to this kind of workflow with different configurations, namely caching on the same nodes that process data or caching on a separate server.

Primary authors: Mr PADULANO, Vincenzo Eduardo (Valencia Polytechnic University (ES)); Dr TEJEDOR SAAVEDRA, Enric (CERN); Prof. ALONSO-JORDA, Pedro (Valencia Polytechnic University (ES))

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

Session Classification: Software

Track Classification: Distributed Computing, Data Management and Facilities