Contribution ID: **33** Type: **Notebook talk**

Computing tag-and-probe efficiencies with Apache Spark and Apache Parquet

Monday 5 July 2021 17:40 (30 minutes)

In this talk we demonstrate a new framework developed by the muon physics object group in CMS to compute tag-and-probe (T&P) efficiencies and scale factors by leveraging the power and scalability of Apache Spark clusters. The package, named "spark_tnp", allows physics analyzers and other users to quickly and seamlessly compute efficiencies for their own custom objects and identification criteria, developed to meet a diverse set of physics goals within the Collaboration. For the backend cluster, we use CERN's Spark and Hadoop services ("analytix"cluster). The ntuples with event information are produced separately in ROOT and converted to Apache Parquet format, which are then stored at CERN's Hadoop filesystem (HDFS) facility. The combined leverage of Spark and Parquet files in HDFS enables a substantial speed-up of T&P computations, with custom scale factors derived in a matter of minutes, compared to days in a previous framework. The tutorial itself will focus on a Jupyter notebook example of a T&P computation, using CERN's SWAN service for easy access to the analytix cluster within an interactive environment (though the package also supports scripted execution for official production).

Author: FRANKENTHAL, Andre (Princeton University (US))

Presenter: FRANKENTHAL, Andre (Princeton University (US))

Session Classification: Plenary Session Monday