Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA
ICSC Centro Nazionale di Ricerca in HPC, Big Data and Quantum Computing

INFN

ICSC
Centro Nazionale di Ricerca in HPC,
Big Data and Quantum Computing

VBS example in RDataFrame

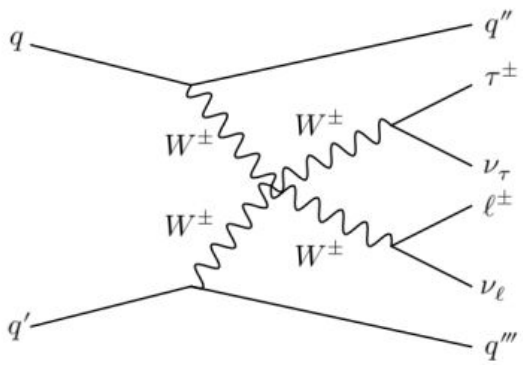
Tommaso Tedeschi

ICSC & Spoke2 Where are we now? - 10/12 Dec 2024 - Catania

RDataFrame's 2017 SSWW VBS with tau

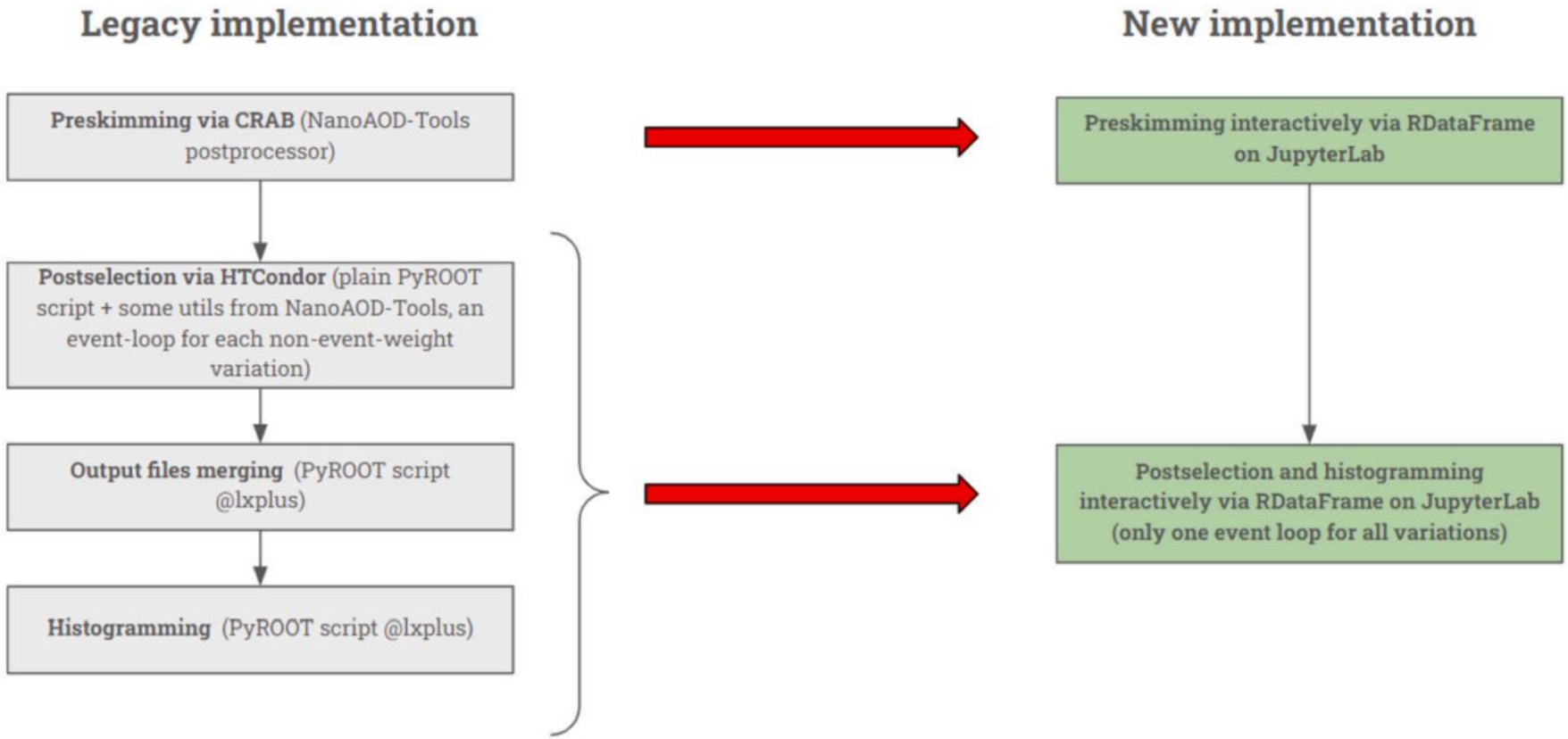
The analysis with CMS detector of the scattering (VBS) of two same-sign W bosons decaying to a hadronic tau and a light lepton was taken as benchmark to test a pure RDataFrame-based approach on the CMS INFN Analysis Facility

$$q\bar{q} \to W^{\pm}W^{\pm}q\bar{q} \to \tau_h^{\pm}\nu_\tau \ell^{\pm}\nu_\ell q\bar{q}$$



The physics analysis was converted from a legacy iterative approach to an RDataFrame-based approach

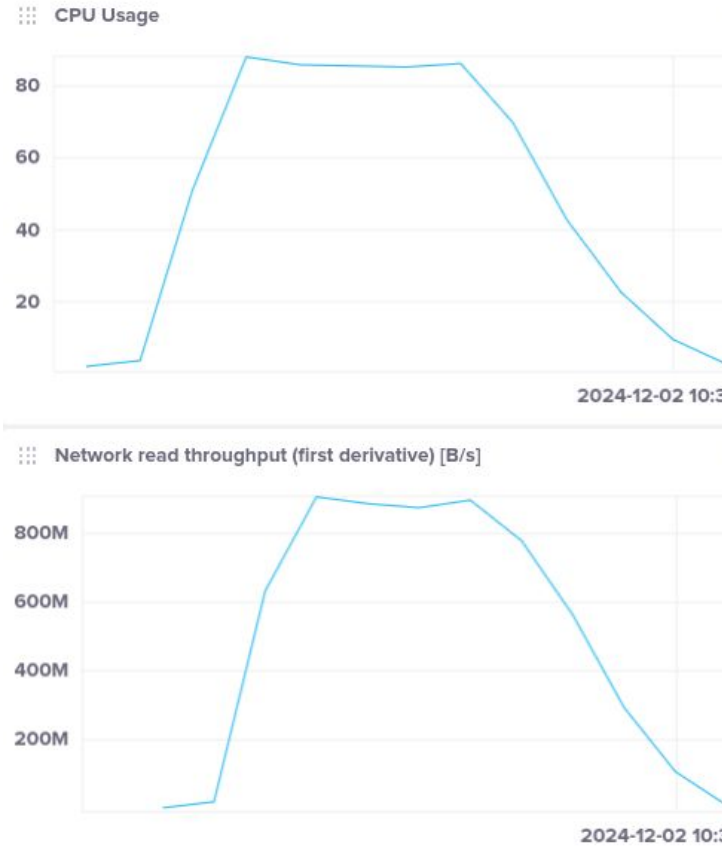
The flow


Legacy implementation
Preskimming via CRAB (NanoAOD-Tools postprocessor)
Postselection via HTCondor (plain PyROOT script + some utils from NanoAOD-Tools, an event-loop for each non-event-weight variation)
Output files merging (PyROOT script @lxplus)
Histogramming (PyROOT script @lxplus)
New implementation
Preskimming interactively via RDataFrame on JupyterLab
Postselection and histogramming interactively via RDataFrame on JupyterLab (only one event loop for all variations)

Benchmarks

We demonstrated a gain of one order of magnitude in terms of time with respect to a batch-like approach, see the dedicated paper [1]

- In particular, preprocessing on ~1 TB 2017 MC samples (filtering and corrections computation) stored at Legnaro Production done with user CPU usage at 80/90 % and network read throughput at 800/900 MB/s



Benchmark run on **96 CPUs at Legnaro production tier2 site**

Demo!

INFN

hub: https://cms-it-hub.cloud.cnaf.infn.it/

image: `ghcr.io/comp-dev-cms-ita/jupyterlab:AF20-alma9-v0.0.10-rc9`

code: https://github.com/ttedeschi/workshop2025_demo/tree/main/RDataFrame/VBS