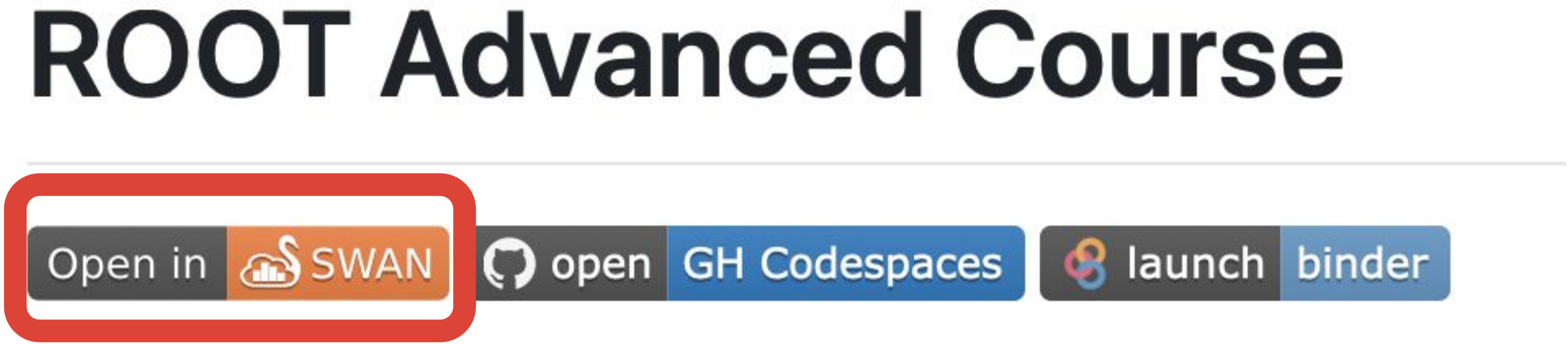
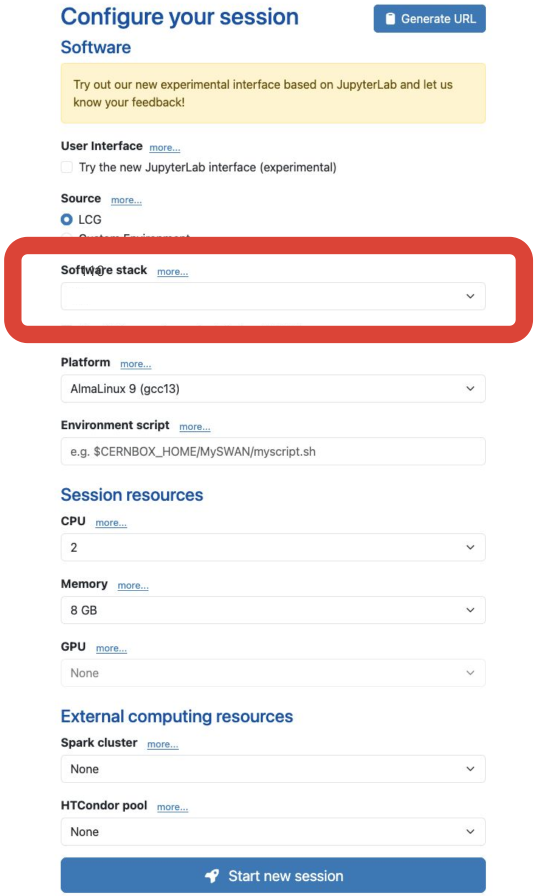
ROOT Advanced Course

15.07.2026

Setup - SWAN

- Go to https://github.com/root-project/ROOTAdvancedCourse
- If you have CERN Computing account → click on the SWAN badge





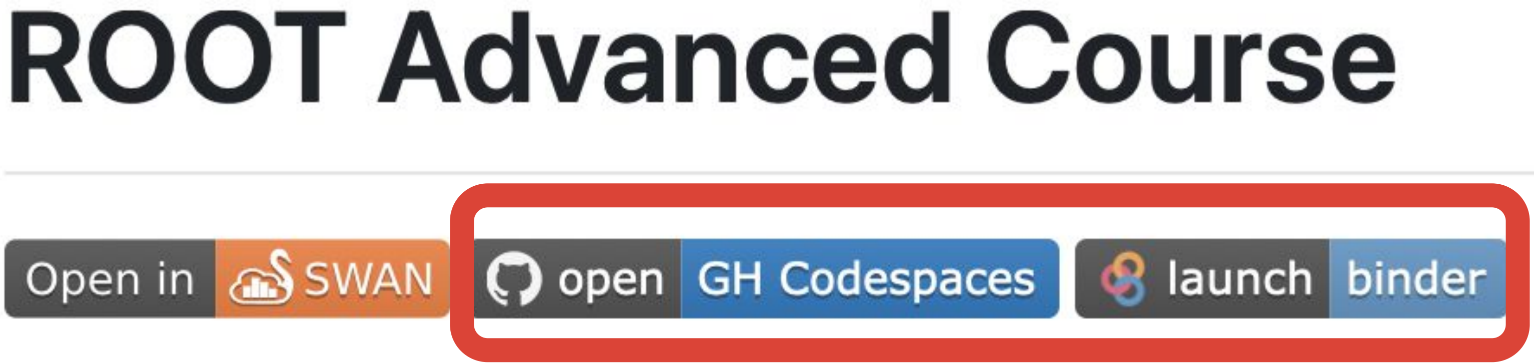
Use the Software Stack LCG 110

Found under “Other releases”

Leave all other parameters as-is!

Setup - no CERN account

- Go to https://github.com/root-project/ROOTAdvancedCourse
- If you don’t have CERN Computing account:
  - Click on either GH Codepaces or binder badges - note binder might take a bit of time to load
  - If you choose GH Codespaces - use Python venv as environment

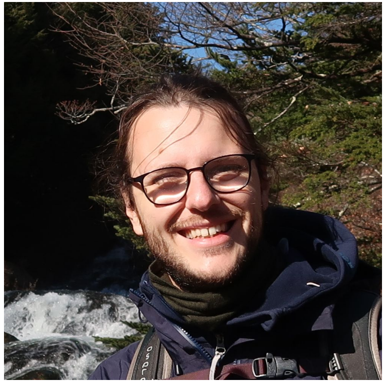


Meet the team

Silia

Florine

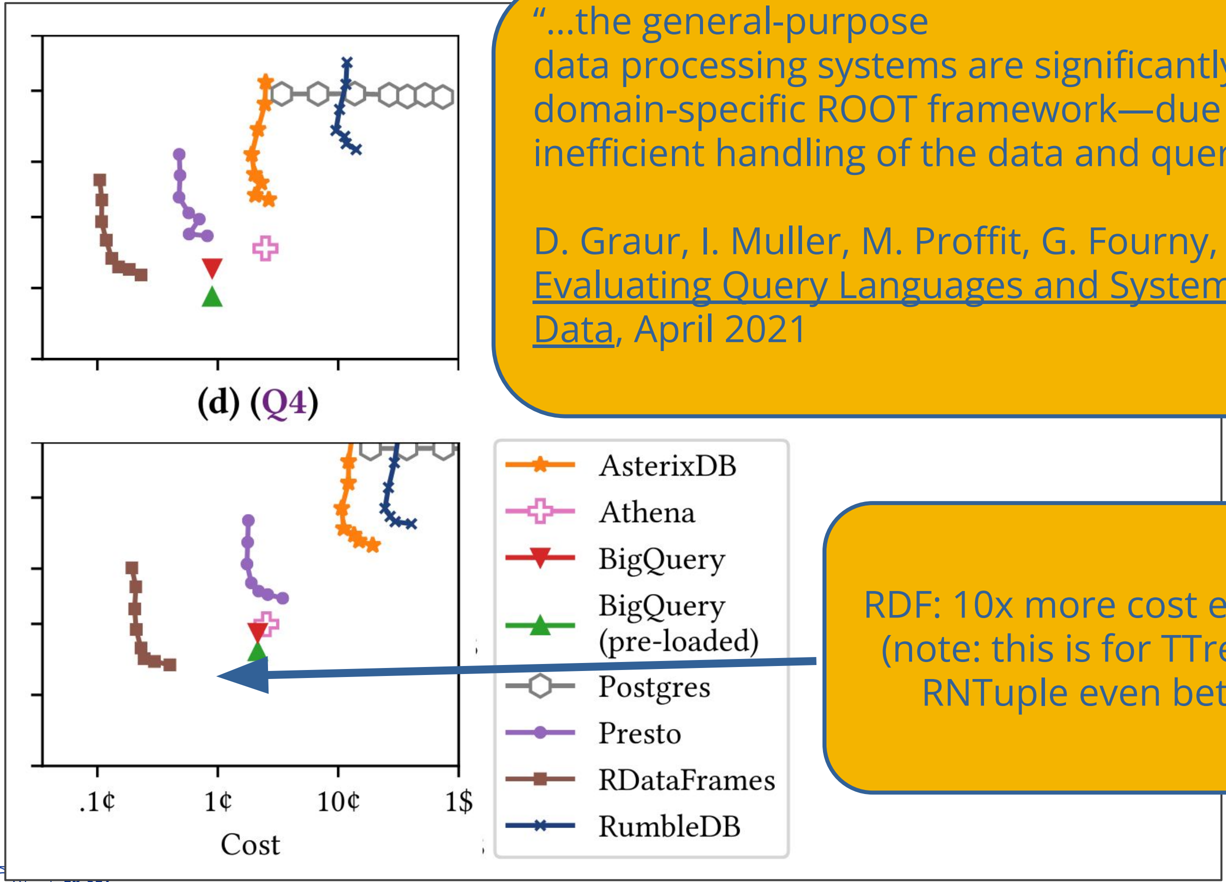
Giacomo



Jonas

Goal

- Run your own analysis leveraging the tools introduced today
- In particular – perform your analysis with RDataFrame



"…the general-purpose
data processing systems are significantly less performant than the domain-specific ROOT framework—due to limited scalability and inefficient handling of the data and queries relevant to HEP"

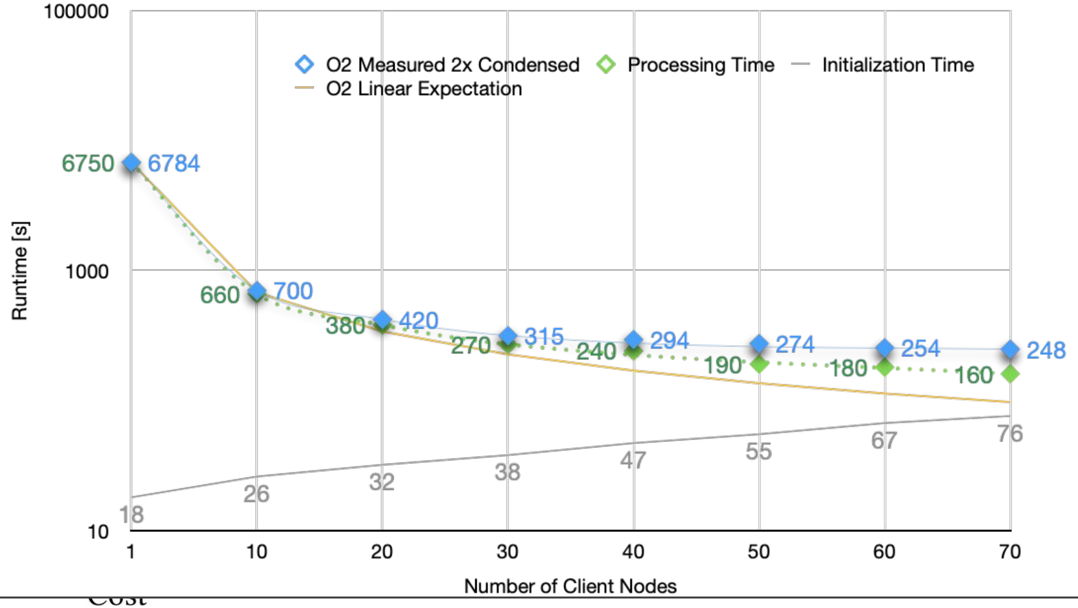
D. Graur, I. Muller, M. Proffit, G. Fourny, G. Watts et al, Evaluating Query Languages and Systems for High-Energy Physics Data, April 2021

RDF: 10x more cost efficient (note: this is for TTree, for RNTuple even better)

Why RDataFrame?

A.J. Peters, J. Blomer
RNTuple Plenary
CHEP24
October 2024

Single Analysis
reaches avg. INGRES
222 GBit/s
during processing
345 GBit/s



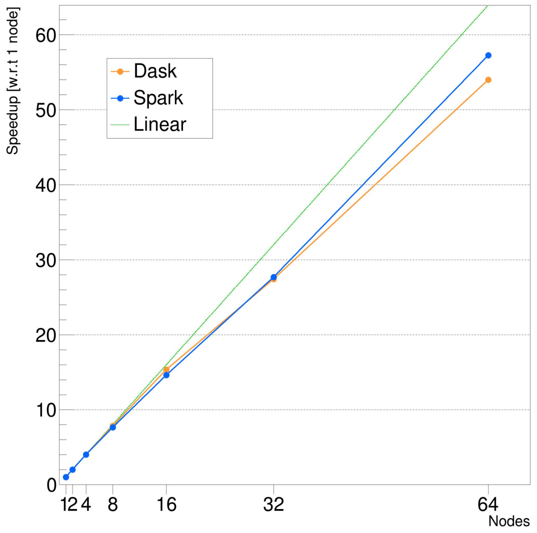
DataFrame?

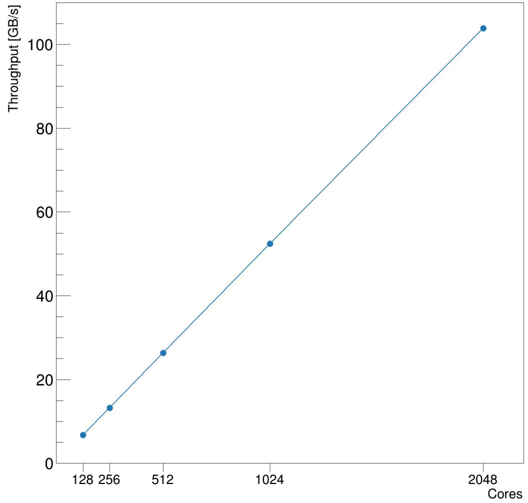
CERN HPC
Slurm jobs (Spark/Dask)
~100 GB/s on 2048 cores

V.E. Padulano et al,
Leveraging State-of-the-Art Engines for Large-Scale Data Analysis in High Energy Physic, February 2023

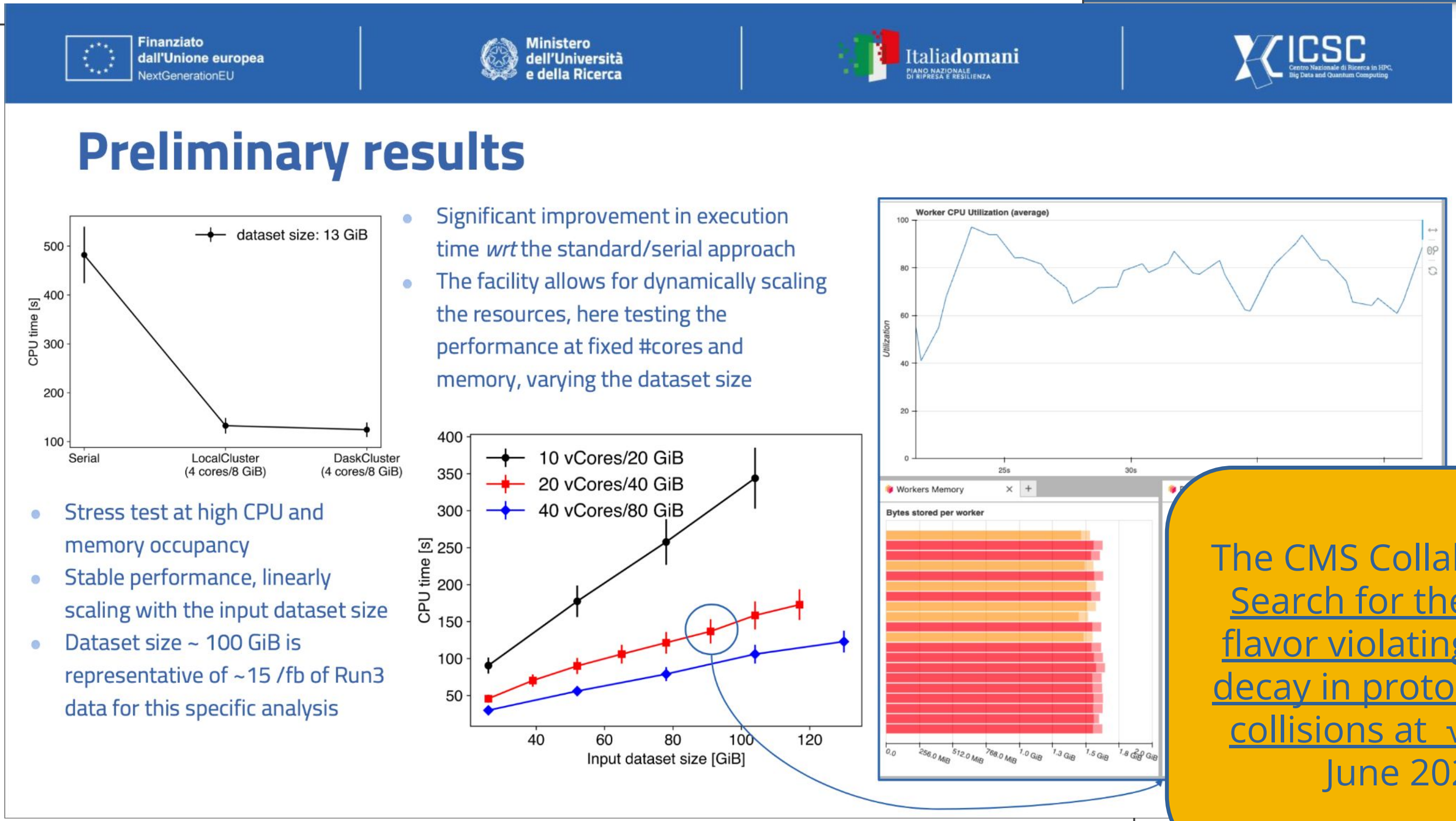
Jülich HPC
Collaboration with OpenLab
Slurm jobs (via Dask)

J. Boulis,
Benchmarking Distributed Analysis at the Jülich HPC Center, August 2023





Why RDataFrame?



The CMS Collaboration, [Search for the lepton flavor violating $\tau \to 3\mu$ decay in proton-proton collisions at $\sqrt{13}$ TeV](), June 2024

Why RDataFrame?

G. Petrucciani, Feedback from CMGRDF and distributed processing,
December 2024

The analysis used for these tests

- A complete CMS analysis on run 2 + 2022 data, at full complexity
  - correctionlib + CMSJMECalculator for applying scale factors, calibrations, uncertainties, …
  - Onnx C++ runtime to evaluate DNNs from RDataFrame
- Measurements in a rare final state (top + Z processes), so only a modest total number of events after preselection, but from many different datasets
  - Often needing different computation graphs depending on data taking era, MC sample, …

Dataset	Graphs (unique)	JITed functions	Files	Evts	Size
Just 2018	70 (43)	138	63	66 M	214 GB
Full dataset (2016-2022)	412 (248) (x 5.8)	260	362 (x 5.7)	273 M (x 4.1)	977 GB (x 4.6)

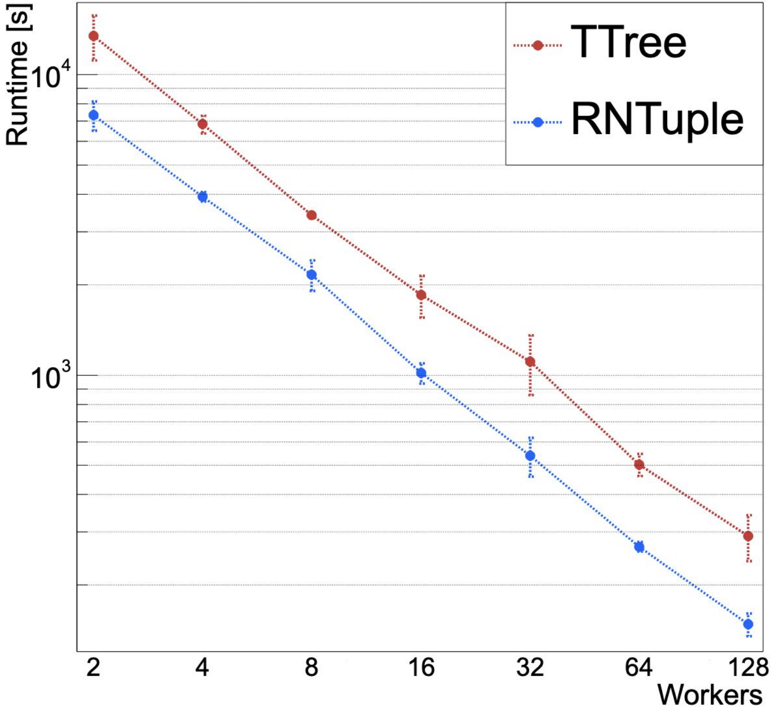
Why RDataFrame?

The Analysis Grand Challenge

RDF implementation: collaboration between ROOT and IRIS-HEP [1],[2]

Linear scaling on SWAN
RNTuple always faster

M. Czurylo et al, Distributed analysis in production with RDataFrame, October 2024



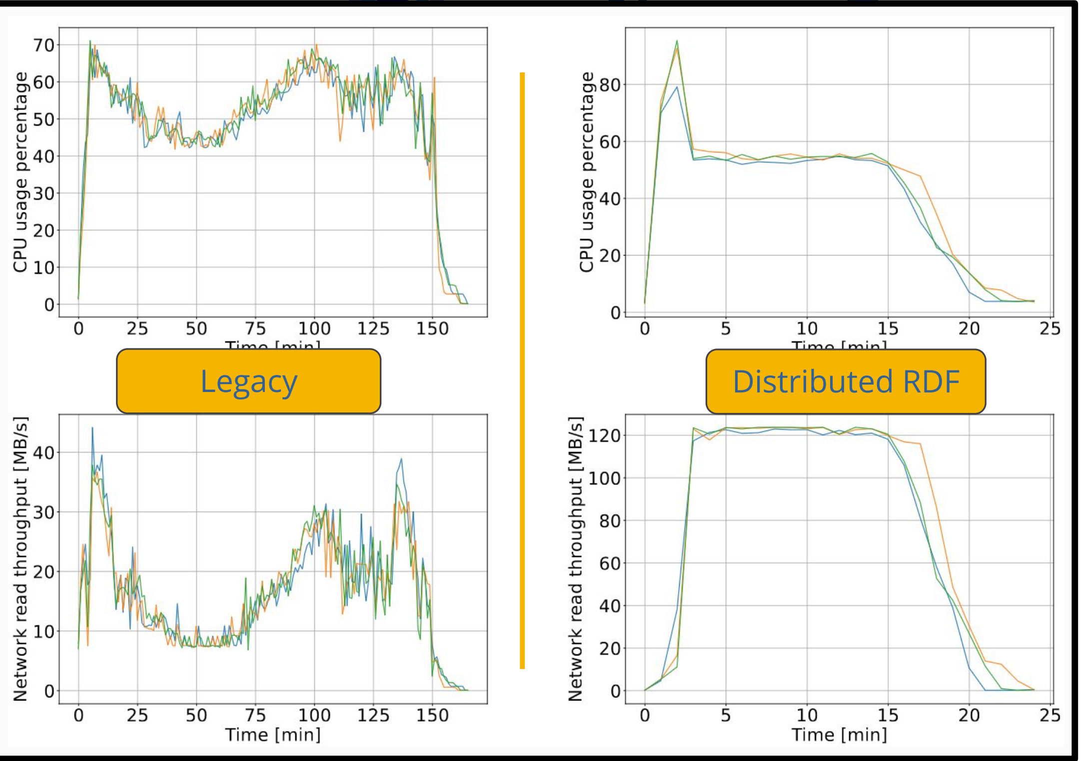
Why RDataFrame?

CMS analysis used on Analysis Facilities

Before: Python for-loop with NanoAODtools, manual job submission

After: Interactive distributed RDataFrame → O(10) speedup

T. Tedeschi et al. Prototyping a ROOT-based distributed analysis workflow for HL-LHC: The CMS use case, February 2024



2 4 8 16 32 64 128
Workers

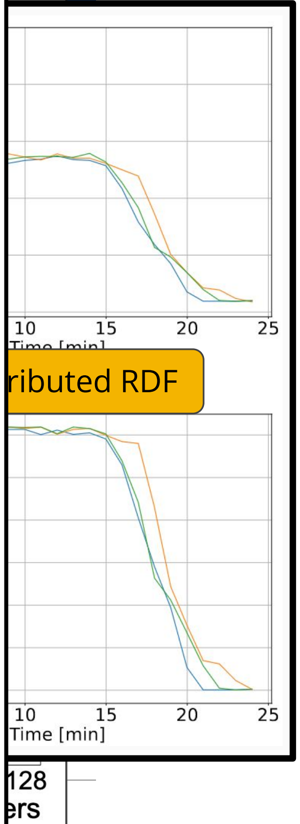
M. D'Alfonso,
CMS Analysis Facility testing with RDataFrame,
March 2025

Status of my tests at various AF

	subMIT (MIT)	SWAN (CERN)	EAF (Fermilab)	Purdue	coffea casa (UNL)
access	MIT account	SSO cern	account + VPN	SSO cern	SSO cern
software stack	CONDA myenvAF	'Bleeding Edge' 'lcg/views/LCG_105a_swan'	CONDA client Singularity on workers	CONDA myenvAF	CONDA myenvAF
xcache		not planned	not planned		
interactive Local Cluster (one node)					
Distributed (multi node)	DASKGateway+ SLURM	Spark + HTcondor	DASKGateway + HTCondor	DASKGateway + K8s	DASkGateway + K8s
	DASK + SLURM	DASK + HTcondor			

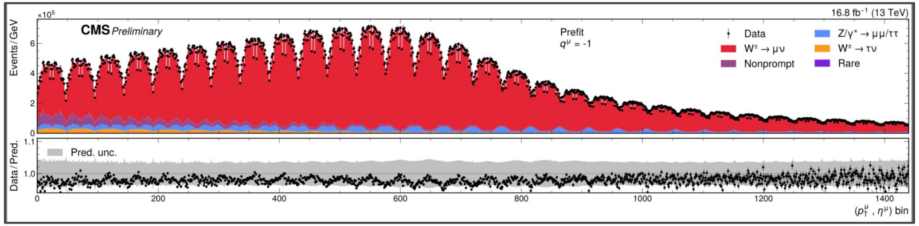
in progress
work
didn't work
not tried

M.D'Alfonso (MIT) 23



Why RDataFrame?

J. Bendavid, CMS W Mass analysis with RDataFrame, October 2024



Summary

Increasing amount of data opens new opportunities

- Software developments must be ahead to fully exploit potential

Fast analysis turnaround was essential for this complex measurement

- RDF provides a convenient and efficient library
  - Initially showstoppers observed in scaling
  - Extensive work on critical parts to improve RDF and histogram implementation
- Full analysis runs in ~hours

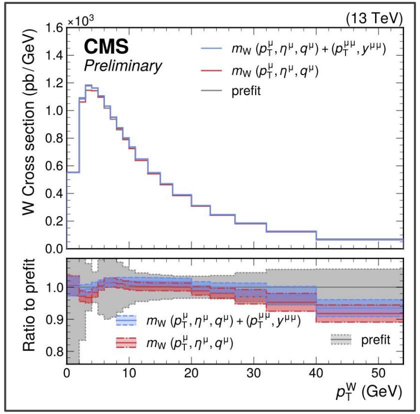
ROOT RDataFrame

RDataFrame → ~1-2 h → Boost Histograms

Critical parts in c++

- Functions: e.g. check if reco muon has a match to any gen muon

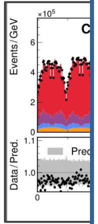
```
bool hasMatchDR2(const float& eta, const float& phi, const Vec_f& vec_eta, const Vec_f& vec_phi, const float dr2 = 0.09) {

  for (unsigned int jvec = 0; jvec < vec_eta.size(); ++jvec) {
    if (deltaR2(eta, phi, vec_eta[jvec], vec_phi[jvec]) < dr2) return true;
  }
  return false;

}
```



Why RDataFrame?

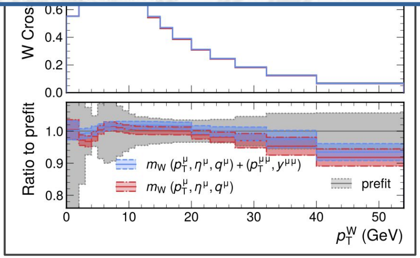
J. Gooding - RDataFrame in LHCb Offline Selection, October 2023

- RDataFrame provides **notable speedup** of offline selection, *if implemented optimally*:
  - *Potential speedup of* $\mathcal{O}(10)$ in some cases.
- uproot+numba **matches RDataFrame performance** in many scenarios:
  - I/O bottleneck identified → performance improvement possible.
- Both tools suitable for **fast parallelised selection**, necessary for upcoming Run 3 analyses.



- Functions: e.g. check if reco muon has a match to any gen muon

```
bool hasMatchDR2(const float& eta, const float& phi, const Vec_f& vec_eta, const Vec_f& vec_phi, const float dr2 = 0.09) {

  for (unsigned int jvec = 0; jvec < vec_eta.size(); ++jvec) {
    if (deltaR2(eta, phi, vec_eta[jvec], vec_phi[jvec]) < dr2) return true;
  }
  return false;

}
```



Why RDataFrame?

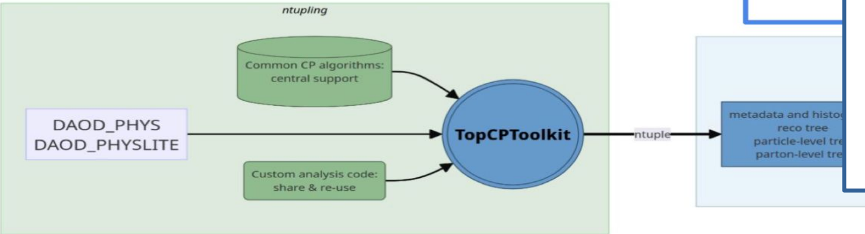
T. Dado, [ATLAS FastFrames](), February 2025

stFrames performance

- **Tested with a realistic example**
  - TTbar single lepton sample with ~ **20M events** (200M events in the DAOD)
    - ~ 400 GBs
  - Producing many control histograms
  - Producing analysis specific histograms
  - Running on all (automatic) systematics: ~400 systematics
  - Total of >25k histograms (total in three regions)
- **Ran on Chicago analysis facility**
  - systematics run, **25k+ histograms**, 10 GB ram requested, **10 CPUs**, running time: **10h 12m**
  - systematics run, **25k+ histograms**, 20 GB ram requested, **50 CPUs**, running time: **2h 52m**
  - systematics run ~**20k histograms**, 20 GB ram requested, **50 CPUs**, running time: **2h 20m**
  - nominal only run, **60 histograms**, 20 GB ram requested, **50 CPUs**, running time: **< 2m**
- **HTCondor**
  - nominal only run, ~**4k histograms**, ~**250 regions**, **30 CPUs**, running time: **< 16 m**
    - ~ **66M events**

Introduction

- **What is FastFrames (FF) ?**

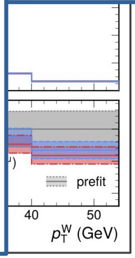


FF

  - Is a framework to produce **histograms/ntuples.**
  - Compatible with the outputs produced by frameworks using **ATLAS format**
  - Build around **ROOT::RDataFrame** technology.
- **What can FF do ?**
  - Run in Linux and MacOS.
  - NTuples → **NTuples.**
  - NTuples → **Histograms.**
    - Supports 1D, 2D, 3D and TProfile
  - Process **reco and truth trees** – including event matching.
  - Prepare inputs for **unfolding** – which can be passed to [TRExFitter]() directly.

s analyses.

26th October 2023 | 12



Schedule

- First part of the course ~ 1h
  - Notebooks 1 and 2 (I/O and RNTuple, UHI)
- Break ~ 30mins
- Second part of the course ~ 1h
  - Notebooks 3, 4 and 5 (RDataFrame, RDataLoader, RooFit)
  - Real analysis example
- Wrap up + time for Q&A ~ 30mins

Course begins

We now move to the course material on Jupyter Notebooks

A look towards the future of ROOT

Improved Python packaging

- [Conda](): all ROOT releases are published on conda-forge, for Linux and MacOS.

```
conda install conda-forge::root
```

- ROOT is published on PyPI, since version 6.38.00
  - Still in alpha mode, not ready for production
  - Try it out and let us know!

```
pip install ROOT
```

ROOT 7 histograms

Jonas Hahnfeld, ROOT PPP188

- Modern histogram package for ROOT 7
  - Well tested and high performance
  - Can be built standalone, without the rest of ROOT
- Key features compared to `TH*`
  - User-defined bin content types → template class
  - Still not go overboard with templates (Boost.Histogram, old prototype)
  - Arbitrary dimensions with dynamic axis types at run-time
    - See presentation at previous PPP: https://indico.cern.ch/event/1403741/
  - Efficient concurrent filling (memory and performance)
  - Histograms as data structures, no drawing and fitting

ROOT 7 histograms

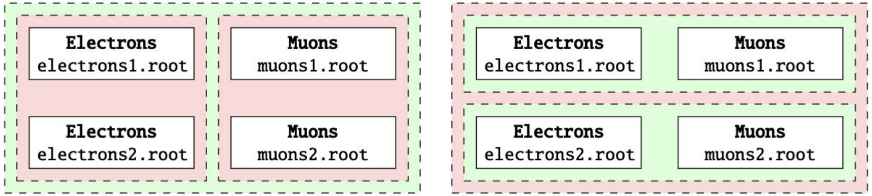
- RHistEngine::FillAtomic
  - Low contention without global statistics
  - No change of bin content type!
  - Keep full performance of sequential filling
- Concurrent filling of RHist:
  - Wrap histogram in RHistConcurrentFiller
  - Create RHistFillContexts that efficiently accumulate RHistStats

Jonas Hahnfeld, ROOT PPP188

Check out the tutorials at https://root.cern/doc/v638/group__tutorial__histv7.html

RNTuple tooling: RNTupleProcessor

- Event iteration API
- Enabling dataset compositions (chains and joins)
- An evolution of TTree friends, fit for more complex analysis requirements
- Tutorials available [1], [2]



F. de Geus, CHEP'24

RFile

- A new ROOT file interface
- No changes to on-disk format
- Modern, minimal interface
- Find tutorials [online]

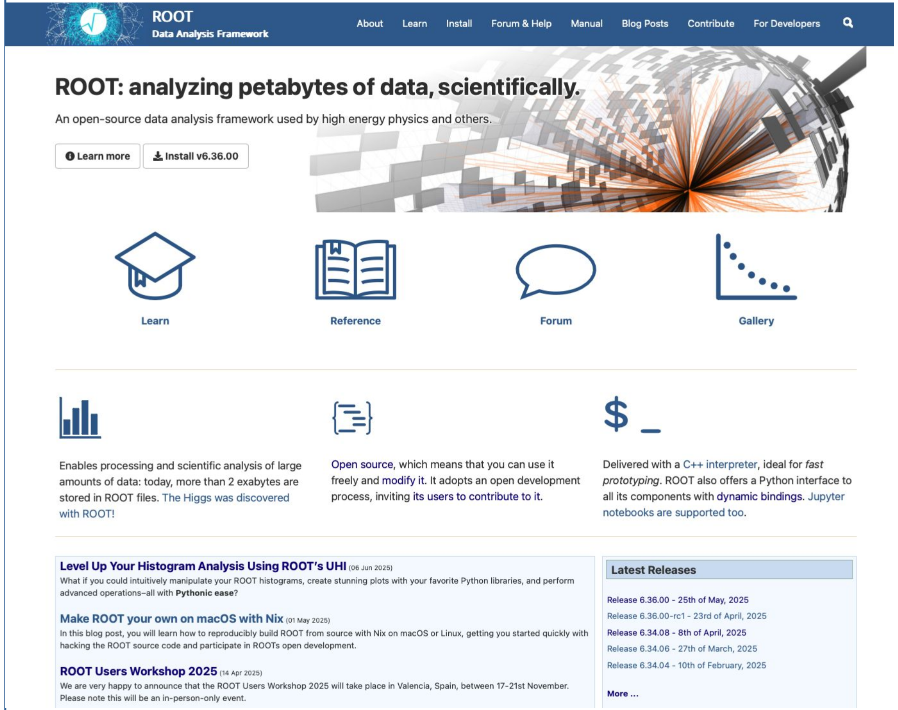
```python
# Reading
with RFile.Open("file.root") as file:
  # From top-level dir:
  hist = file.Get("hist")

  # From sub dir:
  subhist = file.Get("my/path/hist")

# Writing
with RFile.Recreate("file.root") as file:
  # To top-level dir:
  file.Put("hist", hist)

  # To sub dir:
  file.Put("my/path/hist", hist)

# Querying
for key in file.ListKeys("my"):
  print(key.GetPath())
```

Wrap up

Other examples of Analyses with RDF

- RDataFrame implementation of Analysis Grand Challenge (CMS Open Data)
  https://github.com/root-project/analysis-grand-challenge
- LHCb example:
  https://github.com/siliataider/root-lhcb-open-data-analysis

- ROOT web site: **THE** source of information for ROOT users
  - For beginners and experts
  - Installation instructions
  - Class documentation
  - Learn section
  - Access to the forum
  - Blog about recent developments



Resources

- ROOT Website: https://root.cern
- ROOT Forum: https://root-forum.cern.ch
- Materials for learning more: Learn - ROOT
- Documentation: https://root.cern/doc/master/
- Github: https://github.com/root-project/root