Adding RNtuple to the Analysis Grand Challenge

Applicant: Giedrius Juškevičius

Mentor: David Lange

Project Duration: 8 weeks

Proposed start date: 29 July 2024

Project Context and Description

The High-Luminosity Large Hadron Collider (HL-LHC) project aims to crank up the performance of the LHC in order to increase the potential for discoveries after 2029 [1]. However, analysis workflows commonly used at the LHC experiments do not scale to the requirements of the HL-LHC [2]. To address this challenge, the IRIS-HEP software institute started the "Analysis Grand Challenge" (AGC) project. It includes a pipeline including a binned analysis, reinterpretation and end-to-end optimization of a physics analysis use cases. It also includes the development of the required cyber-infrastructure to execute them in order to demonstrate technologies envisioned for HL-LHC [3].

The LHC experiments use detectors to analyze the myriads of particles produced by collisions in the accelerator. The biggest of these experiments, CMS (Compact Muon Solenoid) and ATLAS, use general-purpose detectors to investigate the largest range of physics possible [4]. Due to differences in the sensors, separate data analysis workflows must be used in the AGC software. For this project, we will focus specifically on optimizing the CMS workflow [5]. At the time of this proposal, the most compact data tier format available in Open CMS data is NanoAOD (Nano Analysis Object Data) [6, 7], which will be our main focus for optimization. AOD data is derived from the RECO information to provide data for physics analysis in a convenient, compact format and is usable directly by physics analyses [8].

A NanoAOD file contains a main TTree named events. Since its inception, the ROOT project supports the columnar storage of arbitrary C++ types and collections through TTree. However, the expected increase in the amount of experiments data that needs to be processed and the fact that TTree was not designed to make optimized use of modern hardware and storage systems, called for a new, modernized re-engineering of TTree [9].

RNTuple is the new, still under development, backward-incompatible ROOT columnar I/O subsystem targeting high performance, reliability, and easy-to-use robust interfaces [9]. In its current state, RNTuple can read and write data with an API similar to the TTree one, including bulk reading. RNTuple's supported type system is more limited than TTree's, yet already powerful enough to represent, for instance, CMS NanoAODs, and ATLAS PHYS, PHYSLITE and full AOD files. The ATLAS and CMS frameworks have included experimental RNTuple models in their integration builds [10]. In this project, we will use these models to convert data from TTree format to RNTuple format, aiming to optimize data processing capabilities.

Solution

In this project, our focus will be on analyzing existing Python, ROOT, and C++ code within the AGC/CMS repositories. Our main goal is to transform TTree event data structures into RNTuple formats to enhance pipeline efficiency. By the project's conclusion, optimized pipeline components will be delivered to the AGC repository, complete with thorough documentation.

Goals

1. Study AGC code: understand it and learn how to run it;
2. Convert some NanoAOD input data to RNtuple;
3. Understand and implement what pieces of the AGC need to be changed;
4. Test and evaluate performance to check how faster the pipeline is with RNTuple;
5. Integrate code into AGC repository and write/update documentation.

Project Timeline

Time frame	Objectives
Week 1	• Study AGC code: understand it and learn how to run it.
Weeks 2-4	• Convert some NanoAOD input data to RNtuple. • Understand and implement what pieces of the AGC need to be changed.
Weeks 5-6	• Test and evaluate performance to check how faster the pipeline is with RNTuple.
Weeks 7-8	• Integrate code into AGC repository. • Write/update documentation.

References

[1] CERN Website, “High-Luminosity LHC” [https://www.home.cern/science/accelerators/high-luminosity-lhc].
[2] A. Held, O. Shadura, “The IRIS-HEP Analysis Grand Challenge” (2023), [https://pos.sissa.it/414/235].
[3] IRIS-HEP Website, “The Analysis Grand Challenge”, [https://iris-hep.org/projects/agc.html]
[4] CERN Website, “Experiments”, [https://home.cern/science/experiments].
[5] IRIS-HEP Github Repository, “analysis-grand-challenge/…/ttbar_analysis_pipeline.ipynb”, [https://github.com/iris-hep/analysis-grand-challenge/blob/main/analyses/cms-open-data-ttbar/ttbar_analysis_pipeline.ipynb].
[6] CERN TWiki Website, “The CMS NanoAOD data tier”, [https://twiki.cern.ch/twiki/bin/view/CMSPublic/WorkBookNanoAOD].
[7] M. Peruzzi, G. Petrucciani, A. Rizzi, “The NanoAOD event data format in CMS” (2020), [https://iopscience.iop.org/article/10.1088/1742-6596/1525/1/012038].
[8] CERN TWiki Website, “Analysis Overview: an Introduction-> AOD” [https://twiki.cern.ch/twiki/bin/view/CMSPublic/WorkBookAnalysisOverviewIntroduction].
[9] J. Lopez-Gomez, J. Blomer, “RNTuple performance: Status and Outlook” (2023), [https://iopscience.iop.org/article/10.1088/1742-6596/2438/1/012118/pdf].
[10] J. Blomer, P. Canal, F. de Geus, J. Hahnfeld, A. Naumann, J. Lopez-Gomez, G. Lazzari Miotto, V. E. Padulano, “ROOT’s RNTuple I/O Subsystem: The Path to Production” (2024), [https://www.epj-conferences.org/articles/epjconf/pdf/2024/05/epjconf_chep2024_06020.pdf].