IRIS-HEP Summer 2023 Fellow Program Project Proposal

Analysis Grand Challenge with ATLAS PHYSLITE data

Applicant: Denys Klekots

Mentors: Alexander Held, Matthew Feickert, Vangelis Kourlitis

Project Duration: 12 Weeks

Proposed start date: 3 June 2024

Project description

This project aims to be done in the area of software for high-energy physics at the ATLAS experiment of the Large Hadron Collider. The collider experiments generate an astonishing amount of data providing good statistics and the possibility to detect rare physics events. As the LHC gears up for its High-Luminosity upgrade, a corresponding evolution is essential for the software tools used to analyse ATLAS data. Efficiently processing and storing this massive dataset becomes paramount. In favour of this requirement, the PHYSLITE [1] data format was developed by the ATLAS collaboration and is a fundamental element in reducing data storage resource needs.

This project envisions a student-driven exploration of the ATLAS data measured in 2015, which is going to be openly published in PHYSLITE data format soon. The goal is to develop a version of the Analysis Grand Challenge (AGC) [2], focusing on top quark pair production analysis, using the Scientific Python ecosystem and the Pythonic tool that has been developed as part of IRIS-HEP. This AGC implementation will serve as a benchmark to assess analysis code performance and pinpoint crucial aspects of the scientific Python ecosystem libraries requiring improvements to efficiently analyze PHYSLITE data. A similar AGC implementation has been previously developed using CMS open data [3]. This provides a valuable starting point, allowing the use of the existing analysis as a reference for the ATLAS open data.

The project's nature ensures the uncovering of unpolished aspects and new challenges within the scientific Python ecosystem. These findings will be valuable in pinpointing bottlenecks and guiding future optimisations to achieve high-performing physics analyses in the HL-LHC era.

Software Deliverables

The work will be done in public by making pull requests to the Analysis Grand Challenge GitHub repository [4]. The primary deliverable will be a public and reproducible AGC implementation that runs on ATLAS PHYSLITE open data. Additional deliverables will include documentation that provides clear use instructions and process overview that will exist as Markdown files in the same GitHub directory as the AGC implementation. As the AGC relies on a large collection of scientific tools developed by IRIS-HEP and the broader scientific open-source community, problems encountered with the software will be documented as issues opened on the project's issue tracker on GitHub. Contributions to these software libraries in the form of contributed documentation or code may additionally happen if needed.

Proposed Timeline

Week 1-2: Become familiar with the PHYSLITE data format and the HEP scientific Python ecosystem of the project. Understand the interaction between the data format and the tools and their abilities, limitations and functionalities.

Week 3-4: Study the Analysis Grand Challenge implementation using CMS open data and run the implementation on IRIS-HEP resources.

Week 5-6: Development of short examples that work with the PHYSLITE open data and the IRIS-HEP Pythonic tooling, using the latest version of the data format.

Week 7-8: Implementation of the ATLAS PHYSLITE Analysis Grand Challenge.

Week 9-10: Revising all the documentation and the code to be consistent with the latest releases available and to be presented as if to a new user with no experience yet. Preparing user-friendly example code to be published on GitHub.

Week 11: Wrap up the project, prepare a markdown document in the repository with a clear outline of what issues still exist if any, user experience improvements, and technical improvements in the workflow with corresponding GitHub issues.

Week 12: Preparation of the slides and final presentation of the project.

References

[1] ATLAS collaboration. Catmore James et all. “PHYSLITE - a new reduced common data format for ATLAS”. URL: https://cds.cern.ch/record/2857821

[2] IRIS-HEP. “The Analysis Grand Challenge”. URL: https://iris-hep.org/projects/agc.html

[3] “CMS Open Data $t\bar{t}$: from data delivery to statistical inference”. URL: https://agc.readthedocs.io/en/latest/cms-open-data-ttbar/ttbar_analysis_pipeline.html#CMS-Open-Data-t\bar{t}:-from-data-delivery-to-statistical-inference

[4] “GitHub repository of the Analysis Grand Challenge”**.**
URL: https://github.com/iris-hep/analysis-grand-challenge