IRIS-HEP Project Proposal: Development of Experiment-Specific Data Schemas for Coffea

Fellow: Sam Kelson

Mentors: Lindsey Gray, Nick Smith, Matthew Feickert, Giordon Stark

Project Duration: May 20th - Aug 12th 2024

Proposal

Columnar Object Framework For Effective Analysis (Coffea) is a library designed to integrate various essential components for conducting analysis at high-energy physics (HEP) experiments, utilizing the scientific Python ecosystem. In addition, Coffea allows for analysis workloads to be efficiently scaled across computing resources. By harnessing modern distributed computing technologies, such as Dask, Coffea enables seamless scalability of HEP analysis, from initial testing on a laptop to deployment on large computing clusters, without requiring modifications to the analysis code itself. As Coffea continues to be adopted across a spectrum of HEP experiments that employ diverse data formats, maintaining experiment-specific schemas that prioritize user-friendliness and performance becomes increasingly complex.

The primary objective of this project is to address these challenges by improving the developer and user experience around Coffea schemas. Initially, the focus will be on updating the ATLAS-specific PHYSLITE schema and providing validation tests to make additional changes easier to support. This work will involve communication with the ATLAS Analysis Model Group to ensure coordination of efforts internal to ATLAS. This work will have a significant impact on Coffea-based workflows in ATLAS as well as the IRIS-HEP Analysis Systems ecosystem.

Furthermore, Coffea is focused on scaling analysis workloads across computing resources. These benchmarks will consist of tests for both CMS and ATLAS analysis loads.

Additionally, an assessment of the current Coffea onboarding documentation will be conducted. Time permitting the project will also include the development of a user-friendly system allowing users to easily customize data schemas to suit their specific analysis needs.

Project Deliverables

1. Updates to ATLAS PHYSLITE schema
2. Test suite for future data schemas
3. Community reference benchmarks for Coffea scaling
4. Improved public documentation for new developers
5. General schema improvements (time permitting)

Proposed Timeline

- Week 1 & 2: Familiarize myself with how Coffea is used. Read through the existing user and developer documentation and engage the Coffea team and mentors in discussion on Coffea's design. Run existing examples and set up the development environment.
- Week 3 & 4: Familiarize myself with how Coffea works internally, including data schema design. Additionally, learn how to use Dask to scale computational workflows.
- Week 5: Look at resources produced by the ATLAS Analysis Model Group and learn what PHYSLITE currently looks like. Then develop a test suite that any future schema should be able to pass and add PHYSLITE testing files to the scikit-hep-testdata project
- Week 6-7: Revise the current implementation of the coffea.nanoevents.PHYSLITESchema to be compatible with the implementation in recently produced ATLAS PHYSLITE files. Update and write tests for Coffea based on the new schema. Update documentation on the new PHYSLITE schema.
- Week 8-9: Write user examples for the PHYSLITE schema and write developer notes on schema development.
- Week 10-11: Create benchmark Coffea analysis cases that benefit from scaling and test their scaling performance with Dask under different configurations and compute resources.
- Week 12: Document all work and add notes on developer experience to the Coffea development guidelines and documentation. In addition, work on the final presentation to IRIS-HEP.
    - Time permitting: Begin work on designing and implementing a more general system for user implementation of custom schemas