Matous Adamec
DIANA-HEP Fellowship Proposal

Scaling Data Analysis in Coffea

This project will have two primary goals: testing the scaling capabilities of Coffea for a sample CMS physics analysis, and integrating that Coffea analysis with existing IRIS-HEP software. Coffea is a novel framework that takes a columnar approach to HEP data analysis, which is to say that all relevant data is represented in arrays, and analysis is done on those arrays in Python, pushing cumbersome operations like loops and conditionals into the background. Ultimately, Coffea is meant to function as an end-to-end analysis framework, and, as such, both its ability to scale well on various computing clusters and its ability to meld seamlessly with other analysis tools is important.

Scaling to large sets of data is a critical part of any practical analysis in HEP. For this purpose, a variety of task schedulers for efficiently processing data on clusters exist. It is essential that Coffea is able to execute analyses efficiently on any of them, so that there are no conflicts when users seek to deploy analyses on their own clusters. Currently, Coffea has backends to support clusters which employ Parsl, Spark, and Dask as schedulers. This project will focus on scale-out with Dask, as it is the latest of executors implemented in Coffea.

As part of the scaling portion of the project, the analysis will be made to work with the Dask executor. The analysis will then be optimized specifically for Dask, and performance benchmarks will be made, which shall be presented to the Coffea team to demonstrate the scheduler's strengths and weaknesses. This, in turn, will help guide Coffea's future development, and, in the short-term, may help users decide which backend they wish to use for their own analyses.

The second goal of this project will be to work towards integration of some useful IRIS-HEP tools into the analysis. This would serve as a demonstration of how Coffea can work alongside standard IRIS-HEP and DIANA-HEP software to handle a full analysis from start to finish. It is important that Coffea can cooperate with these tools without too much difficulty or adaptation since the goal of the Coffea project is to encapsulate the whole analysis process; in particular, focus will be placed on integrating data delivery and histogramming tools, both crucial aspects of most HEP analyses.

This project will be worked on from UNL, with Ken Bloom as a mentor. Regular contact with Coffea's development team at Fermilab will be maintained, with Lindsey Gray also serving as a mentor. In addition to collaboration with the development team mentioned previously, relevant milestones will be presented at biweekly Coffea user meetings.

Matous Adamec
DIANA-HEP Fellowship Proposal

Proposed Timeline

Weeks 1-2*: *Ensure analysis is prepared for scaling.

- Debug event selection, make sure all channels output accurate results.
- Normalize data.

Weeks 3-4: Implement scaling.

- Add all signal and background samples.
- Implement analysis to run on Dask executor.

Weeks 5-9: Compare efficiencies of each backend.

- Optimize the analysis for Dask. Play with alternative solutions to analysis sub-problems (i.e., some numba vs. pure columnar analysis) and see which are most efficient.
- Construct performance benchmarks for Dask executor.
- Work with developers to compare Dask, Spark and Parsl.

Week 10-12: Integrate IRIS-HEP tools and report findings.

- Integrate ServiceX and pyhf with analysis.
- Report to Coffea development team.
- Present at IRIS-HEP forum.