Coffea-Casa Analysis Facility IRIS-HEP Fellow Proposal

Goal: Contribute to the further development of the Coffea-Casa Analysis Facility (AF) at University of Nebraska-Lincoln (UNL), to expand a gallery of Coffea-Casa analysis samples with existing analysis from CMS adapted to be executed in AF@UNL. I will facilitate the use of Coffea-Casa AF for Boston University and UNL CMS physicists currently working with NanoAOD datasets and investigation of possibility to use Arrow Dataset API as an input to Coffea for further integration with Skyhook DM.

Motivations: Improving Analysis Facility's stability and efficiency is important for adapting to the new era of HL-LHC with a significant increase of data volume. Currently, I am part of a CMS analysis group led by Indara Suarez (Boston University) and Frank Golf (UNL).

I am interested in working to scale the analysis systems for future broader usage with innovative analysis techniques that are computationally efficient and storage efficient. As a computer science student and member of BU CMS group, I plan to use my knowledge to support physicist colleagues to conduct research on the Coffea-Casa AF, and provide feedback and development support to improve the system for physicists.

Research Plan: I expect to work on this project half-time for the next 6 months (equivalent to 3 full-time months), while attending BU as a full-time student Feburary - May. I am excited to join the IRIS-HEP community and work with Oksana Shadura and Brian Bockelman.

Technically, I'm interested in customizing dask for parallel computing, the integration of services (ServiceX, Skyhook, Columnservice) and the assessment of performances therein. This project would enhance my own understanding of data organization and delivery system for analysis, as well as supporting coffea users (especially those within my group).

I would pivot my deliverables to lie within the quarterly goals of the Coffea-Casa project covering 22-23 weeks in Q1/Q2 2021. During this period I will also be a liaison between UNL+BU coffea users and gather feedback for improvement and maintenance.

- First milestone would be simply converting existing analysis from BU CMS group to a notebook (4 weeks Q1 2021)
- Adopt an analysis notebook to be executed using the Dask backend from `coffea` on Coffea-casa Analysis Facility@UNL and add it in the gallery of examples for Coffea-casa project. Add performance measurements. Adopt notebook to use ServiceX service (4 weeks Q1 2021)
- Investigate a possibility of using Arrow Dataset API (https://arrow.apache.org/docs/python/dataset.html) as input/data delivery for Coffea for further integration with SkyHook DM. Add a simple example and performance measurements (6 weeks Q1/Q2 2021)
- Work on improving the `coffea_casa` python module used to build on top of Dask-Jobqueue to spawn Dask workers in the UNL Tier 2 HTCondor pool (5 weeks Q2 2021)
- Write documentation, finalize to-do items, prepare final presentation (3-4 weeks Q2 2021)