Continuous Testing of Facility's Functionality Including Data Delivery Services Available in Coffea-Casa Analysis Facility

The prototype "Coffea-casa" analysis facility (AF) concept at Nebraska provides novel capabilities for HEP analysts. The current instance, limited to CMS users, has access to the entire CMS data set through access to the global data federation and local caches. It supports the Coffea framework, which provides a declarative programming interface that treats the data in its natural columnar form. An important feature is an access to a "column service"; if a user is working with a compact data format (such as a CMS NanoAOD) that is missing a data element that the user needs, the facility can be used to serve that "column" from a remote site.

We plan to integrate into the Coffea-casa facility different data delivery services: ServiceX and Skyhook DM. ServiceX provides user-level ntuple production by converting experiment-specific datasets to columns and extracts data from flat ROOT files. The service enables simple cuts or simply derived columns, as well as specified fields. The Skyhook DM project can convert ROOT files, such as the NanoAOD format, to the internal object-store format. Instead of providing a file-like interface for reading data, the project is working to implement Ceph APIs to natively filter and project structured data, delivering it to Dask workers through the Arrow format. This object API is expected to eventually offer "join" functionality, allowing users to include their bespoke columns into a shared dataset.

The successful IRIS-HEP fellow candidate will work on the development of a continuous functionality testing procedure (including smoke tests and integration tests) for 'Coffea-casa' analysis facility. The test suite would expect to cover testing of analysis-related components and analysis frameworks deployed in AF as well as data delivery services functionality. The test suite should include but not be limited to an already collected set of available sample physics analyses.

The anticipated duration of the project is the three-month period May - July 2021. During this time, the fellow, who will work at the 0.5 FTE level, will develop skills in Python, Jupyter notebooks, git, and if possible Kubernetes in the course of developing a continuous functionality testing system.

Oksana Shadura, Ken Bloom, and Brian Bockelman will supervise the fellow. A timeline with deliverables is provided on the next page.

Timeline

weeks 1-2 Study the documentation of Coffea-Casa, ServiceX, and Skyhook DM services. Investigate how to use them together within an existing sample physics analysis.

weeks 3-4 Investigate possible mechanisms for continuous functionality testing procedure for AF.

weeks 5-7 Test ServiceX/Skyhook DM data delivery services and check performance measurements. Tune multiple parameters available in ServiceX/Skyhook DM to achieve the best timing results. Communicate results with the ServiceX/Skyhook DM team.

weeks 8-11 Add tests in continuous functionality testing suite for AF and enable testing. Collect feedback from coffea-casa the team about possible improvements.

weeks 12-13 Finish and document work, presenting results publicly.

At the end of the project, the student will present his/her work at an IRIS-HEP topical meeting.

In addition to this project, the student will be working in a chemistry lab at UNL and as such will only be able to work half-time. They anticipate working a minimum of 15 hours, ideally 20, each week.