IRIS-HEP Fellowship Proposal

Durbar Chakraborty
National Institute of Technology, Durgapur

Project: Metrics to define user activities and engagement on the various coffea-casa Analysis Facility deployments
Duration- May, 2022 to September, 2022 (4 Months)
Mentors- Oksana Shadura, Ken Bloom, Brian Bockelman

Proposal:

Coffea is one of the major Python packages developed by the IRIS HEP Analysis Systems focus area, and one of the tools developed, aiming to interact and operate on high energy physics (HEP) data analysis. It is a columnar object framework for analysis that works to improve large data and data analysis tools from Python to generate an array-based syntax for manipulation of HEP event data. Coffea can be deployed and run on the Coffea-Casa Analysis Facility. Coffea-casa serves as a prototype of analysis facility which provides services for 'low latency columnar analysis', enabling rapid processing of columnar data in an easily scalable environment.

My proposed project work would involve defining a set of various user engagement metrics based on the data collected from several platforms including JupyterHub, and similar AF tools. For the purpose of collecting data for future telemetry, we can also make use of the underlying Kubernetes infrastructure. We can also power the Jupyter Notebooks by Elasticsearch for the purpose of data collection of the various metrics. Once we have developed the various user engagement metrics, we will switch to developing a data collection infrastructure for them using telemetry/data-monitoring tools (e.g. Prometheus) which will help us centralize and store the gathered metrics efficiently. Once we have managed to construct such an efficient infrastructure, we will be developing a data visualization dashboard for it using visualization applications (e.g. Grafana, Kibana) for the purpose of easy monitoring. A possible means for the project can be using the ELK/Elastic Stack, but the specific tools are flexible and will be subject to their efficacy under the present constraints for the project.

For the purpose of further improvement on the facility infrastructure, I will be working for a month over the proposed timeline on a part-time basis to ensure the satisfactory completion of the project.

Milestones and Proposed Timeline:

The milestones are essentially flexible and are meant to provide an overall outline which we will try to adhere to, but each specific milestone can get stretched or may shrink, depending on the types of problems faced while tackling them, if any.

1. **Week 1**
   Getting familiar with the Coffea-casa codebase and looking into the JupyterHub Notebooks. Try to get familiar with Prometheus and Grafana documentation.

2. **Week 2-3**
   Test user experience on coffea-casa analysis facility and get familiar with different components from coffea-casa AF infrastructure. Noting the test cases and sample instances which we will be working on, based on the Notebooks and chalking out the set of engagement metrics that we will be developing, which will be what we will work on for the subsequent weeks

3. **Week 4-5**
   Working on defining the user engagement metrics that we had noted in the previous week. Try to understand how to use Prometheus and Grafana hosted at UNL infrastructure.

4. **Week 6**
   Testing the development of coffea-casa AF, collecting feedback and fixing bugs, if any.

5. **Week 7-8**
   Working on developing the data collection infrastructure based on the defined user engagement metrics

6. **Week 9-12**

Working on creating the visualization dashboard for the infrastructure

7. **Week 13-14**

   Working to deploy the service to the production server and move the changes done in local deployment

8. **Week 15**
   This week will be kept as a buffer to allow working on any further developments

9. **Week 16**
   Documenting the entire project and working on the presentation