Translating analyses into prototype analysis systems

Brian O. Cruz Rodríguez

Mentor: Jim Pivarski

Analysis systems are developed tools used to analyze data with the aim of creating a better functionality and streamlining analysis preservation, reproducibility, and reuse.[1] A couple of examples of these tools are coffea and awkward-array. Coffea, short for Columnar Object Framework For Effective Analysis, is a prototype package, HEP community project collaborating with iris-hep, that makes use of uproot and awkward-array to provide an array-based syntax for manipulating HEP event data in an efficient and numpythonic way.[2] Awkward-array is a pure Python+Numpy library for manipulation and accessing complex data structures as columnar data structures, with the efficiency of Numpy arrays.[3]

This projects purpose is to use these tools to translate the Higgs boson decays to four leptons CMS open data physics analysis example [4,5] into a new prototype analysis system. We expect the physics analysis to remain consistent with this prototype, proving as an alternative to the original approach, as an analysis system with more general use, instead of one more aimed to high energy physicists. Once the prototype analysis system is established it will be compared to the original approach to see what benefits its use over the other. Furthermore, we will verify if the prototype has a reduced time-to-insight, which is possible with coffea's horizontal scaling facilities, has a greater functionality than the original approach, and that its reusable.

To balance the work with my classes during the Spring semester, we propose to work on the project starting January 11th, 2021 and finish on June 11th, 2021, working half-time for 6 full-time-equivalent months. The following list divides the work that will be done each month:

1) Familiarize myself with the tools used for the Higgs-to-four-leptons analysis example.
2) Run the Higgs-to-four-leptons analysis example with the original approach.
3) Familiarize myself with Coffea and Awkward-array.
4) Translate the analysis example using Awkward-array and Coffea semantics to produce the analysis system prototype.
5) Run the physics analysis with the prototype analysis system.

6) Compare the prototype to the original, things such as functionality, time-to-insight, workflow, and reproducibility.

References:

[1] IRIS-HEP: https://iris-hep.org/as.html

[2] Coffea-Team: https://github.com/CoffeaTeam/coffea

[3] Jim Pivarski: https://github.com/scikit-hep/awkward-array/blob/master/README.rst

[4] Achintya Rao: https://cms.cern/news/observing-higgs-over-one-petabyte-new-cms-open-data

[5] Jomhari, Geiser, and Bin Anuar: http://opendata.cern.ch/record/5500