IRIS-HEP Fellowship Proposal

Baidyanath Kundu
(Manipal Institute of Technology)

Mentor: Gordon Watts
Duration: 23th Nov 2021 - 14th Feb 2022

Reading CMS Run 1/2 miniAOD files with ServiceX and func_adl

ServiceX is a distributed, cloud-native application that extracts columnar data from HEP event data and delivers it to an analyst. The func_adl data query language is used to tell ServiceX how to extract the data (what columns, what simple cuts, etc.). The func_adl data query language has two backends that are currently part of ServiceX - one based on C++ for ATLAS data and one based on columnar processing using uproot and awkward arrays. The C++ backend currently runs only on the ATLAS binary format, xAOD. The C++ is generated in python.. This project will modify the C++ backend to also run on CMS Mini-AOD binary files (Run 1/Run 2), starting by concentrating on Run 1. The Higgs-Discovery demo will be used as a guide.

The task at hand is to create an interface similar to the existing func_adl_xAOD repository so that the user can send hierarchical SQL-like queries to a Mini-AOD backend. This project is of great importance to ServiceX and func_adl as the Mini-AOD dataset has sufficient information to serve about 80% of the CMS analysis while the disk and I/O requirements are dramatically simplified.[1]

The first step to achieving the goal is to learn more about working with Mini-AOD binary files and to monitor the interactions of func_adl with xAOD backend. This step will begin even before the start of the project. Once the similarity and differences between the two formats are known only then can the next step begin. The next step is to design the interface with minimum changes to func_adl language. This is important so as to not break already existing queries. The final step is to write tests and documentation so that the researchers can work with the Mini-AOD interface with the same confidence as they have been for xAOD. All these steps will be carried out under the supervision of Dr. Gordon Watts at the University of Washington and the timeline is given below.

---

[1] "Mini-AOD: A New Analysis Data Format for CMS." 15 Feb. 2017, https://arxiv.org/abs/1702.04685. Accessed 4 Nov. 2020.

Timeline

Week(s)	Tasks
1	Project Setup. Getting familiar with Mini-AOD. Understanding the variables needed for the Higgs Discovery analysis.
2	Design and discussion of the framework of the new interface that is common with xAOD interface.
3-5	Implementation of the common features between xAOD and Mini-AOD interface.
6	Writing tests for the common features and buffer period for squashing any bugs that might occur.
7	Design and discussion of the framework of the new interface that is dissimilar to xAOD interface.
8-10	Implementation of the said framework.
9-11	Writing the remaining tests and the documentation for the project.
12	Buffer period for squashing any remaining bugs.

The given timeline is very conservative and the project might finish before that in which case the remaining time will be used to add more features and fixing issues of func_adl.