Enabling support for MiniAOD Transformer for ServiceX Data Delivery Service

Haoran Sun
University of Washington

April. 18, 2022

Project Summary

Background

ServiceX[1] 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 (the columns, filters, etc.). The func_adl[4] data query language has two backends that are currently part of ServiceX - one based on C++ for ATLAS data and CMS data, and one based on columnar processing using uproot and awkward arrays. The C++ backend currently runs only on the ATLAS binary format (xAOD) and CMS binary format (CMS AOD).

The MiniAOD transformer is an important ingredient for the physics analysis workflow envisioned in the Analysis Grand Challenge (AGC)[3]. The AGC is a large-scale analysis benchmark which includes the binned analysis, reinterpretation and end-to-end optimization of a physics analysis, with all relevant features encountered in physics analyses at the Large Hadron Collider. The AGC makes use of the rich Python HEP ecosystem and the required cyber infrastructure for its execution, and aims to demonstrate technologies envisioned for the HL-LHC. The MiniAOD format is commonly used for physics analyses with the CMS experiment, and the Run 2 CMS Open Data release[5] made a large amount of MiniAOD files publicly available. These files will be used in AGC demonstrations.

Project Goal

This project aims to create an interface similar to the existing func_adl_xAOD repository[2], where both CMS AOD and ATLAS xAOD backends are available, such that the user can send hierarchical SQL-like queries to a MiniAOD backend. The project also includes the required modifications of the C++ ServiceX backend to allow processing the publicly available CMS MiniAOD binary files. A similar modification for NanoAOD files can be achieved if necessary tools are published by the time of this fellowship.

1 Preliminary Timeline

The anticipated duration of the project is a three-month period, June–Sep 2022, at 100% FTE. Supervision of this project will be provided by Gordon Watts, Ben Galevsky, Alexander Held and Oksana Shadura. A timeline with deliverables is provided below.

Week 1-2 (June 20 - July 3)

Study the documentation of ServiceX, understand the structure and become familiar with its usage. Investigate the structure of MiniAOD files.

Week 3-4 (July 4 - July 17)

Design and start with the implementation of features required for MiniAOD that closely follow the existing AOD implementation.

Week 5-7 (July 18 - Aug 7)

Continue with implementation, and add features specific to MiniAOD without close AOD equivalent.

Week 8-10 (Aug 8 - Aug 28)

Writing tests and finishing documentation for the MiniAOD backend. Test the functionality with existing CMS Opendata (available in MiniAODs) in the context of Analysis Grand Challenge related notebooks.

Week 11-12 (Aug 29 - Sep 11)

Overflow to finish outstanding work and documentation. Develop small notebooks with examples used for demonstration. At the end of the project, present work at an IRIS-HEP topical meeting.

References

[1] KyungEon Choi et al. "Towards Real-World Applications of ServiceX, an Analysis Data Transformation System". In: *EPJ Web Conf.* 251 (2021), p. 02053. DOI: 10.1051/epjconf/202125102053. arXiv: 2107.01789 [physics.ins-det].

[2] IRIS-HEP. *func_adl_xAOD*. https://github.com/iris-hep/func_adl_xAOD/tree/master/func_adl_xAOD. 2022.

[3] IRIS-HEP. *Grand Challenge*. URL: https://iris-hep.org/grand-challenges.html.

[4] Mason Proffitt and Gordon Watts. "FuncADL: Functional Analysis Description Language". In: *EPJ Web Conf.* 251 (2021), p. 03068. DOI: 10.1051/epjconf/202125103068. arXiv: 2103.02432 [physics.data-an].

[5] Achintya Rao. *First CMS Open Data From LHC Run 2 Released*. URL: https://cms.cern/news/first-cms-open-data-lhc-run-2-released.