Verification of the Fidelity of ServiceX Data Set Extractions Using Python and C++

David Liu* and Gordon Watts
Department of Physics, University of Washington
Seattle, WA 98195, USA
(Dated: May 7, 2020)

ServiceX provides delivery of any requested portion of a high energy physics data set for analysis. Typically, particle colliders generate data sets which can be as large as several terabytes; CERN alone generated 88 petabytes of data in 2018. However, physicists are typically only interested in small portions of this data at any one time. ServiceX enables physicists to download only the relevant portions of data sets without the need for downloading the full, cumbersome data, thereby hastening analysis and reducing bandwidth costs for delivery of data worldwide. ServiceX is capable of being driven by any language with a WebAPI on the front end. However, the backend functions by using the func_adl language to extract data from xAOD and TTree ROOT files. It then transforms them for analysis with Python or C++ [1], but the transformers for each type of file were developed independently. These transformers share some common features as well as some distinctions. As a result, it is currently unclear whether, when presented with queries for identical portions of data, the data delivered by ServiceX is identical between the two transformers. This is problematic for analysis, since it suggests calls of the data set may result in different results depending on which transformer was used.

In this work, we propose the development of testing software in order to verify the functionality of the two conversion processes. This work will be conducted from May to August of 2020, with our proposed timeline found in Table 1. The development of this software will be important to the future life of ServiceX, as it will enable researchers to proceed with confidence in the fidelity of the data extractions performed by the software. This project will be remotely conducted and will be overseen by Dr. Gordon Watts at the University of Washington, with the objective of verifying that identical data queries passed through ServiceX using either Python or C++ produce the same outputs by August. If this task is completed before the project end date, secondary objectives will be to begin implementing new features and analysis tools in the Python and C++ versions of ServiceX. Completion of the proposed project will fulfill an IRIS-HEP objective of aligning the func_adl Python and C++ backends.

Week	*Activity*
1	Project setup. Familiarization with using currently existing tests on both backends.
2	Design and discussion of a common test framework compatible with both backends.
3-4	Implementation of common testing suite and validation of its functionality.
5-6	Exploration of current features in backends and implementation of tests.
7-8	Porting features to exist in both backends. Establishment of Github features.

TABLE I. Anticipated timeline for the proposed project.

[1] B. Galewsky, R. Gardner, L. Gray, M. Neubauer, J. Pivarski, M. Proffitt, I. Vukotic, G. Watts, and M. Weinberg, ServiceX: A distributed, caching, columnar data delivery service (2019), 24th International Conference on Computing in High Energy and Nuclear Physics.

* Also at Department of Physics, Ohio State University, Columbus, OH 43210, USA.