Contribution ID: **53** Type: **Oral**

# ServiceX –A Distributed, Caching, Columnar Data Delivery Service

*Monday, 4 November 2019 11:00 (15 minutes)*

We will describe a component of the Intelligent Data Delivery Service being developed in collaboration with IRIS-HEP and the LHC experiments. ServiceX is an experiment-agnostic service to enable on-demand data delivery specifically tailored for nearly-interactive vectorized analysis. This work is motivated by the data engineering challenges posed by HL-LHC data volumes and the increasing popularity of python and Spark-based analysis workflows.

ServiceX gives analyzers the ability to query events by dataset metadata. It uses containerized transformations to extract just the data required for the analysis. This operation is collocated with the data lake to avoid transferring unnecessary branches over the WAN. Simple filtering operations are supported to further reduce the amount of data transferred.

Transformed events are cached in a columnar datastore to accelerate delivery of subsequent similar requests. ServiceX will learn commonly related columns and automatically include them in the transformation to increase the potential for cache hits by other users.

Selected events are streamed to the analysis system using an efficient wire protocol that can be readily consumed by a variety of computational frameworks. This reduces time-to-insight for physics analysis by delegating to ServiceX the complexity of event selection, slimming, reformatting, and streaming.

## Consider for promotion

Yes

**Primary authors:** GALEWSKY, Benjamin; VUKOTIC, Ilija (University of Chicago (US)); WEINBERG, Marc Gabriel (University of Chicago (US)); GARDNER JR, Robert William (University of Chicago (US))

**Presenter:** GALEWSKY, Benjamin

**Session Classification:** Track 4 –Data Organisation, Management and Access

**Track Classification:** Track 4 –Data Organisation, Management and Access