A Functional Declarative Analysis Language in Python

Mason Proffitt, Emma Torró, Gordon Watts
(University of Washington, Seattle)

Introduction

- The High Luminosity LHC (HL-LHC) will collect ~10 times as much data as all the LHC runs in 2010-2023
- New data formats, but we can’t rewrite our software for each new backend
- The challenge that we’re trying to address at UW:
  - proof-of-concept project demonstrating a query-based analysis interface: functional declarative analysis language in Python
  - Github repository: https://github.com/gordonwatts/BDTTrainingAnalysisLanguage/

IRIS-HEP

- Institute for Research and Innovation in Software for High Energy Physics
- Established in 2018 to close the gap between LHC computing needs and expected resources and to optimize the interfaces used by physicists for data analysis
- It incorporates six different focus areas:
  - DOMA (Data Organization, Management and Access)
  - SSC (Software Sustainability Core)
  - SSL (Scalable Systems Laboratory)
  - OSG (Open Science Grid)
  - Innovative Algorithms
  - Analysis Systems

In large HEP experiment, each team must:

- Extract the Analysis Object Data (AOD) from the storage system
- code all the infrastructure to run over it
- plots, systematics, tables
- x several times

Functional-ADL

- Declarative Analysis Language based on queries. Implementation goals:
  1) Design a syntax that matches how physicists think about event data
  2) Run on different back-end formats. Currently implemented to varying degrees for:
    - xAOD’s from ATLAS;
    - flat TTree using columnar data in python (awkward array, uproot) and RDatFrame.
- Model being used in a full Run 2 analysis in ATLAS.
- Benefits for the LHC community:
  - Analysis preservation that goes beyond the lifetimes of experiments
  - Facilitating the abstraction, design, validation, combination, interpretation and overall communication of the contents of LHC analyses
  - SelectMany function changes sequence from event into jets
  - Generate the tuple of data we want to write out
  - Keep only jets with pt > 40 GeV
  - Turn it into a pandas data frame
  - Execute

Analysis Language: from the AOD to the plots and tables

Declarative Analysis Language (ADL):

- The physicist specifies what they want to do rather than how to implement it
- No explicit loops

In [9]:
In [10]:
In [11]:

Analysis Systems Scope

Figure by G. Watts

Analysis Systems, analysis & declarative languages (underlying framework)

Archiving, publication, metadata, etc.

Analysis Database
- Search
- CAYANNE/PERSEUS
- Conda-Forge ROOT

Figure 1: Projected ATLAS computing demands (blue) versus available resources with a flat annual budget (red)

Figure 2: Schematic of prototype analysis interface

https://iris-hep.org/as.html
https://github.com/gordonwatts/BDTTrainingAnalysisLanguage/blob/master/examples/xaod-3-FilteringTracks.py