Analysis benchmarks discussion

Allison Hall (US Naval Academy - CMS)
Nicole Skidmore (Manchester - LHCb)
Teng Jian Khoo (HU-Berlin - ATLAS)

What are the IRIS-HEP benchmarks?

github.com/iris-hep/adl-benchmarks-index

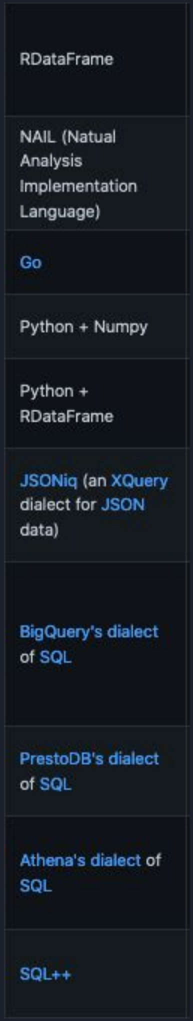
“...list of common agreed-upon benchmark analysis tasks that can be used to exemplify, test, and compare different languages and approaches used for analysis”

Eg. *Plot the E_T^{miss} of all events*

Original benchmark discussions and tasks organised HSF but work picked up and pushed through by IRIS-HEP -> currently 10 implementations ---------->

- Many tools now exist that do similar things
- These benchmarks are an excellent way to advertise new tools/packages
- Can act as a great source of documentation

Consider implementing these benchmarks in your analysis framework and contributing to the GitLab repo!



RDataFrame
NAIL (Natual Analysis Implementation Language)
Go
Python + Numpy
Python + RDataFrame
JSONiq (an XQuery dialect for JSON data)
BigQuery's dialect of SQL
PrestoDB's dialect of SQL
Athena's dialect of SQL
SQL++

Small vs. big benchmarks

Big benchmarks: IRIS-HEP Grand Challenge

- A full analysis chain using OpenData. Looking for:
  - Chaining all pieces of an analysis together including handling of systematics
  - Integration tests for all the software tools required
  - Scalability on analysis facilities
  - Upcoming workshop, Nov 4-5 https://indico.cern.ch/e/agc-tools-workshop

Small benchmarks: IRIS-HEP ADL benchmarks

- Current ADL benchmarks compare how different languages/tools achieve specific, isolated tasks.
  - Simplicity/usability for analyst - how many lines of code are required
  - Timing - CPU/event
  - Qualitative comparisons in https://arxiv.org/pdf/2104.12615.pdf

ADL - Analysis Description Language

Analysis facility - "Tier 2" type site dedicated to analysis (CPU/GPU farm, user disk space, software environments)

Expanding the current ADL benchmarks

Current benchmarks are very ATLAS/CMS centric (for good reason). But it would be good to expand horizons

- Fitting tasks
    - For the b-factories multi-dimensional fitting (eg. amplitude analysis) is the most significant "benchmark-able" task that dictates all other tools used in an analysis
    - Large number of fitters that could be compared, Roo/GooFit, zfit, TensorFlowAnalysis + numerous Minuit based institute-spawned fitters
    - Unit tests *eg. value of PDF(x, y, z)* would be very beneficial
    - LHCb run 1 data to be released at the end of the year
- Testing interface between eg. selection and fitting frameworks. What metric could be used to quantify this?
- Benchmark for updating event database - eg. adding a branch to a tuple
- Systematics benchmarking - is it clear how O(100) uncertainties can be managed in these frameworks?

Discussion

What would you like to see from these benchmarks?

What factors influence the software/tools an analyst chooses to use?

What would you need to see to make you change the software/tools you currently use?

Please also add to the live notes if you have ideas afterwards :)