Analysis Ecosystem Part 2

Context

- Last Analysis Ecosystem Workshop
  - Agenda Development
  - Agenda
  - Report
- Snowmass
  - White paper deadline: May/June/July
  - Inputs to the US process would be great to have!

Plan

The plan for the workshop should revolve around answering a few questions

1. Workshop
2. Write Report

Questions

Questions are meant to drive some of the themes in the workshop and, perhaps, sections or outlines in the report.

How do we make ML a first-class citizen?

Many new tools are coming online to aid in working with Machine Learning. There are two main frameworks for training - PyTorch and TensorFlow - and some new toolkits that are rapidly gaining popularity - JAX; CLAD.
How do we make sure ML is a well supported part of our workflow?

- Workflow: from training to production
  - Access to specialist hardware? It's a coherent platform
- Applications: User Analysis, Reconstruction Programs, Hardware-level Trigger (fastML)
- Easily getting training data into the proper format(s)
- Support of tools like KubeFlow?

What is an Analysis Facility [from the POV of an Analyzer]?

We know a lot more about what we need now. For example, we have information of how data is stored and how much will be stored. We need to answer questions like:

- How many cores to make it reasonable?

    - Access to special resources (GPU farms, etc.)
- Guess of # of users to support for a given size
- Authoring and sharing environments
- Data access model (remote access, local caches)
    - data delivery services
- Scalability for interactive analysis workflows
- How to integrate new interactive analysis workflows with existing one
- Dataset sizes
    - And how to access them efficiently (event stores?)
- Access methods/Interfaces (notebooks, vscode, ssh, “offline” submission)
    - User friendly (including authentication)
- Interaction methods: build software, just run analysis?
- Resource usage: Don’t make wasting resources easy; give feedback about resource usage to users

Need to work with facilities and analysis people to make sure the outcomes are targeted and concise

Way Forward For All Analyses?

There are analyses we can run on nanoAOD and PHYSLITE, and ones we cannot. We should understand what that means.

- What are the data volumes for these analyses?
    - Trains and carousels?
- Are they shared datasets between many analyses (miniAOD and AOD) or specifically created per analysis?
    - Cross reference between data tiers/types? Friend trees.
    - Efficient columnar reading to “mash-up” event content?

Will be extremely experiment specific, so need expt. inputs

Bridges And Ferries

We’ve been working on this for a few years now and made a lot of progress. This is a chance to congratulate ourselves as well as look forward:

- Where have we done it well (and see adoption)?
- What needs to be worked on?
    - RNTuple
    - C++ and Python integration
        - Moving “complex” algorithms in analysis as private functions
    - Do we need additional data formats in the community?
        - People are using hdf5, likely to use parquet
        - In memory formats - Apache Arrow?

Do we remember what “bridges” were and what “ferries” were? The words stayed around, but not the fine distinctions.

- Graeme - AFAICR it was Oli that understood the distinction. The phrase doesn't even appear in the Amsterdam workshop report, only "bridges"! Could we just call this *data format conversion / data access* now?

What about Julia?

- unroot exists, to map ROOT into Julia (without copying?)

Declarative Languages - Where are we? It's UX! Analysis Ergonomics

"Where are we?" and "What does it mean now?" The goals have evolved.

- Are people more comfortable with no event loop these days?
- Missing features - do these developments manage to cover most analyses yet? If not, where are we? 50%, 90%, 99%?
    - Complicated permutations are still difficult
- How do declarative languages mesh with analysis facilities?

(Not declarative, but languages) Python, but also hints of Julia

- The return of the event loop!

Differentiable Analysis - Exploring the Use Case

There's a lot of momentum behind this

- Fold this into the ML sessions above?
- There are technical issues that need to be addressed
    - How do you move grad between tools, etc.

Bookkeeping and Systematics

- Better support for systematic handling
- Improved bookkeeping
- Efficiency considerations: design patterns in systematics handling that efficiently use resources
    - Are analyzers considering this / do they have the relevant information to make informed decisions?
    - Intermediate tools that could help users with this (e.g. translate inefficient approach to efficient approach)?
- Non-event data
- How do the various experiments solve the problem already? What can be lessons can be extracted?

Reproducibility, Reusability and Reinterpretation

- Analysis preservation
- Best practices for development and CI
- Built in from the start, or added at the end?

Visions from the last workshop that didn't happen and why they failed

Because we like to learn from mistakes

- As far as I can see we didn't get anything really wrong last time, but progress was far from even in all areas (e.g., much talk of non-event data, but not much happened after, AFAICT)

Trigger Level / Turbo Stream Analysis

- Analysis data coming directly out of HLT
- Are there common tools or all expt. Specific?
- Implications for the downstream (analysis facility handling)