HEP analysis ecosystem workshop, May 22-24 2017

[Indico page]

[Agenda outline]

Contents

Introduction

Over the past 20 years the HEP community has developed and gravitated around an analysis ecosystem centered on ROOT. ROOT and its ecosystem both dominate HEP analysis and impact the full event processing chain, providing foundation libraries, I/O services etc. that have prevalence in the field.

The analysis tools landscape is however evolving in ways that can have a durable impact on the analysis ecosystem and a strong influence on the analysis and core software landscape a decade from now, a timescale currently in an intensive planning round with the HEP Software Foundation (HSF) Community White Paper process. Data intensive analysis is growing in importance in other sciences and in the wider world. Powerful tools and new development initiatives, both within our field and in the wider open source community, have emerged. Creative developers have an ever more powerful open source toolkit available and are applying it towards innovations that leverage both open source and the ROOT ecosystem. ROOT itself is approaching a major re-engineering with ROOT 7, leveraging the powerful evolution of C++, with a major overhaul in its interfaces as seen both by its users and by satellite tools in the ROOT ecosystem.

With these considerations we are planning an open HEP software community workshop convened by the HSF to examine the analysis ecosystem, currently and in the future with a 5-10 year view. The workshop will be something of a retreat aimed at an inclusive building of consensus among developers, users, projects and their supporters in the HEP analysis ecosystem.

Considerations

- Review or workshop? Workshop, because it is not directed at a particular project, it's a community gathering to examine the analysis ecosystem, aimed at an inclusive building of consensus.
- Use the workshop to complement and build on the fora that already exist
    - ROOT has a regular series of users meetings where users report on how they use ROOT and provides a very valuable feedback to the project
    - This workshop should be oriented more towards development, the ecosystem, future planning/vision
    - Use the workshop to seed more communication, coherence, awareness etc. at the ecosystem level

- The workshop focus is on the HEP analysis ecosystem, while ROOT's scope extends beyond that, in areas that relate closely to analysis as well (e.g. foundation libraries, I/O) and other areas that do not (e.g. geometry). The workshop focus is on the analysis components of ROOT, and touches other components only as they relate to analysis.
- Organize it to deliver a material written outcome: a report
    - To that end, assemble a panel specifically charged to write it *Rethinking this… rather than a pre-defined panel, enlist the do-ocracy of workshop attendees to take notes and produce a report summarizing the workshop and any conclusions and actions arrived at*
    - While the workshop aims at consensus, the report should represent a diversity of views where they exist
    - Report is a narrative structured following the Objectives
    - If people decide to write individual/group white papers or comments arising from the meeting, collect them also as preserved workshop outcomes
- An important outcome will be Actions that are agreed in the workshop, they will be documented in the report and followed up after the workshop
- Support and encourage self-organization in the workshop to seed detailed discussions, formative collaboration, etc.
    - e.g. through parallel session half-days that aren't too pre-programmed in advance

Scope and objectives

- Survey the HEP analysis ecosystem
    - The tools (HEP in-house, open source, industry), their relative roles, interdependencies, trajectories, plans for the future
    - Analysis workflows, how they relate to the tool set and future needs, prospective improvements
    - Analysis foundations
        - Roles of languages, evolution, needs: C++, Python, julia, Go, Rust, …
        - Analysis file formats, data exchange
    - Analysis infrastructure and facilities
    - Build greater awareness
        - among users of the tools, established and emerging, in the ecosystem
        - among developers, current and prospective, on existing and planned projects, to foster collaboration, avoid redundant efforts, and identify capability gaps to be filled
    - Strengthen communication and collaboration
        - ROOT has well developed fora and processes for engaging users, developers, the community. Can they be improved and how?
        - What ecosystem-level communication and collaboration fora should exist? What role should the HSF have?
            - "avoid misconception that HSF would slowly gobble up existing projects removing decision power from them" (Dario Menasce)
    - Build commonality
        - commonality among tools and frameworks
        - leveraging community building/sharing software: github, docker, ...
        - Commonality among tools and frameworks is one way of seeing it. From the other side there are many tools and frameworks around, and we should exploit them. As far as data analysis tools in Python are concerned, for example, the community should exploit them at best. So there is also a need for easy navigation "APIs" between the various tools. Some of us have recently started the Scikit-HEP project for this purpose (among others)
        - development tools and best practices
- Survey ROOT
    - ROOT as the analysis ecosystem hub
        - Modularity, interfaces, packaging, discreteness of components, effectiveness of the ROOT infrastructure and packaging for integrating satellite developments from the community

        - Using and integrating ROOT in C++, in Python, ...
        - ROOT as a Service
    - The ROOT workplan, near and long term, its prioritization, and the input that has driven the plan
        - Use the workshop to gather input on the workplan and its prioritization as input to ROOT's planning
            - Not to replace the ROOT user workshop, but to take opportunity of this gathering to provide complementary (and no doubt overlapping) input
- Survey the future, with a 5-10 year view
    - Many new elements are on the rising edge of their hype cycle. Which should we pay attention to and how much? Which will transform how we do software and analysis?
        - Analytics tools, browser based tools (notebooks, graphics, …), containers, …
        - Machine learning's role and inclusion into the ecosystem
        - Analysis as a Service, including infrastructure aspects
    - How concurrency-aware and -capable is our ecosystem? What work is needed where? Are concurrency implementations and evolutions compatible within the ecosystem and experiment infrastructures? Are they compatible with serial components to allow adiabatic evolution?
    - Where will persistency be in 10 years? What are the requirements? Will it still be ROOT and if so what evolution is needed? Will it be something else and if so, what?
    - What are the big risks and challenges, how do we mitigate and prepare for them?
- Strengthening and growing the ecosystem
    - Where are the gaps and opportunities? Prospective/planned funding proposals, new/extended projects?
    - How do we grow the effort and where do we need new effort the most?

Outcomes

- A report consisting of
    - The workshop's collective input (no doubt varied and far from unanimous) on the ROOT workplan, priorities, technical direction and future plans
        - Complementing and extending the consultation/feedback channels already in place in the ROOT project
    - A survey of the tools and libraries making up the analysis ecosystem today
        - With testimonials, opinions, assessments, advice, input, … ?
    - An assessment (guess) as to the analysis landscape in 5-10 years
        - The requirements
        - The players: what will the ecosystem consist of then
        - The needed work, and in what directions
    - Actions, next steps and outcomes
        - Actions agreed on in the workshop
        - Steps to improve collaboration/communication, opportunities for new/extended projects, …
            - One thing is if projects can be identified, like "a full analysis using this tool chain" - and then find some small source of funding to encourage it. The NSF, for example, will often use this to try to push things... if this is done right, perhaps some support a group might not ordinarily have, might be found to explore a few avenues. (Gordon Watts)
- The report will be useful input to the Community White Paper (CWP)
- First of a regular series if it goes well

Plan

- **Dates are set for May 22-24.**
    - Held near or at CERN
    - Expect considerable attendance! Ask people to register to get a count

    - Vidyo available
- Produce a reasonably curated report out of the workshop with summary, conclusions, actions
- I [Gordon Watts] would love a summary discussing this ended up at ACAT 2017. :-) http://indico.cern.ch/event/567550/

Feedback

- This was sent to the HSF's open forum google group list. Please recirculate it to other relevant lists. If you received it through recirculation, consider joining the HSF forum google group, hep-sf-forum@googlegroups.com; information on this workshop and the HSF meetings where it will be discussed are posted there. More information at http://hepsoftwarefoundation.org.
- To give email input on this proposal, send it either to the forum list, O(200) people, or to the HSF startup team organizing the workshop at hep-sf-startup-team@googlegroups.com, as you prefer.
- Feel free to comment directly in this googledoc.
- This comment is in the context of the HSF workshop, where the horizon is 5-10 year: It might be useful to do an exercise that investigates current bigdata/datascience analysis frameworks out there (spark, pandas, pyspark, rspark, etc.) and see how hard would be to port functionality of ROOT and other HEP analysis tools. This allows the community to NOT have to spend time/money/efforts on maintaining core functionality that is already done by Apache / tech Industry.

March 3 coffee - first pass at agenda

See the agenda outline for the current version.

- Day 1 (10am start)
    - a 'this is today's landscape and toolset' session
        - Elements of the ecosystem to discuss: ROOT, R, TMVA, xrootd, diana, python analysis tools, I/O including its distributed aspects, ...
        - Look at analysis models of the experiments as they relate to a common analysis ecosystem.
            - Experiments and ROOT as a tool box for their analysis systems, looking at this at an architectural level, can we abstract common experiments approaches, can we better integrate in ROOT/across experiments.
        - Identifying commonalities, existing and new opportunities.
        - How to improve sustainability of ROOT and other key components what are the ones that are critically necessary and what are the ones that darwinian evolution is fine.
        - How can contributions large and small to the ecosystem be assimilated.
    - Convener ideas: Markus Elsing, Attila Krasnahorkay, Giovanni Petrucciani, Andrea Rizzi
- Day 2 (full day)
    - a 'vision' session
        - What is our analysis ecosystem model for building an analysis environment composed of modules… R does that, python does that, … how should we do it?
        - ROOT has capabilities today in integrating modules according to defined interfaces, how should this evolve to be more general?
        - Modernization of ROOT; taking account of what is now supported in current C++, obsoleting no longer needed components (TString etc).
        - Analysis as a service, streaming, storage futures, …
        - Convener ideas: Gordon Watts, Axel Neumann, Vincenzo Innocente
    - a 'concrete technologies' session. Languages, coprocessors, …
        - Convener ideas: David Lange, Benedikt Hegner
    - start on missing pieces?

- Day 3 (early afternoon finish)
    - missing pieces, work to be done session
        - Convener ideas: Markus Schulz, James Catmore
    - conclusions session, establishing outcomes, actions, next steps
        - Conveners: Pere, Liz, Torre
- Day 3 afternoon
    - writing the report, for the chairs, the session conveners, and others who want to hang around and write
-
- Presentations need to be concise and following clear guidelines on points of interest