

Analysis Facilities and Use Cases

Next steps after Naples

Oliver Gutsche, Eduardo Rodrigues
Fermilab University of Cincinnati

Summary-ish of session @ Naples Workshop

Session overview

- Workshop = environment for *examples* of new and/or innovative analysis flows techniques and facilities being used/investigated

16:00 → 18:00

Analysis Facilities and Use Cases


Scarlattì (Hotel Vesuvio)

Conveners: Eduardo Rodrigues (University of Cincinnati (US)), Oliver Gutsche (Fermi National Accelerator Lab. (US))

16:00

Overview of CWP paper from WG Data Analysis and Interpretation


Speaker: Mark Neubauer (Univ. Illinois at Urbana Champaign (US))

 Naples HSF Analy...

16:20

Spark-like and query-like analysis systems and tools


Speaker: Jim Pivarski (Princeton University)

 pivarski-querylike.pdf

16:40

HPC analyses

Speaker: Viktor Khristenko (CERN)

 deepest_hpc4hep.pdf

17:00

Real-time analyses - the LHCb case

Speaker: Rosen Matev (CERN)


 slides

17:20

Analyses with SWAN and docker

Speaker: Pere Mato Vila (CERN)

 swan-wlcg-hsf-work...

 swan-wlcg-hsf-work...

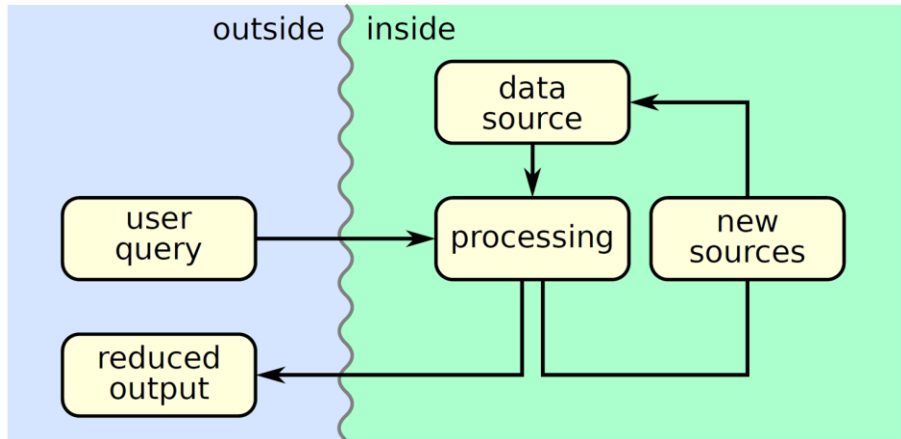
17:40

Open discussion on future avenues and collaboration

Speakers: Eduardo Rodrigues (University of Cincinnati (US)), Oliver Gutsche (Fermi National Accelerator Lab. (US))

Explorations around data querying in analysis flows

- Or: what works, what doesn't, what's needed, what we're building
- Exploratory work around the following grand picture:
 - All boxes deserve dedicated focus
 - Ex.: functional programming style, Dask for distributed processing
- Several proof-of-concept Python packages developed along the way



HEP analysis pipelines on HPC facilities

- DEEP-EST^(*) :
 - EU R&D project for Exascale HPC, in design phase
 - Explore conventional HEP analysis workflows on HPC infrastructure
 - Explore usability of Apache Spark for HEP data analysis

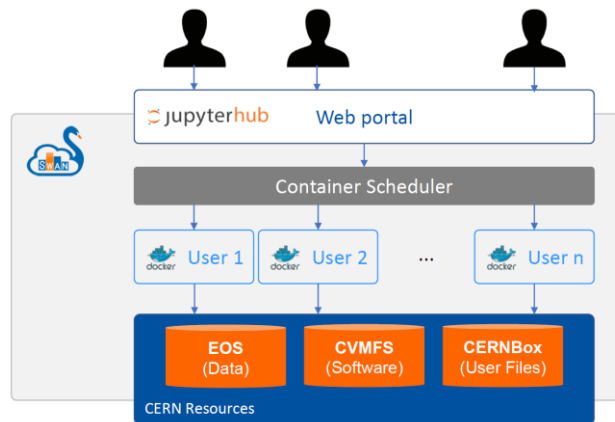
- Some R&D activities:
 - Optimising analysis workflows for novel compute/memory/storage capabilities
 - Utilising industry tools like Apache Spark: enabled to read ROOT files natively
 - Benchmarking with TBs of CMS Open Data
 - In some sense: are general purpose solutions viable for HEP data analysis?

^(*) Dynamical Exascale Entry Platform - Extreme Scale Technologies



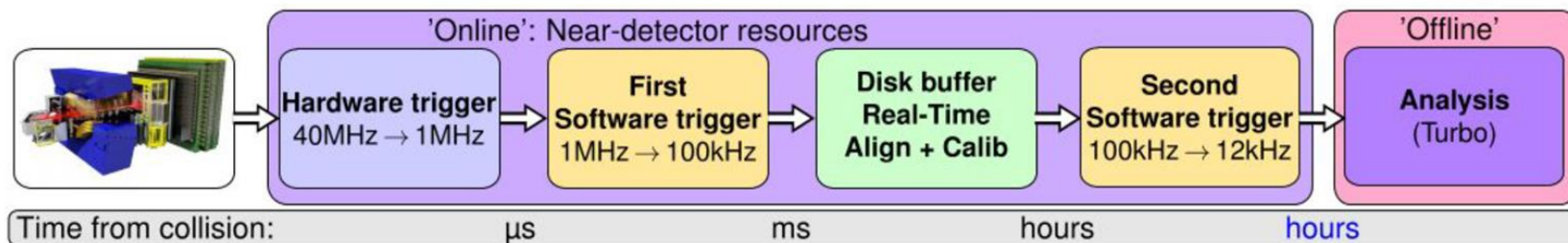
SWAN - service for web-based analysis

- Service integrates experiments/user data and software, and computing resources
- Users can do analysis only with their browser
 - Anywhere, anytime, with no local installation
 - Cloud-based analysis model
- Interface: Jupyter Notebooks on demand
 - Combine code, equations, text and visualisations
 - Most useful for final steps of analysis, teaching, documentation
 - Use personalised docker container instance
 - System can scale to larger resources like Spark clusters
- Developments guided by (heterogeneous) community
- SWAN is first step towards a truly scalable and interactive distributed data analysis environment



LHCb online & high-throughput analysis (flow)

- A reality, in production !



- Boosts up physics reach and allows for otherwise non-doable analyses
- Only saving reconstructed objects (no raw event) in the trigger demands much on online calibration and alignment, hence on a robust and efficient online monitoring of the data quality as well

Data Analysis & Interpretation WG Paper

- In support of Community White Paper
 - Cuts across many other WG areas
 - Data preservation, Machine Learning, ...
 - Submitted to the arXiv yesterday!
-
- Time-to-insight (data) is a crucial metric
 - Think towards a “smart” data analysis system
 - Scope for new ideas of analysis pipelines, collaborating with non-HEP communities

Contents

- 1 Introduction**
- 2 HEP Analysis Software Ecosystem**
- 3 Analysis Languages**
 - 3.1 Python
 - 3.2 Declarative Languages
- 4 Analyzing Data**
 - 4.1 File format and compression algorithms
 - 4.2 Analysis Facilities
 - 4.3 Non-event data handling
- 5 Analysis Models and Future Systems**
 - 5.1 Sequential Ntuple Reduction
 - 5.2 “Spark”-like Analysis
 - 5.3 Query-based Analysis
- 6 Analysis Preservation**
- 7 Analysis Interpretation**
- 8 Analysis Roadmap**
 - 8.1 1-year Time Frame
 - 8.2 3-year Time Frame
 - 8.3 5-year Time Frame

Outlook

- We want, and need(!), to foster collaboration and common projects on analysis-related topics
 - Examples of funded efforts exist (e.g., NSF-funded [DIANA/HEP](#) project) , but need more to make a difference !
- Lots of activities post-CWP but no organisational structure / specific timelines to tackle proposed roadmap ...
... can the HSF facilitate information exchange in the analysis area, Maybe with a dedicated WG ?
 - Let us know about funded (and non-funded) efforts
- Links with communities outside HEP are emerging through the tools explored and used - great !

Next steps beyond Naples

“Random” questions from Naples, to keep in mind ...

- Is more contact information needed for projects already discussed to start more collaboration?
- How do we keep the discussion and exchange of information alive?
- What common projects are identified for this area that will advance?
 - Are the right stakeholders already involved? Should other experiments or organisations be contacted?
 - Are any resource needs covered?
- Are there projects that are not currently advancing?
 - What would allow them to do so?
- Links with communities outside HEP
 - Are any established? Are any links needed that we don't have?
- Are we making best use of community expertise in this area?
 - How best do we advance collaborations that will be of benefit more widely and improve our software and maximise the effectiveness of effort

Further engage with the open-source community ?

- For what concerns data analysis, there is much to be learned/used from the non-HEP community
 - Software - huge (Python) scientific software ecosystem
 - Training - lots of conferences/workshops with tutorials and hands-on (see next slides for examples)
 - Exchanges - effectively done via open-source events and contributions to OS software
- Many non-HEP open-source conferences / workshops we should try to have a more “active” presence in (again, see next slide for a few examples)
 - HEP community presence is not overwhelming
 - We do have our own dedicated conferences/workshops, which work very well
 - Ex.: ACAT, CHEP, DS@HEP, ROOT Users
 - But much to be gained with a stronger exchange, building relations & contacts: attend a broader range of events, further invite non-HEP speakers to our events
 - The HSF may help as a “visiting card”
 - Obvious challenge for event participation: financial support ...

Further engage with the open-source community ?

Non-HEP open-source SW&C / Statistics / Data Science / Machine Learning Events

Event	Site	Scope
AnacondaCON	https://anacondacon.io/	Real World Data Science, Anaconda Enterprise Open Source Technology
JupyterCon	https://conferences.oreilly.com/jupyter/jup-ny	Bring together data scientists, business analysts, researchers, educators, developers, and core Project contributors and tool creators.
IEEE BigData	https://bigdata.ieee.org/conferences	Around Big Data – analysis, computing, cloud, ...
PyData	https://pydata.org/	International community of users and developers of data analysis tools to share ideas and learn from each other.
NIPS	https://nips.cc/	Multi-track ML and computational neuroscience. Invited talks, oral/poster presentations, demos.
GTC	https://www.nvidia.com/en-us/gtc/	GPU Technology Conference From DL to Gaming, via AI, VR and DevTools
EuroSciPy	https://www.euroscipy.org/	Cross-disciplinary. Focused on use and development of the Python language in scientific research.
SciPy	https://conference.scipy.org/	Participants from academic, commercial, and governmental organizations to (1) showcase their latest Scientific Python projects, (2) learn from skilled users and developers, and (3) collaborate on code development.
Strange Loop	https://www.thestrangeloop.com/	Multi-disciplinary conference bringing together developers and thinkers building tomorrow's technology in fields such as emerging languages, alternative databases, concurrency, distributed systems, security, and the web.
Strata Data Conference and other O'Reilly conferences such as AI Conf.	https://www.oreilly.com/conferences/	

(Non comprehensive!)

Promoting Python as a 1st-class analysis language

- PyHEP “Python in HEP” workshops
 - Initiative strongly welcomed by HSF
- In longer term will hopefully enable stronger bridges with the Data Science, Machine Learning, SW&C communities
 - E.g. inviting speakers from interesting projects, tools, etc.
 - Hopefully that may result in more of those projects/companies becoming aware and interested in our use cases ... they may then invite us too!

PyHEP 2018 Workshop

7-8 July 2018
Sofia, Bulgaria

Search...

- Overview
- Timetable
- Contribution List
- Participant List
- Registration
- Travel Info
- Accommodation
- Proceedings
- Code of conduct
- Contact us

The **PyHEP workshops** are a new series of workshops initiated and supported by the [HEP Software Foundation](#) (HSF) with the aim to provide an environment to discuss and promote the usage of Python in the HEP community at large. Further information is given [here](#).

This first workshop, held just before the start of the CHEP 2018 conference, will focus on a review of where and how Python is used in our community, and what the future will hold. The workshop will also be a forum for the participants, representatives of the community, to discuss topics around the areas of work identified in the HSF Community White Paper. There will be ample time for discussion.



A keynote presentation on [JupyterLab](#) will be given by [Vidar Tonaas Fauske](#), a member of the developers team.

Organising Committee

Eduardo Rodrigues - University of Cincinnati (Chair)
Graeme Stewart - CERN-HSF
Jeff Templon - Nikhef (Co-chair)

Sponsors

The event is kindly sponsored by

CWP roadmap - 1-year time frame (1/2)

- Enable new open-source tools to be plugged in dynamically in the existing ecosystem and mechanisms to dynamically exchange parts of the ecosystem with new components
 - Ongoing projects: e.g. the [Scikit-HEP project](#)
 - Work by Giulio Eulisse on initial Apache Arrow TDataSource for ROOT, as starting point to deliver Arrow interoperability to ROOT
- Develop requirements and design a next generation analysis facility concept, incorporating fast caching technologies to explore a query-based analysis approach and open-source cluster management tools
 - See e.g. Jim's and Viktor's work presented @ Naples

CWP roadmap - 1-year time frame (2/2)

- Finalise full support of Python in our ecosystem including long-term maintenance
 - Need strong commitment and effort from ROOT team
 - Would profit from more community engagement
 - See initiative of PyHEP workshop series
- Evolve policies to minimise this effort by retiring less used components from the integration and validation efforts
- Establish a schema for the “analysis database”
 - Obvious link to groups on Data Preservation & Interpretation
- Interpretation Gateway: conceptualisation integrating the analysis facility, analysis preservation infrastructure, data repositories, and recasting tools

Note: the WG paper contains R&D items on the 3- and 5-year time frame, but not so important to list them here at this stage (many are follow-ups and finalisations of the above)

In short

- Lots of activities post-CWP and many people involved across the various experiments
 - We need to keep these activities alive, “connected” and, hopefully, focused, with an effective communication across experiments in particular
- Challenging to keep a broad overview of what’s happening
 - No organisational structure so far
 - This may be a role for the HSF
 - Post-CWP WLCG/HSF workshops might be a good thing, say once a year
- Links with communities outside HEP are emerging through the tools explored and used - great !
 - Trend is established and we should encourage it further