

Live notes: Analysis workflow management tools

[Workflow management tools \(December 8, 2021\) · Indico \(cern.ch\)](#)

Attending: Allie Hall, Nicole Skidmore, TJ Khoo, Valeriia Lukashenko, Alba Vendrell, Andrew Melo, Dana Seman Bobulska, Davide Valsecchi, Graeme Stewart, Jonas Eschle, Jonathan Davies, Jordy Butter, Marcel Rieger, Mason Proffitt, Mindaugas Sarpis, Nicolas Lurkin, Sebastien Wertz, Tibor Simko, Adam Morris, Brett Viren, Nick Manganelli, Elena Gazzarrini, Valeria D'Amante, Johannes Lange, Matteo Marchegiani, Fidan Suljik, Igor Kostiuik (25 participants :))

Snakemake presentation by Valeriia Lukashenko

- Workflow management tools document HOW to run jobs. Helpful for others and yourself
- Snakemake - Scalability, readability, portability, modularisation, transparency.
- Extremely collaborative
- Easy to learn and adopt after the fact
- Built in preservation tools eg. can tag a workflow
- Originally from Biology research - wide usage. StackExchange etc., very good documentation
- Simple syntax, driven by contents of a single *Snakemake* file
- Supports multithreading over multiple cores
- Snakemake identifies dependency tree and what rules need to be run. Will rerun any rules where input has changed
- Can use XRootD for files hosted centrally
- Can see graphical representation of workflows and drill down to see attributes of each rule in browser. Shows time taken for each job in workflow for optimising resources
- LHCb gives tutorials on Snakemake in their StarterKit and this is now part of HSF tutorials
- Snakemake part of LHCb default Conda environment

Note: snakemake.utils link seems to be broken, change '%23' to '?'

Discussion

- TJ Khoo: How big is the scope of the workflows that you can run with Snakemake? E.g. Grid submission, job monitoring and error handling, resubmission, ...
 - Grid should be supported
 - Error handling on cluster similar to local, would want to test locally first. Does not necessarily transmit errors very transparently.
 - Large output handling – can mark outputs as transient, to be cleared up automatically.
- Nicole: If there is a problem with an intermediate stage, are outputs cached so they can be resumed?
 - (Non-transient) outputs will be produced, and should not need to be rerun.

LAW presentation by Marcel Rieger

- Extension of Luigi, currently used at CMS, Belle
- Analysis workflow management becomes necessary with high scale and complexity
- Motivation - portability, reproducibility, preservation - “daily working environment should provide preservation features out of the box”
- Python based, originally from Spotify but now community driven. 5-6 years old
- Building blocks are *tasks* and *targets*
- Web UI detailing tasks with automatic error handling
- Dependency is based on *tasks* rather than the *targets* (unlike Snakemake)
- *make*-like execution model
- LAW has 3 goals - experiment agnostic (not even-HEP), scalability on HEP infrastructure, decouple execution location, storage location and environments
- Job submission built into tasks
- Supports ALL WLCG protocols - work with files as if they were local
- Env sandboxing - Singularity and Docker

Discussion

- Tibor: Hurdles with moving analyses to LAW
 - Use of Luigi is first threshold
 - Workflow management is tricky at first
 - Large community to give support
- Sebastien: how are jobs batched?
 - Semi automated
 - Can specify how to split jobs eg. tasks per job
- SLURM integration by end of the year
- TJ: Grid submission and access?
 - CMS uses CRAB to submit to tier 2 / 3 sites
 - Can use gLite and ARC

REANA presentation by Tibor Simko

- Based on FAIR principle - Findable, Accessible, Interoperable, Reusable
- Four pillars - Input data, analysis code, computing environment, computational recipes
- Supports Serial, Yadage, CWL, Snakemake
- *Preproducibility* - analyses should be born reproducible - too late at publication. Achieved through use of GitLab CI for instance
- Live demo - see recording

Discussion

- Nicole: Analysis preservation steps becoming standard in LHCb, how about ATLAS & CMS?

- ATLAS BSM (SUSY, Exotics) have mandated RECAST in analysis approval
 - CMS not as organised
- Nicole: How does RECAST relate to REANA?
 - Similar systems, encodes more information about the analysis itself. Can use REANA as a workflow backend for RECAST, but Docker containers e.g. could be used directly