Data & Analysis Preservation: status update

Maxim Potekhin
Nuclear and Particle Physics Software Group

BROOKHAVEN NATIONAL LABORATORY

PHENIX DAP Meeting
05/27/2021

reana

Overview

- Zenodo+Website
- HEPData
- Docker
- REANA
- DAP@PHENIX School
- DPHEP Collaboration - workshop in June 2021
- Open Data

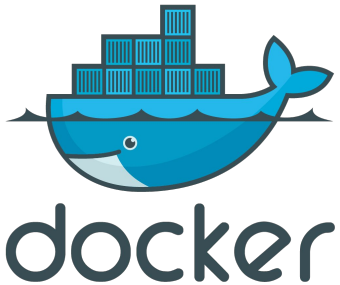
Zenodo+Website

- Uploads ongoing (thanks Gabor)
- REANA pages updates
  - Corrections
  - Reference to tutorials on GitHub (new)

HEPData

- Master spreadsheet updated
- PPG115, PPG234 close to final
- PPG081 - done
- PPG139 added to repo, review pending
- PPG201 added to repo, preparing upload

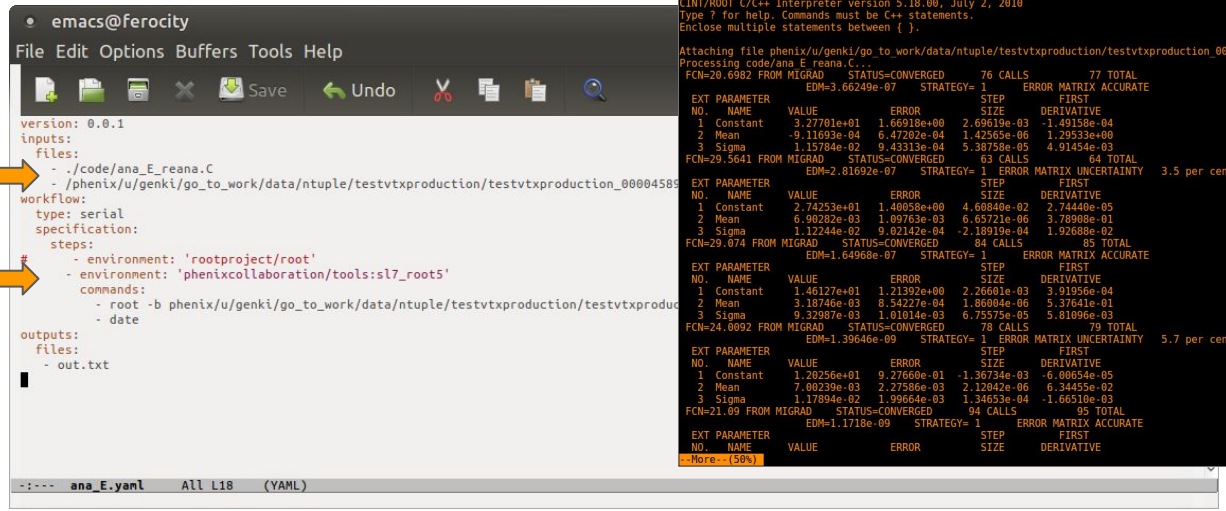
Docker: the two options

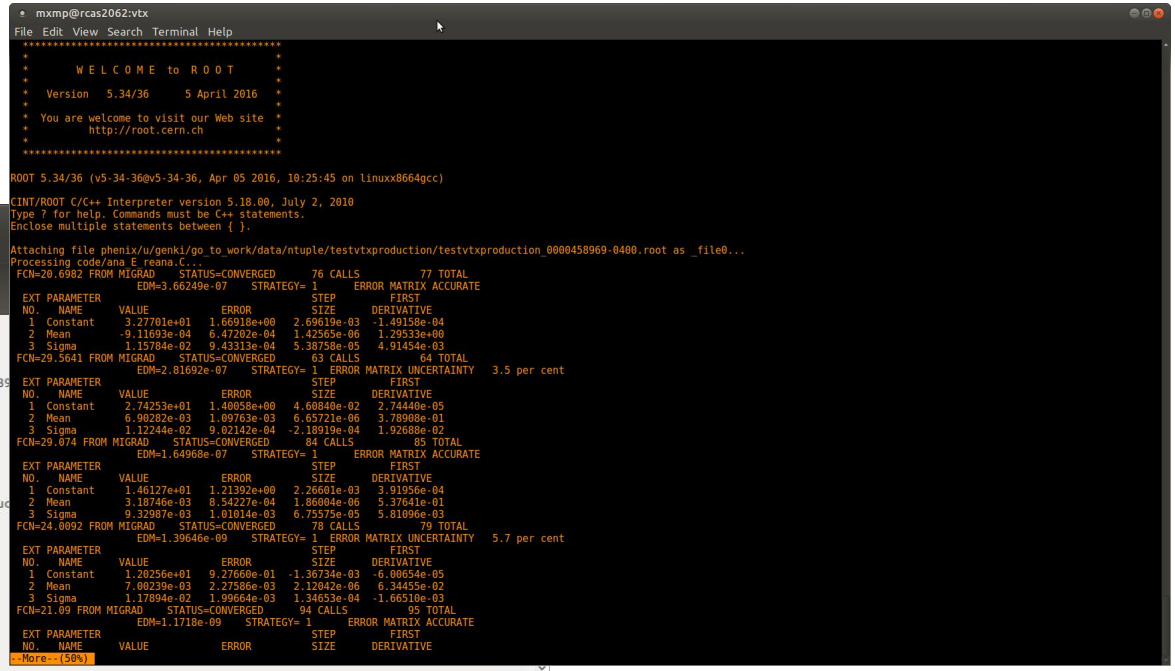


- *Custom 32-bit Docker image (collections of binaries)*
  - *Cleanup of the Dockerfile, improved directory structure*
  - *Initial simple testing done*
  - *Validation with complex macros still on the to-do list - please help with a few examples*
- *Use SDCC-provided images*
  - *Previous problem - very large size of all the libraries combined*
  - *Solution - offload most libraries to a network file system - CVMFS*
  - *Some progress in understanding proper CVMFS configuration, testing under way*
  - *In a nutshell, the idea is to copy over most of the software stack to CVMFS*

REANA

- Tutorials (next slides)
- Genki's VTX macro tested with real data and our sl7_root5 image
- Non-zero exit code, will investigate - but it works.





REANA: misc items of note

- As shown in the previous slide, rean'izing analysis based on macros can be trivial
  - But data cataloging and managing may still be a challenge
- Please give Chris H enough lead time to create accounts when prepping for school
- Current storage cap for all workflows: 200GB
  - Workable, with active disk space management
  - NB. The user does not have to re-upload all the data from scratch when starting a job in REANA
- Can be increased to 8TB (by formal request to SDCC)
- Potential scaling-up of the compute resource by using SLURM to incorporate "many" worker nodes on the farm

School

- REANA Tutorials
  - Created two basic “hello world” tutorials, materials are on GitHub
  - Added links to the REANA page on our website
  - Gabor/Maxim’s macros for pi0/gamma analysis will serve as the next step in the tutorials
  - Will see how much more material it will be possible to create
- Looked into CWL (graph-like description of workflows in REANA)
  - Complex workflows (i.e. non-linear) will be hard to demo due to lack of accessible documentation e.g. how to use CWL (the workflow description language), will take another look.

School (cnt'd)

- Optimal way to run the REANA client software
  - Using rcas interactive nodes as client machines for the School exercise may be optimal
  - Environment more predictable compared to individual users' machines
  - The instructors will be able to inspect users' folders if necessary
  - Easy to share files in real time if necessary
- The Web GUI should still be run on the users' machines via a SSH tunnel or VPN
  - NX should be an equally good option

DPHEP Collaboration and Workshop

- DPHEP is the CERN-led collaboration on DAP, of which BNL is a stakeholder
- PHENIX (3 members) participated in the DPHEP workshop in Fall'19
- Participation is important since we use crucial CERN services which are under the DPHEP umbrella and need access and support
  - Zenodo
  - HEPData
  - OpenData
  - REANA
- We are indeed 100% invested in the best community tools and practices
- Invited to present at the next workshop in late June 2021
  - Preparations are underway
  - Will probably have a scheduling conflict with the School on 06/22, need to address

Open Data

- Last(?) round of corrections

Plans

- Docker images - ongoing work
- Additional REANA tutorials
  - Please suggest real macros/examples similar to Genki's
- HEPData, steady state effort
- Presentations for AUM and DPHEP
- ...both will take effort
- AUM meeting - we should probably skip our DAP meeting on June 10th