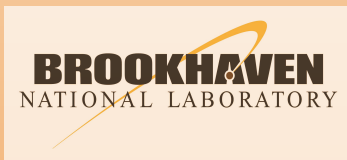


Data & Analysis Preservation: status update

Maxim Potekhin

Nuclear and Particle Physics Software Group



PHENIX DAP Meeting

05/27/2021

The reana logo, consisting of the word "reana" in a lowercase, sans-serif font. The "re" is orange and the "ana" is dark brown.

Overview

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

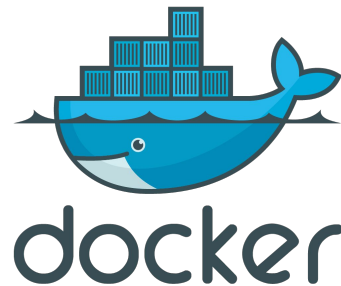
Zenodo+Website

- 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.

The screenshot shows two terminal windows. The left window is an emacs editor showing a workflow configuration file named `ana_E_reana.C`. The workflow is of type `serial` and includes environment settings for `rootproject/root` and `phenixcollaboration/tools:sl7_root5`. The right window is a terminal running the ROOT interpreter, displaying a welcome message and the output of a macro execution. The output shows the processing of a file, followed by a table of parameters and their values, errors, and derivatives. The table is divided into three sections, each corresponding to a different macro call (FCN=20.6902, FCN=29.5641, and FCN=24.0092). Each section includes a table of parameters (Constant, Mean, Sigma) and their values, errors, and derivatives. The output also shows the status of the macro execution (e.g., STATUS=CONVERGED) and the total number of calls (e.g., 76 CALLS).

```
version: 0.0.1
inputs:
  files:
    - ./code/ana_E_reana.C
    - /phenix/u/genki/go_to_work/data/ntuple/testvtxproduction/testvtxproduction_00004589
workflow:
  type: serial
  specification:
    steps:
      - environment: 'rootproject/root'
      - environment: 'phenixcollaboration/tools:sl7_root5'
    commands:
      - root -b phenix/u/genki/go_to_work/data/ntuple/testvtxproduction/testvtxproduction_00004589
      - date
outputs:
  files:
    - out.txt
```

```
mxmp@rcas2062:~$
File Edit View Search Terminal Help

*****
* WELCOME to ROOT *
* Version 5.34/36 5 April 2016 *
* You are welcome to visit our Web site *
* http://root.cern.ch *
*****

ROOT 5.34/36 (v5-34-36@v5-34-36, Apr 05 2016, 10:25:45 on Linuxx86_64gcc)
CINT/ROOT C/C++ Interpreter version 5.18.00, July 2, 2010
Type ? for help. Commands must be C++ statements.
Enclose multiple statements between { }.

Attaching file phenix/u/genki/go_to_work/data/ntuple/testvtxproduction/testvtxproduction_0000458969-0400.root as _file0...
Processing code/ana_E_reana.C...
FCN=20.6902 FROM MIGRAD STATUS=CONVERGED 76 CALLS 77 TOTAL
EDM=3.66249e-07 STRATEGY= 1 ERROR MATRIX ACCURATE
EXT PARAMETER NO. NAME VALUE ERROR STEP SIZE DERIVATIVE
1 Constant 3.27701e+01 1.66919e+00 2.69619e-03 1.40158e-04
2 Mean -9.11693e-04 6.47202e-04 1.42565e-06 1.29533e+00
3 Sigma 1.15784e-02 9.43313e-04 5.38759e-05 4.91454e-03
FCN=29.5641 FROM MIGRAD STATUS=CONVERGED 63 CALLS 64 TOTAL
EDM=2.01692e-07 STRATEGY= 1 ERROR MATRIX UNCERTAINTY 3.5 per cent
EXT PARAMETER NO. NAME VALUE ERROR STEP SIZE DERIVATIVE
1 Constant 2.74253e+01 1.40958e+00 4.60849e-02 2.74440e-05
2 Mean 6.90202e-03 1.09763e-03 6.45721e-06 3.78900e-01
3 Sigma 1.12244e-02 9.02142e-04 2.18919e-04 1.92688e-02
FCN=29.074 FROM MIGRAD STATUS=CONVERGED 84 CALLS 85 TOTAL
EDM=1.64968e-07 STRATEGY= 1 ERROR MATRIX ACCURATE
EXT PARAMETER NO. NAME VALUE ERROR STEP SIZE DERIVATIVE
1 Constant 1.46127e+01 1.21392e+00 2.26601e-03 3.91956e-04
2 Mean 3.18746e-03 8.54227e-04 1.86804e-06 5.37641e-01
3 Sigma 9.30237e-03 1.01014e-03 6.75757e-05 5.81096e-03
FCN=24.0092 FROM MIGRAD STATUS=CONVERGED 78 CALLS 79 TOTAL
EDM=1.39646e-09 STRATEGY= 1 ERROR MATRIX UNCERTAINTY 5.7 per cent
EXT PARAMETER NO. NAME VALUE ERROR STEP SIZE DERIVATIVE
1 Constant 1.20256e+01 9.27669e-01 1.36734e-03 6.00654e-05
2 Mean 7.00239e-03 2.27586e-03 2.12042e-06 6.34455e-02
3 Sigma 1.17894e-02 1.99664e-03 1.34653e-04 1.66510e-03
FCN=21.09 FROM MIGRAD STATUS=CONVERGED 94 CALLS 95 TOTAL
EDM=1.1718e-09 STRATEGY= 1 ERROR MATRIX ACCURATE
EXT PARAMETER NO. NAME VALUE ERROR STEP SIZE DERIVATIVE
<More--(50%)>
```

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