Data & Analysis Preservation: status update

Maxim Potekhin

Nuclear and Particle Physics Software Group

PHENIX DAP Meeting

04/22/2021

Overview

- Zenodo, Website
- HEPData
- Docker
- REANA
- DAP@PHENIX School
- Open Data

Zenodo + Website

- Added more conferences
  - Maxim and Gabor
  - GHP17&19, Moriond 19&21, QAT21
- Incremental updates on the website (keywords), minor fixes
  - ...work in progress, catching up with recent work done by Gabor
  - Healthy level of effort/activity
- EMCAL paper added/linked

HEPData

- There is now a solution to the long standing problem of harmonizing the accuracy of values and errors (i.e. decimal places) in the HEPData submission packages - a Python script was shared by STAR for better control over notation
  - Further improved, committed to our repository
  - Currently in use by a few members
- For example PPG147 HEPData package has been updated (Takashi)
  - Ready to upload
  - Can Ron please be the official reviewer so the submission can be finalized?
- READMEs etc have been updated, spreadsheet up to date
  - Some items stalled, lack of available effort in groups
- PPG201 in development
- *PPG071 successfully revised (an old item where inaccuracies were discovered) - thanks Krista*

HEPData - Christine’s idea

- There are ~150 papers yet to be committed to HEPData (ballpark)
- This is important to ensure availability of PHENIX data in the long term
- Although the system is reasonable and user-friendly, preparation of HEPData material still takes non-trivial effort (some entries are genuinely complex)
- Due to volume this cannot be addressed by assigning this type of work to interns, summer students or other participants since it goes beyond educational purposes and is effectively a sizeable technical assignment
- Idea - hire undergrads at $12-15 an hour
- Estimated cost of processing ~150 papers is approx. $45k
- Students would work at U of Tennessee and perhaps other locations
- Not necessarily in the scope of BNL-funded work
- What are the possibilities for a proposal? Talk to the DOE Office of Science?

REANA - quick recap (more to follow)

- REANA: captures the workflow, the software, the data
  - Workflows described as graphs in a YAML format
  - Software captured in containers and/or CVMFS (defined network storage)
  - Data can be uploaded to the sandbox or accessed via the network
- REANA - since it’s run at BNL - can help solve the local database access issues (when needed by analyses)
- Containerization is on the critical path

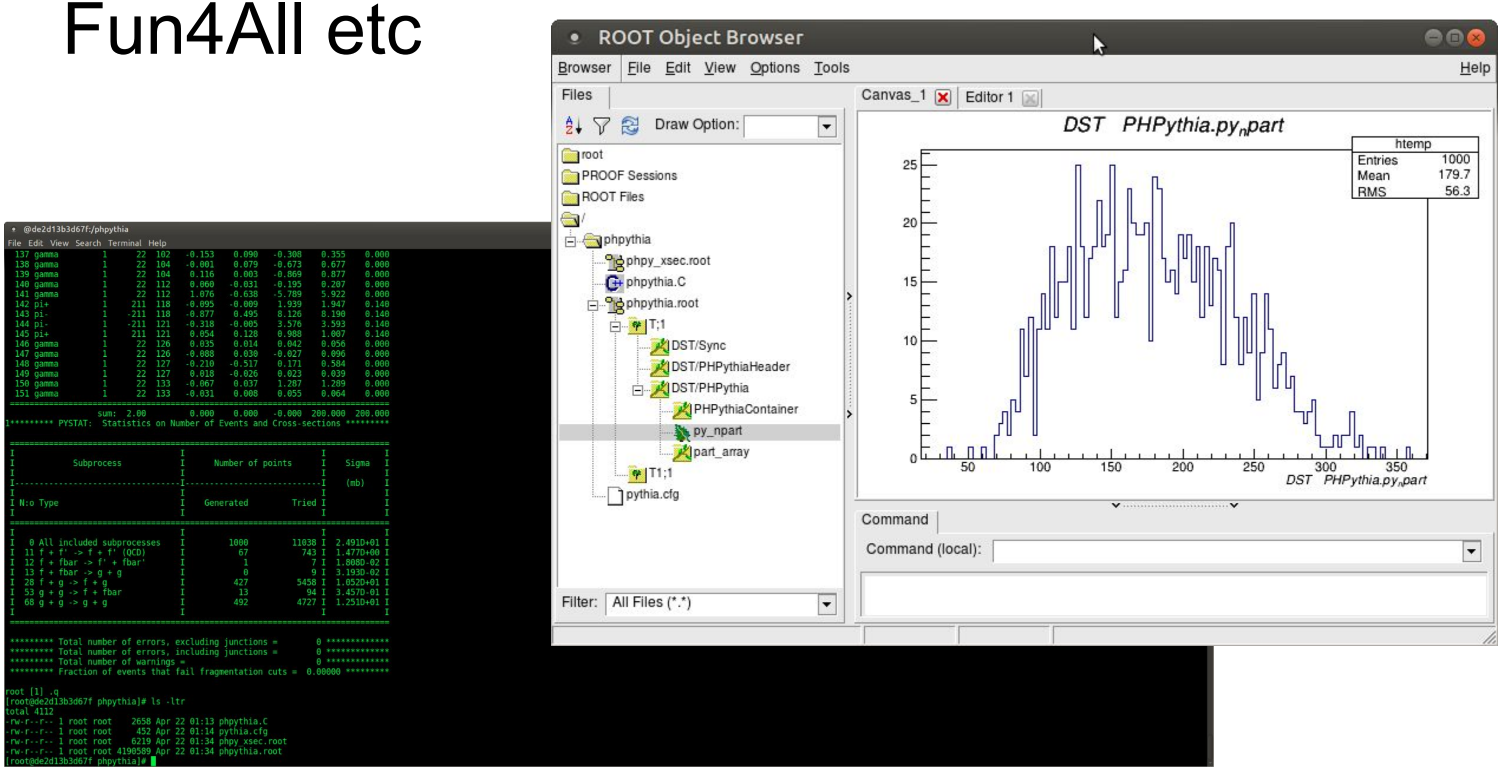
Docker: status and outlook

- Two approaches are being tested:
  - ***1. Copying 32-bit binaries/dependencies to a Docker image***
    - Can be committed to Docker Hub
    - ...see next slide
  - ***2. Use SDCC-provided images***
    - *The images have been made available recently, not tested yet -* ***on the to-do list***
    - *Can’t be made public but we will provide them on request or from the Docker registry at BNL*

Docker “1”: collection of 32-bit binaries

- *Goes way beyond copying a few libraries into the image*
- O(100) dependencies of both ROOT and PHENIX libraries
- ...not obvious (since libraries are loaded dynamically), dependencies on local installs of various packages - need to be collected, mostly by hand, labor-intensive
  - PHENIX libraries are 1.2GB, root5 is 240MB, package libs are tiny but numerous, with the base SL7 the total is ~2.5GB
  - Having a Docker-capable machine at BNL would have been of help
- **Created a working image that successfully runs PHPythia**
  - AFS mounted on a laptop to collect software
  - Need to test other use cases to validate dependencies
  - See next slide for screenshots
- If problems continue to pop up will probably shift focus to option “2”
- If successful, the image can go on Docker Hub

Docker “1”: 32-bit PHENIX container on a 64-bit laptop, Fun4All etc



```
 137 gamma          1        22  102   -0.153    0.090   -0.308    0.355    0.000
 138 gamma          1        22  104   -0.001    0.079   -0.673    0.677    0.000
 139 gamma          1        22  104    0.116    0.003   -0.869    0.877    0.000
 140 gamma          1        22  112    0.060   -0.031   -0.195    0.207    0.000
 141 gamma          1        22  112    1.076   -0.638   -5.789    5.922    0.000
 142 pi+            1       211  118   -0.095   -0.009    1.939    1.947    0.140
 143 pi-            1      -211  118   -0.877    0.495    8.126    8.190    0.140
 144 pi-            1      -211  121   -0.318   -0.005    3.576    3.593    0.140
 145 pi+            1       211  121    0.054    0.128    0.988    1.007    0.140
 146 gamma          1        22  126    0.035    0.014    0.042    0.056    0.000
 147 gamma          1        22  126   -0.088    0.030   -0.027    0.096    0.000
 148 gamma          1        22  127   -0.210   -0.517    0.171    0.584    0.000
 149 gamma          1        22  127    0.018   -0.026    0.023    0.039    0.000
 150 gamma          1        22  133   -0.067    0.037    1.287    1.289    0.000
 151 gamma          1        22  133   -0.031    0.008    0.055    0.064    0.000
 ==============================================================================
                  sum:  2.00          0.000    0.000   -0.000  200.000  200.000
1********* PYSTAT:  Statistics on Number of Events and Cross-sections *********

 ==============================================================================
 I                                  I                            I            I
 I            Subprocess            I      Number of points      I    Sigma   I
 I                                  I                            I            I
 I----------------------------------I----------------------------I    (mb)    I
 I                                  I                            I            I
 I N:o Type                         I    Generated         Tried I            I
 I                                  I                            I            I
 ==============================================================================
 I                                  I                            I            I
 I   0 All included subprocesses    I         1000         11038 I  2.491D+01 I
 I  11 f + f' -> f + f' (QCD)       I           67           743 I  1.477D+00 I
 I  12 f + fbar -> f' + fbar'       I            1             7 I  1.808D-02 I
 I  13 f + fbar -> g + g            I            0             9 I  3.193D-02 I
 I  28 f + g -> f + g               I          427          5458 I  1.052D+01 I
 I  53 g + g -> f + fbar            I           13            94 I  3.457D-01 I
 I  68 g + g -> g + g               I          492          4727 I  1.251D+01 I
 I                                  I                            I            I
 ==============================================================================

 ********* Total number of errors, excluding junctions =       0 *************
 ********* Total number of errors, including junctions =       0 *************
 ********* Total number of warnings =                          0 *************
 ********* Fraction of events that fail fragmentation cuts =  0.00000 *********

root [1] .q
[root@de2d13b3d67f phpythia]# ls -ltr
total 4112
-rw-r--r-- 1 root root    2658 Apr 22 01:13 phpythia.C
-rw-r--r-- 1 root root     452 Apr 22 01:14 pythia.cfg
-rw-r--r-- 1 root root    6219 Apr 22 01:34 phpy_xsec.root
-rw-r--r-- 1 root root 4190589 Apr 22 01:34 phpythia.root
[root@de2d13b3d67f phpythia]#
```

Docker

- ***Important - need simple yet realistic use cases/analyses to test these containers in operation***
  - Can the DAP team help? - install Docker, run a container with some macros?
  - Essentially need samples of simple macros, based on the PHENIX software stack
- Having a valid image is only 20% of work, need input from experts to package analyses for Docker (may be as simple as creating folder with all necessary components)
- Should pick those macros not requiring DB access at first

REANA and Docker

- Progress with REANA will depend on successful creation of 32-bit images (PHENIX software including fun4all, root etc)
- Minimalistic ROOT macros (e.g. such as created by Gabor) can be used to demo REANA at the PHENIX School right away
- As mentioned earlier, a better estimate in ~2 weeks
- Please suggest analyses to be tested in REANA

School

- Leveraging the PHENIX website to support School activities
  - Makes it easier to stay organized, increases visibility of the site, encourages contributions
- Curated School material should probably be pre-uploaded to Zenodo
- Agendas can easily link to items committed to GitHub, Zenodo, the website
  - Materials no longer trapped in CDS and instead are easily discoverable
- Giving a simple REANA demo at the School makes a lot of sense
  - Currently there isn’t a lot of REANA awareness, and it will be useful for students even beyond PHENIX
  - Accounts need to be created prior to School
- Content and complexity of containerized use cases is TBD
  - Need participation of the DAP group
- At least an intro to HEPData makes 100% sense - accounts need to be created prior to School

Open Data

- Uploaded to CERN, admins notified, awaiting response
- No news for weeks now, will give them more time

Plans

- Docker images
- Ongoing HEPData work
- PHENIX School