Data & Analysis Preservation: status update
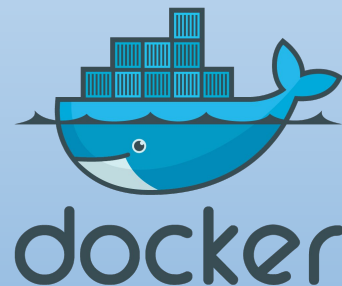
Maxim Potekhin
Nuclear and Particle Physics Software Group

BROOKHAVEN
NATIONAL LABORATORY

PHENIX DAP Meeting
05/06/2021

Overview

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

Zenodo+Website

- More uploads of conference materials (thanks Gabor)
- Keywords added, conference page updated
- Added info on running Docker containers in **Singularity** (on SDCC nodes at BNL)
- Consolidated sparse pages in the Analysis section to save space in the drop-down menu
- Expanded section on REANA
  - Improved client installation instructions
  - Corrections in the procedure

HEPData

- The spreadsheet has been updated
- Some older entries revisited, teams notified
- PPG234 close to final
- PPG147 pending assignment of a reviewer by the IRC

BROOKHAVEN
NATIONAL LABORATORY

Docker: the two options

- ***Copying 32-bit binaries/dependencies to a Docker image***
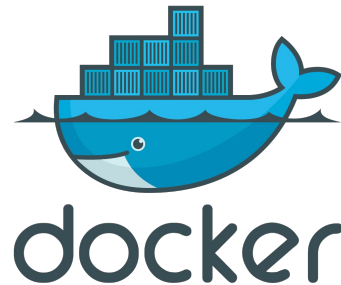- ***Use SDCC-provided images***
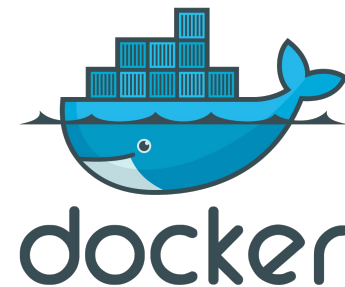
Docker: custom 32-bit image

- ***Copying 32-bit binaries/dependencies to a Docker image***
- In addition to PHPythia, assembled many libraries from the PHENIX software stack
- **Basic testing** done with existing macros - doesn't crash
  - But didn't validate physics yet
- Stay tuned
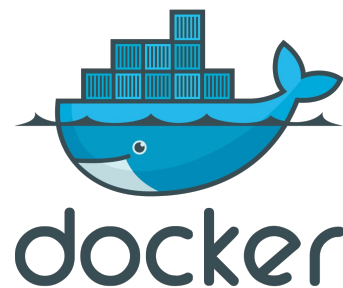
Docker: SDCC

- **Use SDCC-provided images**
  - *The image built in early April turned out to be incomplete*
  - *Awaiting news from our SDCC colleagues*
  - *Identified challenge - a total size of ~10GB of the total RACF+PHENIX stack*
    - *Pushing various limits including Docker itself*
    - *Could be optimized (a guess)*
  - *Due to the very large cumulative size of the stack may have to use CVMFS*
    - *That's actually a good option but **requires more work***
    - *Phased in for the REANA EIC activity elsewhere*
    - *A network file system allowing transparent access to libraries*

Docker

- We don't have (yet) machines on the SDCC farm that can run Docker interactively
- However, it is possible to make use of compatibility between Singularity and Docker and run images on rcas nodes where Singularity is available by default
- Documentation page has been updated on our dev website
- Works reasonably well with X11 even without NX

```
singularity exec --bind /phenix/u/phnxuser:/user \
docker://phenixcollaboration/tools:sl7_root5 bash
```

REANA accounts

- Received an OK from admins to create accounts
  - For the DAP team members
    - Akiba-san obtained an account and started using REANA
    - Stacyann will apply for an account shortly
  - For participants of the PHENIX School in June 2021
    - Around 16 people
    - Having a SDCC account is a prerequisite
- DAP team is very much welcome to try using REANA (it's fun, actually)
- In addition to the password giving you access to the very basic GUI you get an access token which is mandatory for access to the service
  - Tokens are not sent via e-mail, you need access to site to get the value

REANA access

- Connect to the landing page of the REANA server, "sing-up"
- Receive a confirmation from Chris Hollowell
- Log-in
- Get a working example from the RANA page and try it out
- You'll need client software
- Note: REANA is not a part of the general RACF cluster, it's a completely separate entity
  - All interaction via clients (Web+CLI)
  - Data (and code, if needed) is staged in and staged out - no folders are shared with your machine
- Workflows can be restarted at will, not necessarily uploaded and started from scratch (this also relates to data which can be updated)
- **Can use your machine at home or an interactive rcas node for access**

BROOKHAVEN
NATIONAL LABORATORY

REANA client

- Installation of the REANA client is essential
  - cf. the GUI provides only a subset of functionality
    - e.g. impossible to initiate workflows
  - Python-based (now Python 3.6 is highly recommended, Python 2 deprecated)
  - Trivial on a Linux machine (and presumably WSL2)
  - Tested on the rcas interactive nodes (Python 3.8 "virtual environment")
- Note that the client handles all aspects of REANA operation
  - Managing workflows
  - Staging in and staging out of the data
    - There is no connection between the machines in the REANA cluster and the shared file system like other RCAS systems
- Fairly complete automation of analyses is possible by using the client in scripts

REANA and Docker
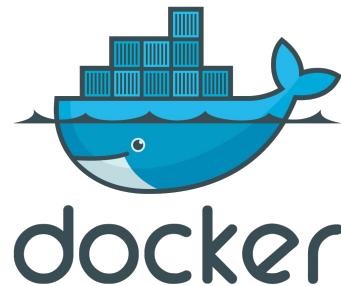
- Images can be pulled from Docker Hub **OR** the local BNL registry (only available within the perimeter) - both tested
- Not sure if the full PHENIX image won't materialize in time for the School, also a substantial work-up needed for the content
- However in addition to the EMCal Ntuple analysis macros (Gabor) there may be other candidates which we can "reanize"
  - e.g. Genki's recent VTX analysis, testing by Akiba-san
  - Encouraging - we can hopefully produce more of these materials
- In cases where it's optimal, can use Singularity to run Docker containers w/o REANA
  - certainly a good way to test, perhaps also a legitimate way to package a few analyses.
  - cf. it's easy to preserve image e.g. as .tar files

School

- Two DAP-related items agreed upon in a recent prep meeting:
  - General DAP overview and DAP status in PHENIX
  - Hands-on REANA session
- Christine - HEPData (hopefully)
- Comments?

Open Data

- Uploaded an updated version of the EMCal "Ntuple analysis" package to CERN ~2 months ago, admins notified
- **Just got a response (**admins were not available before), action forthcoming
- Note one recent VTX analysis based C++ macros in its final stages
  - As mentioned, also tested by Akiba-san
  - Can this be another Open Data candidate?

BROOKHAVEN
NATIONAL LABORATORY

Plans

- Docker images - ongoing work
  - The simple SL7/ROOT5 image works perfectly with PHENIX ROOT macros
  - Encountered challenges with the full RCAS/PHENIX build, work in progress
  - Likely will acquire a more limited "custom" image functionality in foreseeable future, with some (most) fun4all libraries and functions
- General REANA onboarding
- HEPData, steady state effort
- PHENIX School
  - Development of REANA tutorials and other materials