The MC Data Catalog - a possible starting point

Maxim Potekhin (BNL/NPPS)

May 12th 2021

The problem

- *In the long term* a data catalog will likely become necessary
- Need to keep the data *discoverable and accessible*
- During the YR process data discoverability was identified as an issue - but this didn’t have much on of impact due to scale
    - People relied on personal knowledge of data location and characteristics and a combination of Wikis, Google Drive folders and GitHub
- With detector simulations coming into the scope of the MC effort the data complexity will increase, bringing about
    - Issues of managing software configuration, both for MC generators and detector simulation, geometry management etc
    - Increased computational costs e.g. “losing” data becoming more expensive

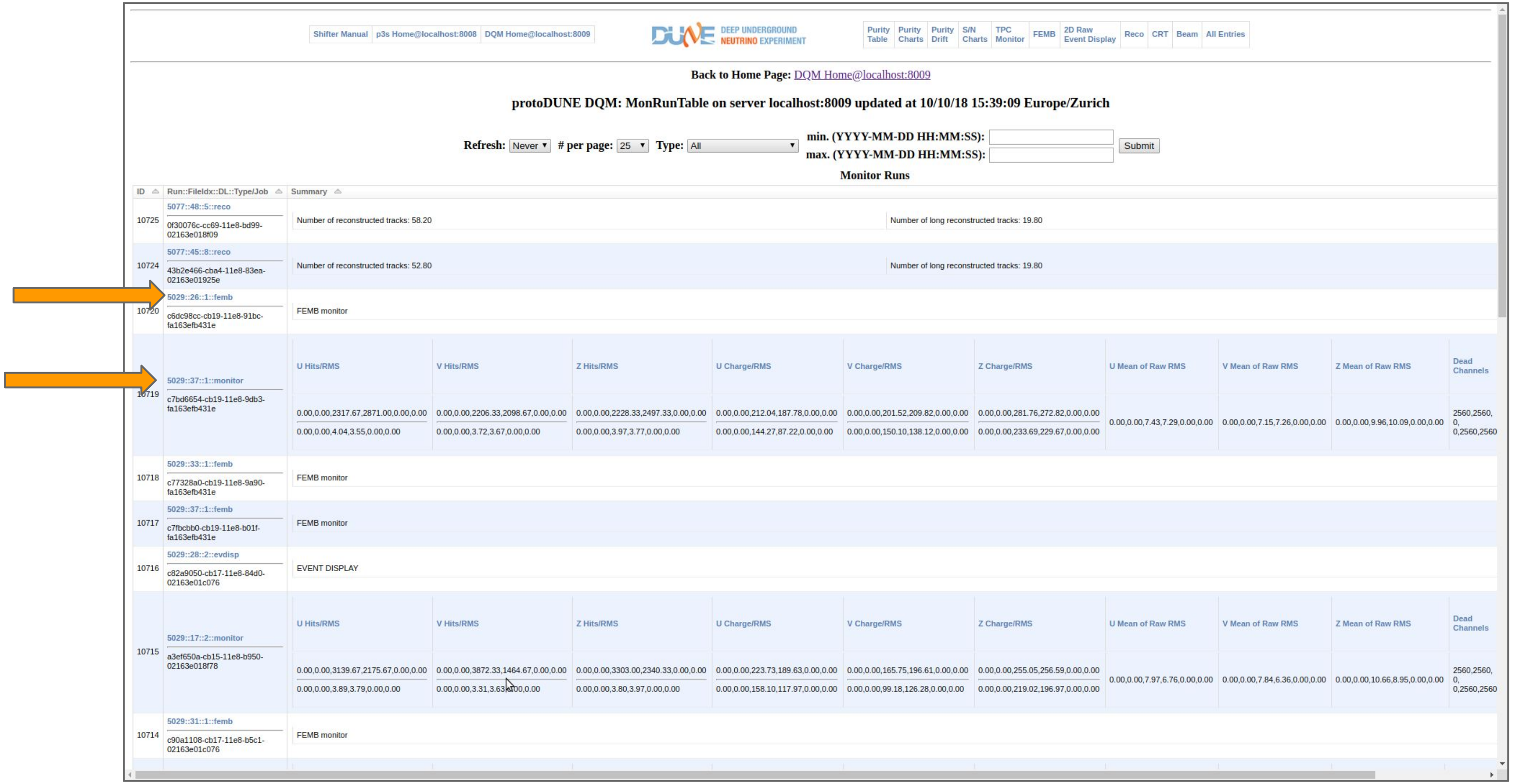
Modus Operandi

- If a *Workload Management System* (PanDA, Dirac etc) is implemented there is a need for integration and support for automated data management (including metadata)
- If simpler options are used i.e. direct submission to batch systems there is still a need to maintain records
- At the present stage in the development of the proto-collaborations and detector proposals for the EIC there is no immediately available solution to adopt without substantial effort upfront
    - However the need for such functionality may emerge in the next months - a problem of timescale
- Need a solution with a low cost of entry but extensible and forward-looking

Experience in protoDUNE-SP

- protoDUNE-SP was an experiment which ran at CERN in 2018-2020
  - A large-scale LAr TPC - a prototype of the Far Detector of the DUNE experiment
- Data Quality Monitoring: prompt processing i.e. several distinct types of jobs
  - Challenge: keep track of diverse data and their provenance, discover and navigate the data
- Solution
  - Save job configuration in a small JSON file
  - Establish a convention whereby each job (which can be of any variety) produces JSON files describing its outputs
  - Can be done in ROOT macros or in wrapper scripts
  - Result: automatic classification of the outputs and straightforward access in the UI
- Comment: in hindsight, using YAML instead of JSON would have been a functional equivalent but would provide much better readability.

protoDUNE-SP DQM screen: job summary



Back to Home Page: DQM Home@localhost:8009

protoDUNE DQM: MonRunTable on server localhost:8009 updated at 10/10/18 15:39:09 Europe/Zurich

Refresh: Never ▾ **# per page:** 25 ▾ **Type:** All ▾ **min. (YYYY-MM-DD HH:MM:SS):** **max. (YYYY-MM-DD HH:MM:SS):** Submit

Monitor Runs

ID	Run::FileIdx::DL::Type/Job	Summary	
10725	5077::48::5::reco 0f30076c-cc69-11e8-bd99-02163e018f09	Number of reconstructed tracks: 58.20	Number of long reconstructed tracks: 19.80
10724	5077::45::8::reco 43b2e466-cba4-11e8-83ea-02163e01925e	Number of reconstructed tracks: 52.80	Number of long reconstructed tracks: 19.80
10720	5029::26::1::femb c6dc98cc-cb19-11e8-91bc-fa163efb431e	FEMB monitor	
10719	5029::37::1::monitor c7bd6654-cb19-11e8-9db3-fa163efb431e		
10718	5029::33::1::femb c77328a0-cb19-11e8-9a90-fa163efb431e	FEMB monitor	
10717	5029::37::1::femb c7fbcbb0-cb19-11e8-b01f-fa163efb431e	FEMB monitor	
10716	5029::28::2::evdisp c82a9050-cb17-11e8-84d0-02163e01c076	EVENT DISPLAY	
10715	5029::17::2::monitor a3ef650a-cb15-11e8-b950-02163e018f78		
10714	5029::31::1::femb c90a1108-cb17-11e8-b5c1-02163e01c076	FEMB monitor	

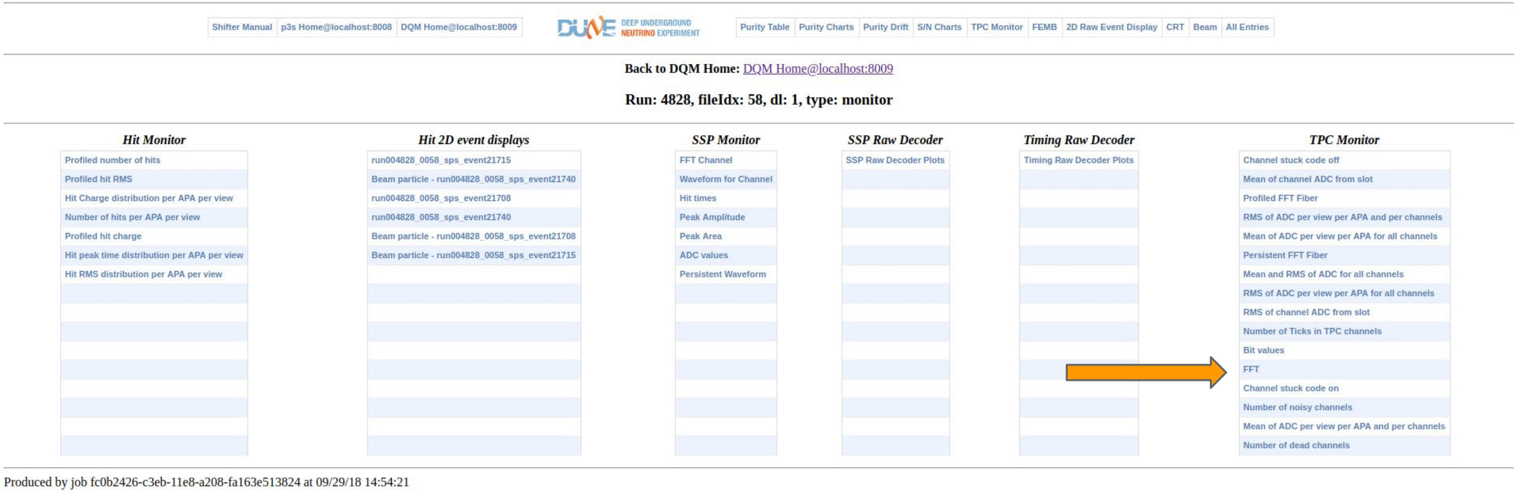
Summary of 10719 (5029::37::1::monitor):

U Hits/RMS	V Hits/RMS	Z Hits/RMS	U Charge/RMS	V Charge/RMS	Z Charge/RMS	U Mean of Raw RMS	V Mean of Raw RMS	Z Mean of Raw RMS	Dead Channels
0.00,0.00,2317.67,2871.00,0.00,0.00 0.00,0.00,4.04,3.55,0.00,0.00	0.00,0.00,2206.33,2098.67,0.00,0.00 0.00,0.00,3.72,3.67,0.00,0.00	0.00,0.00,2228.33,2497.33,0.00,0.00 0.00,0.00,3.97,3.77,0.00,0.00	0.00,0.00,212.04,187.78,0.00,0.00 0.00,0.00,144.27,87.22,0.00,0.00	0.00,0.00,201.52,209.82,0.00,0.00 0.00,0.00,150.10,138.12,0.00,0.00	0.00,0.00,281.76,272.82,0.00,0.00 0.00,0.00,233.69,229.67,0.00,0.00	0.00,0.00,7.43,7.29,0.00,0.00	0.00,0.00,7.15,7.26,0.00,0.00	0.00,0.00,9.96,10.09,0.00,0.00	2560,2560, 0, 0,2560,2560

Summary of 10715 (5029::17::2::monitor):

U Hits/RMS	V Hits/RMS	Z Hits/RMS	U Charge/RMS	V Charge/RMS	Z Charge/RMS	U Mean of Raw RMS	V Mean of Raw RMS	Z Mean of Raw RMS	Dead Channels
0.00,0.00,3139.67,2175.67,0.00,0.00 0.00,0.00,3.89,3.79,0.00,0.00	0.00,0.00,3872.33,1464.67,0.00,0.00 0.00,0.00,3.31,3.63,0.00,0.00	0.00,0.00,3303.00,2340.33,0.00,0.00 0.00,0.00,3.80,3.97,0.00,0.00	0.00,0.00,223.73,189.63,0.00,0.00 0.00,0.00,158.10,117.97,0.00,0.00	0.00,0.00,165.75,196.61,0.00,0.00 0.00,0.00,99.18,126.28,0.00,0.00	0.00,0.00,255.05,256.59,0.00,0.00 0.00,0.00,219.02,196.97,0.00,0.00	0.00,0.00,7.97,6.76,0.00,0.00	0.00,0.00,7.84,6.36,0.00,0.00	0.00,0.00,10.66,8.95,0.00,0.00	2560,2560, 0, 0,2560,2560

protoDUNE-SP DQM screen: auto-generated menus



Back to DQM Home: DQM Home@localhost:8009

Run: 4828, fileIdx: 58, dl: 1, type: monitor

Hit Monitor

- Profiled number of hits
- Profiled hit RMS
- Hit Charge distribution per APA per view
- Number of hits per APA per view
- Profiled hit charge
- Hit peak time distribution per APA per view
- Hit RMS distribution per APA per view

Hit 2D event displays

- run004828_0058_sps_event21715
- Beam particle - run004828_0058_sps_event21740
- run004828_0058_sps_event21708
- run004828_0058_sps_event21740
- Beam particle - run004828_0058_sps_event21708
- Beam particle - run004828_0058_sps_event21715

SSP Monitor

- FFT Channel
- Waveform for Channel
- Hit times
- Peak Amplitude
- Peak Area
- ADC values
- Persistent Waveform

SSP Raw Decoder

- SSP Raw Decoder Plots

Timing Raw Decoder

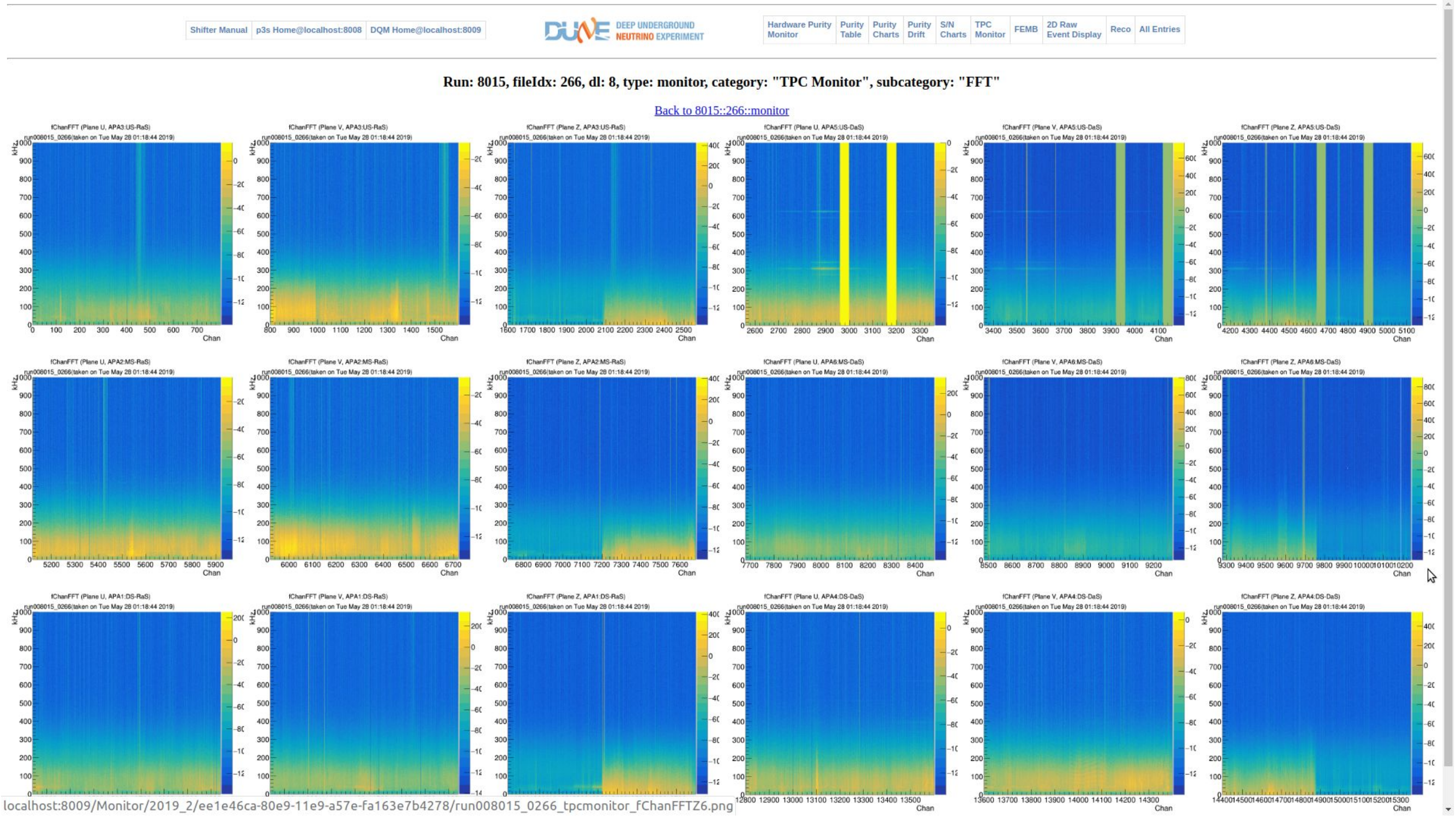
- Timing Raw Decoder Plots

TPC Monitor

- Channel stuck code off
- Mean of channel ADC from slot
- Profiled FFT Fiber
- RMS of ADC per view per APA and per channels
- Mean of ADC per view per APA for all channels
- Persistent FFT Fiber
- Mean and RMS of ADC for all channels
- RMS of ADC per view per APA for all channels
- RMS of channel ADC from slot
- Number of Ticks in TPC channels
- Bit values
- FFT
- Channel stuck code on
- Number of noisy channels
- Mean of ADC per view per APA and per channels
- Number of dead channels

Produced by job fc0b2426-c3eb-11e8-a208-fa163e513824 at 09/29/18 14:54:21

protoDUNE-SP DQM screen: data products


Shifter Manual
p3s Home@localhost:8008
DQM Home@localhost:8009
DUNE DEEP UNDERGROUND NEUTRINO EXPERIMENT
Hardware Purity Monitor
Purity Table
Purity Charts
Purity Drift
S/N Charts
TPC Monitor
FEMB
2D Raw Event Display
Reco
All Entries
fChanFFT (Plane U, APA3:US-RaS)
fChanFFT (Plane V, APA3:US-RaS)
fChanFFT (Plane Z, APA3:US-RaS)
fChanFFT (Plane U, APA5:US-DaS)
fChanFFT (Plane V, APA5:US-DaS)
fChanFFT (Plane Z, APA5:US-DaS)
fChanFFT (Plane U, APA2:MS-RaS)
fChanFFT (Plane V, APA2:MS-RaS)
fChanFFT (Plane Z, APA2:MS-RaS)
fChanFFT (Plane U, APA6:MS-DaS)
fChanFFT (Plane V, APA6:MS-DaS)
fChanFFT (Plane Z, APA6:MS-DaS)
fChanFFT (Plane U, APA1:DS-RaS)
fChanFFT (Plane V, APA1:DS-RaS)
fChanFFT (Plane Z, APA1:DS-RaS)
fChanFFT (Plane U, APA4:DS-DaS)
fChanFFT (Plane V, APA4:DS-DaS)
fChanFFT (Plane Z, APA4:DS-DaS)
run008015_0266(taken on Tue May 28 01:18:44 2019)
kHz
Chan

Run: 8015, fileIdx: 266, dl: 8, type: monitor, category: "TPC Monitor", subcategory: "FFT"

Back to 8015::266::monitor

localhost:8009/Monitor/2019_2/ee1e46ca-80e9-11e9-a57e-fa163e7b4278/run008015_0266_tpcmonitor_fChanFFTZ6.png

The descriptor

- A possible solution with a low cost of entry - yet hopefully future-proof - could be to establish a convention where files are accompanied with short descriptors/metadata records formatted in YAML (or JSON)
  - e.g. myFile.root comes with myFile.root.yaml (or, myFile.txt with myFile.txt)
  - The volume/cost of extra data is “small” (technically depends on implementation)
  - Far superior to using filenames and/or folder names as metadata - this is hard to extend and it may not scale well
  - Schema can be updated/augmented at a later time
  - The process is asynchronous i.e. no DB is updated in real time or at all
- If a folder is moved or copied, the descriptors remain with the data

The contents

- The descriptor would keep information on the file provenance and various aspects of configuration of the software used (all TBD)
  - Type (e.g. MCEG vs G4 etc)
  - Version, references to configuration files (tags)
  - Number of events
  - md5, sha-1 or other hash
  - …
- Configuration files for MCGENs can be referred to by their SHA-1 hash in git/GitHub which guarantees non-ambiguity and audit trail, also is compact

The catalog

- YAML provides a high degree of (organic) compatibility with the current technology used on the ECCE Software Documentation site (Liquid/Jekyll)

- A master catalog can be compiled and/or recompiled by skimming the data descriptors
  - Can be hosted as YAML or a simple Web app (e.g. Django-based) can be created
  - A few different search mechanisms are available

- **Not locking into any technology at this point**

- Since parsing YAML is “almost free” these metadata can be ingested in future databases chosen by the collaboration
  - Granted, metadata still needs thought and design
  - Future system can be RDBMS or noSQL-like in their properties

Questions, other considerations...

- Files vs datasets?
- Tags
- XRootD
- Capabilities of Rucio
  - Widely used in HEP - at scale
  - “Extensible Metadata” feature has become available in the past year
  - Deployment for EIC requires effort (TBD)