LBL Reco 2 Physics datasets

Physics datasets

- A physics dataset is defined as a collection of files. In the context of MetaCat, a collection is understood as a group of files derived from an MQL query. Conversely, in the SAM framework, it refers to a collection of files from a dataset definition
- We have defined a policy for datasets, see docdb 29787
- Are we ready to provide official dataets?

Physics datasets

===== Neutrinos input dataset
https://metacat.fnal.gov:9443/dune_meta_prod/app/gui/dataset?namespace=higuera&name=fardet-hd__fd_mc_2023a__mc__hit-reconstructed__prodgenie_nu_dune10kt_1x2x6.fcl__v09_78_01d01__preliminary
Input Files: 20784

metacat query -s " files where core.application.version=v09_81_00d02 and core.application.name=reco2 and core.data_stream=out1 and core.data_tier='full-reconstructed' and core.file_type=mc and core.run_type='fardet-hd' and dune.campaign=fd_mc_2023a_reco2 and dune.config_file=standard_reco2_dune10kt_nu_1x2x6.fcl and dune.requestid=ritm1780305 and dune_mc.detector_type='fardet-hd' and dune_mc.gen_fcl_filename=prodgenie_nu_dune10kt_1x2x6.fcl and dune.output_status=confirmed and core.group=dune "

Files: 18443

We are missing 2341… not bad but…

Physics datasets

===== electron neutrinos input dataset
https://metacat.fnal.gov:9443/dune_meta_prod/app/gui/dataset?namespace=higuera&name=fardet-hd__fd_mc_2023a__mc__hit-reconstructed__prodgenie_nue_dune10kt_1x2x6.fcl__v09_78_01d01__preliminary
Input Files: 20826

metacat query -s " files where core.application.version=v09_81_00d02 and core.application.name=reco2 and core.data_stream=out1 and core.data_tier='full-reconstructed' and core.file_type=mc and core.run_type='fardet-hd' and dune.campaign=fd_mc_2023a_reco2 and dune.config_file=standard_reco2_dune10kt_nu_1x2x6.fcl and dune.requestid=ritm1780305 and dune_mc.detector_type='fardet-hd' and dune_mc.gen_fcl_filename=prodgenie_nue_dune10kt_1x2x6.fcl and dune.output_status=confirmed and core.group=dune "

Files: 40713

No quite ready….did not we nuke duplicates?
Let's look at one file

https://metacat.fnal.gov:9443/dune_meta_prod/app/gui/show_file?show_form=yes&namespace=&name=&did=fardet-hd%3Anue_dune10kt_1x2x6_1407_19_20230826T073429Z_gen_g4_detsim_hitreco__20240218T034332Z_reco2.root&fid=

| fardet-hd | nue_dune10kt_1x2x6_1407_19_20230826T073429Z_gen_g4_detsim_hitreco__20240218T034332Z_reco2.root | 2.037 GB | 5a4d8f93 | DUNE_CERN_EOS: root://eospublic.cern.ch:1094//eos/experiment/neutplatform/protodune/dune/fardet-hd/c3/be/nue_dune10kt_1x2x6_1407_19_20230826T073429Z_gen_g4_detsim_hitreco__20240218T034332Z_reco2.root

Physics datasets

https://metacat.fnal.gov:9443/dune_meta_prod/app/gui/show_file?show_form=yes&namespace=&name=&did=fardet-hd%3Anue_dune10kt_1x2x6_1407_19_20230826T073429Z_gen_g4_detsim_hitreco__20240218T034332Z_reco2.root&fid=

core.first_event_number	1901
core.group	dune
core.last_event_number	2000
core.parents	[{'file_name': 'fardet-hd:nue_dune10kt_1x2x6_1407_19_20230826T073429Z_gen_g4_detsim_hitreco.root'}]
core.run_type	fardet-hd

```
<dunegpvm11.fnal.gov> metacat query "files where core.parents['file_name']= 'fardet-
hd:nue_dune10kt_1x2x6_1407_19_20230826T073429Z_gen_g4_detsim_hitreco.root'"
fardet-hd:nue_dune10kt_1x2x6_1407_19_20230826T073429Z_gen_g4_detsim_hitreco__20240218T034332Z_reco2.root
fardet-hd:nue_dune10kt_1x2x6_1407_19_20230826T073429Z_gen_g4_detsim_hitreco__20240221T015902Z_reco2.root
```

Files was produced in workflow_id = 1568

From Jake logs https://docs.google.com/spreadsheets/d/1lec1kMoZFgyzzQDk9Vjmy3h0VbRjDsp_AXrTERtWxkE/edit#gid=811082660 workflow_id 1600 “50 files Failed, resubmitted with more memory and to replace missing files original submission - 1568”

One file corresponds to workflow_id 1568 and the other to 1600

Physics datasets

===== electron neutrinos input dataset
https://metacat.fnal.gov:9443/dune_meta_prod/app/gui/dataset?namespace=higuera&name=fardet-hd__fd_mc_2023a__mc__hit-reconstructed__prodgenie_nue_dune10kt_1x2x6.fcl__v09_78_01d01__preliminary
Input Files: 20826

```
metacat query -s " files where core.application.version=v09_81_00d02 and core.application.name=reco2
core.data_stream=out1 and core.data_tier='full-reconstructed' and core.file_type=mc and
core.run_type='fardet-hd' and dune.campaign=fd_mc_2023a_reco2 and
dune.config_file=standard_reco2_dune10kt_nu_1x2x6.fcl and dune.requestid=ritm1780305 and
dune_mc.detector_type='fardet-hd' and dune_mc.gen_fcl_filename=prodgenie_nue_dune10kt_1x2x6.fcl and
dune.output_status=confirmed and core.group=dune and (dune.workflow['workflow_id']=1599 or
dune.workflow['workflow_id']=1600 or dune.workflow['workflow_id']=1601 or
dune.workflow['workflow_id']=1602) "
Files:          19890
```

There are 4 workflow_id that Steve reported to be ready, this will allow to provide “clean datasets”

However, duplicate files are still available in MetaCat and Rucio

I guess we can nuke all the data from failing workflow_ids (1567, 1568, 1569, 1575…)

Discussion

How does the resubmission procedure work?
“50 files Failed, resubmitted with more memory and to replace missing files original submission - 1568”

It seems that we should develop a procedure for only resubmitting failing jobs,
However, it seems that we are resubmitting the whole workflow again, which is causing duplicates

We need to find a solution now… otherwise we are run out of storage

The current queries my result in a different list of files
“files from higuera:fardet-hd__fd_mc_2023a__mc__hit-reconstructed__prodgenie_nue_dune10kt_1x2x6.fcl__v09_78_01d01__preliminary skip 0 limit 5000”

We should add “ordered” to be safe
“files from higuera:fardet-hd__fd_mc_2023a__mc__hit-reconstructed__prodgenie_nue_dune10kt_1x2x6.fcl__v09_78_01d01__preliminary skip 0 limit 5000 **ordered**”