protoDUNE 2 GeV MC Official Datasets

Context

It was requested additional data for the MC 2 GeV protoDUNE analysis, this includes files with SCE ON and SCE OFF

The original production was done in summer 2021

A new production (with the same old software) was complete early this year and it was done in POMS

The final step of a production is to generate official physics datasets by data collection manager

An official datasets is a collection of files from a MQL query in MetaCat

Context

The original 2021 dataset was declared with the legacy FTS and SAM system

The 2024 dataset was declared via the DeclaD.

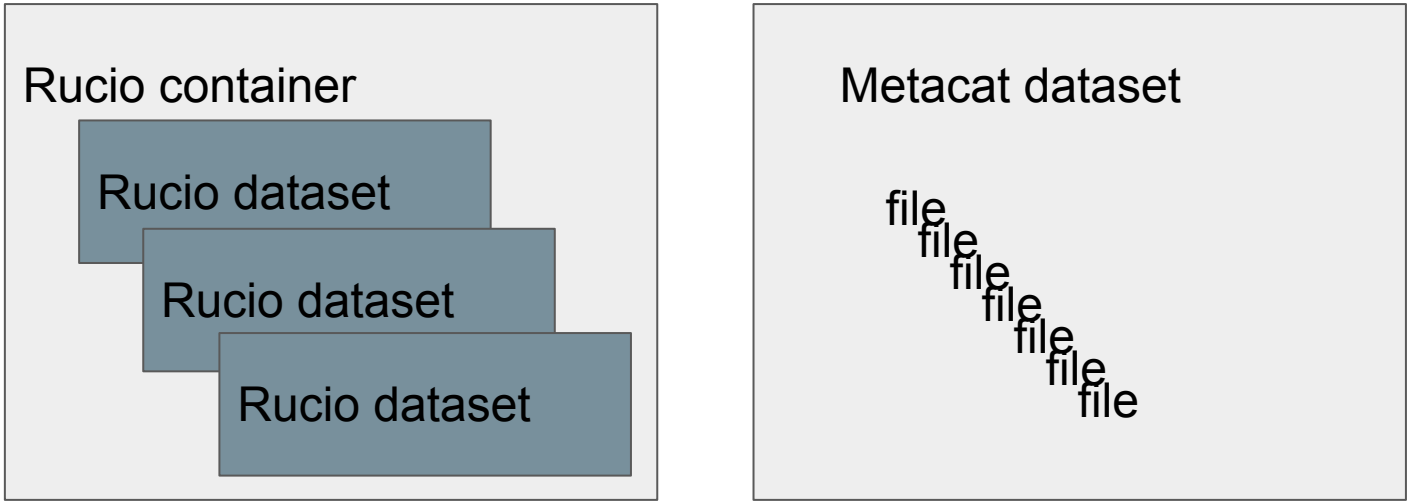
When generating the official datasets we observed a few “features”

- Set 1, production from 2021 was declared in rucio scope `pdsp_mc_reco`
- Set 2, production from 2024 was declared in rucio scope `protodune-sp`

DCM proceed to create two set of official datasets

Rucio and MetaCat

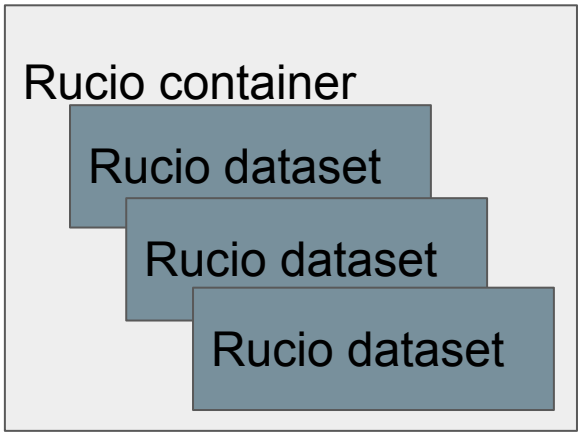
To accommodate best practices when dealing with official data a one-to-one correspondence between Rucio container and MetaCat dataset is imposed and we have been able to achieve that for the past production (e.g. FD reco2)



New rules a.k.a transferring data

During the creation of Set1 Rucio container a few files where found to have not Rucio replicas (handful of files) these files were retired from MetaCat and the dataset was produced and one-to-one correspondence was achieved

Then it was suggested to make an extra copy of the data in another RSEs, meaning creating a new rules for the data



SURFSARA (sce ON)
&
NIKHEF (sce OFF)

Some rules got status stuck

Solution to complete new rules at SURFSARA/NIKHEF

Or maybe there is no solution and we will NUKE the files

78 files out of each of the pdsp_mc_reco (off and datadriven) didn’t make it to SURFSARA and NIKHEF. Those which have been examined thus far have been found to have file size and checksum different than the checksum and file size at the time the file was declared to SAM and written to tape. It is likely that this was due to a failure mode in FTS which allowed files to be written twice.

Rucio

2021 data set was declared under fts/ 2024 was declared under declad.

when generating the official datasets we observed a few “features”

Set 1, production from 2021 was declared in rucio scope `pdsp_mc_reco`

Set 2, production from 2024 was declared in rucio scope `protodune-sp`

2024 Production (Set2) produced both SCE ON and SCE OFF at the same time and they end up being declared in the same rucio dataset

This is a feature of DeclaD–stuff in the same run will end up in the same data set, also dataset name is hardwired to core.run_type:core.run_type_core.run_number –we know how to change this now.

Proposed solution to DM, split them up such that we have a one-to-one correspondence between Rucio and MetaCat official datasets==should be possible but not first priority.

Summary

Requester needs official datasets to proceed with their analysis

This was a unique situation and should not happen again (at least not for an MC production)

2 GeV SCE OFF DONE!

Done