CAF production: examples of files processed twice

Workflow 2977

- 10 files/job

Stage ID	Files	Finding	Unallocated	Allocated	Outputting	Processed	Not found	Failed
1	2000	0	0	0	0	2000	0	0

- output dataset: usertests:fardet-hd-caf_2977

```
-bash-4.2$ metacat query  -s "files where dune.workflow['workflow_id'] in (2977)"
Files:          252
Total size:     333988500 (333.988 MB)
```

- Some input files (20) have been processed twice → output files having the same parents
- Three examples in next slide →

1) fardet-hd:atmnu_max_weighted_randompolicy_dune10kt_1x2x6_50581281_1000_20231204T053514Z_gen_g4_detsim_hitreco__20240509T212141Z_reco2.root

usertests:caf_20240831T075822Z.root (1k events) ⬅
usertests:caf_20240831T083248Z.root (1k events)

2) fardet-hd:atmnu_max_weighted_randompolicy_dune10kt_1x2x6_50574864_780_20231203T161822Z_gen_g4_detsim_hitreco.root

usertests:caf_20240831T075736Z.root (1k events)
usertests:caf_20240831T083245Z.root (100 events)

3) fardet-hd:atmnu_max_weighted_randompolicy_dune10kt_1x2x6_6472782_45_20231208T091351Z_gen_g4_detsim_hitreco.root
usertests:caf_20240831T083257Z.root (500 events)
usertests:caf_20240831T075822Z.root (1k events) ⬅

- In case a workflow has duplicates the “simplest” solution is to mark the output dataset as bad, and resubmit the workflow...but this requires additional work from DM team

- Check done on 13 workflows, duplications found in 4 workflow

- We need to understand

  1. why this happens
  2. how to minimize the double processing