Derived data for heavy-flavour analyses

Fabrizio Grosa
CERN

O^2 Tutorial
CERN | 8th November 2023

Derived data: definition

Derived data:

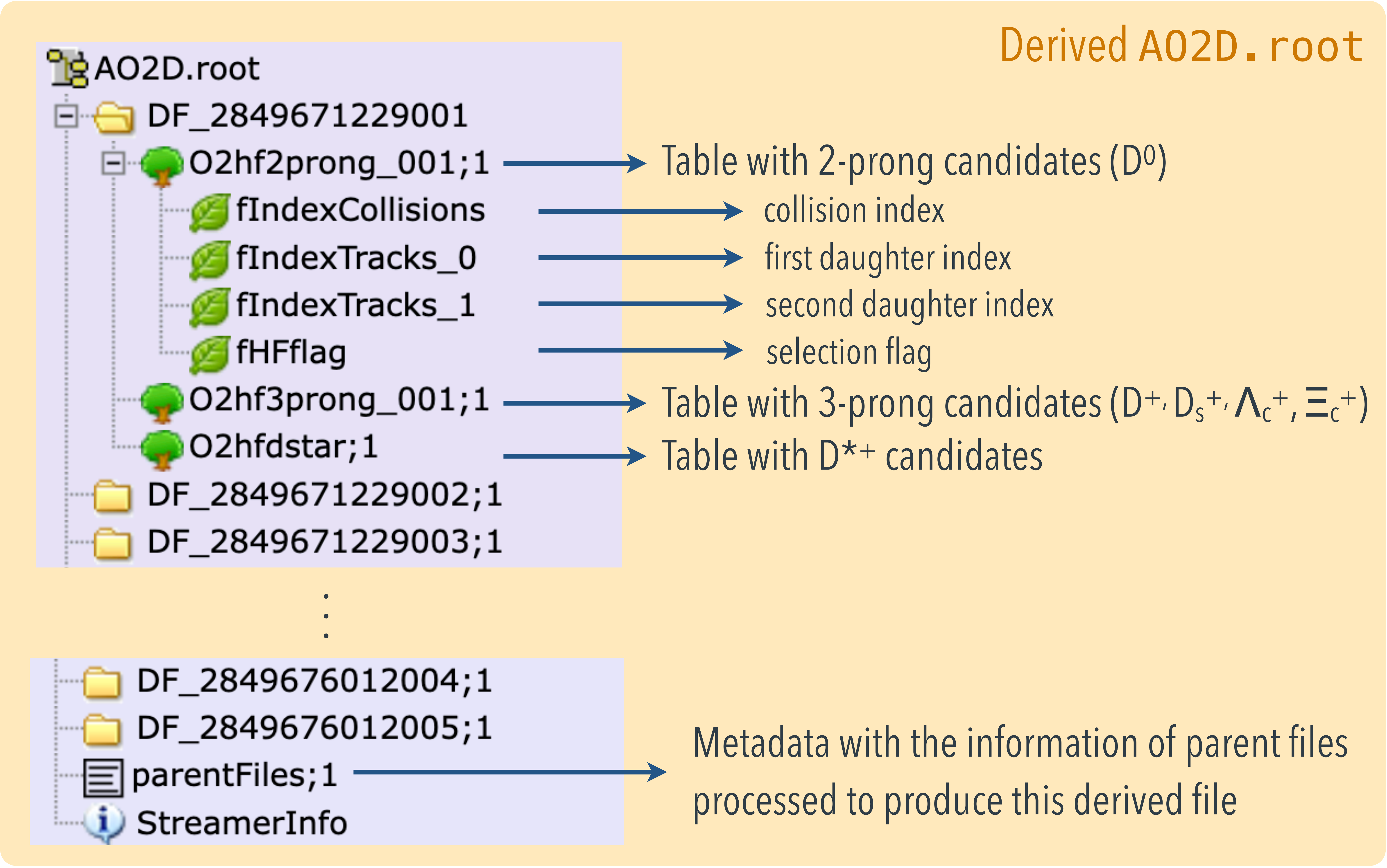
- What are they? `AO2D.root` files produced with a task that creates given tables by processing other `AO2D.root` files

- Why are they useful? By storing in the derived `AO2D.root` only the information needed for your analysis, you reduce the size of `AO2D.root` files to analyse and speed up the execution of your analysis code by skipping at least part of the workflows of your analysis

- Types of derived data
  - ➡ Self contained: derived `AO2D.root` files that contain all the information needed for your analysis that hence do not require to access the original `AO2D.root` files that were used to produce them
  - ➡ Linked: derived `AO2D.root` files that contain additional information with respect to the original `AO2D.root` files that were used to produce them and hence require access to the parent `AO2D.root` files

Derived data for HF from `trackIndexSkimCreator.cxx` task

- Charm-hadron decays are not found in the reconstruction step (you don't find them in the `AO2D.root` files), but at the analysis level with the [`trackIndexSkimCreator.cxx`]() task
  - ➡ It produces tables filled per candidate with indices and selection flags



Derived `AO2D.root`

```
AO2D.root
  DF_2849671229001
    O2hf2prong_001;1
      fIndexCollisions
      fIndexTracks_0
      fIndexTracks_1
      fHFflag
    O2hf3prong_001;1
    O2hfdstar;1
  DF_2849671229002;1
  DF_2849671229003;1
      ⋮
  DF_2849676012004;1
  DF_2849676012005;1
  parentFiles;1
  StreamerInfo
```

O2hf2prong_001;1 → Table with 2-prong candidates (D^0)

fIndexCollisions → collision index

fIndexTracks_0 → first daughter index

fIndexTracks_1 → second daughter index

fHFflag → selection flag

O2hf3prong_001;1 → Table with 3-prong candidates (D^{+}, D_s^{+}, Λ_c^{+}, Ξ_c^{+})

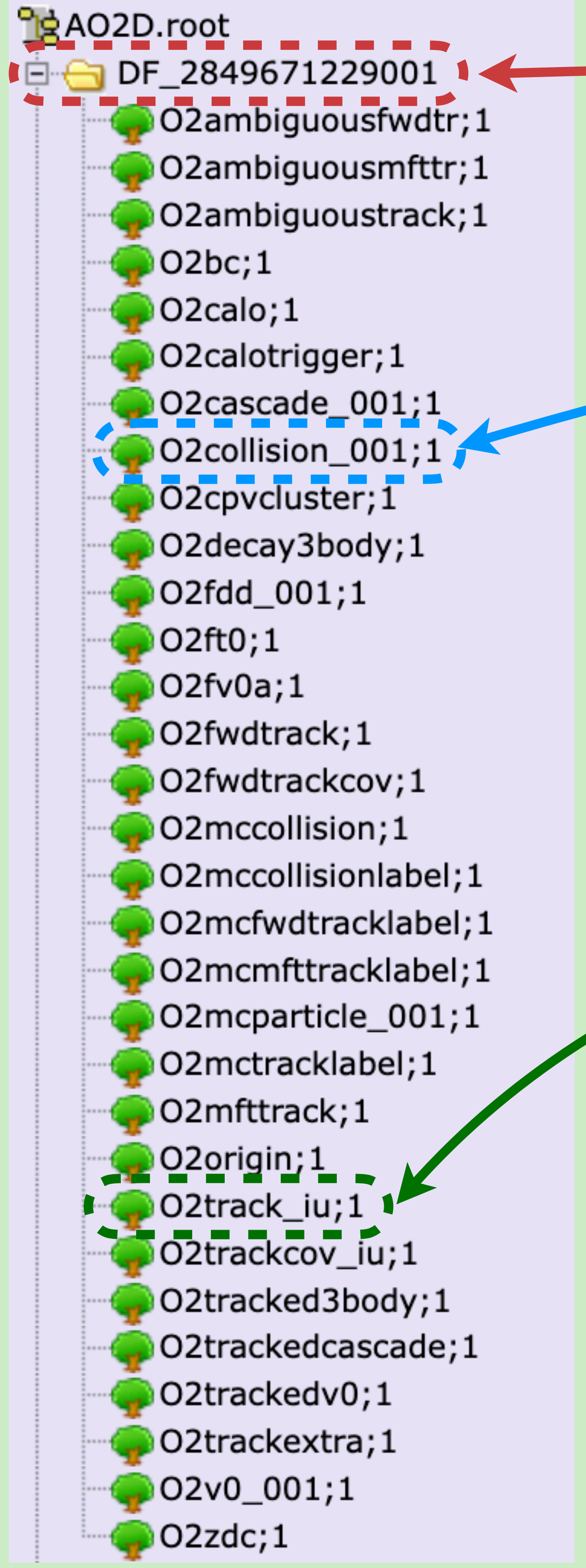
O2hfdstar;1 → Table with D^{*+} candidates

parentFiles;1 → Metadata with the information of parent files processed to produce this derived file

Derived data for HF from `trackIndexSkimCreator.cxx` task

- Charm-hadron decays are not found in the reconstruction step (you don't find them in the `AO2D.root` files), but at the analysis level with the [`trackIndexSkimCreator.cxx`]() task
  - ➡ It produces tables filled per candidate with indices and selection flags → linked derived data

Parent `AO2D.root`



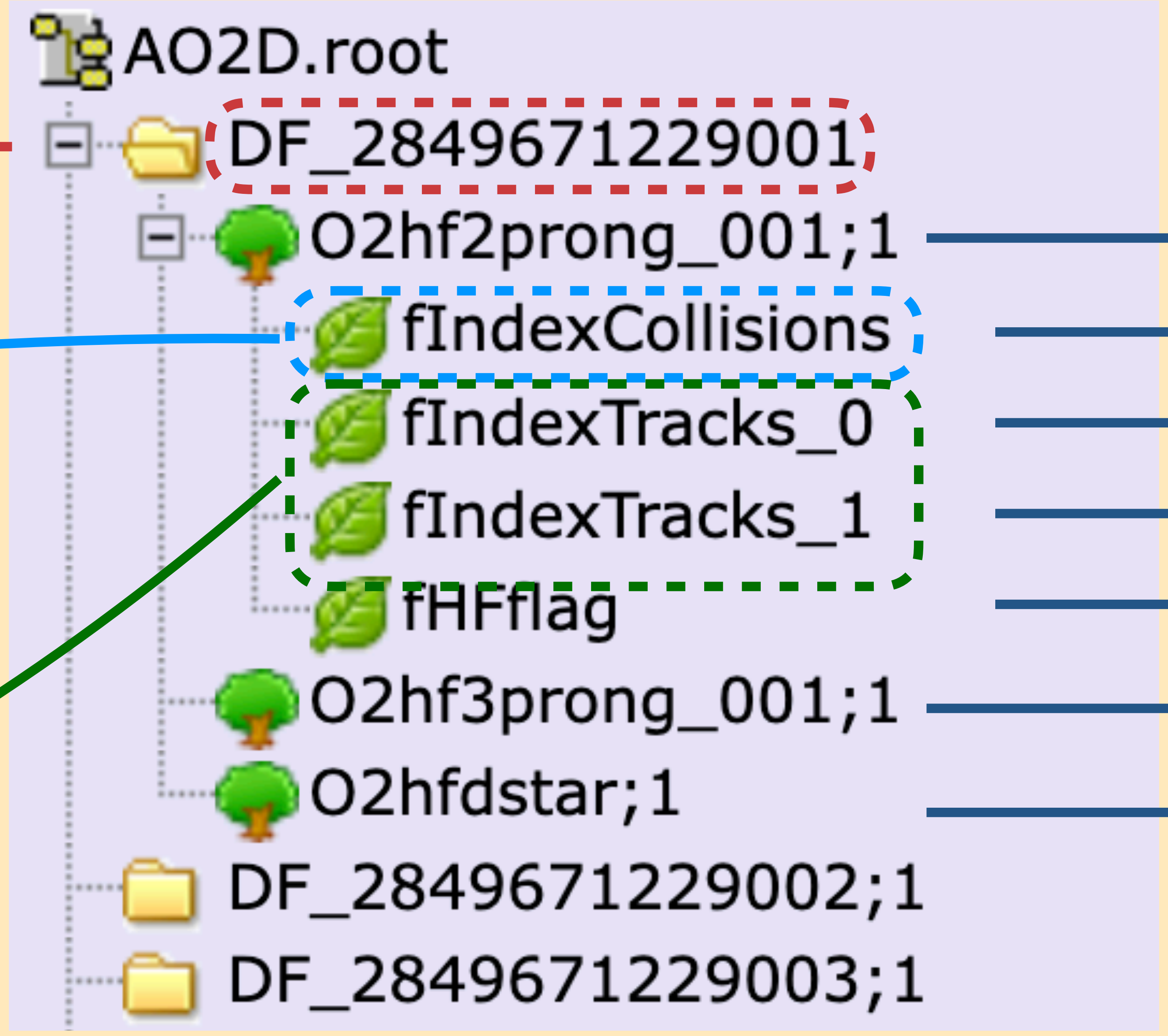
```
AO2D.root
⊟ DF_2849671229001
    O2ambiguousfwdtr;1
    O2ambiguousmfttr;1
    O2ambiguoustrack;1
    O2bc;1
    O2calo;1
    O2calotrigger;1
    O2cascade_001;1
    O2collision_001;1
    O2cpvcluster;1
    O2decay3body;1
    O2fdd_001;1
    O2ft0;1
    O2fv0a;1
    O2fwdtrack;1
    O2fwdtrackcov;1
    O2mccollision;1
    O2mccollisionlabel;1
    O2mcfwdtracklabel;1
    O2mcmfttracklabel;1
    O2mcparticle_001;1
    O2mctracklabel;1
    O2mfttrack;1
    O2origin;1
    O2track_iu;1
    O2trackcov_iu;1
    O2tracked3body;1
    O2trackedcascade;1
    O2trackedv0;1
    O2trackextra;1
    O2v0_001;1
    O2zdc;1
```

Same DF name

Index corresponding to entries in collision table

Indices corresponding to entries in track table

Derived `AO2D.root`



```
AO2D.root
⊟ DF_2849671229001
  ⊟ O2hf2prong_001;1
      fIndexCollisions
      fIndexTracks_0
      fIndexTracks_1
      fHFflag
    O2hf3prong_001;1
    O2hfdstar;1
  DF_2849671229002;1
  DF_2849671229003;1
```

O2hf2prong_001;1 → Table with 2-prong candidates (D^0)

fIndexCollisions → collision index

fIndexTracks_0 → first daughter index

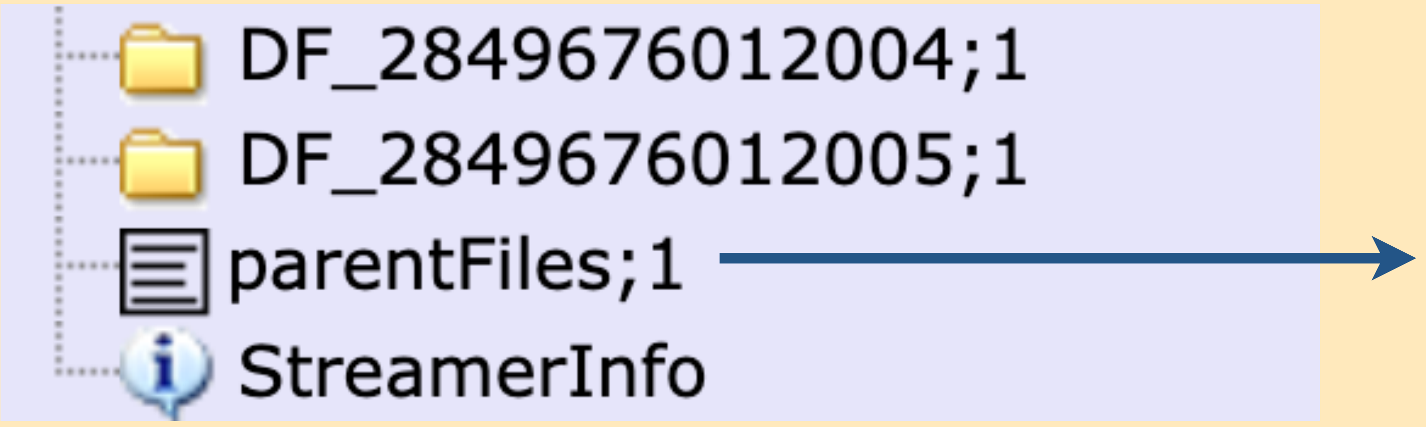
fIndexTracks_1 → second daughter index

fHFflag → selection flag

O2hf3prong_001;1 → Table with 3-prong candidates (D^+, D_s^+, Λ_c^+, Ξ_c^+)

O2hfdstar;1 → Table with D^{*+} candidates

⋮



```
DF_2849676012004;1
DF_2849676012005;1
parentFiles;1
StreamerInfo
```

parentFiles;1 → Metadata with the information of parent files processed to produce this derived file

How to run on linked derived datasets?

- Hyperloop: just treat them as any other dataset, the parent access is automatically managed by hyperloop
- Locally:
  - ➡ run your workflows setting the derived `AO2D.root` files as input files
  - ➡ set the parent access and the path of parent files

```
o2-analysis-timestamp -b --configuration json://configuration.json |
o2-analysis-bc-converter -b --configuration json://configuration.json |
o2-analysis-event-selection -b --configuration json://configuration.json |
o2-analysis-ft0-corrected-table -b --configuration json://configuration.json |
o2-analysis-track-propagation -b --configuration json://configuration.json |
o2-analysis-tracks-extra-converter -b --configuration json://configuration.json |
o2-analysis-pid-tpc-full -b --configuration json://configuration.json |
o2-analysis-pid-tpc-base -b --configuration json://configuration.json |
o2-analysis-pid-tof-full -b --configuration json://configuration.json |
o2-analysis-pid-tof-base -b --configuration json://configuration.json |
o2-analysis-hf-candidate-creator-2prong -b --configuration json://configuration.json |
o2-analysis-hf-candidate-selector-d0 -b --configuration json://configuration.json |
o2-analysis-hf-task-d0 -b --configuration json://configuration.json --aod-file @input_data.txt --aod-parent-access-level 1 --aod-parent-base-path-replacement "alien://path"
```

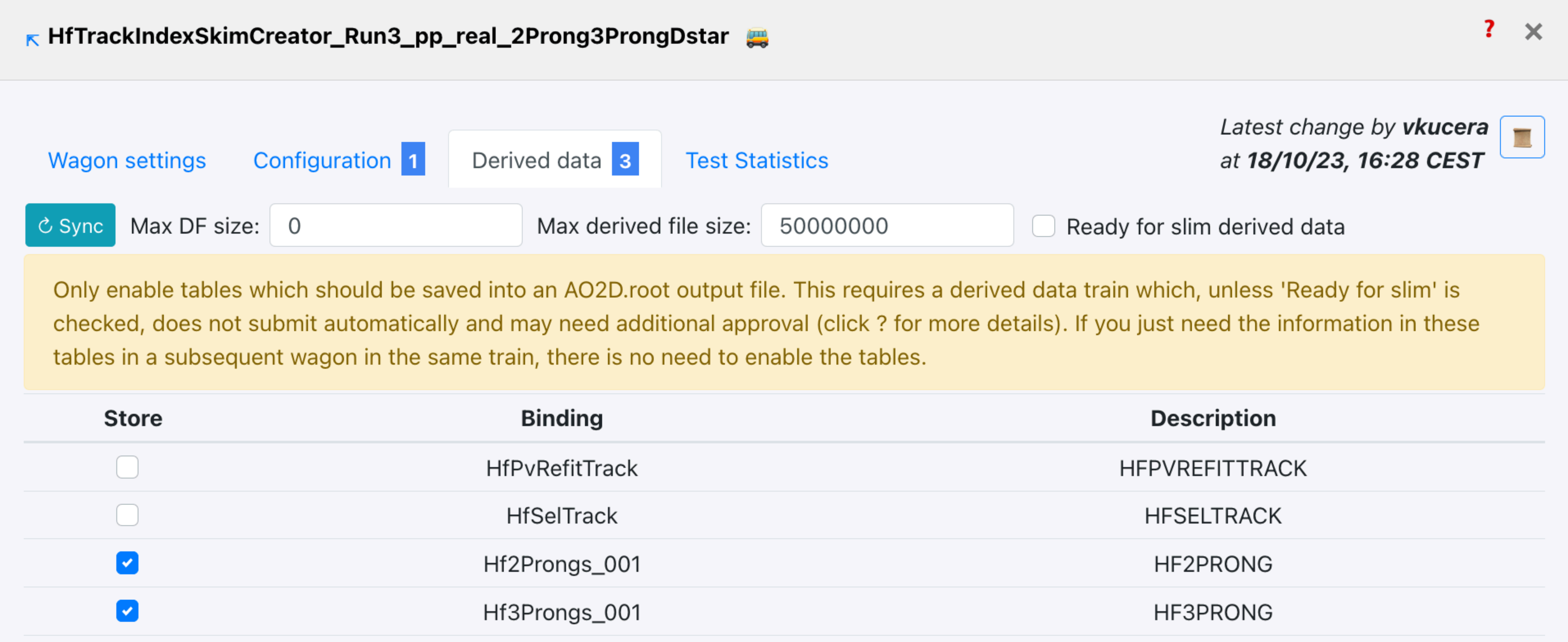
Text file containing the paths to your derived `AO2D.root` files (either local or on alien)

Argument to set parent access level

Argument to set path for parent `AO2D.root` files

How to produce derived datasets?

- Hyperloop:
  - ➡ select the tables that you want to save in your derived data from the configuration of the wagon
  - ➡ if the derived data requires parent access, MaxDF must be 0
  - ➡ inform the train operator that your derived data must be linked to the parent dataset



HfTrackIndexSkimCreator_Run3_pp_real_2Prong3ProngDstar

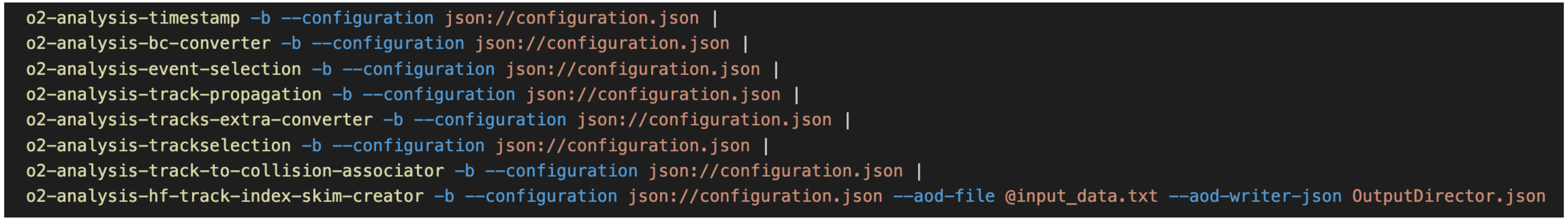
Latest change by ***vkucera*** *at* ***18/10/23, 16:28 CEST***

Sync Max DF size: 0 Max derived file size: 50000000 □ Ready for slim derived data

Only enable tables which should be saved into an AO2D.root output file. This requires a derived data train which, unless 'Ready for slim' is checked, does not submit automatically and may need additional approval (click ? for more details). If you just need the information in these tables in a subsequent wagon in the same train, there is no need to enable the tables.

Store	Binding	Description
□	HfPvRefitTrack	HFPVREFITTRACK
□	HfSelTrack	HFSELTRACK
☑	Hf2Prongs_001	HF2PRONG
☑	Hf3Prongs_001	HF3PRONG

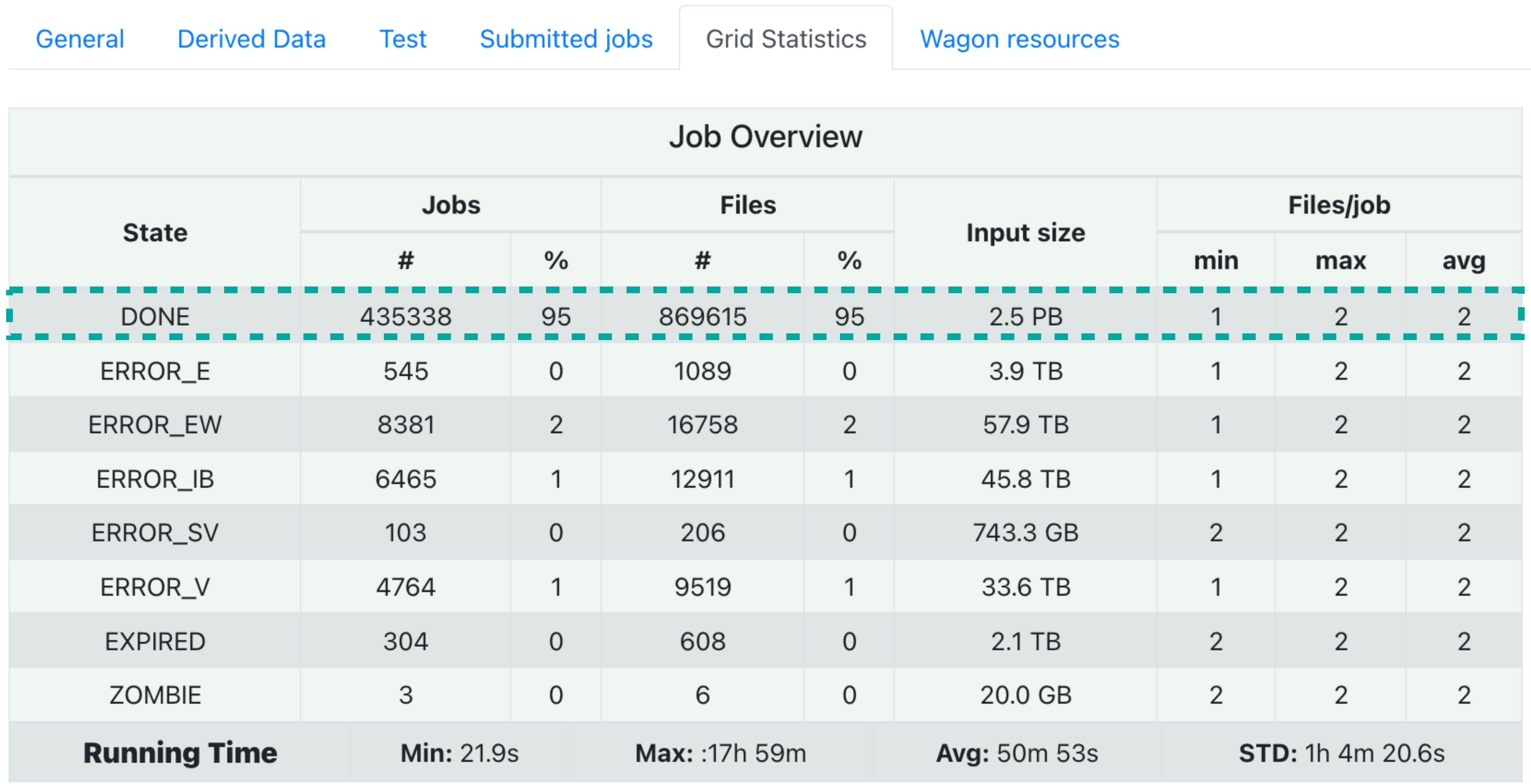
- Locally:
  - ➡ Run the workflow that produces the tables that you want as derived data and specify them in the `OutputDirector.json` file



```
o2-analysis-timestamp -b --configuration json://configuration.json |
o2-analysis-bc-converter -b --configuration json://configuration.json |
o2-analysis-event-selection -b --configuration json://configuration.json |
o2-analysis-track-propagation -b --configuration json://configuration.json |
o2-analysis-tracks-extra-converter -b --configuration json://configuration.json |
o2-analysis-trackselection -b --configuration json://configuration.json |
o2-analysis-track-to-collision-associator -b --configuration json://configuration.json |
o2-analysis-hf-track-index-skim-creator -b --configuration json://configuration.json --aod-file @input_data.txt --aod-writer-json OutputDirector.json
```

Size of HF derived datasets in pp collisions

- The reduction factor (i.e. size of parent dataset divided by the on of the derived dataset) appears in the test output and then it can be seen in the Grid Statistics tab once the train that produces the derived data is done

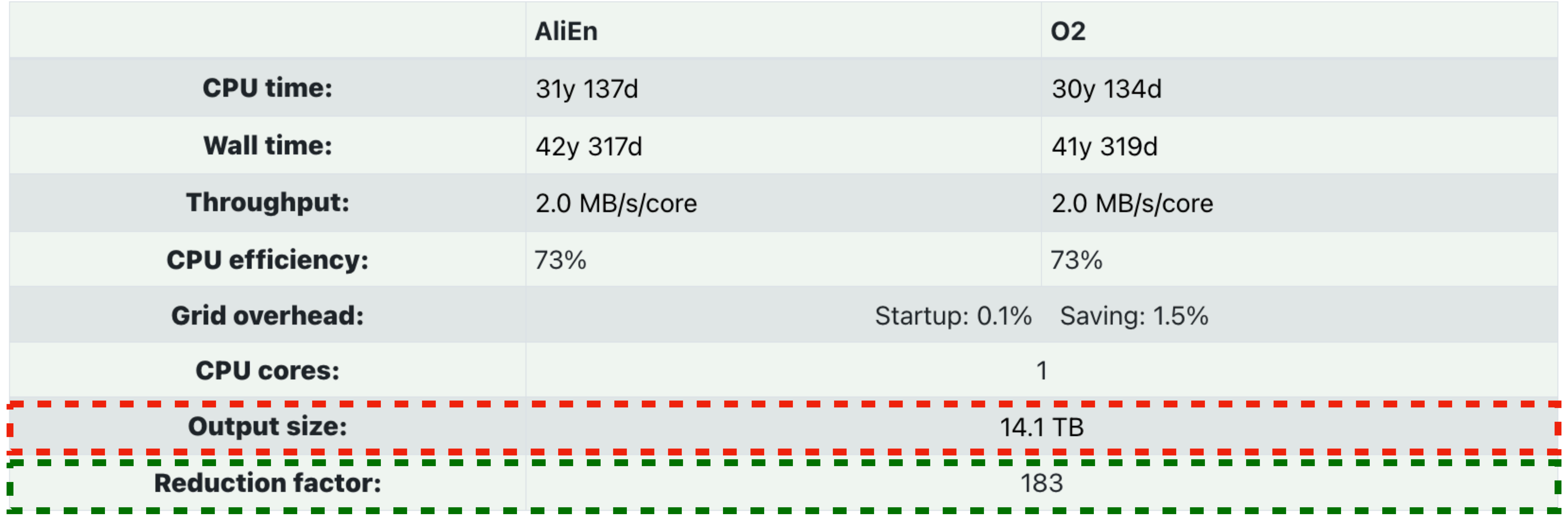


General | Derived Data | Test | Submitted jobs | Grid Statistics | Wagon resources

Job Overview

State	Jobs		Files		Input size	Files/job		
	#	%	#	%		min	max	avg
DONE	435338	95	869615	95	2.5 PB	1	2	2
ERROR_E	545	0	1089	0	3.9 TB	1	2	2
ERROR_EW	8381	2	16758	2	57.9 TB	1	2	2
ERROR_IB	6465	1	12911	1	45.8 TB	1	2	2
ERROR_SV	103	0	206	0	743.3 GB	2	2	2
ERROR_V	4764	1	9519	1	33.6 TB	1	2	2
EXPIRED	304	0	608	0	2.1 TB	2	2	2
ZOMBIE	3	0	6	0	20.0 GB	2	2	2

Running Time **Min:** 21.9s **Max:** :17h 59m **Avg:** 50m 53s **STD:** 1h 4m 20.6s

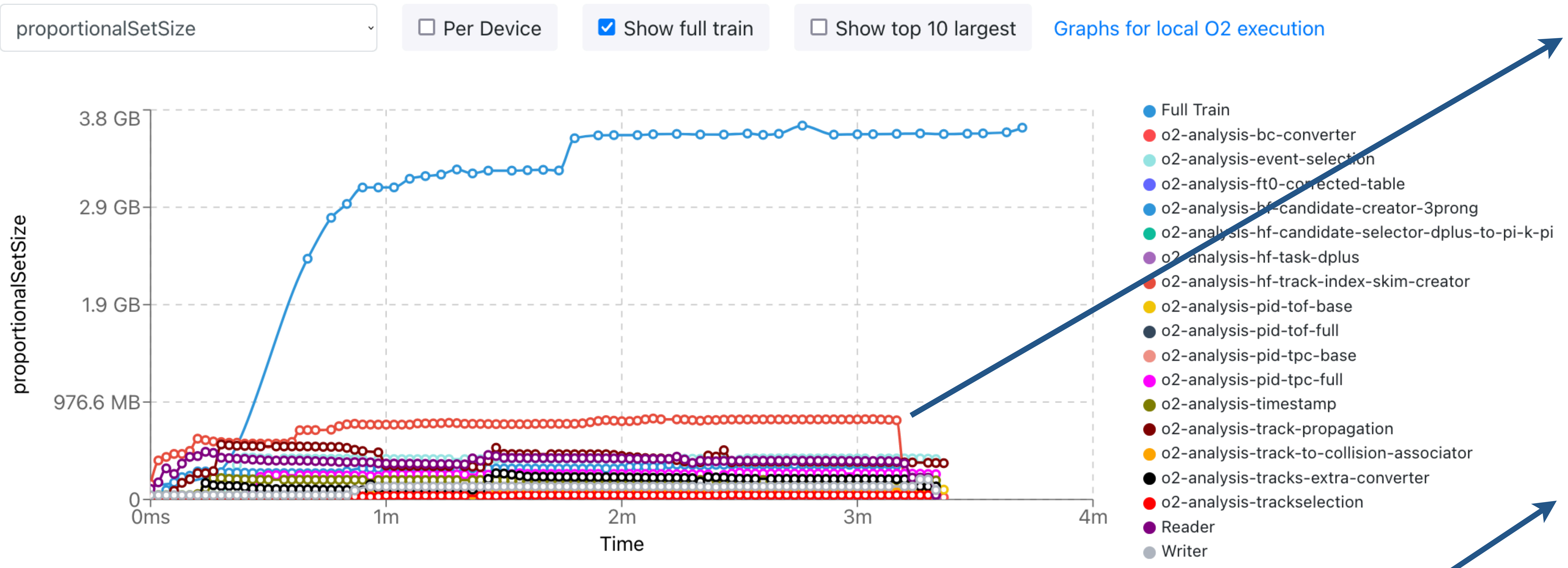


	AliEn	O2
CPU time:	31y 137d	30y 134d
Wall time:	42y 317d	41y 319d
Throughput:	2.0 MB/s/core	2.0 MB/s/core
CPU efficiency:	73%	73%
Grid overhead:	Startup: 0.1% Saving: 1.5%	
CPU cores:	1	
Output size:	14.1 TB	
Reduction factor:	183	

- For pp collisions, the reduction factor is around 180, meaning that the current HF derived `AO2D.root` files occupy ~0.6% of the disk space occupied by the parent `AO2D.root` files
  - ➡ This depends on the selections applied and the colliding systems (e.g. in Pb–Pb we expect many more candidates per event)

	Reduction factor	Links to train outputs
Data	~140–180	128492, 127820, 127451, 126921
MC	~610–680	129264, 129265, 129266

Performance without derived data in pp collisions

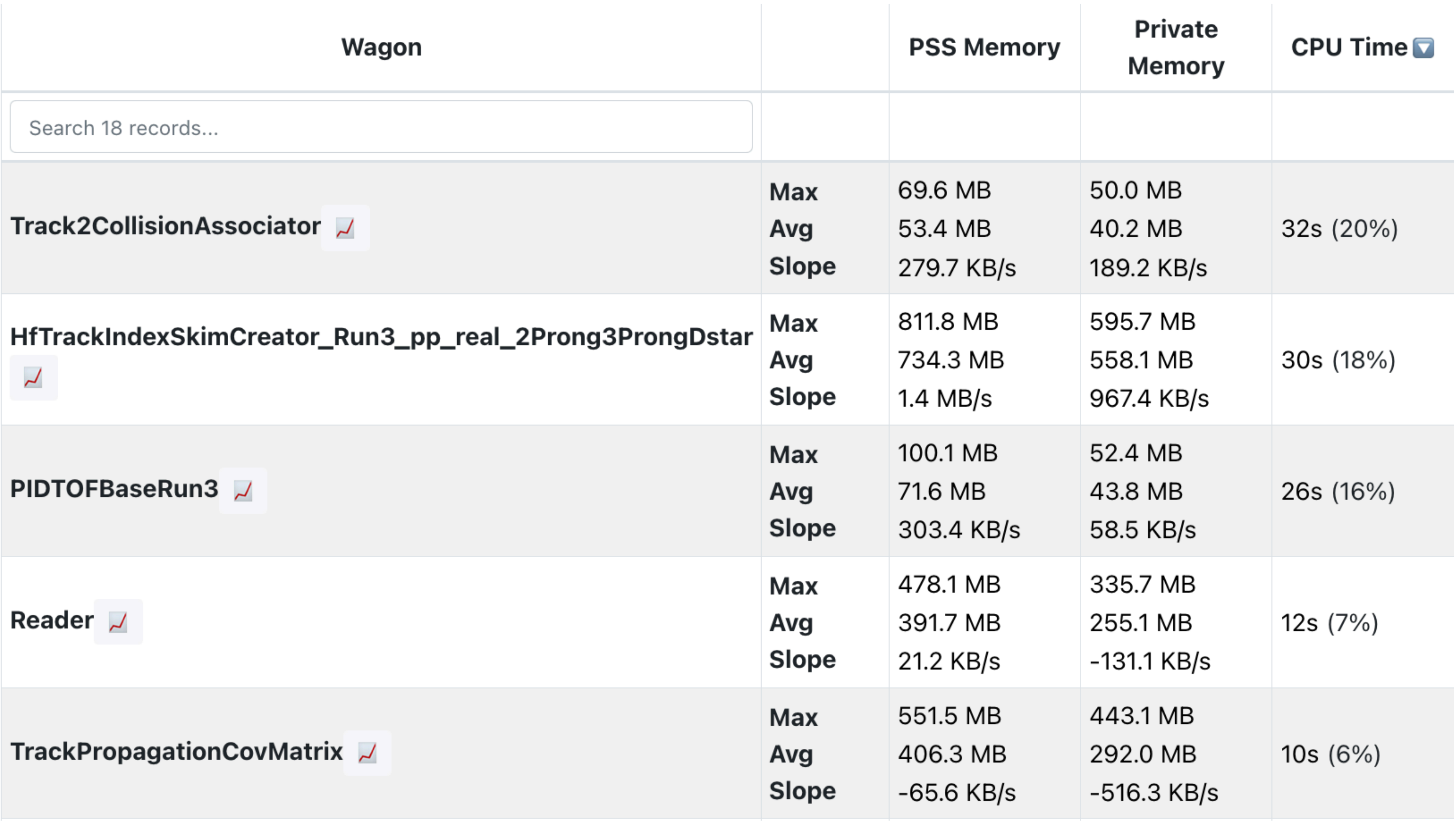


proportionalSetSize ⌄ ☐ Per Device ☑ Show full train ☐ Show top 10 largest Graphs for local O2 execution

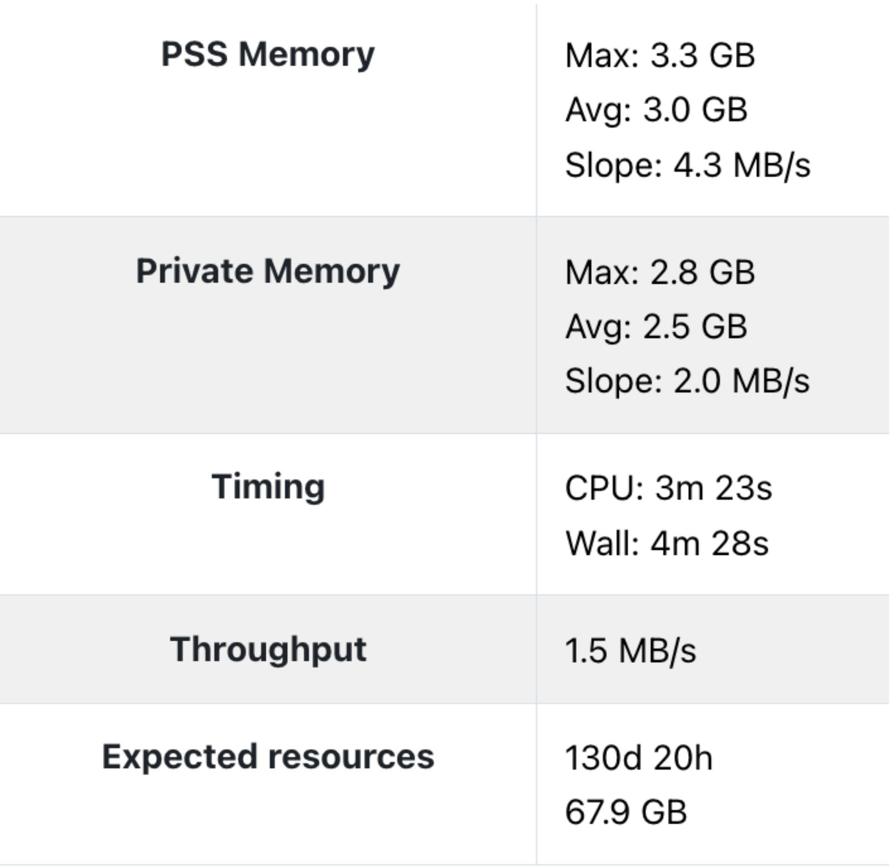
3.8 GB
2.9 GB
1.9 GB
976.6 MB
0
proportionalSetSize
0ms
1m
2m
3m
4m
Time
● Full Train
● o2-analysis-bc-converter
● o2-analysis-event-selection
● o2-analysis-ft0-corrected-table
● o2-analysis-hf-candidate-creator-3prong
● o2-analysis-hf-candidate-selector-dplus-to-pi-k-pi
● o2-analysis-hf-task-dplus
● o2-analysis-hf-track-index-skim-creator
● o2-analysis-pid-tof-base
● o2-analysis-pid-tof-full
● o2-analysis-pid-tpc-base
● o2-analysis-pid-tpc-full
● o2-analysis-timestamp
● o2-analysis-track-propagation
● o2-analysis-track-to-collision-associator
● o2-analysis-tracks-extra-converter
● o2-analysis-trackselection
● Reader
● Writer

The largest memory consumption is for the `hf-track-index-skim-creator` workflow

40% of the CPU time taken by the `hf-track-index-skim-creator` and `track-to-collision-associator` (needed because of the `hf-track-index-skim-creator`) workflows

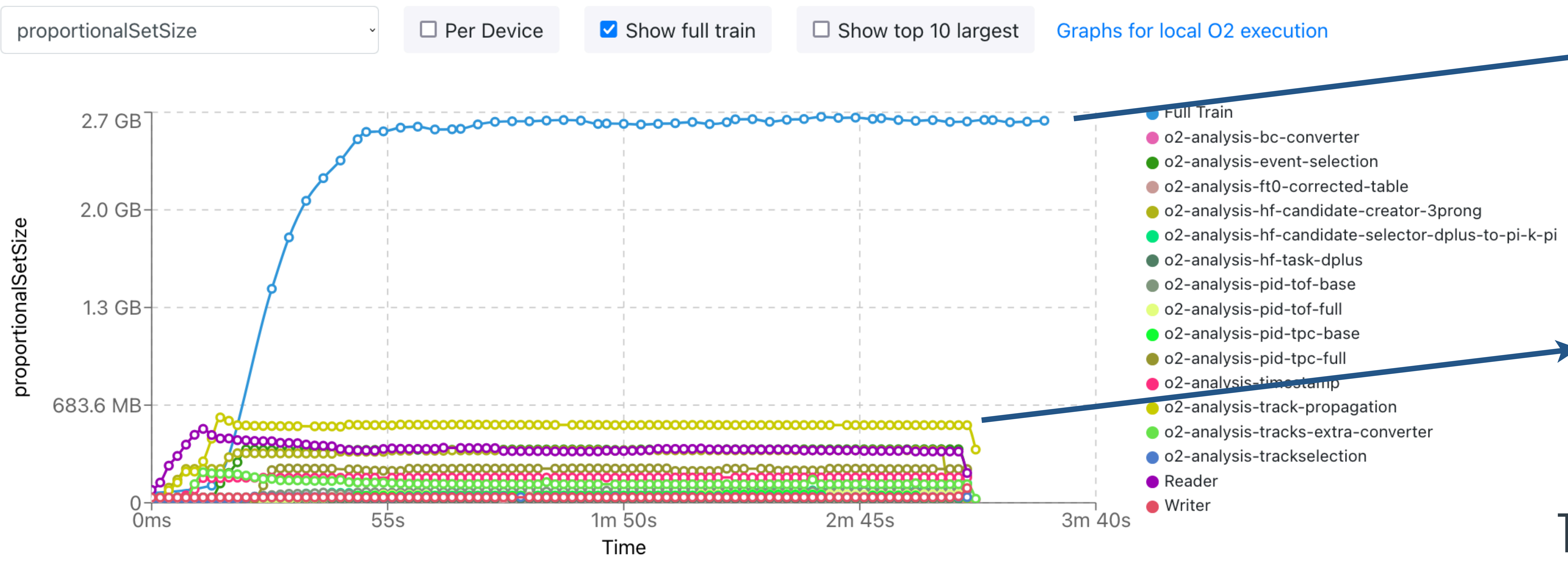


Wagon		PSS Memory	Private Memory	CPU Time ▾
Search 18 records...				
Track2CollisionAssociator	Max Avg Slope	69.6 MB 53.4 MB 279.7 KB/s	50.0 MB 40.2 MB 189.2 KB/s	32s (20%)
HfTrackIndexSkimCreator_Run3_pp_real_2Prong3ProngDstar	Max Avg Slope	811.8 MB 734.3 MB 1.4 MB/s	595.7 MB 558.1 MB 967.4 KB/s	30s (18%)
PIDTOFBaseRun3	Max Avg Slope	100.1 MB 71.6 MB 303.4 KB/s	52.4 MB 43.8 MB 58.5 KB/s	26s (16%)
Reader	Max Avg Slope	478.1 MB 391.7 MB 21.2 KB/s	335.7 MB 255.1 MB -131.1 KB/s	12s (7%)
TrackPropagationCovMatrix	Max Avg Slope	551.5 MB 406.3 MB -65.6 KB/s	443.1 MB 292.0 MB -516.3 KB/s	10s (6%)



PSS Memory	Max: 3.3 GB Avg: 3.0 GB Slope: 4.3 MB/s
Private Memory	Max: 2.8 GB Avg: 2.5 GB Slope: 2.0 MB/s
Timing	CPU: 3m 23s Wall: 4m 28s
Throughput	1.5 MB/s
Expected resources	130d 20h 67.9 GB

Performance with derived data in pp collisions

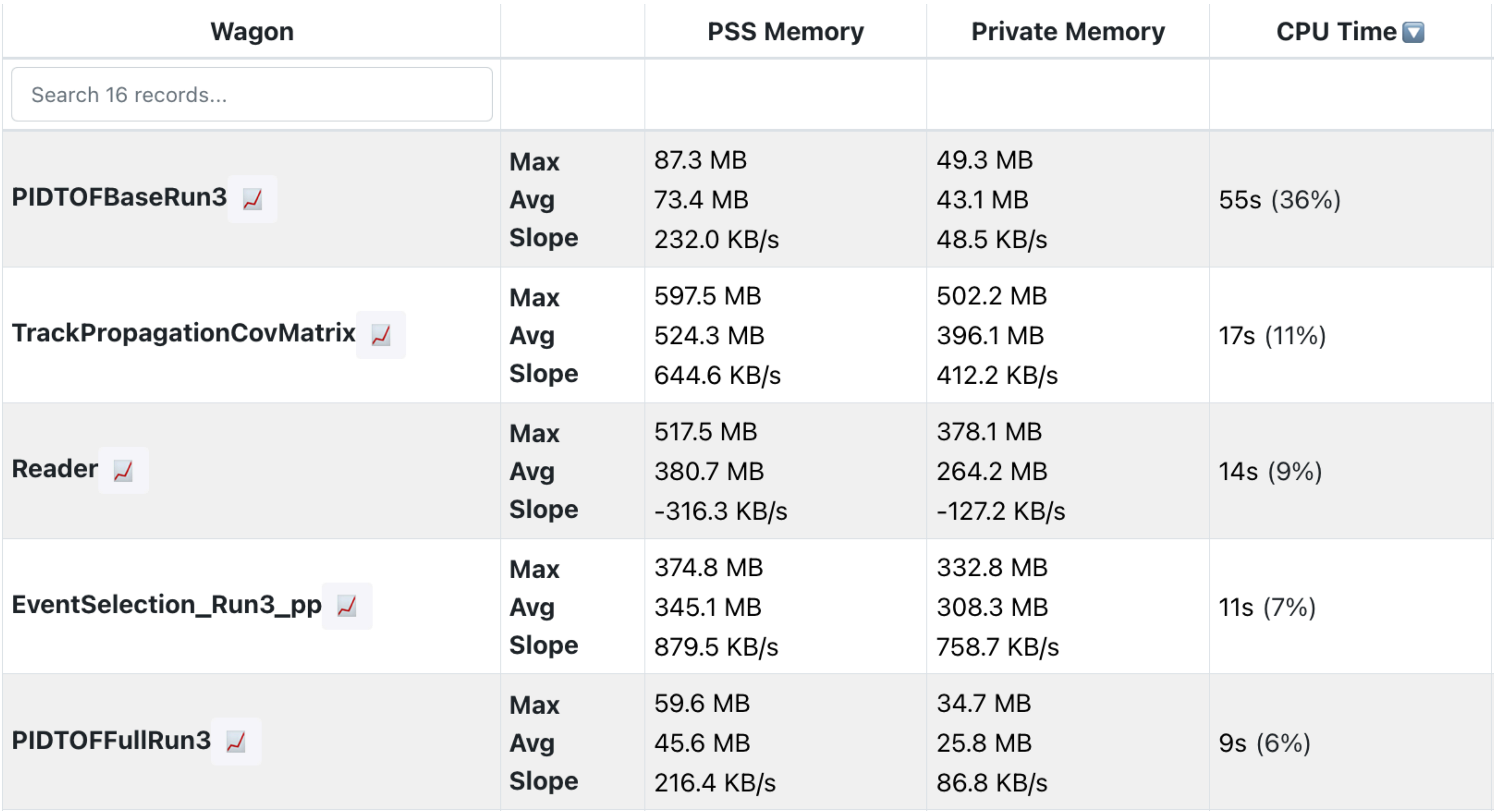


proportionalSetSize ⌄ ☐ Per Device ☑ Show full train ☐ Show top 10 largest Graphs for local O2 execution

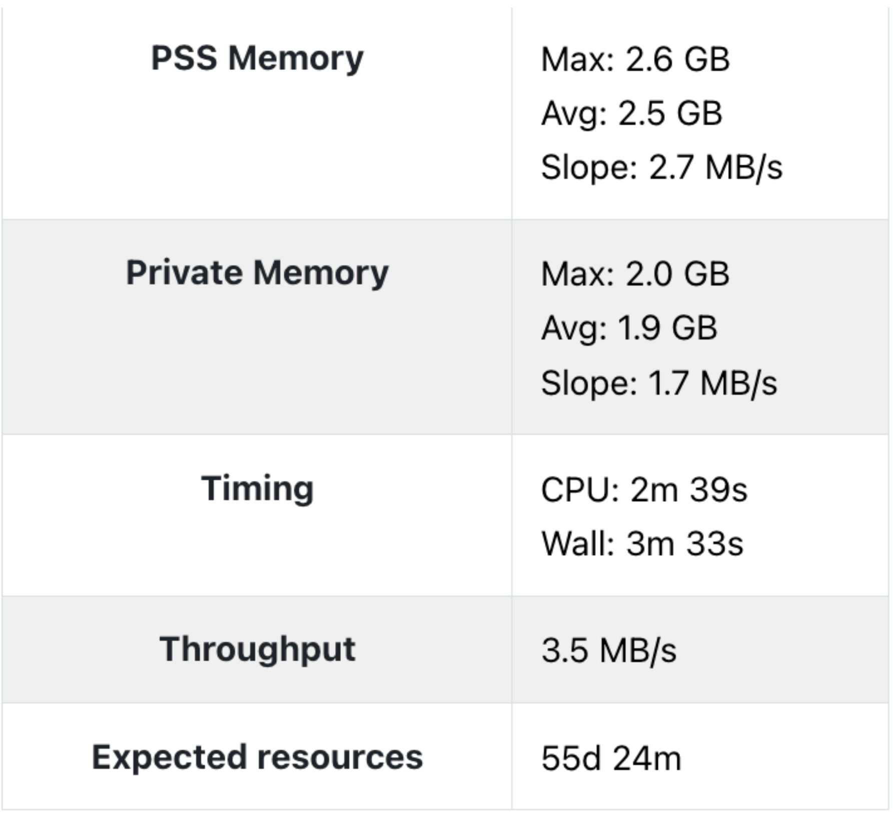
Reduced memory consumption from 3.8 GB to 2.7 GB

Largest memory consumption is for the `track-propagation` workflow

Top 70% of the CPU time taken by non-HF workflows



Wagon		PSS Memory	Private Memory	CPU Time
Search 16 records...				
PIDTOFBaseRun3	Max Avg Slope	87.3 MB 73.4 MB 232.0 KB/s	49.3 MB 43.1 MB 48.5 KB/s	55s (36%)
TrackPropagationCovMatrix	Max Avg Slope	597.5 MB 524.3 MB 644.6 KB/s	502.2 MB 396.1 MB 412.2 KB/s	17s (11%)
Reader	Max Avg Slope	517.5 MB 380.7 MB -316.3 KB/s	378.1 MB 264.2 MB -127.2 KB/s	14s (9%)
EventSelection_Run3_pp	Max Avg Slope	374.8 MB 345.1 MB 879.5 KB/s	332.8 MB 308.3 MB 758.7 KB/s	11s (7%)
PIDTOFFullRun3	Max Avg Slope	59.6 MB 45.6 MB 216.4 KB/s	34.7 MB 25.8 MB 86.8 KB/s	9s (6%)



PSS Memory	Max: 2.6 GB Avg: 2.5 GB Slope: 2.7 MB/s
Private Memory	Max: 2.0 GB Avg: 1.9 GB Slope: 1.7 MB/s
Timing	CPU: 2m 39s Wall: 3m 33s
Throughput	3.5 MB/s
Expected resources	55d 24m

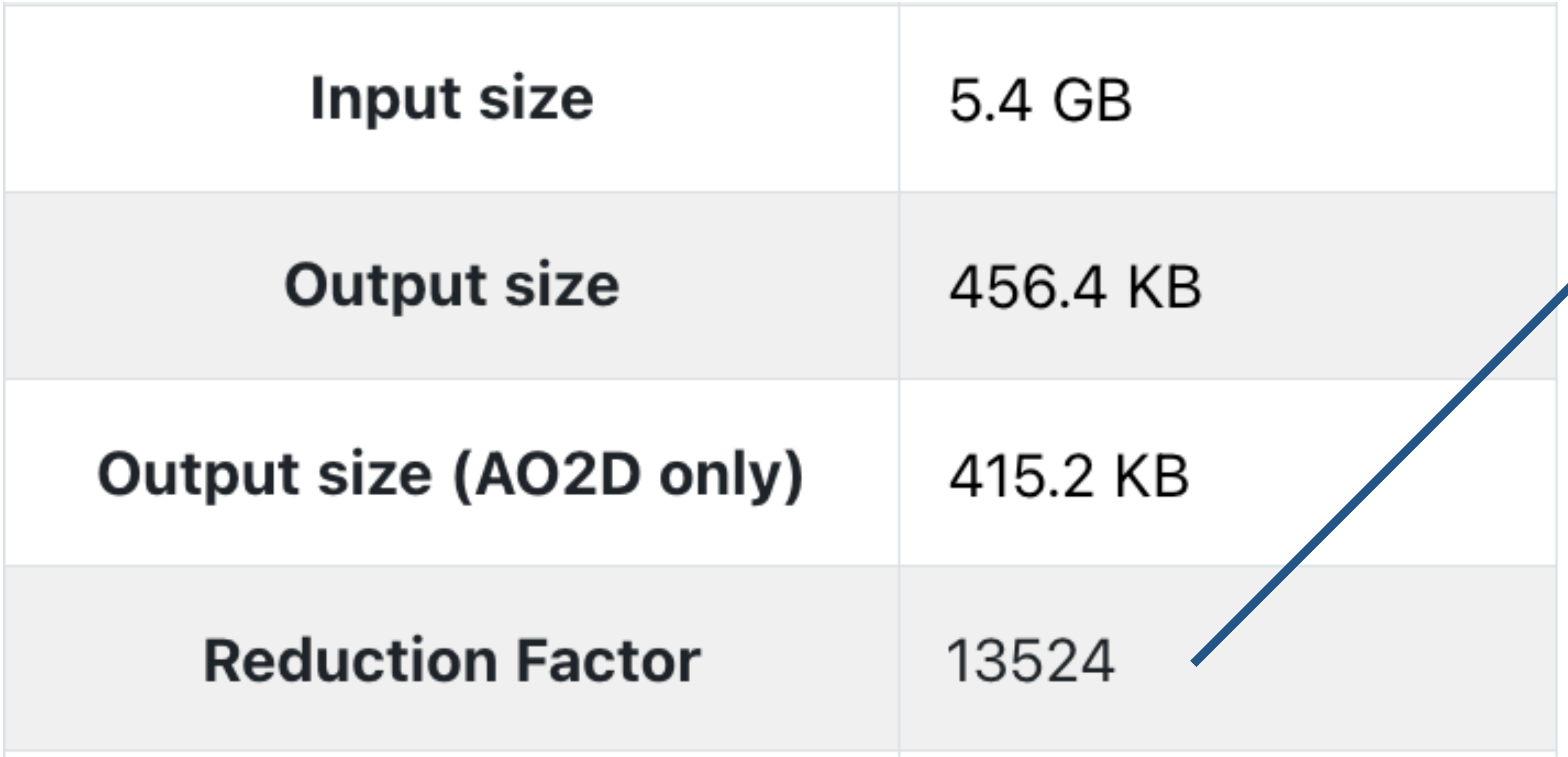
Overall resources needed reduced

➡ Your code runs faster on hyperloop and takes less memory

Next step: self-contained derived data

- Weak point: being linked, the HF derived datasets still require the access to the parent `AO2D.root` files
  - ➡ Still access to large datasets needed
  - ➡ Especially in periods before approval sessions, this could be problematic because many analyses will access the same data files

- Next step? Produce a derived dataset containing only the information needed for a specific analysis
  - ➡ E.g. analyses of $B^0 \rightarrow D^-\pi^+$ and $B^+ \rightarrow \bar{D}^0\pi^+$ can run on linked derived data and produce self-contained derived `AO2D.root` that have tables for preselected D mesons and pions as well as collisions that contain a B candidate (see dataCreatorDplusPiReduced.cxx and dataCreatorD0PiReduced.cxx)



Input size	5.4 GB
Output size	456.4 KB
Output size (AO2D only)	415.2 KB
Reduction Factor	13524

Very large reduction factor implies very small datasets that can be analysed very quickly since no access to the parent dataset is needed

Example test: 130390 produced derived data of a total of 2.6 GB starting from a dataset of 12.6 TB (LHC23c1)

- ➡ Derived data can even be analysed locally in few minutes

Summary and useful links

- Summary
  - ➡ If your analysis uses D^0, D^+, $D^{+,}$ $D_s^{+,}$ Λ_c^+, Ξ_c^+, or D^{*+} candidates and the linked derived datasets are available, use them to avoid the dependency on the `trackIndexSkimCreator` task to reduce the resources needed
  - ➡ Linked derived data for charm cascades will be produced soon as well
  - ➡ Studies for the production of derived datasets for Pb–Pb data will start soon
  - ➡ The goal for all the analyses should be to produce self-contained derived data (easier for "rare" observables)

- Useful links:
  - ➡ Derived data for 2022 pp sample (apass4) already available for D^0, D^+, $D^{+,}$ $D_s^{+,}$ Λ_c^+, Ξ_c^+, or D^{*+} candidates. Spreadsheet with available derived datasets and corresponding selections used https://docs.google.com/spreadsheets/d/1khi-SB0wpVkEymv6UJ2brXD6RhcPFG5TsOsTGHdzxhE/edit#gid=277044673
  - ➡ More general information about derived data in Hyperloop documentation https://aliceo2group.github.io/analysis-framework/docs/hyperloop/operatordocumentation.html#derived-data