WAYNE STATE UNIVERSITY

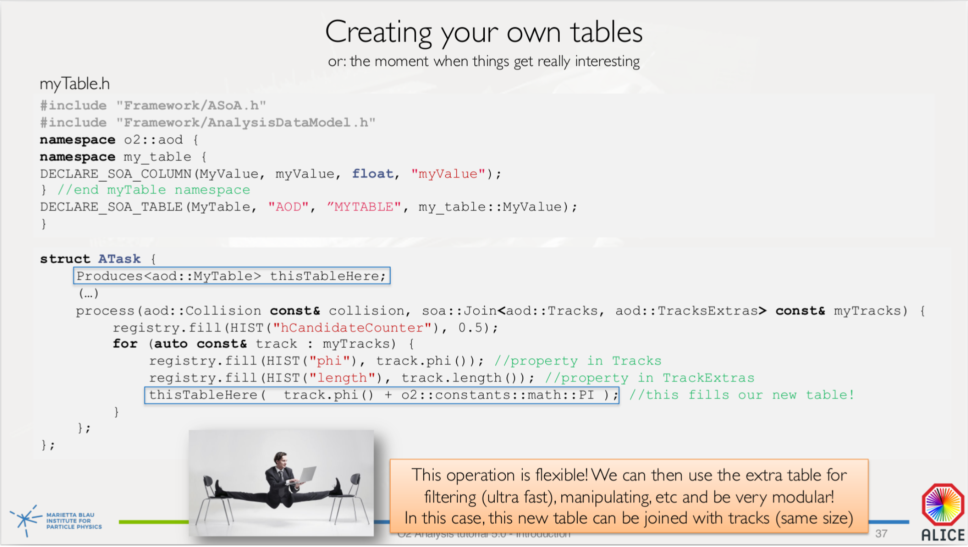
ALICE

Using derived data in O2/O2Physics analyses

O2 Analysis tutorial 5.0, 11th November 2025

A few things already said

How to produce derived data



Creating your own tables

or: the moment when things get really interesting

myTable.h

```
#include "Framework/ASoA.h"
#include "Framework/AnalysisDataModel.h"
namespace o2::aod {
namespace my_table {
DECLARE_SOA_COLUMN(MyValue, myValue, float, "myValue");
} //end myTable namespace
DECLARE_SOA_TABLE(MyTable, "AOD", ”MYTABLE", my_table::MyValue);
}
```

```
struct ATask {
    Produces<aod::MyTable> thisTableHere;
    (…)
    process(aod::Collision const& collision, soa::Join<aod::Tracks, aod::TracksExtras> const& myTracks) {
        registry.fill(HIST("hCandidateCounter"), 0.5);
        for (auto const& track : myTracks) {
            registry.fill(HIST("phi"), track.phi()); //property in Tracks
            registry.fill(HIST("length"), track.length()); //property in TrackExtras
            thisTableHere(  track.phi() + o2::constants::math::PI ); //this fills our new table!
        }
    };
};
```

This operation is flexible! We can then use the extra table for filtering (ultra fast), manipulating, etc and be very modular! In this case, this new table can be joined with tracks (same size)

Why you would need derived data

A use case

- **Say you need to run over tracks**
  - You want to extract two-particle correlations
    - You need two nested loops
  - Need particle identification
  - Need some filtering
  - Constraints on execution time

Why you would need derived data
A use case - implementation

- **You define/use two tasks**
  - First one classifies the tracks – **the classifier**
  - Second one processes the classified tracks – **the consumer**
    - and extracts the two-particle correlations
- **Tracks classification in a new table**
  - Just one single column
  - Produced by the classifier
  - Joined to the `Tracks` table
    - in the consumer `process...` subscription

A use case – Table declaration

```
#include "Framework/ASoA.h"
#include "Framework/AnalysisDataModel.h"

namespace o2::aod {
namespace myTable {
DECLARE_SOA_COLUMN(TrackCode, trackCode, int, "trackCode");
} //end myTable namespace
DECLARE_SOA_TABLE(MyTable, "AOD", "MYTABLE", myTable::TrackCode);
} //end o2::aod namespace
```

A use case – The producer

```
DECLARE_SOA_COLUMN(TrackCode, trackCode, int, "trackCode");
DECLARE_SOA_TABLE(MyTable, "AOD", "MYTABLE", myTable::TrackCode);
```

```
struct producer {
  Produces<aod::MyTable> thisTableHere;
  ...
  process(soa::Join<Tracks, TracksExtras> const& myTracks) {
    for (auto track : myTracks) {
      int thetrackcode = -1;
      ...
      thisTableHere(thetrackcode); //this fills our new table!
    }
  }
};
```

A use case – The consumer

```
DECLARE_SOA_COLUMN(TrackCode, trackCode, int, "trackCode");
DECLARE_SOA_TABLE(MyTable, "AOD", "MYTABLE", myTable::TrackCode);
```

```
struct consumer {
  ...
  process(o2::aod::Collision const& collision,
          soa::Filtered<soa::Join<Tracks, TracksExtras, MyTable>> const& myTracks) {
    ...
    for (auto track1 : myTracks) {
      for (auto track2 : myTracks) {
        ...
        myHist[track1.trackCode][track2.trackCode]->Fill(getDeltaPhi(track1,track2));
      }
    }
  }
};
```

Are these derived data?

The described use case

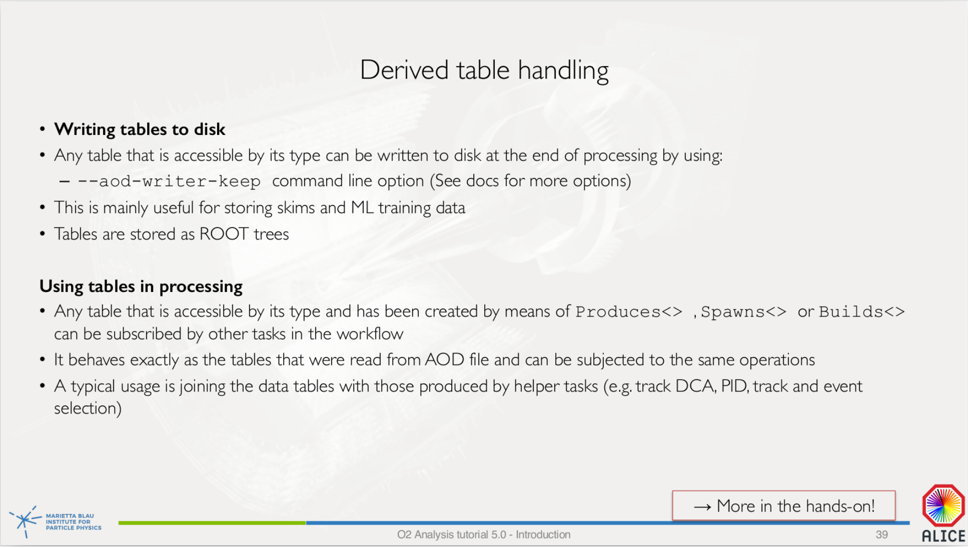
- **Actually, yes**
  - You produce a table from the processing of other tables
- **You benefit from the SOA approach**
  - Faster access
  - Bulk processing
  - Zero copy

Are these derived data?

The described use case

- **But we will not refer to them as derived data**
  - You process them on the fly
  - You don't store them
  - You shouldn't / cannot store them
  - You should use them as much as you can!!!

Storing and using derived data



Derived table handling

- **Writing tables to disk**
- Any table that is accessible by its type can be written to disk at the end of processing by using:
  - `--aod-writer-keep` command line option (See docs for more options)
- This is mainly useful for storing skims and ML training data
- Tables are stored as ROOT trees

Using tables in processing

- Any table that is accessible by its type and has been created by means of `Produces<>` , `Spawns<>` or `Builds<>` can be subscribed by other tasks in the workflow
- It behaves exactly as the tables that were read from AOD file and can be subjected to the same operations
- A typical usage is joining the data tables with those produced by helper tasks (e.g. track DCA, PID, track and event selection)

→ More in the hands-on!

MARIETTA BLAU INSTITUTE FOR PARTICLE PHYSICS
O2 Analysis tutorial 5.0 - Introduction
39
ALICE

Saving and retrieving derived data

- **Saving tables to a file**
  - `OutputDirector` configuration file with `--aod-writer-json`
  - https://aliceo2group.github.io/analysis-framework/docs/basics-usage/SavingTablesToFile.html
- **Reading tables from files**
  - `InputDirector` configuration file with `--aod-reader-json`
  - https://aliceo2group.github.io/analysis-framework/docs/basics-usage/ReadingTablesFromFile.html

But that is for your local tests

How to do it

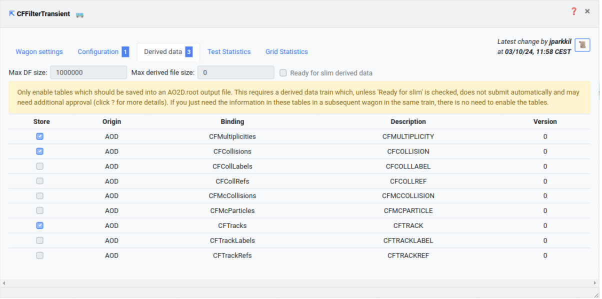
```
namespace cfskim
{
DECLARE_SOA_COLUMN(CFCollisionFlags, selflags, uint64_t);
DECLARE_SOA_INDEX_COLUMN(CFCollision, cfcollision);
DECLARE_SOA_COLUMN(CFTrackFlags, trackflags, uint64_t);
DECLARE_SOA_COLUMN(CFPidFlags, pidflags, uint64_t);
DECLARE_SOA_COLUMN(Pt, pt, float);
DECLARE_SOA_COLUMN(Eta, eta, float);
DECLARE_SOA_COLUMN(Phi, phi, float);
DECLARE_SOA_DYNAMIC_COLUMN(Sign, sign,
                           [](uint64_t mask) -> int8_t
                           { return ((mask & 0x1L) == 0x1L) ? 1 :
                                ((mask & 0x2L) == 0x2L) ? -1
} // namespace cfskim
DECLARE_SOA_TABLE(CFCollisions, "AOD", "CFCOLLISION",
                  o2::soa::Index<>,
                  collision::PosZ,
                  bc::RunNumber,
                  timestamp::Timestamp,
                  cfskim::CFCollisionFlags);
DECLARE_SOA_TABLE(CFTracks, "AOD", "CFTRACK",
                  o2::soa::Index<>,
                  cfskim::CFCollisionId,
                  cfskim::CFTrackFlags,
                  cfskim::Pt,
                  cfskim::Eta,
                  cfskim::Phi,
                  cfskim::Sign<cfskim::CFTrackFlags>);
DECLARE_SOA_TABLE(CFTrackPIDs, "AOD", "CFTRACKPID",
                  cfskim::CFPidFlags);
```

```
{
  "OutputDirector": {
    "debugmode": false,
    "resfile": "AnalysisResults_trees",
    "resfilemode": "RECREATE",
    "ntfmerge": 1,
    "OutputDescriptors": [
      {
        "table": "AOD/CFCOLLISION/0",
        "treename": "O2cfcollision",
        "columns": [
          "fPosZ",
          "fRunNumber",
          "fTimestamp",
          "fCFCollisionFlags"
        ]
      },
      {
        "table": "AOD/CFTRACK/0",
        "treename": "O2cftrack",
        "columns": [
          "fIndexCFCollisions",
          "fCFTrackFlags",
          "fPt",
          "fEta",
          "fPhi"
        ]
      },
      {
        "table": "AOD/CFTRACKPID/0",
        "treename": "O2cftrackpid",
        "columns": [
          "fCFPidFlags"
        ]
      }
    ]
```

On hyperloop it is easier

Derived data settings

- Displays the tables which are produced by the task
- Here you can 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***
- ***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***
- ***For derived data of small output size, you can enable the slim derived data option***



CFFilterTransient

Latest change by ***jparkkil*** *at* ***03/10/24, 11:58 CEST***

Max DF size: 1000000 Max derived file size: 0 □ Ready for slim derived data

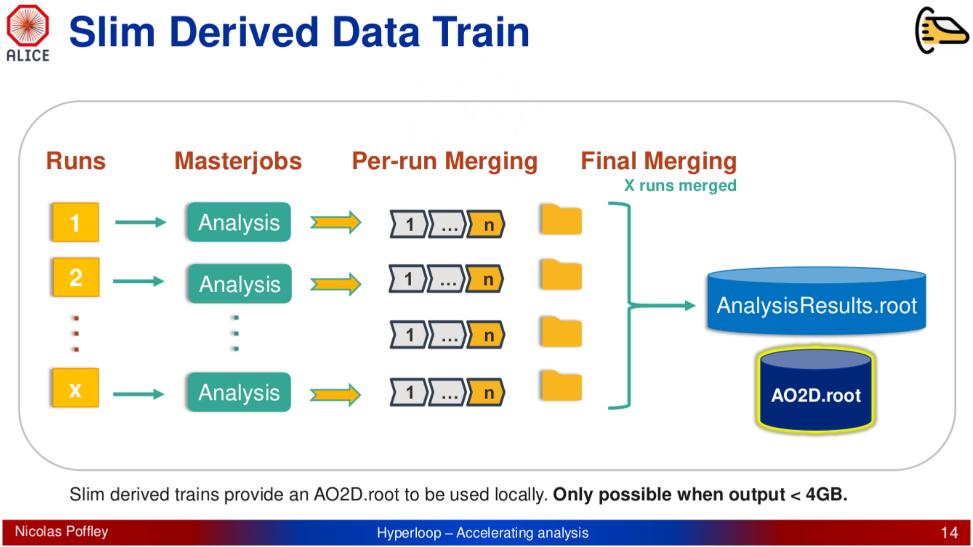
Only enable tables which should be saved into an AO2D.root output file. This requires a derived data train which, unless 'Ready for slim' is checked, does not submit automatically and may need additional approval (click ? for more details). If you just need the information in these tables in a subsequent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

Store	Origin	Binding	Description	Version
☑	AOD	CFMultiplicities	CFMULTIPLICITY	0
☑	AOD	CFCollisions	CFCOLLISION	0
□	AOD	CFCollLabels	CFCOLLLABEL	0
□	AOD	CFCollRefs	CFCOLLREF	0
□	AOD	CFMcCollisions	CFMCCOLLISION	0
□	AOD	CFMcParticles	CFMCPARTICLE	0
☑	AOD	CFTracks	CFTRACK	0
□	AOD	CFTrackLabels	CFTRACKLABEL	0
□	AOD	CFTrackRefs	CFTRACKREF	0

➢ In order to ***update*** the derived data configuration with the latest O2Physics version of the workflow, click on the **↻sync button**

➢ By synchronizing the derived data, the tables which no longer belong to the workflow will be removed, and the values of the tables will be updated

But a more varied zoo



Slim Derived Data Train

Runs

Masterjobs

Per-run Merging

Final Merging

X runs merged

1 → Analysis → 1 ... n

2 → Analysis → 1 ... n

1 ... n

x → Analysis → 1 ... n

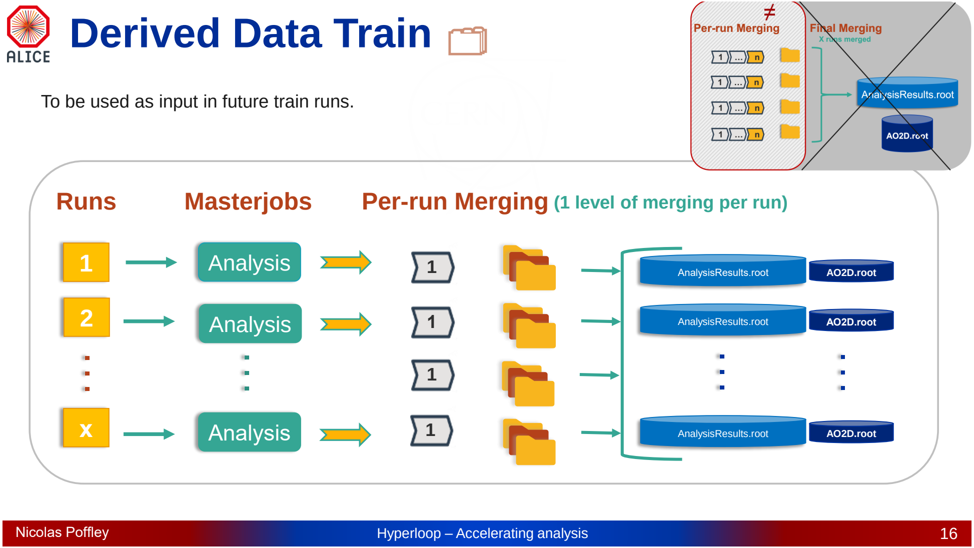
AnalysisResults.root

AO2D.root

Slim derived trains provide an AO2D.root to be used locally. **Only possible when output < 4GB.**

Nicolas Poffley Hyperloop – Accelerating analysis 14

But a more varied zoo



Derived Data Train

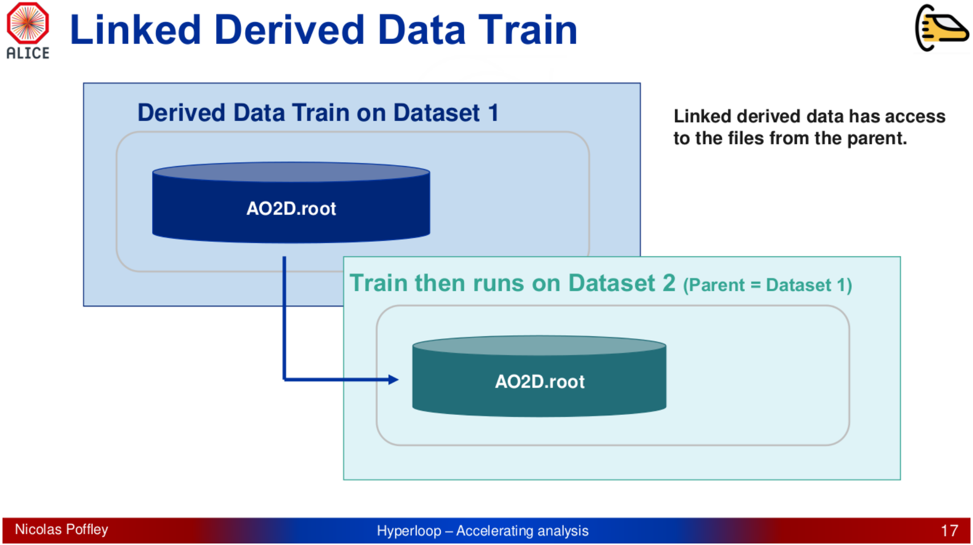
To be used as input in future train runs.

Nicolas Poffley

Hyperloop – Accelerating analysis

16

But a more varied zoo



Linked Derived Data Train

Derived Data Train on Dataset 1

AO2D.root

Train then runs on Dataset 2 (Parent = Dataset 1)

AO2D.root

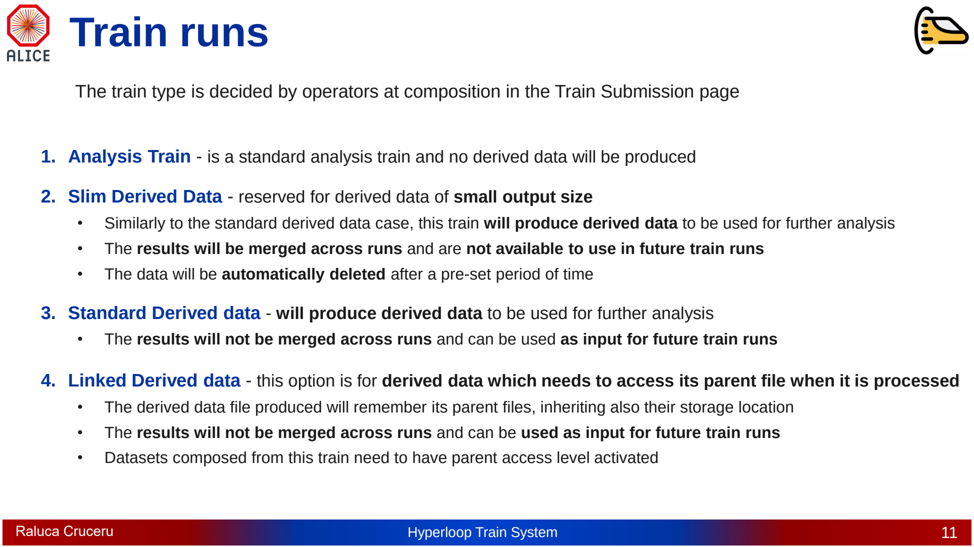
Linked derived data has access to the files from the parent.

Nicolas Poffley

Hyperloop – Accelerating analysis

17

Ask the train operators



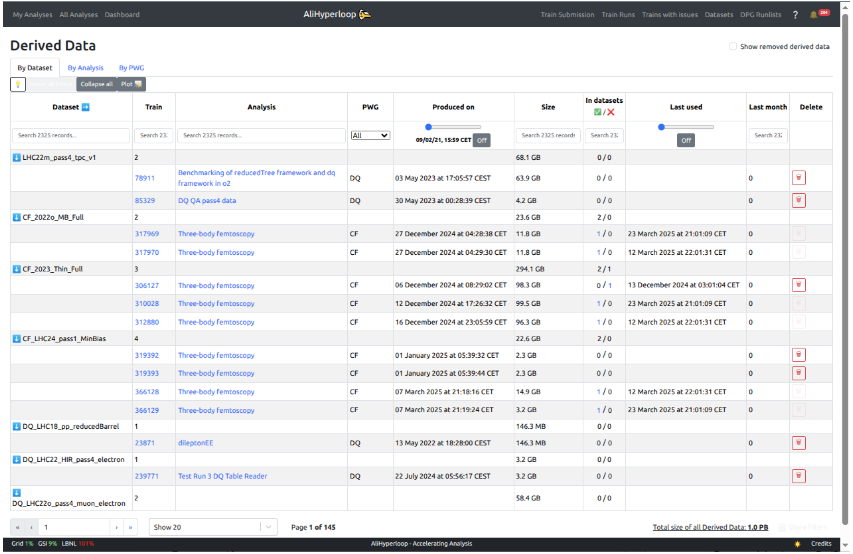
Train runs

The train type is decided by operators at composition in the Train Submission page

1. **Analysis Train** - is a standard analysis train and no derived data will be produced
2. **Slim Derived Data** - reserved for derived data of **small output size**
   - Similarly to the standard derived data case, this train **will produce derived data** to be used for further analysis
   - The **results will be merged across runs** and are **not available to use in future train runs**
   - The data will be **automatically deleted** after a pre-set period of time
3. **Standard Derived data** - **will produce derived data** to be used for further analysis
   - The **results will not be merged across runs** and can be used **as input for future train runs**
4. **Linked Derived data** - this option is for **derived data which needs to access its parent file when it is processed**
   - The derived data file produced will remember its parent files, inheriting also their storage location
   - The **results will not be merged across runs** and can be **used as input for future train runs**
   - Datasets composed from this train need to have parent access level activated

Raluca Cruceru Hyperloop Train System 11

Productified derived data



Derived Data

□ Show removed derived data

Dataset	Train	Analysis	PWG	Produced on	Size	In datasets ✅/❌	Last used	Last month	Delete
Search 2325 records...	Search 23:	Search 2325 records...	All	09/02/21, 15:59 CET Off	Search 2325 records	Search 23:	Off	Search 23:	
LHC22m_pass4_tpc_v1	2				68.1 GB	0 / 0			
	78911	Benchmarking of reducedTree framework and dq framework in o2	DQ	03 May 2023 at 17:05:57 CEST	63.9 GB	0 / 0		0	
	85329	DQ QA pass4 data	DQ	30 May 2023 at 00:28:39 CEST	4.2 GB	0 / 0		0	
CF_2022o_MB_Full	2				23.6 GB	2 / 0			
	317969	Three-body femtoscopy	CF	27 December 2024 at 04:28:38 CET	11.8 GB	1 / 0	23 March 2025 at 21:01:09 CET	0	
	317970	Three-body femtoscopy	CF	27 December 2024 at 04:29:30 CET	11.8 GB	1 / 0	12 March 2025 at 22:01:31 CET	0	
CF_2023_Thin_Full	3				294.1 GB	2 / 1			
	306127	Three-body femtoscopy	CF	06 December 2024 at 08:29:02 CET	98.3 GB	0 / 1	13 December 2024 at 03:01:04 CET	0	
	310028	Three-body femtoscopy	CF	12 December 2024 at 17:26:32 CET	99.5 GB	1 / 0	23 March 2025 at 21:01:09 CET	0	
	312880	Three-body femtoscopy	CF	16 December 2024 at 23:05:59 CET	96.3 GB	1 / 0	12 March 2025 at 22:01:31 CET	0	
CF_LHC24_pass1_MinBias	4				22.6 GB	2 / 0			
	319392	Three-body femtoscopy	CF	01 January 2025 at 05:39:32 CET	2.3 GB	0 / 0		0	
	319393	Three-body femtoscopy	CF	01 January 2025 at 05:39:44 CET	2.3 GB	0 / 0		0	
	366128	Three-body femtoscopy	CF	07 March 2025 at 21:18:16 CET	14.9 GB	1 / 0	12 March 2025 at 22:01:31 CET	0	
	366129	Three-body femtoscopy	CF	07 March 2025 at 21:19:24 CET	3.2 GB	1 / 0	23 March 2025 at 21:01:09 CET	0	
DQ_LHC18_pp_reducedBarrel	1				146.3 MB	0 / 0			
	23871	dileptonEE	DQ	13 May 2022 at 18:28:00 CEST	146.3 MB	0 / 0		0	
DQ_LHC22_HIR_pass4_electron	1				3.2 GB	0 / 0			
	239771	Test Run 3 DQ Table Reader	DQ	22 July 2024 at 05:56:17 CEST	3.2 GB	0 / 0		0	
DQ_LHC22o_pass4_muon_electron	2				58.4 GB	0 / 0			

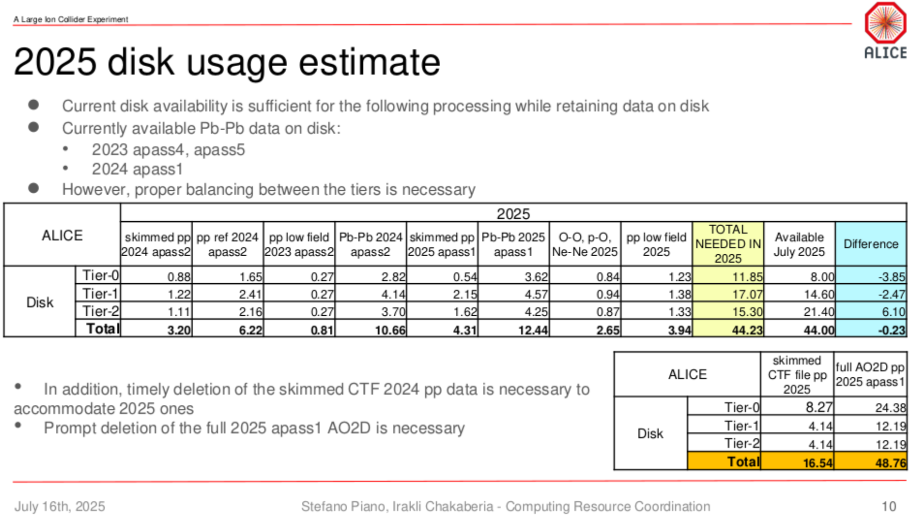
Page 1 of 145

Total size of all Derived Data: 1.0 PB

Now we are talking!

In Run 3 you cannot walk alone But that's why we are a collaboration

Huge amount of collected data



A Large Ion Collider Experiment

ALICE

2025 disk usage estimate

- Current disk availability is sufficient for the following processing while retaining data on disk
- Currently available Pb-Pb data on disk:
  - 2023 apass4, apass5
  - 2024 apass1
- However, proper balancing between the tiers is necessary

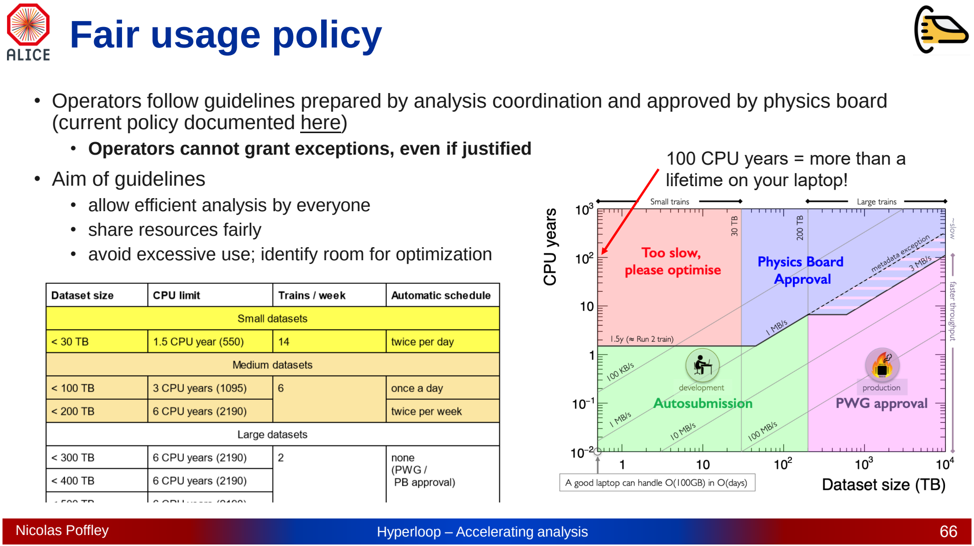
ALICE		2025										
		skimmed pp 2024 apass2	pp ref 2024 apass2	pp low field 2023 apass2	Pb-Pb 2024 apass2	skimmed pp 2025 apass1	Pb-Pb 2025 apass1	O-O, p-O, Ne-Ne 2025	pp low field 2025	TOTAL NEEDED IN 2025	Available July 2025	Difference
Disk	Tier-0	0.88	1.65	0.27	2.82	0.54	3.62	0.84	1.23	11.85	8.00	-3.85
	Tier-1	1.22	2.41	0.27	4.14	2.15	4.57	0.94	1.38	17.07	14.60	-2.47
	Tier-2	1.11	2.16	0.27	3.70	1.62	4.25	0.87	1.33	15.30	21.40	6.10
	Total	**3.20**	**6.22**	**0.81**	**10.66**	**4.31**	**12.44**	**2.65**	**3.94**	**44.23**	**44.00**	**-0.23**

- In addition, timely deletion of the skimmed CTF 2024 pp data is necessary to accommodate 2025 ones
- Prompt deletion of the full 2025 apass1 AO2D is necessary

ALICE		skimmed CTF file pp 2025	full AO2D pp 2025 apass1
Disk	Tier-0	8.27	24.38
	Tier-1	4.14	12.19
	Tier-2	4.14	12.19
	Total	**16.54**	**48.76**

July 16th, 2025 Stefano Piano, Irakli Chakaberia - Computing Resource Coordination 10

Limited processing capacity



Fair usage policy

- Operators follow guidelines prepared by analysis coordination and approved by physics board (current policy documented here)
  - **Operators cannot grant exceptions, even if justified**
- Aim of guidelines
  - allow efficient analysis by everyone
  - share resources fairly
  - avoid excessive use; identify room for optimization

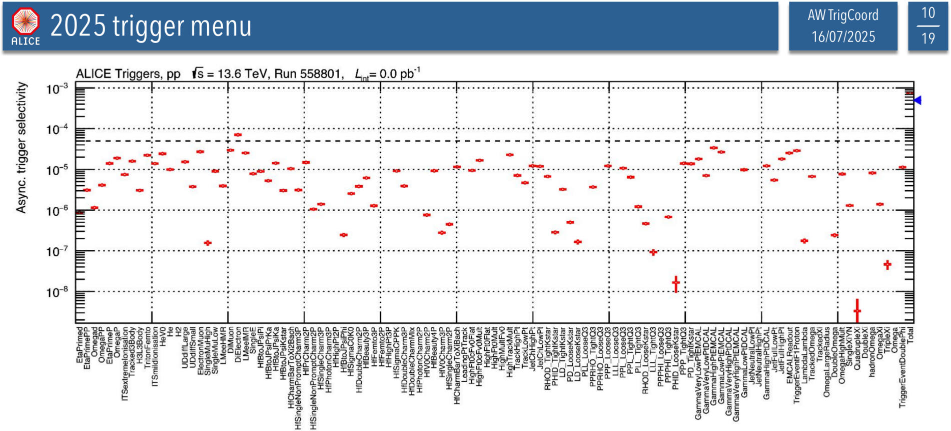
Dataset size	CPU limit	Trains / week	Automatic schedule
Small datasets			
< 30 TB	1.5 CPU year (550)	14	twice per day
Medium datasets			
< 100 TB	3 CPU years (1095)	6	once a day
< 200 TB	6 CPU years (2190)		twice per week
Large datasets			
< 300 TB	6 CPU years (2190)	2	none (PWG / PB approval)
< 400 TB	6 CPU years (2190)		
< 500 TB	6 CPU years (2190)		

100 CPU years = more than a lifetime on your laptop!

A good laptop can handle O(100GB) in O(days)

Nicolas Poffley — Hyperloop – Accelerating analysis — 66

Relaying on derived/skimmed data



ALICE

2025 trigger menu

AW TrigCoord
16/07/2025

10
19

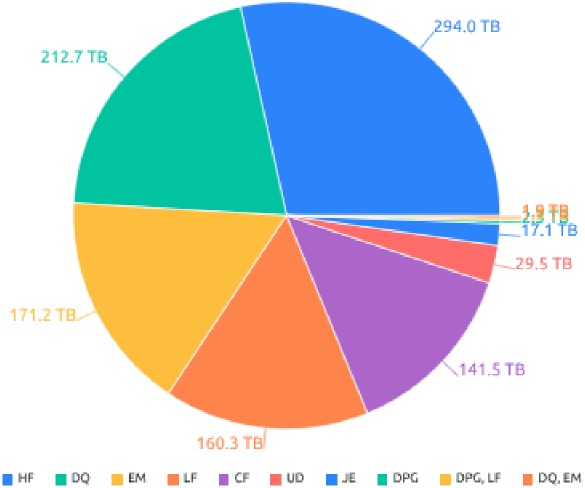
- Even more extended menu with 110 triggers!
  - ➡ Adding more femto, beauty-hadron decays, exotic tetraquark/hexaquark searches
  - ➡ Selectivity of the similar magnitude compared to 2024 (< 0.1%)

Derived/skimmed data

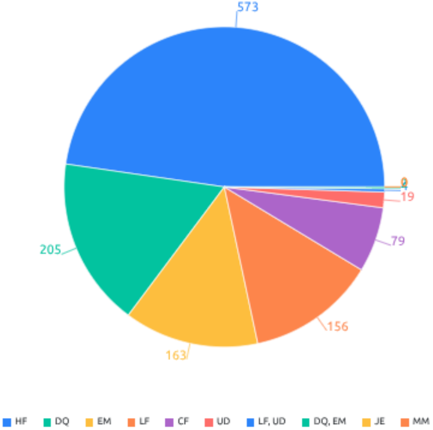
- **Statistics demanding analyses**
  - Will only be able to be run on derived data
  - Derived data stored and productified as actual data
  - Amount of stored derived data limited at PWG level
- **Only golden periods will be available for analyses**
  - Derived data concept able to be used
  - Derived data will not be stored (size on pair of actual data)

Unless we act as a collaboration

We are doing really well!



Trains last month



Provided we keep cleaning!

Derived/skimmed data

- First rule: don't create your own stored derived data
- Second rule: don't create your own stored derived data

- Present your needs in your PAG
- Be ready to discuss them in your PWG
- Be ready to incorporate others' needs into your schema

- Familiarize with the derived data data model

- The more we share the larger our reach

Derived/skimmed data

- First rule: don't create your own stored derived data
- Second rule: don't create your own stored derived data

- Present your needs in your PAG
- Be ready to discuss them in your PWG
- Be ready to incorporate others' needs into your schema

- Familiarize with the derived data data model

- The more we share the larger our reach

– THANK YOU –