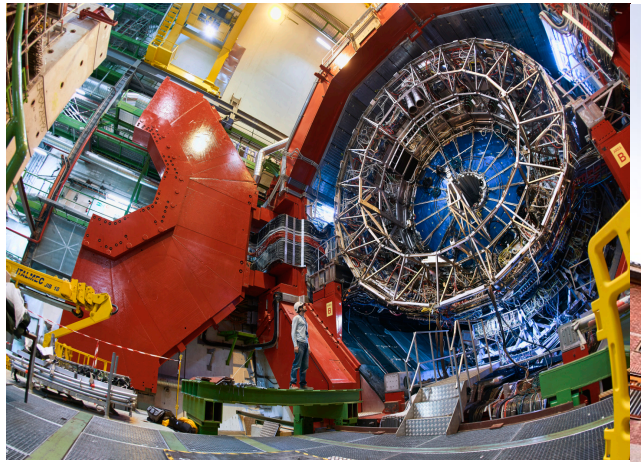
Enabling distributed analysis for ALICE Run 3

Raluca Cruceru (CERN) on behalf of the ALICE Collaboration



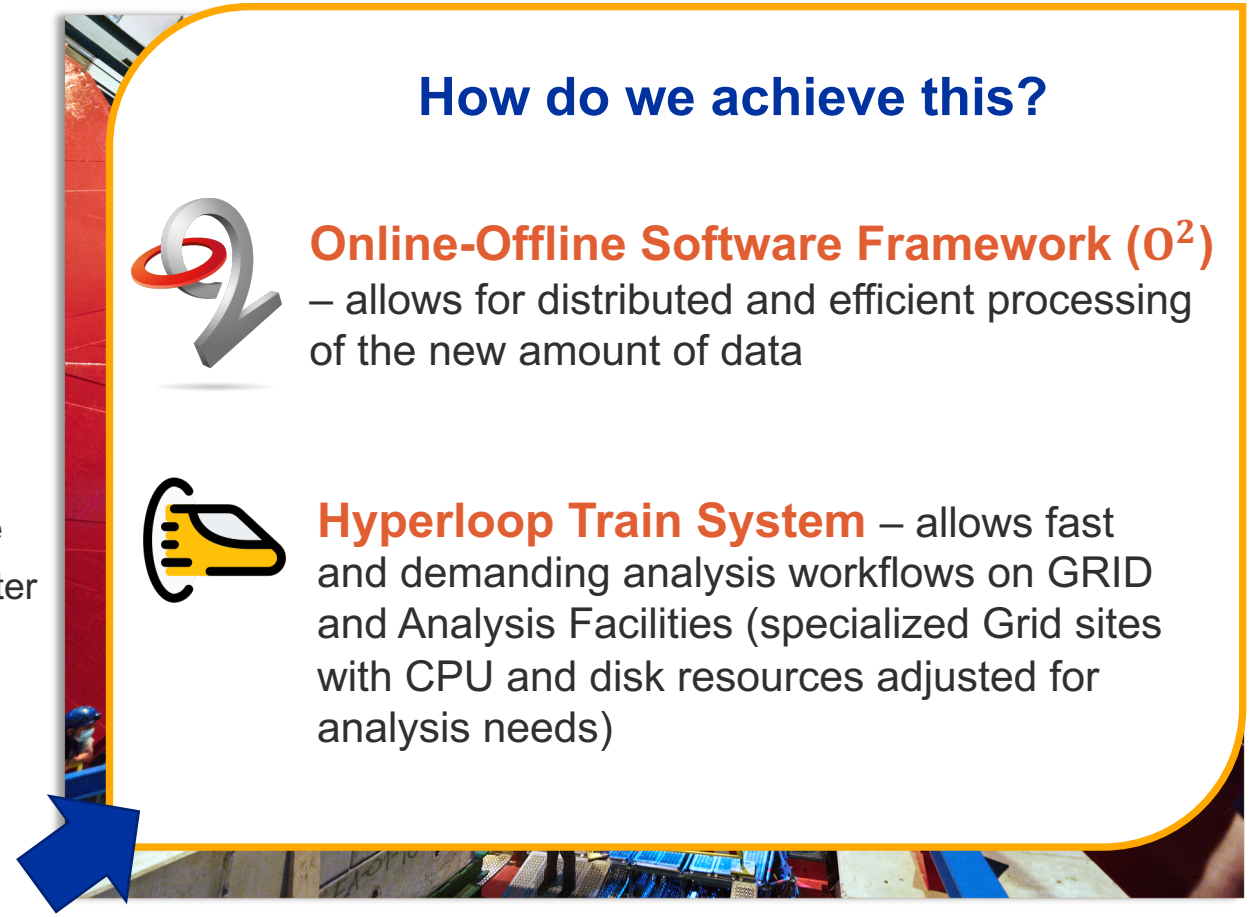
The 41st International Conference on High Energy Physics, 6-13 July 2022

ALICE Detector Upgrades for Run 3

- **Main detector parts replaced / upgraded, such as:**
  - New Inner Tracking System (**ITS2**)
  - Upgraded Time Projection Chamber (**TPC**)
  - New Fast-interaction Trigger system
  - $\mathbf{O^2}$(Online-Offline) - Data taking and Processing Infrastructure

- **What does this mean?**
  - ✓ **100 times more recorded collisions** compared to Runs 1 and 2
  - ✓ An increased data-taking capability **by two orders of magnitude**
  - ✓ Resulting data throughput from the detector estimated to be greater than **1 TB/s for Pb-Pb collisions**
  - ✓ 1 month of Pb-Pb data would create ~ 4 PB of AODs

- **What are the needs for analysis in Run 3?**
  - Process this unprecedented amount of data
  - Analysis infrastructure needs to cope with 100 times more data with more efficient algorithms and techniques



How do we achieve this?

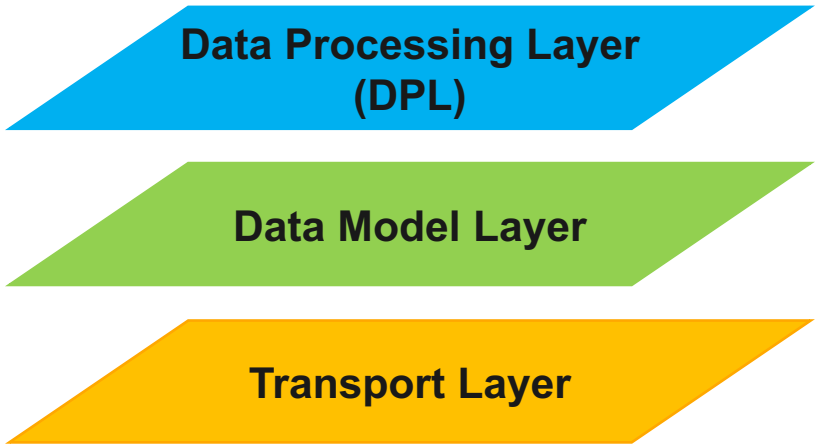
Online-Offline Software Framework (O^2) – allows for distributed and efficient processing of the new amount of data

Hyperloop Train System – allows fast and demanding analysis workflows on GRID and Analysis Facilities (specialized Grid sites with CPU and disk resources adjusted for analysis needs)

See talk by Aimeric Landou on Physics Performance - 8 July 2022, 09:18
Operation, Performance and Upgrade (Incl. HL-LHC) of Present Detectors

O^2 Framework

- **Dedicated Framework to tackle the challenges of LHC Run 3**
- Derived from ALICE high-level Trigger Architecture – message-passing multiprocess system used in Runs 1 and 2
- Processing stages:
  - **Synchronous processing** *(Online)* → resulting in the **CTF** (Compressed Time Frames) – stored on disk buffer
  - **Asynchronous stage** *(Offline)* → reconstruction with final calibration – produces the final **AOD** (Analysis Data Object)
    → CTF + AOD permanent storage
- The **O^2 Analysis Framework** hides the complexity of the underlying distributed framework, which allows analyzers to focus on physics
- **Consists of three main components:**



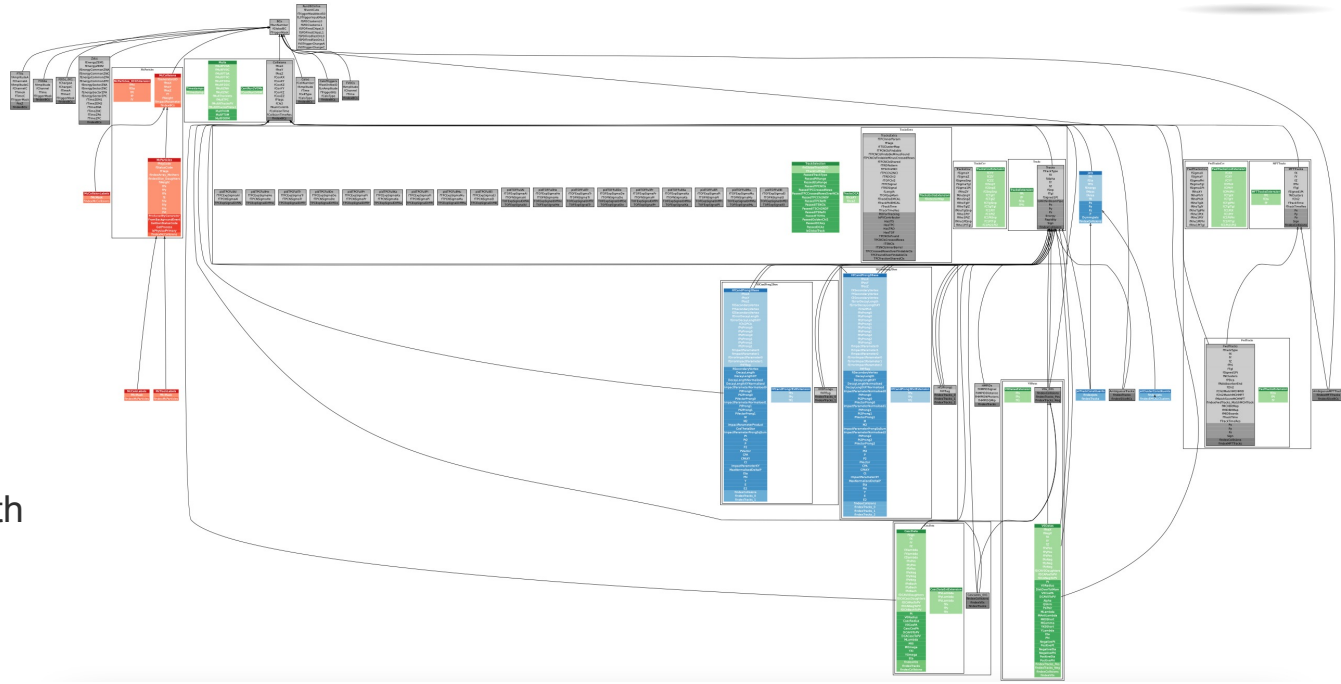
Translates user's computational problem in a low-level topology of devices exchanging messages

Allows logical description of messages and their interconnections, computer language agnostic, multiple data formats

FairMQ message passing architecture, standalone general processes (devices), shared memory backend

O^2 – Data Model

- **Collisions and tracks are represented in trees (flat tables)**
  - Connected through indices passed through shared memory
  - Minimize I/O cost and improve vectorization/parallelism
- **Columnar Format / Flat tables provided by Apache Arrow**
  - No nesting – similar to relational databases
  - Represented by C++ basic types
  - Fast and efficient operations – applied per large data blocks with 1000s of collisions
- **Complexity of representation is shielded from the user**
  - Object “look and feel”: track.pt()



Collisions

VertexPosX	VertexPosY	…

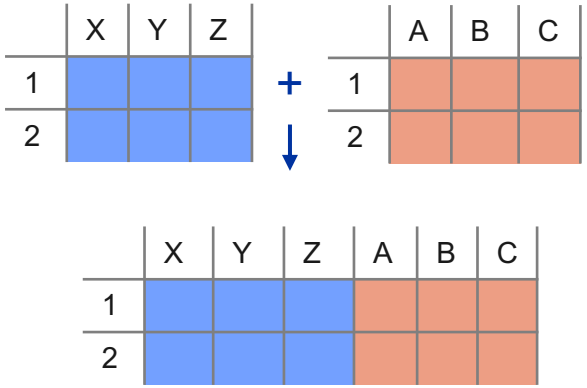
CollisionID	eta	phi	…
0			
0			
1			
1			
1			
2			

Tracks

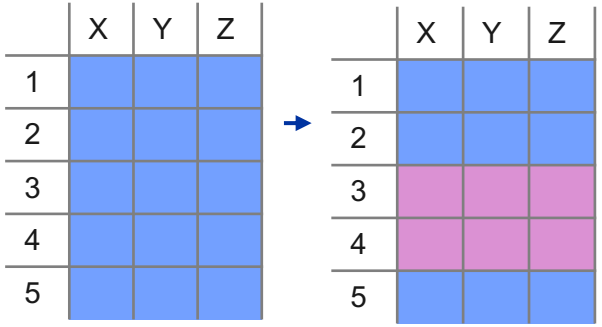
On Disk
In Memory
Monte Carlo
HF & Jets

O^2 – Table Manipulation

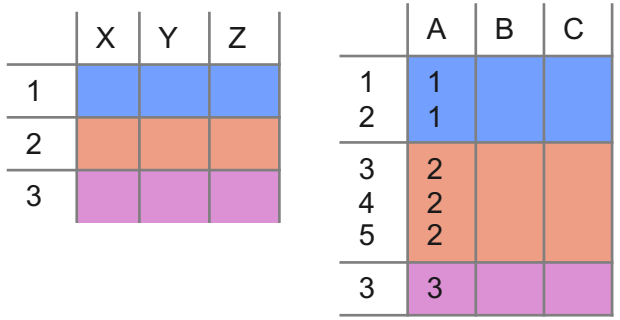
JOIN



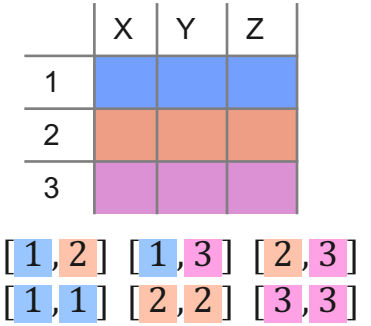
FILTERING



GROUPING



COMBINATIONS



Database-like Operations

- A Table is defined as a unique C++ type, templated on Columns
- *Join, grouping*, *partitioning* and *filtering* are requested by the analyzers (can be combined if needed)
- Apache Arrow ensures **zero-copy operations** (underlying contiguous arrays in memory are immutable - no data is removed or copied in common operations)
- The analyzer can define and create new Tables for
  Derived Data/Skims

```
60  namespace collision
61  {
62  DECLARE_SOA_INDEX_COLUMN(BC, bc);
63  DECLARE_SOA_COLUMN(PosX, posX, float);
64  DECLARE_SOA_COLUMN(PosY, posY, float);
65  DECLARE_SOA_COLUMN(PosZ, posZ, float);
66  DECLARE_SOA_COLUMN(CovXX, covXX, float);
67  DECLARE_SOA_COLUMN(CovXY, covXY, float);
68  DECLARE_SOA_COLUMN(CovXZ, covXZ, float);
69  DECLARE_SOA_COLUMN(CovYY, covYY, float);
70  DECLARE_SOA_COLUMN(CovYZ, covYZ, float);
71  DECLARE_SOA_COLUMN(CovZZ, covZZ, float);   ...
```

```
DECLARE_SOA_TABLE(Collisions, "AOD", "COLLISION", //! Time and vertex information of collision
                  o2::soa::Index<>, collision::BCId,
                  collision::PosX, collision::PosY, collision::PosZ,
                  collision::CovXX, collision::CovXY, collision::CovXZ, collision::CovYY, collision::CovYZ, collision::CovZZ,
                  collision::Flags, collision::Chi2, collision::NumContrib,
                  collision::CollisionTime, collision::CollisionTimeRes);

using Collision = Collisions::iterator;
```

O^2 – Data Processing

- **Multiprocess capability** – multiple operations run in parallel
- The user defines tasks that include callbacks and declarations of input/outputs (table subscription)
- Tasks are combined into **workflows** representing a particular analysis or a group of analyses

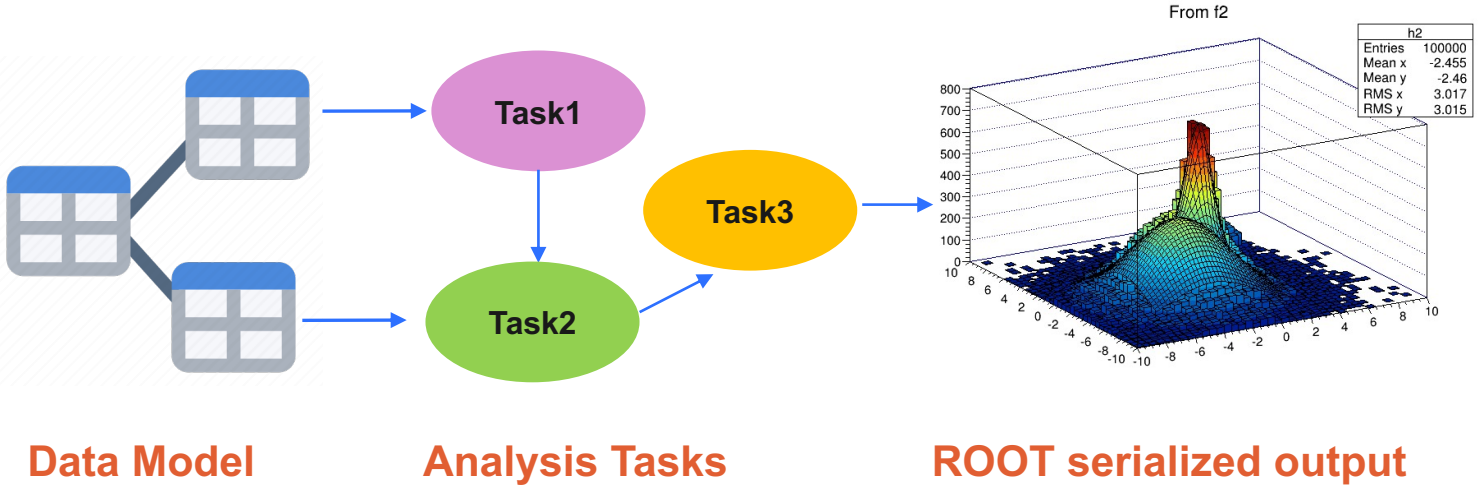
O^2 Analysis model:

Declarative: user defines filters, grouping

- Clear and compact definition of analysis cuts and flow
- Can be vectorized and parallelized by framework

Imperative: user corrections/modifications

- Certain conditions cannot be implemented in a declarative way
- Flexible, limited implicit parallelization



Data Model | Analysis Tasks | ROOT serialized output

Data Processing Layer

- Hides complexity of the abstract transport layer and low-level data model
- Builds the **workflow topology** based on interdependencies between tasks
- Translates the defined workflows to an actual **FairMQ** topology of devices

Hyperloop Train System

- Concept of **analysis trains** to optimize the usage of computing resources (built upon tools used in Runs 1 and 2)
- Workload for Run 2 analysis:
  - **40 000 cores utilized**
  - Used by more than 90% of the analysis
- Re-written with a modern reactive front-end technology: **React**
- Allows **organized analysis** on the **GRID** and **Analysis Facilities**
- Fully integrated with $\mathbf{O^2}$, allowing task configuration
- Individual workflows - known as wagons - are combined into trains
- **Skimmed / Derived data** stored for further processing in subsequent trains
- Dedicated views for regular **users** and **operators**
- **Full bookkeeping**, changelog and several comparison tools
- **Data available**: converted Run 2 data, Run 3 data and MC, Derived data

USER

MyAnalyses

AllAnalyses

Dashboard

OPERATOR

Train Submission

Train Runs

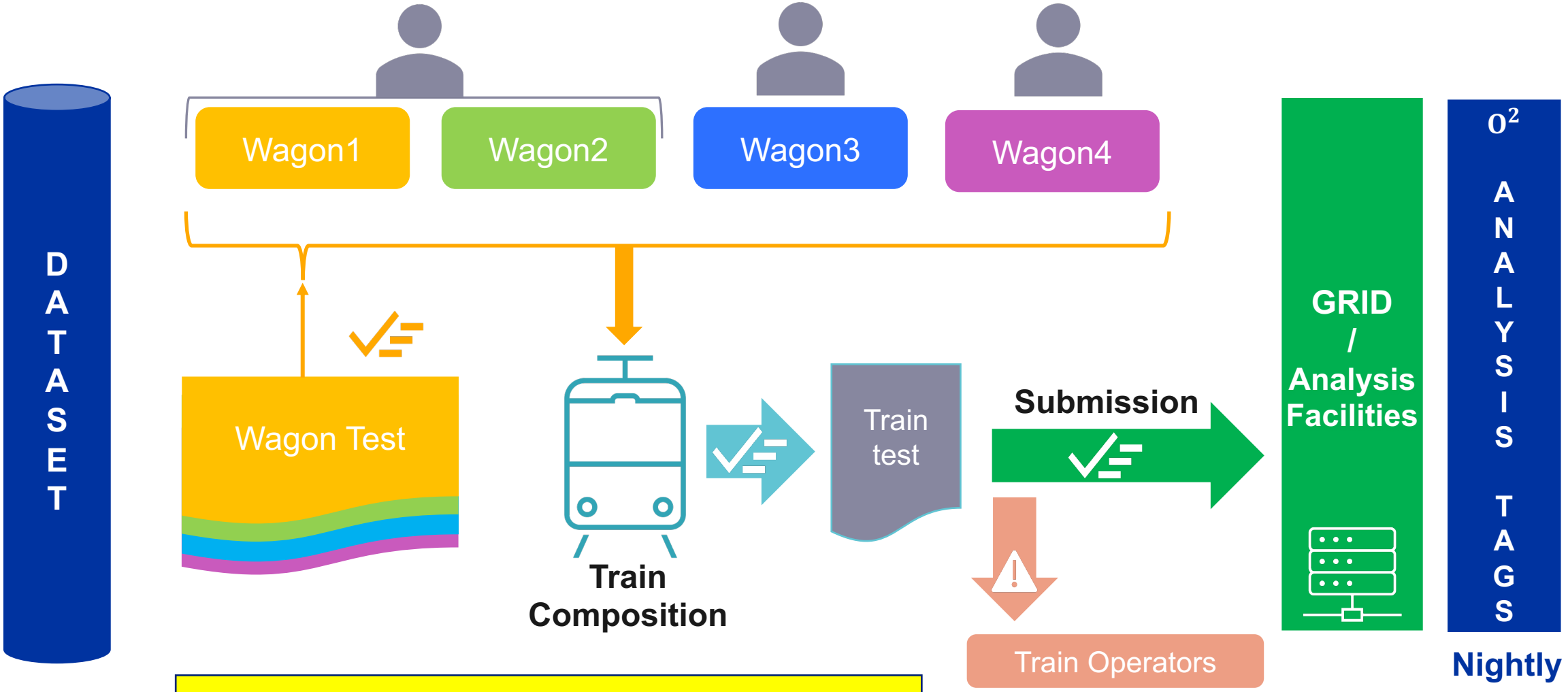
Datasets

Derived Data

DPG runlists

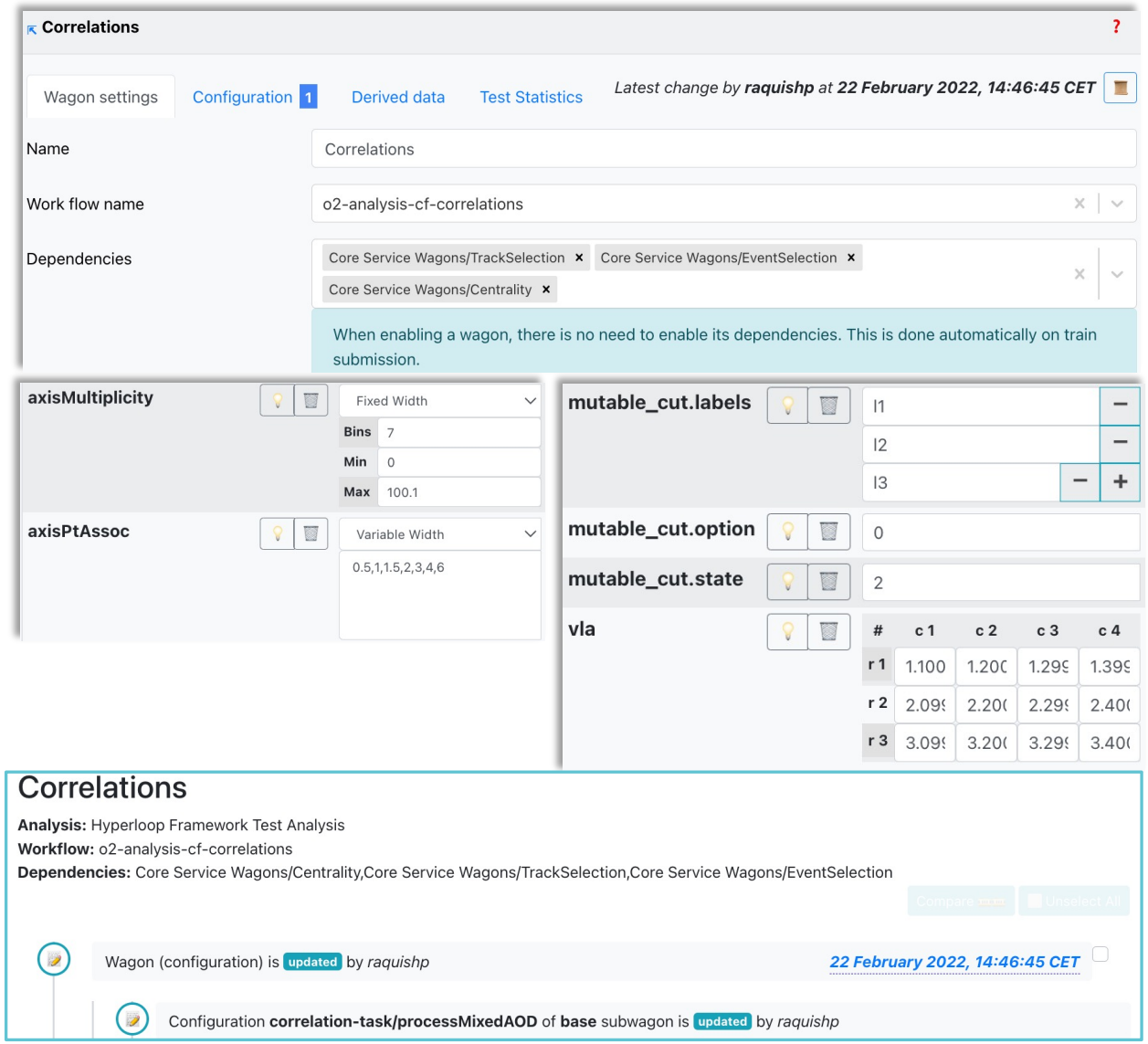
Trains with issues

Hyperloop Train System



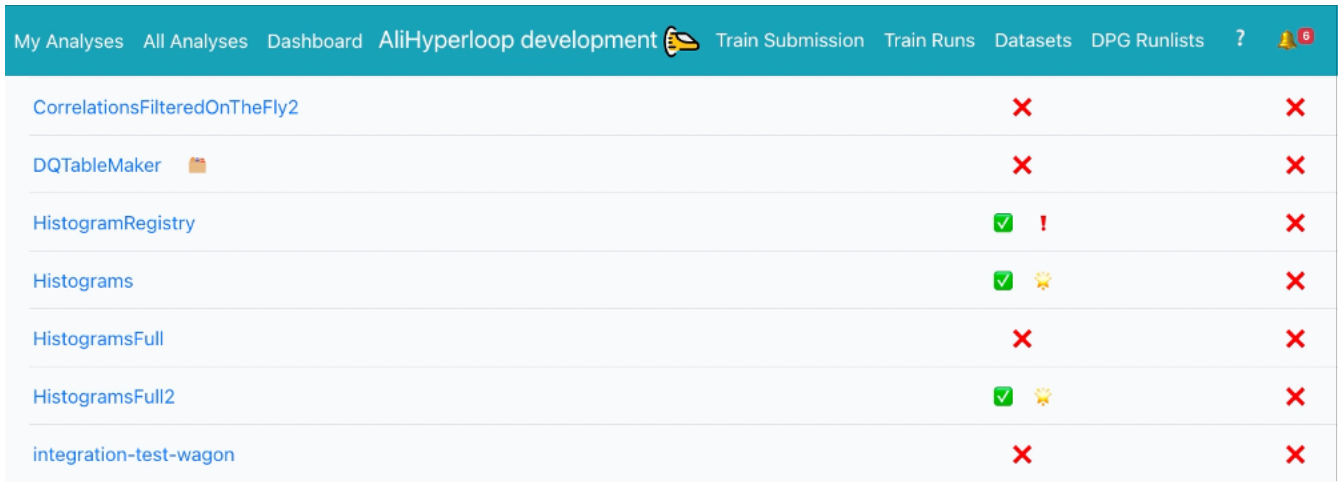
Powerful backend machine (50 cores) for train tests

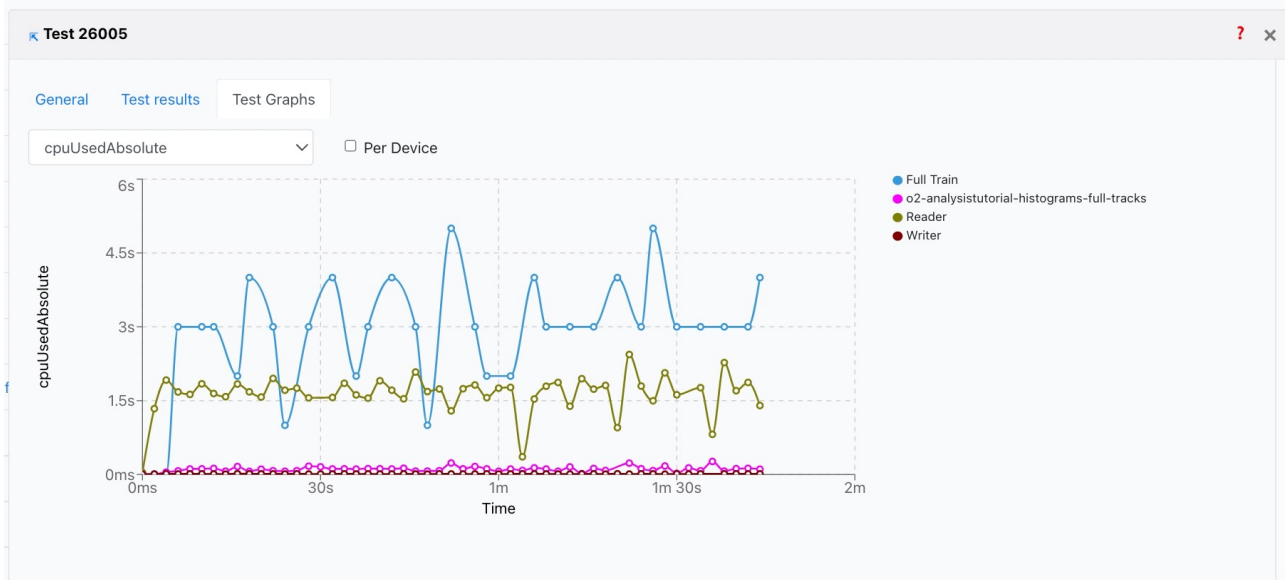
Hyperloop – Wagons and tests



- **Analyses defined in JIRA, shared by users**
- **Datasets are enabled per analysis**
- **The user can add, configure, clone or remove wagons**
- **Wagons created from available O^2 workflows or pull request**
  - Supports a variety of parameter types such as primitive types, array, matrices, labelled matrices, histogram binning
  - Allows subwagons creation – these will run the same task with different parameter values
  - Output can be stored as derived data

Hyperloop – Wagons and tests





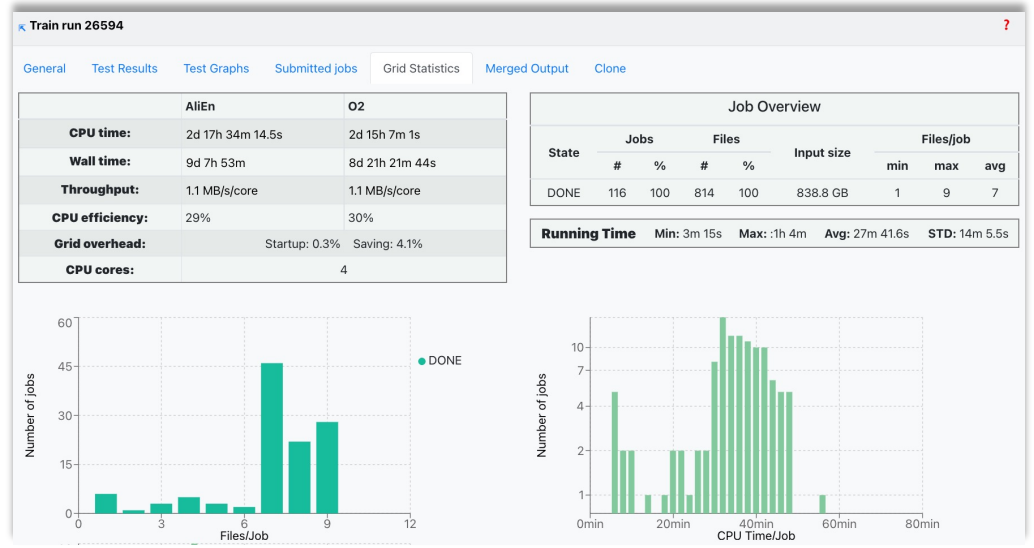
- **Analyses defined in JIRA, shared by users**
- **Datasets are enabled per analysis**
- **The user can add, configure, clone or remove wagons**
- **Wagons created from available O^2 workflows or pull request**
  - Supports a variety of parameter types such as primitive types, array, matrices, labelled matrices, histogram binning
  - Allows subwagons creation – these will run the same task with different parameter values
  - Output can be stored as derived data
- **Analyzers can enable / disable wagon tests**
- **Immediate testing and overview**
- **Per device (reader, workflows, writer) and total performance metrics**
- **Expected resources and Interactive graphs**

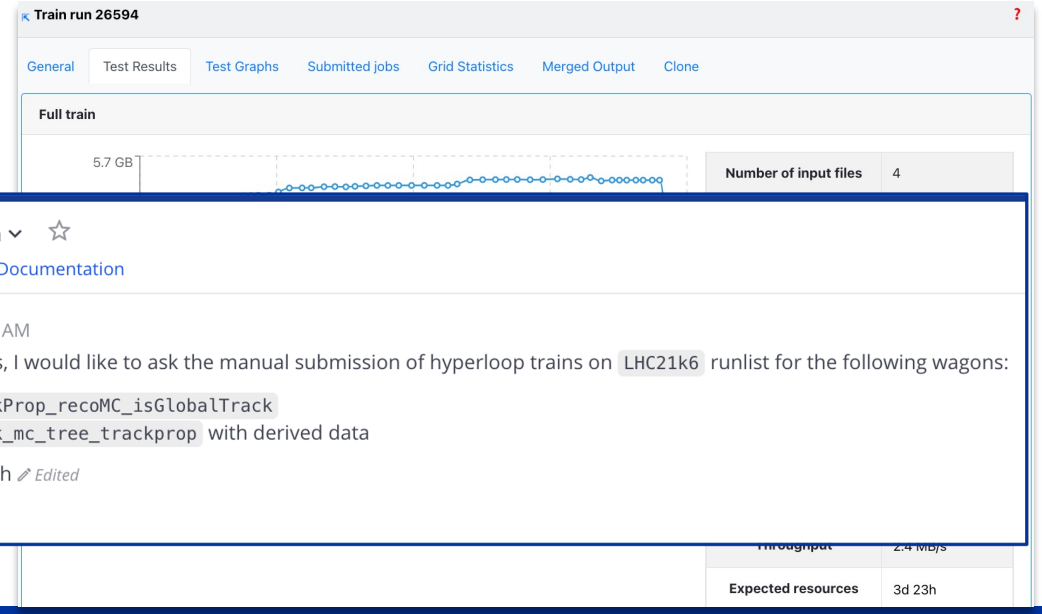
Hyperloop – Train runs

- Automatic **train composition** can be scheduled per dataset – composition based on target memory, wagon configuration and dependencies
- Automatic **train submission** to **GRID** or **Analysis Facilities**
- Train test results, submitted jobs and Grid statistics
- **Clone train** – runs with the same wagon timestamp but updated dataset configuration
- **Staged submission** for large data samples – run first on a smaller dataset before approval to run on bigger dataset (approved by PWGs)

Train Support

- 24/5 Operation (different timezones)
- Institutes: 1 in Americas, 2 in Europe,1 in Asia
- Dedicated channel for users' request and issues
- Shift-type support during working hours
- Organized feedback sessions

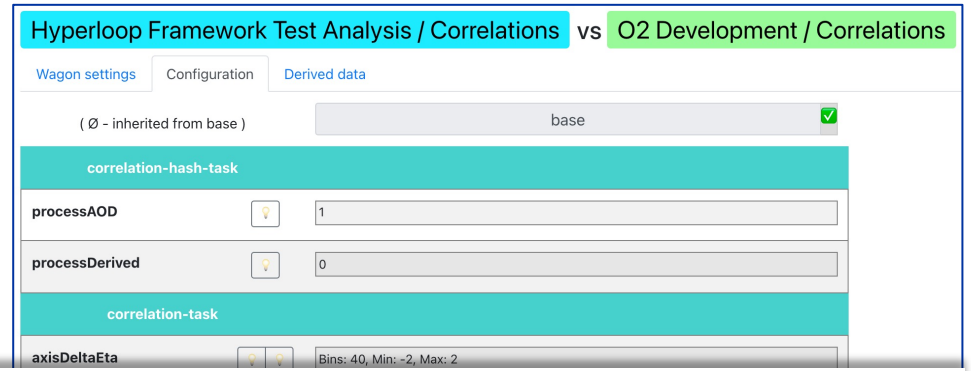




Hyperloop – Monitoring and status

Monitoring and Bookkeeping

- Wagon, dataset and runlist bookkeeping – **comparison tools**
- Derived data recording – track usage and mark for deletion
- **Dashboard** - up to date information about the system
  - Previous week summary, job status overview per site
- **Unit and web testing** (**Jest**, **Puppeteer**)
- WebSocket implementation
- **Personalized notifications** (also by email)
- Extensive documentation accessible from the UI



Status

- Already in production - currently **140 users**
- 100 analyses, 95 datasets, 3200 train runs
- **Converted data to O^2 format** – the new framework supports analysis on old data
- **Reconstructed Pilot beam data** available on Hyperloop (5 dedicated datasets)
- Ongoing QA of the MC data

Datasets

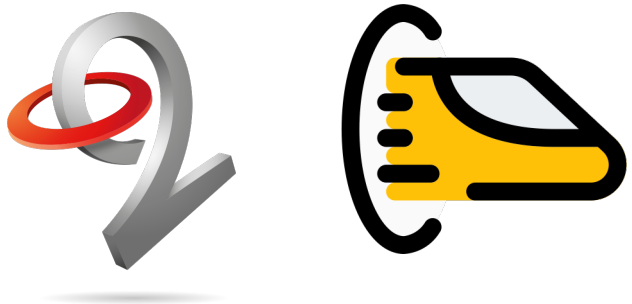
Derived data | Show removed datasets

Clear all filters

Name	Description	Type	Production name	DPG runlist
PilotBeam	Search 5 records...	Se	Search 5 records...	Search 5 records...
PilotBeam_EMC	Runs with EMC from Pilot Beam (O2-2768)	DATA	NOV_ctf_emc AODmerge_OCT_99 OCT_ctf_emc	all all all
PilotBeam_Unaligned	First reconstruction, not aligned	DATA	AODmerge_OCT_10	all
PilotBeam_pass2	Async pass2 : https://alice.its.cern.ch/jira/browse/O2-2726	DATA	AODmerge_OCT_20	all
PilotBeam_pass3	async pass 3	DATA	AODmerge_OCT_30	all
PilotBeam_pass4	Async pass4	DATA	AODmerge_OCT_40	all

Summary

- **ALICE detector received several major upgrades, allowing for 100 times more data in Run 3**
- **New tools were developed: O^2 Framework and Hyperloop Train System**
- **O^2 Framework allows for distributed and efficient processing of the unprecedented data**
  - The three main layers (transport, data model and data processing) will secure a uniform complete framework for distributed and efficient processing of the new amount of data
  - Hides underlying complexity from the users
- **Hyperloop enables analysis workflows to be run on the Grid and Analysis Facilities**
  - Automatic activity (e.g. train composition and submission) → less actions to be taken by the operators
  - Modern interactive User Interface
  - Easy access to status overview and documentation
- **User support 24/5**
- **ALICE is ready for analysis in Run 3**



Thank you!

https://alimonitor.cern.ch/hyperloop/