Hyperloop Datasets

Jan Fiete Grosse-Oetringhaus (CERN)

08.12.2020

Hyperloop Datasets

3 Types

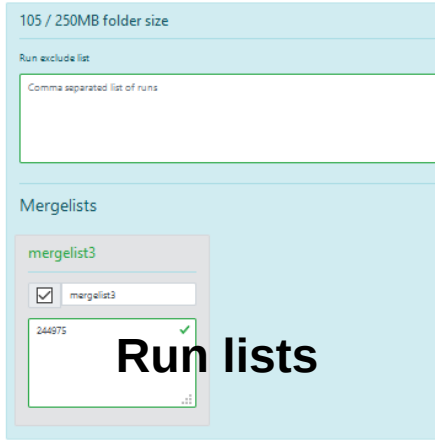
- Data
- MC
- Derived data
- Converted Run 2 data

Production — **Hyperloop train** — **LEGO train (train: Run3_Conversion)**

- **Each dataset, can contain several elements (of the same type)**
- **Each element processes either all runs, or a subset**

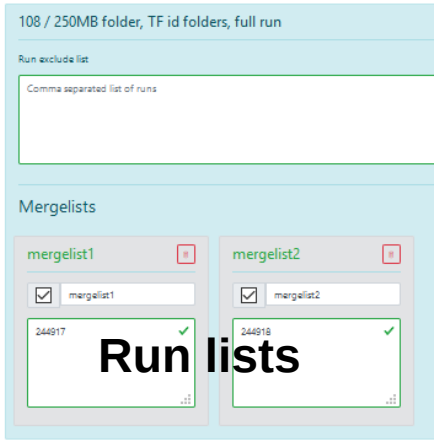
Dataset

Element 1



Run lists

Element 2



Run lists

- **List of directories**
  (code that extracts directory list from dataset exists in ML)
- **File pattern**
  (for now fixed to: AO2D.root)

PID	Run no.	Output directory
Search	Search	Search 3 records...
2058123240	244918	/alice/data/2015/LHC15o/000244918/pass5_lowIR/PWGZZ/Run3_Conversion/108_20201117-1536_child_1/hy_1776
2058123233	244917	/alice/data/2015/LHC15o/000244917/pass5_lowIR/PWGZZ/Run3_Conversion/108_20201117-1536_child_1/hy_1775
2058123231	244975	/alice/data/2015/LHC15o/000244975/pass5_lowIR/PWGZZ/Run3_Conversion/105_20201021-0845_child_1/hy_1774

NB. The hyperloop database is change aware (we store the full history, and the timestamp of the last change)

Requirements

- Staging should be for entire Hyperloop datasets
- Staging status of a Hyperloop dataset needs to be available machine readable (we have to restrict submission to where the dataset is)
- Changes of Hyperloop datasets should be taken into account
  - Database contains already history and last change data of a dataset
  - Typical use case (Run 2 example): Dataset “LHC15o” always traces the last related production
- There can be overlaps between datasets (data should only be staged once)
  - Example:
    - LHC15o_test (few runs for pilot trains)
    - LHC15o_af (10% of the runs of the full production, each run is fully available)
    - LHC15o (all runs)

Example

- Yesterday: Dataset A : 1 2 3 | Pin (set by op.): **Yes**/No | Status: Ready/**In Progress**
- Today: Dataset A : 1 2 3 | Pin (set by op.): **Yes**/No | Status: **Ready**/In Progress
  - Train run 7, 8, 9
- Operator change dataset A → A': (add run 4) | Pin: **Yes** | Status: **In Progress**
- Tomorrow: Dataset A': 1 2 3 4
  - Train run 10, 11, 12

Some notes on staging tool

- Datasets retrieved from hyperloop database (current state)
- Each dataset
  - Can be pinned to AFs (one pin per AF)
  - Total size displayed
  - Status of staging (one status per AF)
- Available space displayed per AF
- Staging tool
  - Generates the necessary (copy, remove) operations to achieve the pinning status
    - Take into account that a directory can be pinned by several datasets
  - Processes them in the order of request
  - Updated the pinning status when transfers are done

- Display what a change of pin status will imply wrt available space