Data Carousel and Archival Metadata

Xin Zhao (BNL), Alexei Klimentov (BNL), Mario Lassnig (CERN), Misha Borodin(UIowa)
Technical Exchange Meeting (TIM), January 21st, 2025

Start with a question from a recent (DOMA) meeting –

“Before you go on with all the R&Ds (in Data Carousel), have you done any mathematical or theoretical studies that prove the goals are achievable ?”

– “the goals” refer to not only the various R&Ds, but also in general if Data Carousel is the answer to our HL-LHC data challenges.

Straight answer is No we haven’t. But …

- Do we (ATLAS) have plan-B ? Tape stays as the **affordable** way to address the HL-LHC data challenge.
- We have to use tape efficiently, or the cost saving won’t be as much (as it can be).
- Success stories – BNL RHIC experiments/SDCC, TRIUMF/INFN, KIT …

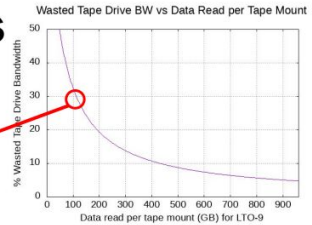
What’s happening in Data Carousel ?

- Unifying Data Carousel machinery for both production and user analysis jobs
  - Stream 4 in the PanDA project management organization
- Better control of tape writing streams, to help release pressure on tape buffer at sites
- Two HL-LHC demonstrators – results reported at CHEP’24
  - DAOD-on-demand
  - Tape smart writing with KIT Tier-1
- Smart writing – file grouping on tape
  - Will work with more sites on smart writing demo, when they are ready.
  - CTA/CERN has started their own study on this topic (for CTA sites)
    - Collected archival metadata for RAW in the recent HI run
    - Simulation/modeling between different approaches and metrics
  - Evaluations at other sites (e.g. ongoing internal discussions at BNL SDCC)
  - ***Archival metadata***
- Expected data volume/dataset size/throughput targets for tape for Run4?
- ***ADC to evaluate tape workflows and possible improvements*** for more efficient tape usage (e.g. event index/picking jobs)
  - Sporadic effort so far, need more systematic studies

Archival Metadata & Data Grouping Unit(s)

- Archival metadata is the hints ATLAS provides to sites for grouping ATLAS files on tape.
- Dataset is a natural grouping unit
  - The grouping level used by all site smart writing solutions so far.
- As tape capacity and speed continue to grow in the future, grouping levels above dataset will become necessary, in order to keep the bandwidth utilization high
  - c.f. BNL studies (Shigeki)

Optimizing for Small Datasets



- Most ATLAS datasets are small relative to size needed to efficiently use tape drive
  - ~70% of DATATAPEdatasets < 100 GB
  - ~80% of MCTAPE datasets < 100 GB
- Co-location of datasets in a “retrieval group” could increase efficiency by increase data retrieved per tape mount [1]
  - Requires identifying these groups of datasets and getting them on a common set of tapes

Archival metadata

- A generic solution being developed
  - Using HTTP header (in json format) in the transfer request
  - A format proposed by CTA/dCache group (1KB size limit enforced)
  - Very flexible format
    - Level names just numbers “0”,”1”,..., can be associated with any attributes (stream, data type,...)
    - No need to fill in all levels if not needed
    - Higher level grouping should keep the lower level units intact
  - Experiments need to fill in the contents of the metadata

```
archive_metadata = {
    "scheduling_hints": {
        "archive_priority": "100"           # highest priority
    },
    "collocation_hints": {
        "0": "data23_13p6TeV",                                      # project
        "1": "RAW",                                                 # datatype
        "2": "physics_Main",                                        # stream_name
        "3": "data23_13p6TeV.00452799.physics_Main.daq.RAW",        # dataset
    },
    "additional_hints": {
        "activity": "T0 Tape",          # Tier-0/DAQ
        "3": {                          # dataset level
           "length": "19123",               # total number of files at specified level
           "size": "80020799318456"         # total size of files at specified level
        }
    },
    "file_metadata": {                  # file content metadata
        "size":"193734404",
        "adler32":"379ebf71",
        "md5":"952c4c0dabc622a94f09b053d71d0dfb"
    }
}
```

Open Questions about Archival Metadata Templates

- What are a good grouping hierarchy for a data type ?
  - Ask experts (ADC experts, production managers, physics groups …)
    - Sometimes not easy to converge among experts
  - Ask data …
    - Analyze historical recall logs
  - Ask machine …
    - Train AI with our historical recall logs, let AI/ML learn recall patterns (e.g. what datasets are likely to be recalled together ?)
- It's hard (if not impossible) to know the size of a grouping unit above dataset level
  - Size info is important, refer to the KIT implementation
  - Ideas floating around …
    - No need to know the real size of all RAW datasets belonging to a particular stream collected during 2024 run. Our purpose is to find grouping units that's big enough to ensure good bandwidth utilization in recall campaign
    - Rucio can create artificial retrieval groups within a level, e.g. put several AOD datasets having the same tid into one container, and tell sites to co-locate them together.
      - we can call them "tape containers", a container type solely for tape grouping purpose
    - Definition of a "good size" is expected to grow as tape technology evolves, and may even be different per site.

A proposed project …

Analysis of ATLAS tape access patterns

- Objectives
  - To figure out the archival metadata templates for the various data types (RAW, AOD, etc) on tape
  - To analyze how ATLAS recalls data from tape today, recommend improvements for more efficient tape usage
- Phases
  - Collect recall history
    - Sources:
      - ProdSys2 DB/PanDA DB/Rucio DB
      - Or, ELK stack having them all?
    - Outcome:
      - a data sample that contains the history of ATLAS recalls from tape, since the beginning of Run3.
        - Information to collect → to be defined (one sketchy idea to the right)
      - By what percentage that ATLAS recalls data from tape by partial dataset

```
{
  "dataset name": [
    "real data or MC",
    "tag1 (e.g. project)",
    "tag2 (e.g. data type)",
    "tag3 (e.g. task ID)",
    "tag4 (...)",
    "# files (original)",
    "size (GB)",
    "issuer (PS2 or PanDA)",
    "physics groups/campaigns",
    "when requested",
    "when DDM started recall",
    "when DDM finished recall",
    "# files (recalled)",
    "where recalled (src site)",
    "destination site"
  ]
}
```

Analysis of ATLAS tape access patterns

- Phases (continued …)
  - Analysis of the recall history
    - Tools/platforms to use ?
      - Analytics, AI/ML, … ?
    - Outcome:
      - Recommended hierarchy of grouping levels for all data types on tape
      - Categorized use cases that lead to partial dataset recalls (e.g. individual users, special workflows, disaster recovery etc), and their percentage
  - Discussion of the preliminary results with relevant expert groups
    - Present the above analysis results to the physics groups, production managers etc, for their experts' feedback on :
      - Do such grouping levels make sense?
      - Can we improve some tape workflows ?
    - Several iterations of analysis-discussion may be needed.

Analysis of ATLAS tape access patterns

- Final results
  - Archival metadata templates for various data types on tape.
    - Rucio team will code them into the metadata, to be passed to site storage when writing files to tape.
  - Possible improvements recommended in ATLAS tape workflow, leading to less partial dataset recall cases and others, for better tape usage.
    - Guidance for future evolution of the Data Carousel machinery in PanDA.
- Questions –
  - Shall we also study how ATLAS writes to tape ?
    - Makes more sense **after** we understand better about the read pattern.
  - Archival metadata support in dCache ?

Timeline ?

- Run4 is around 2030, but R&D has a shorter timeline
  - There is a regular tape system procurement and deployment cycle, which varies from site to site.
  - For BNL SDCC, the next tape procurement cycle will start in 2027, any HL-LHC oriented tape R&D and prototyping should wrap up before then.

Backup slides

Other open questions/discussions

- Tape simulator
  - Proposed and planned by some sites
    - For example, one proposal is to replay tape write history, through a particular file placement scenario; then replay tape read history, and tell what's the expected (theoretical) tape drive bandwidth utilization and overall throughput
  - Answer questions like :
    - which grouping scenario is better, under a certain condition, e.g. one dataset on one (or few) tape or stripped grouping among multiple tapes ?
    - how much performance improvements (theoretically) is expected from one grouping scenario over the others ?
    - what's the ideal size of grouping units, assuming certain conditions and tape technology ?
    - may point out things to improve also on the way tape write/read requests are sent to sites
- Tape monitoring
  - Overall throughput delivered from tape
  - Bandwidth utilization
  - …