Some Notes on Metadata

Jack Cranshaw

August 20, 2014

General Status

- No major changes to metadata content from Run 1.
- How that metadata is collected and used has evolved and is continuing to evolve.
- It is one of the last elements of the software for Run 2 reaching a real validation stage.

A Different Data Processing Model

Run I

- > 5 RAW streams + express
- Nominal trigger rate of 0.2 kHz.
- Backward compatibility through schema evolution.
- Derived streaming (DESD, DAOD) based on needs of specialists.
- Massive repetition of data at D3PD level.
- Physicists run on ntuples.

Run II

- One RAW stream + express
- Nominal trigger rate of 1 kHz.
- Backward compatibility through reprocessing.
- Derived streaming includes needs of physicists as well.
- Monitored and managed data at DxAOD level.
- Physicist run on ntuples.

Derivation Framework

- Athena based
  - All standard tools available.
- XAOD in, XAOD out.
  - Format may change, output may be more dynamic aux store (quasi-ntuple) like.
- Train structure still being worked out.
  - Possible: big jobs together, similar tools together.
  - Probable: one train per physics group.
  - Configs defined by release. New config = new release.
- *Could* run every two weeks.
  - Run I use cases place this at more like every 4-6 weeks.
- Uses standard athena output streams
  - *DecisionSvc*: monitors event overlap
  - *ItemListSvc*: monitors content overlap
- Tool content assurance
  - Tools are run during release build/test, and a list of the expected inputs is generated and made available.
- Status of stream definitions
  - Good contact (ASG) with physics groups.
  - Around 30 streams currently being defined.
  - Many of them have <5% of input events and <10% of input data containers retained.
- Status of production definitions
  - Plans in place. Initial, small scale tests.
  - Metadata and merging still have details to be worked out.
  - Still thinking about collapsible work flows.

Data Preservation Concerns

- This really hit the radar (or the fan) a few years ago.
- ATLAS concerns
  - Access to full data sample for lifetime of experiment.
  - Focus on RAW and reproducible processes, but be flexible. Resource intensive.
- Public concerns
  - Provide tangible return on investment for public resources.
  - Focus on analysis/publication stage. Emphasize metadata and subsetting to reduce resource intensity.
  - Should exist independent of collaboration.
- For both of them, the archiving of processes is a new approach whose implications, IMO, have not been fully worked out. The implication is:

Upgrading data is easy, upgrading software is hard.

Luminosity Accounting

- Situation for most of Run 1
  - File boundaries = luminosity boundaries
- For late Run 1 and in Run 2, this is not true.
  - Dataset boundaries = luminosity boundaries
- The previous default was to just make sure that all files were processed and use a prepared lumiblock list to calculate luminosity. No accounting needed.
- More scalable. Do accounting at file level, but when dataset is finished, run a job which compiles the lumiblock list for that dataset and stores it as dataset metadata.
- A better accounting of expected events in a given lumiblock for an input dataset would be useful (lumi meets cutflow).
- Work in progress.

File Peeking

- Serious performance problems.
  - https://indico.cern.ch/event/325424/contribution/5/material/slides/0.pdf
- File contents may be out of date.
- Improved access and integration with external metadata sources?
- Difficult because this is so pervasive within ATLAS.
- There were differences in file metadata architectures between D3PD and AOD in Run 1.

Data Location

- TAGS
  - Still around, but now a monitoring ntuple.
  - Will probably continue to function as grid resident pre-reco navigation tool.
- Event Index
  - Factor mutable data such as physics quantities into rump TAG data product. Concentrate on supporting event picking and static/immutable metadata such as trigger decisions.
  - Two primary improvements in the pipeline
    - Fast access: Jobs will be able to send packets of data directly to a server at CERN at job completion. No small files or extra data transfers for DDM.
    - All data products indexed: Job level data collection allows it to be extended to post-reco data products.

Analysis Metadata

- In Run 1 there was a lot of activity where physicists couldn't be sure what was done to the data, so they simply ran and re-ran comparisons of final data sample.
- Further discussion after Nils' talk.