Metadata

Heidi Schellman, Oregon State University

Heidi Schellman, Sept 7, 2020

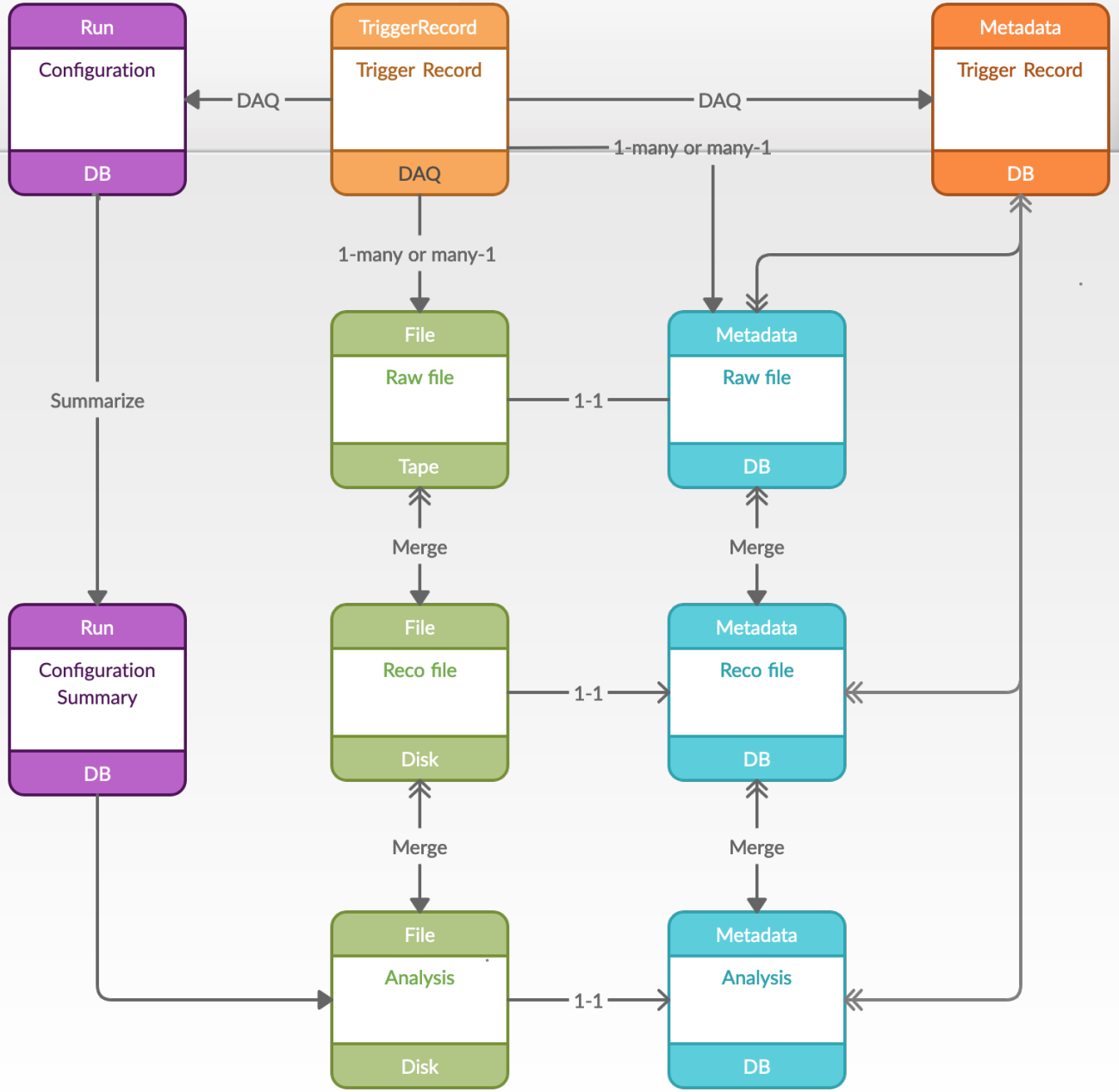
Metadata

- Data can either be in a file or in a DB
- Expect that a file will be on tape and may have access latency of > 1 day
- Basically, if you want it fast, it's metadata

Need to have

- Data file
- File metadata to describe file content
- Trigger record metadata to describe mapping from Trigger record to file (and vice versa)
- Record of configuration

Each processing stage generates file metadata



File description metadata content

- Run, Subrun, Event or….
- Trigger record ID << not here yet.
- Creation timestamp
- Data stream (what type of trigger is it?)
- Data tier (what level of processing is it?)
- Version of code used to generate it?
- What input files were used to generate it?
  - What output files have been associated with it? (updates)
- Other info about calibrations, processing parameters …
- Who dunnit?

https://wiki.dunescience.org/wiki/SAM_Data_Catalog

- Information on file location will be stored separately.

Use case - HPC

- Configuring a processing job
  - Find the right data
  - Identify calibration and conditions db info for possible local caching
  - Get the data to the processing system
- Running a processing job
  - Understand internal data structure of file
  - Access local copy of calibration/conditions
  - Produce output file and consistent output metadata