

IceCube File Catalog

CHEP 2018 - 12 July 2018

Patrick Meade

IceCube Neutrino Observatory



Metadata Catalog

- Files from various sources:
 - Filtered Experimental Data (Daily Satellite)
 - Raw Experimental Data (Yearly Shipment)
 - Simulation Data
 - Data Analysis (Level 2, Level 3, ...)
 - Misc (Backups, Debug data, Media)
- Almost everything has XML metadata attached to it
- Archives (tar, zip) make it difficult to access
- Need a simple service to make this data accessible

Flexible Schema

- File Catalog uses MongoDB for backend storage
 - Metadata is just a plain old JSON document
- A very few top-level fields are required:
 - uuid, logical_name, locations, file_size, checksum
- Some other top-level fields are usually populated:
 - start_datetime, end_datetime, run_number, subrun_number, first_event, last_event
- Remaining fields are flexible; by discretion of populating application

CRUD via REST API

- GET /api/files - Query the catalog
- POST /api/files - Add a new metadata record for a file
- DELETE /api/files/{uuid} - Delete a metadata record for a file
- GET /api/files/{uuid} - Query a metadata record for a file
- PATCH /api/files/{uuid} - Update a metadata record for a file
- PUT /api/files/{uuid} - Replace a metadata record for a file

Collections and Snapshots

- Collections are defined by a QUERY
 - GET /api/collections - Query collections
 - POST /api/collections - Create a new collection
 - GET /api/collections/{uuid} - Query a collection
 - GET /api/collections/{uuid}/files - Query a collection's current files
- Snapshots are defined by a Collection at a point in time
 - GET /api/collections/{uuid}/snapshots - Query snapshots of a collection
 - POST /api/collections/{uuid}/snapshots - Create a new snapshot of a collection
 - GET /api/snapshots/{uuid} - Query a snapshot
 - GET /api/snapshots/{uuid}/files - Query the files of the snapshot

More Possibilities

- Current entities are:
 - File
 - Collection
 - Snapshot
- Considering new entities:
 - Analysis_Sample
- Web interface does not exist yet:
 - Popular queries
 - Well known events
 - Can make use of flexible REST API query

Interesting Reflections

- File create performance is 200 docs/sec
 - 10M documents ~= 14 hours
 - No exploration of scale yet; 1 FC + 1 MongoDB
- Metadata schema is flexible
 - Applications need only meet basic requirements
 - Fixed schema communicates meaning between applications
- Simplicity is a boon
 - Clients in Java, JavaScript, and Python
 - Eating our own dog food (Final Analysis Sample)

Thank You ^_^

- Wisconsin IceCube Particle Astrophysics Center (WIPAC)
- Patrick Meade
patrick.meade@icecube.wisc.edu
- Thank you for your kind attention! ^_^