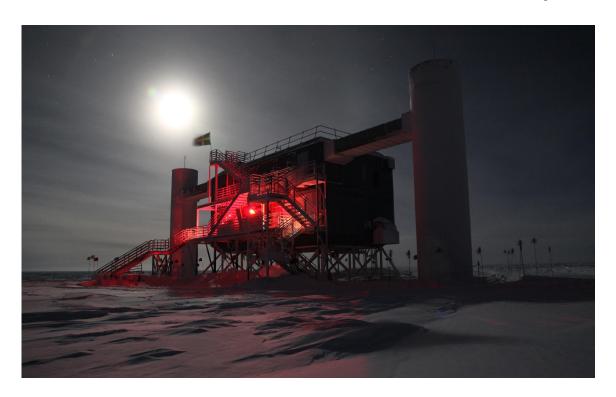
# IceCube File Catalog

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:
  - o 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
- POST /api/files
- DELETE /api/files/{uuid}
- GET /api/files/{uuid}
- PATCH /api/files/{uuid}
- PUT /api/files/{uuid}

- Query the catalog
- Add a new metadata record for a file
- Delete a metadata record for a file
- Query a metadata record for a file
  - Update a metadata record for a file
- Replace a metadata record for a file

### Collections and Snapshots

- Collections are defined by a QUERY
  - GET /api/collections
  - POST /api/collections
  - GET /api/collections/{uuid}
  - GET /api/collections/{uuid}/files

- Query collections
- Create a new collection
- Query a collection
- 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}
  - GET /api/snapshots/{uuid}/files

- Query a snapshot
- 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! ^\_^