

JADE Long Term Archive

CHEP 2018 - 09 July 2018

Patrick Meade

IceCube Neutrino Observatory



IceCube Data Outlook

- Data Warehouse is ~6 PB
- RAW data is large (~1 TB/day)
Filtered data is ~75 GB/day
- We want to retain older data, but not keep it hot
- We want space for new data, new analysis

JADE Long Term Archive

- Custom software to package data for long term archival
- Indexer and Bundler components process a single day of data:
`/data/exp/IceCube/{YYYY}/unbiased/PFRaw/{MMDD}`
- Bundles are 500-1000 GB
Our collaborators at DESY and NERSC want large files for tape
- Bundles are manually staged and queued for transfer

Take 1: Globus Online

- Globus Online manages the transfer between UW-Madison and our collaborators DESY and NERSC
- Up to 3 transfers at once; individual files gave poor performance
- Directory transfers allowed Globus to do internal concurrency (3x4); performance increased to acceptable levels after tweaking
- About 1 PB moved to NERSC so far

Drawbacks and Limitations

- GlobusOnline has decided to go commercial and closed-source
- Significant operator effort
 - Data directories to be bundled are manually selected
 - Archive bundles are manually staged
 - Transfers are manually initiated
- Limited tools for reporting
 - Reporting tools were buggy at first
 - Reporting is not configured to consider new directories

Take 2: Rucio

- Rucio is a Data Management tool born at ATLAS
<https://rucio.cern.ch/>
- Development by OSG is quite active (~10-20 commits/week)
- Declarative policy makes old and new data easy
- Reporting tools are robust and sophisticated

Interesting Reflections

- Globus was a good service, but we like open
- Rucio gives us a good replacement, not only for Globus but also for a significant chunk of JADE itself.
- Rucio integration is on-going.
Preliminary tests looks good.

Thank You ^_^

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