Data Abstraction October 2023 Status & Plans

Tim Jenness & K-T Lim &
Gregory Dubois-Felsmann

Retrospective (Aug-Oct 2023)

Pipeline Middleware

- Hired David Irving into middleware team. Full NOIRLab ops person based in Tucson.
- Released “Quantum Backed Butler” for batch execution as the default.
- Many cleanups to prepare for client/server future.
- Had Butler client/server design meeting in Princeton at start of October 2023.
- Supporting BPS development for htcondor and PanDA.

Build Engineering

- Received in-kind contribution from Australia (Arianna Ranabhat).
- Lost AWS Jenkins server. Migrated to USDF. Added a dev instance; improved some parts of the system in the process.
- Evaluated cloud-based Apple Silicon Macs as build hosts; ordered Chile-hosted Macs instead.

Release Engineering

- Matthias released pipelines v24.1, v25.0, and v25.0.1, and started on v26.

Data Engineering

- Hired Jeremy McCormick (SLAC) as Data Engineer.
  - Developed a long list of issues associated with the Felis tooling and the sdm_schemas data model

- Started a spreadsheet for dataset types to publish to data rights holders.
  - Trying to understand what incorrect assumptions may have been made by users because of the availability of “extra” dataset types in DP0.2

Construction Architecture

Requirements & Design

- DMTN-227 was updated to better describe the Consolidated Database.
- Transfer mechanisms for guider "postage stamps" and LSSTCam shutter motion profiles were decided.
- Refined the requirements and design of the Rucio/Butler "merge job" integration component.
- Refined DMTN-199 embargo diagram to reflect design changes.

Investigation & Coding

- A short-term plan was finalized to deliver initial components of the ConsDB, and initial code skeletons for two components were released along with sample outputs from one.
- The s3daemon for efficient image transfers was documented and released to the Camera Team.
- Wrote code for controlling automated transatlantic transfer of raw images via Rucio, demonstrating it with HSC data.

Future Plans to end 2023

Pipeline Middleware

- Butler client/server Minimal Viable Product
  - Read only: butler.get(), butler.query.datasets().
  - Deployed using standard SQuaRE tooling.

Pipeline Middleware Short Term Concerns

- Andy Salnikov leaves team end of December to work on APDB/PPDB.
  - Hoping he returns at some point. Without him there is only David and Tim working on client/server and general middleware support. (Jim's and Nate's effort is fluid).
- Scaling client/server for DP1. How many users? How many servers? How many postgres databases? Who is administering and deploying the postgres databases in the cloud? Is DP1 hybrid or like DP0.2?

Not in immediate plan:

- Provenance tracking. Depends on Jim Bosch deciding to work on it (this impacts DMS-REQ-0386: priority 1b).

Build Engineering

- Enable more pipelines to be run (at least partially) in dev instance.
- Automate deployment of dev and prod instances.
- Complete modernization of prod based on current dev.
- Fix disappearing worker problem (likely USDF networking).
- Improve pipelines.lsst.io build.
- Deploy Apple Silicon workers (USDF-based Linux workers are a possibility, but probably later).

Release Engineering

- Make v26 release

Data Engineering

- Jeremy, Gregory, and Tim will attend the IVOA Interop in Tucson. (Tim is LOC chair)
- We will significantly improve the felis and sdm_schema tooling
  - Will focus on two tracks: technical debt cleanup and delivering new features
  - Major goals:
    - Make it (much) easier to move what we learned from DP0.2 / .3 back into the reference data models (imsim and hsc)
    - Deploy new metadata-driven features in the data services and Portal

Construction Architecture

Requirements & Design

- Continue cross-subsystem architecture meetings.

Investigation & Coding

- Complete initial ConsDB components (flexible metadata insert, batch EFD exposure/view summarization, low-latency Summit exposure/visit tables).
- Continue assisting with Rucio/Butler integration.

Questions

Boundaries

- What is the boundary between Pipeline Middleware team and Rucio team?
  - Is Steve doing the ingest work for US Data Facility or Middleware team?
- What is the boundary between Pipeline Middleware team and embargo team?
- For hybrid cloud who is working on the caching we say we need when someone gets a dataset from USDF but we want them to prefer to read it from Google?
- Build Engineering and Release Engineering are currently deemed to be part of the Pipeline Middleware team (which makes no sense but can't be changed for another year).
- OODS and OCPS support is meant to transition to Data Abstraction team at some point.
- Alert Production Pipeline Infrastructure is nominally part of Data Abstraction once delivered but is it going to be supported by AP team as for construction?

Butler Client/server next steps

- Butler client/server Phase 2
  - Group/user management with signed URLs.
  - Missing query APIs (dimension records, dataIDs) and getURI.
  - Pagination of query responses.
- Butler client/server Phase 3:
  - Butler.put
  - Butler.transfer_from
  - Butler ingest
  - Butler import/export