# User Batch meeting-2022-07-13

## July 13 at 14:00 PT  on ls.st/wom

## Attendees

Open to all

## Discussion Items

https://dmtn-223.lsst.io/

Wil wants to know what do we need to declare we meet the requirements?  Does DMTN-202 introduce new requirements we can not afford?

- Primary conflict - feeling that community was "promissed ad hoc user batch "
  - could wrap any code in a pipeline task which grabs from butler puts fits on disk and runs the user code.
  - it does not have to python if the binary is in the container.
- Colin would seperate out bullet 4 from 223.
  - for provenance you have to use pipeline frame work (Tim agrees)
  - if you want to run over bunch of files just use slurm - no provenance and stuff
  - Richard PanDA can do this and things like simulation (that will use all the processing).
  - All user batch must go through TAC .. this leads to better support in our framework - they need SLAC account etc. Frossie - small birthright is not useful. So all you get be default is notebook.
- KT arbitrary code requires arbitrary access to data
  - we could force them to wrap in pipeline task
  - Frossie points out we do not have support for arbitrary stuff - should prototype pipeline task as a wrapper.
  - staff tools may not be portable to external users
  - provide fund for googel resource ..
  - bulk export to another center - fairly rigid
  - forcing to pipeline task does not save us much in terms of the arbitrary code and authentication ...
- SRP - whats the requirement - only one seems to be user provided code
  - lots of problems with user provided code - more restrictive better ..
  - Support is actually not that high for big stuff - the users do this. (May not be the case for first timers from undrepresented institutes)
    - the documentation is really really good though ..
- Frossie - TAC will decide what the user can do even simulations.
  - need to separate what allows us return the requirements
  - impression on the community (the slide decks of promises)
  - should go for the pipeline wraper ..
  - would be great to get the data to TACC or some place.
  - 10% and relative breakdown .. not fungible in hybrid model .
  - We provide adhoc arbitrary RSP type acess birtghright
    - user batch has to be in our DRP framework AT SLAC
    - access to outputs should be provided

- [x] run sextractor inside pipeline task 21 Sep 2022 Wil O'Mullane ?
- [x] Present PST to agree on User Batch uses data holdings and pipeline task ... 28 Sep 2022 Wil O'Mullane

Tim: For image processing increasing levels of difficulty:

- Users submit pipelines with bps using Rubin code but their own user queries and pipeline configuration. Essentially trivial.
- Users write their own PipelineTask and want it included in a pipeline submitted with bps. How do they create a container? Does their code need to be audited? If they have a container that includes all our code plus theirs then this is fairly straightforward. Products would be written using butler to wherever we want them to go and made available like any other butler products.
- Users want to use their own code that is not using butler. Their own graph builder? Writing to a per-user file system at USDF? Using butler retrieve-artifacts to get the files? They provide a container that includes butler code and their own binaries? How are those files visible to the CloudDF? Do they write to their VOSpace? This third option is clearly the hardest.

KT questions:

1. If we use BPS, are we forcing people to write PipelineTasks? Should we demonstrate how to wrap Astromatic in such? Can people use IVOA interfaces instead of Butler?
   a. DMTN-202 1.1 says "community-standard bulk-numeric-data-processing frameworks"
   b. DMTN-202 1.2 says "Others will wish to use community image processing tools"
2. If batch is at USDF, how is it authenticated/authorized? Long-lived submission token? Login-session-length token? Does BPS or the underlying workflow system need to understand refresh tokens?
3. If batch is at USDF do we still use the Butler server in CloudDF or is there something local? How is authentication/authorization for the Butler handled from batch jobs?
4. Should we mention GPUs? (At CloudDF only?)

5. Is single-ADQL-query use of external-catalog matches and data sufficiently motivating to bring them into Qserv, or are identifier lists for remote, user-programmed joins adequate?
6. The Resource Allocation Committee should determine which external catalogs deserve space at USDF. Should they define scientifically-useful cross-match processing?
7. The document (in §4.3) still seems to say Qserv is meeting all catalog processing requirements.
8. 9% (81 of 892) of NCSA home directories had more than 10 GB on 2022-02-07.

SRP: users are going to user. If they have an alternate way of doing something, which they can work around to access it, they will.

- IS this a problem - you are on your own for support at this point.

Richard:

- is there any requirement for arbitrary batch use? eg running simulations? This would speak to direct access to a batch system.
- is all batch provided by the USDF? We have not budgeted for any/much in the cloud.

## PCW - LINCC https://project.lsst.org/meetings/rubin2022/program/agenda?field_day_value=04

Need some DM people at this - someone for light curves, Science Platform and User batch.