User batch - possibilities and plans.

William O'Mullane

2023-07-10

1 Introduction

We have some requirements on user batch as discussed in section 2. We also have certain resources available at USDF and development effort as discussed in section 3. DMTN-202 gives some desirable use cases.

In this note we rendezvous all of this information and clarify what DM could and will do during construction.

2 DM requirements on user batch

The DM requirements LSE-61 which flow down from the System level requirements [LPM-17; LSE-29; LSE-30], has a few requirements which require some sort of user driven processing. Many of these pertain to *user generated data products*[1]. In the era of pre-operations many of these user generated products will be created and may exist at IDACs. The a broad summary of requirements pertaining to user batch and more broadly to the topic of user generated products and services is covered in Level 3 Definition and Traceabilityi on confluence.

The requirements most pertinent to user batch from LSE-61 are :

- DMS-REQ-0119: DAC resource allocation for Level 3 processing
- DMS-REQ-0121: Provenance for Level 3 processing at DACs
- DMS-REQ-0125: Software framework for Level 3 catalog processing
- DMS-REQ-0128: Software framework for Level 3 image processing
- DMS-REQ-0123: Access to input catalogs for DAC-based Level 3 processing

[1] Previously known as Level 3 data products.

- DMS-REQ-0127: Access to input images for DAC-based Level 3 processing
- DMS-REQ-0124: Federation with external catalogs
- DMS-REQ-0106: Coadded Image Provenance

A set of potential use cases for user batch have been proposed in DMTN-202. This document also provided a summary list which is not dissimilar to the requirements above:

1. The user computing capability should allow running in bulk over catalog data.
2. The user computing capability should allow running in bulk over image data.
3. The system capacity is defined as an "amount of computing capacity equivalent to at least userComputingFraction (10%) of the total LSST data processing capacity (computing and storage) for the purpose of scientific analysis of LSST data and the production of Level 3 Data Products by external users".
4. We have to provide a software framework to facilitate both catalog- and image-based user computation, which has to support systematic runs over collections of data and has to preserve provenance.
5. The framework(s) has/have to support re-running standard computations from the pipelines in addition to running more free-form user jobs.
6. There has to be a resource allocation mechanism to allow users to be given quotas, which can be modified per-user. The association of quotas with defined groups of users (e.g., ad-hoc collaborations and/or formal Science Collaborations) would be a useful further capability.

3 DM resources for user batch

The requirements and this allocation for this are not very large. At the time of requirement writing Jupyter nor the notion of a Science Platform existed. The provision of the RSP has to be done within the allocation of user compute.

3.1 Compute

As pointed out in DMTN-202 the DM requirements stipulate at least 10% of compute for user batch. This is included in the USDF sizing model DMTN-135. Looking at table 21 in DMTN-135 we see a total cores in USDF at around 15K. This includes 10% for users so around 1.3K for user. The RSP assuming a peak of 500 simultaneous users would need 250 or so of those. User batch would therefore have about 1K cores depending on load. This implies the need for a controlled sharing mechanism (subsection 3.3).

It would also be excellent to pursue a research project with an NSF facility (e.g. TACC) into what could be done with Rubin data at such a facility in terms of arbitrary compute. TACC have expressed interest in this.

3.2 Storage

Table 40 of DMTN-135 list all datasets used for sizing, any user batch processing would have to fit in the science user home space. There is a margin for other data sets as should be clear from section 2 agreed user defined products will have to be stored at, and served, from the USDF. No requirement was laid down as to what data volume should be allocated. The sizing model covers about 10% (as stipulated for compute) in the Other/Misc data set provision.

3.3 Quotas and allocation

As users come online in RSP we will have to enforce quotas.

Nominally a notebook user requires a virtual core though a 0.5 slice is actually sufficient in most cases. In subsection 3.1 we suggested 250 cores for notebook users with a peak of 500. If we assume the average is more like 100 users we could allow a few more cores per user for e.g. DASK execution. If we assume that only fraction, say 10%, of users would run DASK we could allocate give a 10 core quota for users. It would have to be a zero sum game though not exceeding the total allocation more than briefly. If were not using the batch allocation we could use more of that on DASK type tasks. Here we will clearly have to see what users do and want.

Restriction of the core per Kubernetes pod and hence per user is doable. This however needs

more work.

In operations we propose catalog processing with Dask/Spark goes to Google; any processing (images or, less often, catalogs) done with BPS goes to USDF batch.

We will send all batch jobs to USDF. We have some capacity at USDF for this and potentially some spare cycles in the shared data facility.

DMTN-135 table 40 allocates 2PB in year 1 growing to 9PB in year 10 for science user home space. This is still driven by the original requirement to support 7500 users. Table 40 assumes 5K users in year 1 yielding a limit of 40GB per user. Of course usage will not be even so we could set the limit higher like 60GB and monitor what happens. On SDSS and BaBar the quota was kept low and justified requests to get more were accepted - most people kept within the quota by keeping their user space. Its not obvious what a good number is here but perhaps we should start the RSP users on 10GB quota.

KT points out on NCSA "9% (81 of 892) of NCSA home directories had more than 10 GB on 2022-02-07." This would indicate starting allocation on RSP of 10GB may work for 90% of users - we can then increase on request for some users.

Power users who would be some of that other 10% may be more likely to seek batch production at SLAC in any case.

It is not clear we have the technology in place to enforce disk quotas yet.

3.3.1 Peak times

There are going to be known events such as a Data Release which we know will put pressure on the system. For these event we would want to scale back to basic quotas like 0.5 core per user and potentially disabling DASK. We should also experiment with paying for the surge and see if subsequent lulls even out the bill. The latter approach would allow us to maintain a good service level and would be a more equitable approach since it would not penalize the users who have no other form of access.

3.3.2 Unused home space

Space on cloud means cost. If a user is not using their space for long period it should be migrated to cheaper storage. Along period could be a year but six months seems long enough. Again we have no mechanism for this - Richard even suggested taking it to USDF but cold storage on google might work.

3.4 Other sites

SLAC have some potential cores available on site to give more batch possibilities.

Some science collaborations, most notably DESC, will have their own resources at some center (NERSC for DESC). In the DESC case there are good links between NERSC and SLAC making this easy to manage.

No NSF facility is currently planned to hold the data. TACC are interested to experiment with us.

These sites are true batch sites with Slurm workload management. Users would submit jobs with BPS allowing it to figure out if the Parsl, PanDA for job submission on the back end.

4 Proposal

Let us assume the resources available are fairly fixed in line with section 3. Hence we will need a resource allocation committee subsection 4.4 to arbitrate requests.

We promised some form of user batch which it seems likely will be needed at least initially as many users understand batch processing of images say. By batch here we should be clear we mean a BPS command line interface to something like PanDA. subsection 4.3

These two items with the RSP fulfill our requirements. RSP covers a bunch of other requirements on queries and image access subsection 4.1.

4.1 Science Platform

The RSP is now well established and is following the vision LSE-319. Data Preview 0 has introduced an early version of this to the community.

DMTN-202 point 4 (section 2) in respect of catalogs we now propose to be satisfied by some form of Dask/Spark access to the catalogs as spatially sharded Parquet files. The previous baseline design for this was a layer on top of Qserv, but in practice we do not have this "next to the database processing" implemented, and the RSP and Data Access teams feel that, with the advance of data-science software, a Parquet-based solution is now a good solution. There is some effort available in FY24 to prototype and implement this. We shall consider carefully what the minimum system we need here is to cover our requirements.

We will not at this point promise extensive Dask/Spark like services. The minimum needed to meet requirements will be done on construction and we shall work further on it in the background with LINCC. We all agree this will be scientifically useful, but we need to finish construction as a priority and this is not a requirement. We will work with LINCC on it certainly and something will be available but we are not promising this and we are not accepting requirements on it.

The standard/default allocation for any data rights user will be RSP access. Users with bulk needs will use BPS (subsection 4.3) at a Data Facility.

4.2 Butler Gen3

Whether Batch or RSP a butler is available with provenance information which should cover DMS-REQ-0106 and DMS-REQ-0121. If users do not use the Butler they will not have provenance.

4.3 BPS Batch

Our baseline is user batch based on Gen3 Task framework. Users needing batch will need to go via the Resource Allocation Committee (subsection 4.4) and will be given SLAC accounts to run user Batch jobs at SLAC. Those who do not wish to use the Rubin pipelines could be facilitated see subsubsection 4.3.2.

We set DP0.2 as an initial test of seeing how we could work with PanDA RTN-013 for DRP. Potentially it could also be used for Prompt Processing and user batch processing. In fact we use the BPS front end to submit jobs and middleware which to keep BPS working with at least on other back end as well as PanDA.

It seems then the minimum promise to our batch users would be a BPS front end for job submission. This would make it look the same as any of our internal processes. In the end if we use Condor, Slurm, the submission would be BPS. The job tracking of course would depend on the back end. We should not necessarily commit to the same back end in all locations - hence USDF may be PanDA but on Google or an processing IDAC we may have a different back end.

This BPS fronted system would meet DMTN-202 points 1-5 (section 2). The Point 4 on catalogs can be met with Qserv. This would also cover DMS-REQ-0119, DMS-REQ-0125, DMS-REQ-0128, DMS-REQ-0123, DMS-REQ-0127.

4.3.1 Interactive vs Batch

User will need to consider when they should switch to batch. This will largely be a function how long the user is willing to wait,' how big their job is and how repeatable it is. Processing say a full focal plane image on RSP with interactive processing via DASK should take some minutes and should be fine. Should one wish to process ten focal planes (hence > 1K images) it may become tedious and be more suited to batch. That of course means one is not tweaking the code each time. We will have to settle on a guide number for where switching to batch makes more sense but between 500 and 1K images would seem to be the point where a user is falling more in the batch realm. Also if one is processing 1K images it will probably soon be 2 or 3K or more.

4.3.2 Non Rubin code

Users may wish to run arbitrary code on Rubin images. It should be possible to provide a butler recipe to get an image or set of images to the users space and allow them to run any code which takes FITS images. Done in a container this could be done for a large volume of images - it would have to get past the resource allocation committee of course. This would require custom work by the user and would not be a general facility. Run in this mode little

provenance would be kept. (See subsection 4.2)

Some provenance could be garnered by ingesting the generated files which could at least give the execution graph of their creation. If users wish only to write files they would have to go in their user space or potentially a VOSpace end point - the user would need to organize and track such files themselves.

4.4 Resource Allocation Committee

As section 3 points out we will need an arbitration committee of some sort. It seems this should also favor individuals from under represented and under resourced groups though that is being proposed here for the first time. We note also proposals committing to releasing Level 3 data products that are made available to all data rights holders will also be allowed to be given a preference by the RAC. The committee is called out in the operations plan [RDO-018] Section 12.2 but we have as yet not created a charge for it. This committee would at least be responsible for creating, implementing and updating policies on at least:

- Requests for storage over the default allocation, either temporary or permanent.
- Requests for large batch compute accounts at SLAC.
- Requests for extra batch compute cycles on Google.
- Bringing compute from another Google funded account to operate on the Rubin Object Store Cache or Science Platform.
- Cache Optimization on Google wrt. which data sets may live entirely on Google (e.g. coadds).
- ...

This would meet DMTN-202 point 6 (section 2).

4.5 Other catalogs

DMS-REQ-0124: Federation with external catalogs, is a bit open ended. Theoretically we already meet this with IVOA protocols since we could match to any IVOA service. DAX will pro-

vide a small user catalog match to Qserv. Other catalogs could be loaded and matched however getting a list has proved fairly inconclusive. LINCC may again help here by coordinating IDACs to provide neighbor tables/services for other catalogs. One notion would be to have a the Object catalog at IDACs matched to their local holdings. We could consider migrating some of those neighbor catalogs back to USDF Qserv. Gaia will be provided as part of the processing - there is no requirement listing catalogs so from a verification perspective we can put this on IDACs

A References

[RDO-018], Blum, R., 2021, *PLAN for the OPERATIONS of the VERA C. RUBIN OBSERVATORY*, RDO-018, URL https://docushare.lsstcorp.org/docushare/dsweb/Get/RDO-18

[LSE-29], Claver, C.F., The LSST Systems Engineering Integrated Project Team, 2017, *LSST System Requirements (LSR)*, LSE-29, URL https://ls.st/LSE-29

[LSE-30], Claver, C.F., The LSST Systems Engineering Integrated Project Team, 2018, *Observatory System Specifications (OSS)*, LSE-30, URL https://ls.st/LSE-30

[DMTN-202], Dubois-Felsmann, G., 2021, *Use cases and science requirements on a user batch facility*, DMTN-202, URL https://dmtn-202.lsst.io/,
Vera C. Rubin Observatory Data Management Technical Note

[LSE-61], Dubois-Felsmann, G., Jenness, T., 2019, *Data Management System (DMS) Requirements*, LSE-61, URL https://lse-61.lsst.io/,
Vera C. Rubin Observatory

[LPM-17], Ivezić, Ž., The LSST Science Collaboration, 2018, *LSST Science Requirements Document*, LPM-17, URL https://ls.st/LPM-17

[LSE-319], Jurić, M., Ciardi, D., Dubois-Felsmann, G., Guy, L., 2019, *LSST Science Platform Vision Document*, LSE-319, URL https://lse-319.lsst.io/,
Vera C. Rubin Observatory

[RTN-013], O'Mullane, W., Dubois, R., Chiang, H.F., 2022, *Near term workflow for pre-operations with PanDA*, RTN-013, URL https://rtn-013.lsst.io/,
Vera C. Rubin Observatory Technical Note

[DMTN-135], O'Mullane, W., Dubois, R., Butler, M., Lim, K.T., 2023, *DM sizing model and cost plan for construction and operations.*, DMTN-135, URL https://dmtn-135.lsst.io/, Vera C. Rubin Observatory Data Management Technical Note

B Acronyms

Acronym	Description
B	Byte (8 bit)
BPS	Batch Production Service
DAC	Data Access Center
DAX	Data Access Services
DESC	Dark Energy Science Collaboration
DM	Data Management
DMS	Data Management Subsystem
DMS-REQ	Data Management System Requirements prefix
DMTN	DM Technical Note
DP0	Data Preview 0
DRP	Data Release Production
FITS	Flexible Image Transport System
FY24	Financial Year 24
GB	Gigabyte
IDAC	Independent Data Access Center
IVOA	International Virtual-Observatory Alliance
LINCC	LSST Interdisciplinary Network for Collaboration and Computing
LPM	LSST Project Management (Document Handle)
LSE	LSST Systems Engineering (Document Handle)
LSST	Legacy Survey of Space and Time (formerly Large Synoptic Survey Telescope)
NCSA	National Center for Supercomputing Applications
NERSC	National Energy Research Scientific Computing Center
NSF	National Science Foundation
PanDA	Production ANd Distributed Analysis system
Parsl	Parallel Scripting Library http://parsl-project.org/
RAC	Resource Allocation Committee

RDO	Rubin Directors Office
RSP	Rubin Science Platform
RTN	Rubin Technical Note
SDSS	Sloan Digital Sky Survey
SLAC	SLAC National Accelerator Laboratory
TACC	Texas Advanced Computing Center
USDF	United States Data Facility