LSST DESC Needs for USDF to NERSC Data Transfers

This is a draft document that will eventually become a DESC note or Rubin technote

Version, date

v0.9 May 21 2025

Authors

DESC and Rubin people

Purpose

In this document, we provide a compilation of the data sets and data transfer methods that the Dark Energy Science Collaboration (DESC) will need for its analyses at NERSC, and possibly other sites, at each data release.

References

RTN-011 v6.1: Rubin Plans for an Early Science Program
Some Rubin documentation on data releases, DPDD, etc
Any relevant DESC documentation, DESC Computing Model, etc

Introduction

The Vera Rubin Observatory will process LSST data at the US Data Facility (USDF) at SLAC. Given that computing resources for the scientific community will be limited at the Rubin Data Access Centers (DACs), DESC decided early on to adopt a model by which it would apply annually for computing and storage allocations for its work at NERSC. This model implies that a non-trivial fraction of Rubin data will need to be transferred to NERSC from the USDF. In this note we detail the specific data set transfers required at different stages:

- Data Preview 1, which will include ComCam data taken over a few selected fields, but including full coadd information over part of it (a few square degrees).
- Data Preview 2, in which a substantial amount of commissioning and science verification and validation data will be released, with a certain expectation that many of the final processes (coadds, photo-zs) will be run on it by DESC, even if the data set sizes are not representative of a full data release.

- Data Release 2, corresponding to one year of LSST data, is deemed as the first globally relevant release for the DESC for cosmology analyses, due to the extent and expected uniformity of the data set (for static probes).

NOTE: that all of the relevant data sets for transfer being considered here are for the static analyses (i.e., those based on data releases), as the time domain science will be based on alerts and data being processed via the the FAST database (ref).

In every instance, we have divided the data sets into *catalogs*, *images* and *ancillary information*.

Data Preview 1

Given that the size of DP1 is relatively small (~25 TB, very probably <5 TB of non-intermediate data), we plan to transfer all of it to NERSC, even PVIs (processed visit images). The storage requirements come from an assessment of an early DP1 run (by Jim Chiang) who compiled the results into a parquet file listing the total sizes of each dataset after each step. Object catalogs and coadds are <0.2 TB each.

Plan as of February 2025 is to transfer the complete DP1 from USDF to NERSC at a date TBD, starting at some point after the First Look event (planned for June 2025) when Rubin releases the data for data rights holders, or before that if it is deemed technically feasible and desirable. We will use Globus for this transfer as opposed to Rucio (the tool used by Rubin itself and the IDACs). Plans for earlier than public release transfer are in discussion but would require adequate data restrictions. One option is to funnel the data into a UNIX protected directory (against r/w/x so not even listing). DESC has a special very limited UNIX group of < 5 people including Computing Coordinators, SLAC liaison and DESC Operations personnel. This would be obscured further with soft measures such as a low profile for the data transfer (no announcements), a very limited list of people to be able to access, and some directory name opaqueness. This would be a one-off strategy, policy to be discussed again with LSSTCam data (DP2 onwards).

Data Preview 2

In the case of DP2 (projected for early 2026), we would like to test a more realistic data transfer with Rucio over FTS, similar to what IDACs will be doing after each release. This would allow NERSC and the DESC to learn from the IDACs experience and get support from NERSC to implement Rucio, and ramp up Computing personnel expertise in the DESC.

The following items have been collected through explicit information gathering across the DESC. Numbers come from Data Preview 0.2, which provides a rough approximation (300 square degrees at 5 years depth).

Catalogs

Description	Butler collection name	Approximate size	Number of files and format
Coadd object catalogs	objectTable	1 TB	O(200) Parquet files
PSF star catalogs	TBD	<< 1 TB	TBD

Notes:

- There is an uncertainty here regarding format and a possible separate table for Rubin photoz table. Alternatively, DESC photozs for DP2 will be DESC run at NERSC or elsewhere (e.g. IDACs).
- There is an uncertainty about whether metadetect (shear) catalogs will be made available at this time, effectively multiplying by 5 these estimates.
- Possible smaller external catalogs that Rubin validation will use (to avoid the hassle of downloading them or looking them up, format TBD)

Images

Description	Butler collection name	Approximate size	Number of files and format
Astrometric solution for PVIs	TBD	<< 0.1 TB	TBD
PSF (Piff) models for PVIs	TBD	few MB on their own, but bundled with the PVIs	Same as number of PVIs
Cell based coadd images: WFD and DDF	deepCoadd_calexp (TBD new name for cell based coadds)	O(100 TB) TBD	TBD
Small number of raw images for reprocessing experimentation		O(10 TB)	TBD

Notes:

- Astrometric solutions are WCS solutions for every PVI..

- We will likely need to work with Rubin to extract the astrometric solutions and Piff model for each PVI into a separate data structure suitable for transfer, querying, etc. since these items may be stored in the same file as the PVI images. Further, transfers are large numbers of small files will be very inefficient.
- WFD and DDF will have a different meaning than in Operations, in this case corresponding to a hypothetical wide and shallow science validation survey, plus some field(s) in which 10+ year depth is achieved. TBD how (if) these will be differentiated at butler level.
- A small amount of raw images could be requested, but this could be small and transferred on an ad-hoc basis (reprocessing experimentation), and associated visit level metadata.

Ancillary information

Description	**Butler collection name**	**Approximate size**	**Number of files and format**
Survey property maps	collection name in Butler TBD (see Jim's comment for some examples)	< 0.1 TB	O(100) in healsparse format

Notes:

- A survey mask could be included in the above as well.

Data Release 1 and onwards

TBD