Vera C. Rubin Observatory
Data Management

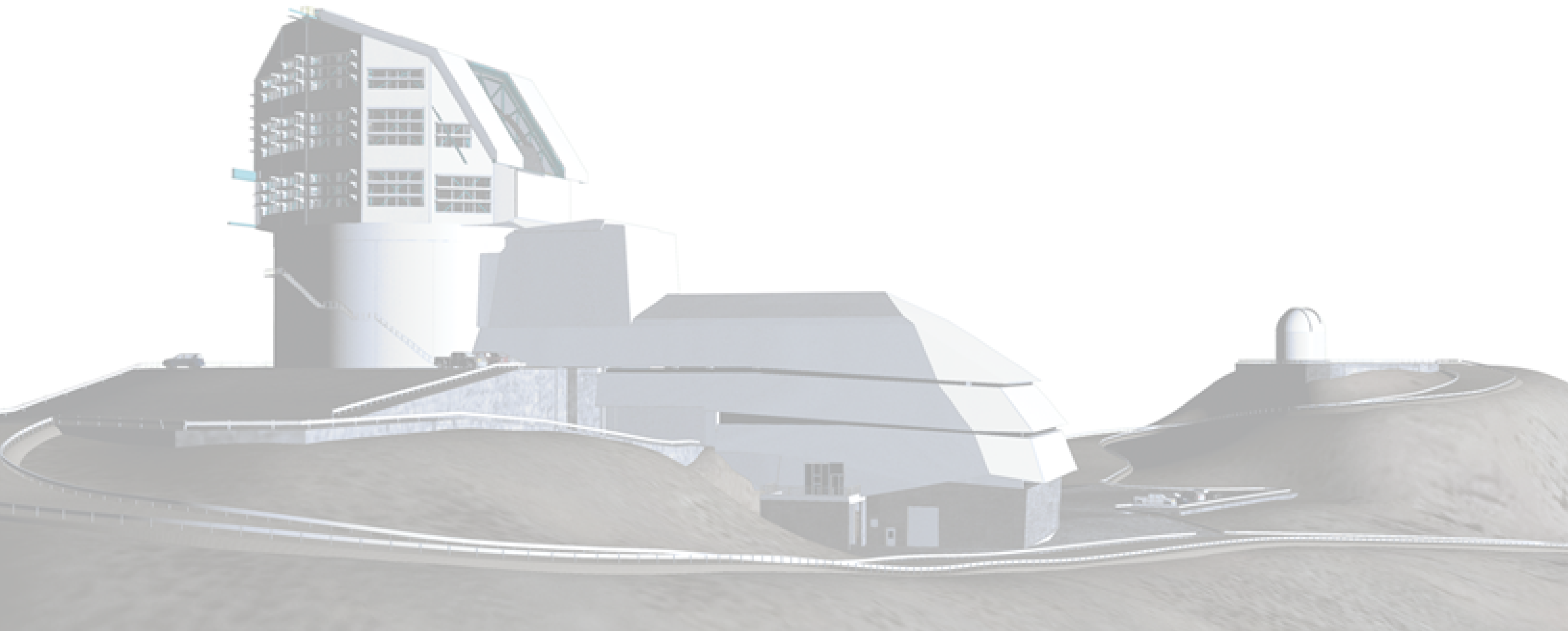
Data Facilities Transition Plan

Richard Dubois, William O'Mullane

RTN-021

Latest Revision: 2023-07-10

DRAFT



Abstract

This document outlines the plan to get the USDF fucntional for operations.

Change Record

Version	Date	Description	Owner name
1	YYYY-MM-DD	Unreleased.	William O'Mullane

Document source location: `https://github.com/lsst/rtn-021`

Contents

Data Facilities Transition Plan

1 Introduction

This document describes the preparations that are required to be ready to process telescope data into science-ready products in Operations and in the run-up to operations (e.g., during Commissioning). This is commonly referred to as Data Release Processing.

The document covers a period of transition, for the telescope, from Construction into Operations. While on the high-level Project roadmap the move from Construction to Operations is reasonably discrete, in practice it is a relatively complicated with a need to seamlessly maintain operational capabilities that have been implemented during Construction – such as, for the processing of telescope data.

Data Release Processing is intended to be completed across three independent Data Facilities:

- The United States Data Facility (USDF), hosted at the SLAC National Accelerator Laboratory in California.
- The French Data Facility, at the Institut national de physique nucléaire et de physique des particules (IN2P3) in Lyon.
- The UK Data Facility hosted by the IRIS distributed infrastructure.

These Data Facilities are in different states of preparation, and also have different responsibilities during – and in the run-up to – Operations.

In the next section, we explain in high-level terms the key elements of DRP and the strategy that has been adopted for addressing them.

Then, in Section ???, we describe the plan for identifying, testing, and incorporating the technologies that are needed to support a three-facility DRP, and define the timeline on which this plan needs to be completed.

In Section ???, we consider the preparations that are in place at each, individual Data Facility —

typically within the control of local staff — which are required to ensure each Facility achieves various stages of technology and infrastructure readiness, as the Observatory scales up to full operations.

2 Motivation

The Legacy Survey of Space and Time (LSST) will be delivered to the science community as a collection of products, including a (typically) annual update to the survey, called a Data Release (DR) and a pseudo-realtime alert stream. Some or all of the three Data Facilities will play a role in the delivery of both of these products. To this end, this document considers:

- the preparation and publication of Data Releases (DRP)
- the reception of the data from the telescope,
- the preparation and dissemination of the alert stream

All three data facilities are involved in the first of these (DRP), though only the USDF is involved in the other elements.

2.1 Evolution from Construction to Commissioning to Operations

The Construction of Rubin is winding down and Commissioning of systems is underway. Support of Commissioning will be a function of the USDF: transfer of image, calibration and time-series data from the summit, as well as support of the Rubin staff and commissioners in the processing of the data. Along with the commissioning of the telescope, this process will also act as a pathfinder for activities needed in operations.

2.2 Data Reception and Alert Processing

There is a tight time budget for delivering data to the USDF and generating alerts: the goal is to issue alerts within one minute of closing the shutter. Parallel transfers will be required for each visit, broken down by CCD, and will be encrypted. A prompt processing pipeline will be executed upon receipt of the data, resulting in the initial alert stream. Some seven brokers are anticipated to be receiving the stream, in addition to sending data to the Minor Planet Center.

2.3 Data Release Processing

Each Data Release is a substantial, multi-Petabyte resource consisting of various science-ready datasets:

Processed Visit Image Processed versions of telescope images, corrected for instrumental signatures, etc.

Deep Coadds Stacked images of the same region of the sky, to create an image of increased sensitivity and detail.

science catalogues Collections of metadata that document detected astronomical objects along with standard measurements on those objects (location, colour, flux, etc.).

Ancillary and intermediate products Additional outputs that support different science use cases with different requirements, and enable downstream processing and the generation of derived datasets.

The preparation of each DR is a substantial piece of work involving significant computational and storage infrastructure. It is a multi-stage process in which the LSST Science Pipeline (sometimes called, the LSST Stack) progresses through a *campaign* of image-manipulation and analysis processes. There are science-relevant decisions needed for many of these processes, meaning there may be a need to vary the configuration (branches in the pipeline) used at specific points, to support different science use cases. Thus, in addition to the end products noted above, it is expected that (at least) some of the intermediate products will be preserved (or should be readily reproducible).

DRP progress is recorded in a database registry called the Data Butler. This is intended to track the locations, provenance and contents of datasets as they are processed. It is heavily used by the LSST Stack and, on the completion of a DR, becomes the primary metadata enabling science users to interact with the survey images and coadds.

The Operations Plan (** ref **) details how the resources required for each DRP will be contributed by three facilities. The overall responsibility for DRP is with the Rubin Observatory.

They are responsible for developing the pipeline software, selecting appropriate middleware on which to deploy the processing workflow, and for undertaking quality assurance on the output data products. The UK (and French) Data Facilities are responsible for completing their agreed share of the workload, as follows:

- The US DF will complete 25% of the processing
- The French DF will complete 50% of the processing
- The UK DF will complete 25% of the processing

At a high-level, DRP involves the following workflow:

- Raw images, captured at the telescope, are transferred to the USDF over a dedicated network link, with very low latency.
- The portion of the raw images to be processed in France and the UK are then transmitted (on a timescale to be determined) on to those facilities, over the public Internet, along with calibration images and other ancillary products.
- Once each facility has the raw image data it is to process, it may proceed with processing (possibly after some form of data validation, or similar). Processing can effectively be completed independently at each facility, though at several stages, there will be a need to exchange intermediate products between Data Facilities. Further, there will be a requirement for the USDF to have an overview of fine-grain progress at each facility. Finally, there should be a capability to reassign processing work on the granularity of days/ weeks, in response to processing problems.
- Once processed, output data is transferred back to the USDF (again, over the public Internet) to be assembled into a Data Release.
- Processed data needs to be validated (by Rubin Observatory staff) before it is confirmed to be ready for publication. It is still to be determined, at the time of writing, if validation work will be undertaken when data is returned to the USDF or can be done earlier, at each data facility.

A Data Facility cannot work completely independently on their raw data and initial calibration data alone. Some processing steps aggregate data from across the DR. Because these steps are not necessarily at the end of the processing, and for other reasons, some data products (which may be part of the final DR or may be intermediates) will need to be distributed between Data Facilities during a campaign.

- It is also intended that a full copy of all DR-related output data will be held at the French DF. This may serve as an online replica, for disaster recovery, though that may/ may not be the primary motivation for doing this.
- The UK DF requires a full copy of the output data from a DRP campaign, so it can serve science-ready data to a subset of the Rubin science community from a UK-based Independent Data Access Center. It is not anticipated, however, that the UK DF will need or want all the raw images: just those images that are to be processed in the UK.

It is intended that each DRP campaign will reprocess all images to date (that is, going back to the beginning of the survey). This is required to ensure consistent processing of all images (the LSST Stack will is likely to change as the survey progresses), and means that the volume of DRP-related computing, storage and data transfers will grow year on year.

The preparation of a Data Release – that is, a DRP campaign – should be completed within twelve months of the end of observing period (six month for DR1, which will be based on the first six months of observations). This includes assembly of complete datasets in both the US DF and French DF. It may not include the distribution of products to Independent Data Access Centers, though that should also be done in a timely manner, so as not to delay the release of a DR to the science community.

2.4 Timeline

The timeline for setting up the three-DF infrastructure and operation is driven by the requirements of Commissioning and early operations, which are documented in **RDO-011**. That document describes the production of a number of pre-operational and early-operations data products, including three Data Previews (denoted DP0.2, DP1, and DP2) and data releases (denoted by DR1, DR2, and DR3).

The Data Previews (and early Data Releases) serve a number of purposes – for example, testing the LSST Stack and informing science users of LSST capabilities. They are also a convenient framework on which to build three-DF capabilities.

At the time of writing, RDO-011 was last updated in April 2022. However, a newer timeline was presented at the Project and Community Workshop in August 2022, and RDO-011 will be updated to reflect this update in the near future:

- **AuxTel/LATISS (in operation since 2021)** – single CCD camera run in imagine or spectroscopic mode for measuring atmospheric effects. It has been used as a pathfinder for several systems.
- **DP0.2 (released June 2022)** – a DRP experiment to reprocess data from DESC DC2 [REF] using the current LSST Pipeline software. This DRP experiment only involves a single data facility (the Interim Data Facility on Google Cloud), but is a vital source of information for the three-site DRP capability, as it exercises the fundamental elements of DRP.
- **3DF Rehearsal - Spring 2023** – a ramped re-run of the DRP workflow used to create DP0.2, though completed across the three Data Facilities, rather than on the IDF. It will be worthwhile to compare the properties of the DP0.2 and rehearsal campaigns, considering efficiency, resilience, and effectiveness, for example.
- **ComCam commissioning - engineering first light (July 2023)** Nine-CCD camera used as a pathfinder for the full system. Commissioning of the device will start 6 months earlier (Jan 2023).
- **DP1 (April–June 2024)** – data from the Commissioning Camera (ComCam), installed at the Observatory, will be processed twice: first, at SLAC (???) (Construction DF); and, then, across the three Data Facilities.
- **LSSTCam commissioning - engineering first light (Mar 2024)** Commissioning of the full system. Commissioning of the device will start 6 months earlier (Sep 2023).
- **DP2 (December 2024–March 2025)** – Commissioning Data from the full LSST Camera, installed at the Observatory, will be processed across the three Data Facilities, aiming to mimic the conditions and timeline of a production DRP campaign as closely as possible.
- **DR1 (October 2025–January 2026)** – The first production release, containing data from the first six months of observing.
- **DR2 (July–October 2026)** – The second production release, containing data from the first twelve months of observing.

- **DR3 (July–October 2027)** – The second production release, containing data from the first twelve months of observing.

This timeline is used for planning throughout this document.

3 Organisation

The infrastructure group is within Data Production lead by Richard Dubois, there are a number of teams and leads withing this with given responsibilities. These teams are:

- Leadership (Richard Dubois): Provide technical and project management for the group.
- Data Curation (Brandon White): Maintain Rucio based data backbone system for support of data transfers and retention; data backbone functions, such as data parity, replication and Rucio management of the backbone; bulk downloads and LFA destination.
- Advanced Databases (Fritz Mueller): Provide database services on top of database hardware provisioning for numerous databases for Chile + the US Data Facility (DF). Includes: cassandra, Qserv, user databases, EFD, A&A, ETL, workflow management, butler. Provide database admin functions such as schema evolution, backup and replication.
- Wide Area Networking: (Phil Demar): Ongoing collaborative network architecture to support evolving networks and to sustain a monitoring interface appropriate to Rubin Observatory operations, from BDC in Chile to the US DF and across the world to other data centers.
- Processing Execution (Hsin-Fang Chiang): Evolve and verify the data release and prompt scientific processing pipelines. Data release processing is applied to batches of multiple visit images grouped together, in contrast to prompt processing, which is applied to each visit as it is received from the telescope. Data release processing includes annual data releases (including single frame processing, coadd generation, coadd processing, object characterization, and annual solar system processing). It also includes periodic regeneration of templates for use in the prompt processing system.
- Infrastructure (TBD): Provide configurable hardware upon which are layered the required science platform, prompt processing and Data Release Production (DRP) pipelines, including compute nodes, various performance levels of nearline storage and tape backup;

account management related to A&A; data transfer capabilities to the summit, other Data Facilities, IDACs and users; batch processing capabilities. Provide performance monitoring, uptime and usage statistics of the Data Facilities. Enforce network security.

- Alert Vetting System (Michael Schneider): Maintains the Alert Vetting System (code and processes), and monitors it during operations. The Alert Vetting System (AVS) monitors alerts (of satellite nature) and may embargo some. It may also put a hold on specific Charge-Coupled Device (CCD) images. This is supplied by LLNL with oversight for integration from the Rubin Data Production(Obsolete use Rubin Data Management (RDM)) (RDP) A&P alerts team.

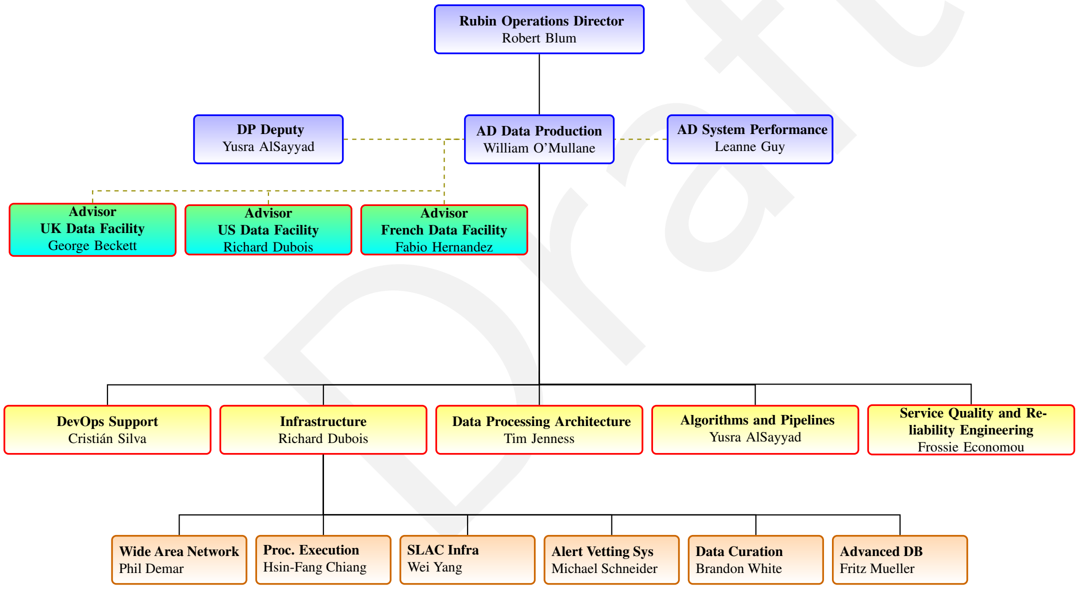


FIGURE 1: Rubin Data Production Org Chart from O'Mullane (RTN-001)

4 Implementation

4.1 Commissioning of AuxTel, ComCam and LSSTCam

During the commissioning phase, the USDF will be the primary resource for personnel to interact with data coming off the summit. This is to minimize connections to the summit and ensure that summit computing resources are reserved for direct support of the summit team. As part of regular commissioning and operations observations, a "Rapid analysis" of the data will be performed in Chile. This is a pared down version of the Single Frame Measurement Pipeline that is configured to run quickly at the expense of skipping and/or reducing the quality of the reduction. The data products from this reduction are stored in a database and used for helping observers address issues that may arise during the night and need to be immediately addressed. However, this data reduction is not sufficient for science analysis, and therefore within 12-24 hours of observation, the full data reduction will be performed using the production-level single frame measurement pipeTask.

Another application where the USDF will assist in summit operations is in the analysis of the data acquired by pistoning the camera in and out of focus. This procedure essentially uses the entire camera as a curvature wavefront sensor, then the data is analyzed using the Active Optics System analysis code (another pipeTask). This dataset can be reduced on the summit but due to the superior computing resources it will be faster to perform the analysis at the USDF. The triggering mechanism is to be determined, but is expected to utilize the methods developed by the prompt processing team.

The data products from all reductions will be stored in a database that allows their results to be compared, often referred to as the consolidated database. As with many of the operational routines, the systems will be first implemented using the Auxiliary Telescope, then ComCam, and ultimately LSSTCam.

4.1.1 AuxTel

- start date: Jan 2021
- end date: xx 202?
- Target: use as pathfinder for processing, handling of data products and analysis prior to ComCam on-sky operation.
- Functionality
  - automatically receive data at the USDF
  - infrastructure in place to analyse the data: DM stack, functioning butler and repo, RSP
  - live EFD feed
  - workflow tools to process segments of data (eg an entire night) or specific datasets
  - modest compute and disk storage

4.1.2 ComCam

- start date: Jan 2023
- end date: Jul 2023
- Target: prepare for processing, handling of data products and analysis prior to ComCam on-sky operation.
- Functionality
  - automatically receive data at the USDF
  - infrastructure in place to analyse the data: DM stack, functioning butler and repo, RSP
  - Portal function of RSP, with TAP service backed by a Qserv instance
  - live EFD feed
  - workflow tools to process segments of data (eg an entire night) or specific datasets
  - 1000 compute cores and 1 PB disk storage

4.1.3 LSSTCam

- start date: Sep 2023
- end date: Mar 2024
- Target: prepare for processing, handling of data products and analysis prior to LSSTCam on-sky operation.
- Functionality
  - automatically receive data at the USDF
  - infrastructure in place to analyse the data: DM stack, functioning butler and repo, RSP
  - Portal function of RSP, with TAP service backed by a Qserv instance
  - live EFD feed
  - workflow tools to process segments of data (eg an entire night) or specific datasets
  - 5000 compute cores and101 PB disk storage

4.2 Data Release Productions

The decision to distribute Data Release Processing across three facilities increases the complexity of the task, but yields significant opportunities in the form of enhanced data-processing capacity (infrastructure) and improved resilience (that is, no reliance on a single data center).

Compared to a single-DF approach, the following additional challenges are envisaged:

- **Software synchronisation** – to ensure that each site undertakes precisely the same processing, involving not just the same versions of the LSST stack, but also the same (or guaranteed compatible) versions of third-party supporting libraries, tools, and supporting services.
- **Distributed campaign management** – is required to ensure that the many tasks involved to turn a collection of raw images into a full and complete Data Release is completed in an efficient, reliable, and timely manner.
- **Data movement and staging** – significant additional data movement is required to stage input files to the relevant Data Facility and receive back outputs into the master archives held at the USDF and in France. This movement needs to happen in a timely manner, even though it is taking place on public Internet capacity with consequent variations in performance and capacity.
- **Campaign monitoring** – to ensure that each facility progresses as expected through its portion of the DRP campaign, but also to provide contingencies in the event that one or more of the facilities falls behind with processing (maybe because of infrastructure issues or because of unforeseen complexities in image data or DRP).
- **Data aggregation** – the output data produced at each DF needs to be assembled into a consistent Data Release (along with required intermediate products). In particular, a Data Butler instance needs to be created that captures the provenance of the processing campaign, in a way that is indistinguishable from a DRP undertaken at a single facility.
- **Data publication** – possibly an advantage of distributed DRP, rather than a challenge but, at the end of each DRP campaign, the Data Release (either in part of whole) needs to be distributed to a number (around 10) Data Access Centres, internationally, from where it is made available to the Rubin Science Community.

- **Authentication and authorisation** – despite the heterogeneity of the infrastructure, across the three facilities, it is necessary for DRP staff from across the facilities to contribute to the DRP campaign – e.g., for USDF staff to have seamless access to data products being created in the UK and even to intervene in processing, should the need arise.
- **Accounting** – the three-site configuration is underpinned by an in-kind agreement which translates contributions to the DRP (in France and the UK) into data rights for the relevant France-based and UK-based science communities. It is therefore likely that the DFs will need to be able to record and present evidence that they have contributed resources in line with the intent of DRP.

4.2.1 Candidate Technologies

DRP has been observed to share several analogues with large-scale data processing undertaken for LHC experiments (e.g., ATLAS). That observation has motivated consideration of tools, technologies, and processes from the LHC community.

A number of specific tools and technologies have been highlighted, though more may be required:

Rucio A data management system from the ATLAS experiment at the LHC, now used widely in particle physics (and possibly SKA). This is a policy driven system for making (multiple) copies of data at any number sites globally, whilst maintain a coherent global catalogue. Sites are registered as a Rucio Storage elements (RSEs). Typically, data is identified, moved and accessed by a url-like string.

PanDA large-scale, distributed workflow orchestration is likely to fulfil many of the requirements of processing campaign management.

FTS The File Transfer Service that takes care of reliable (used in the sense of TCP reliable packet delivery) end-to-end file transfer, including third party file transfer. FTS guarantees a file will be delivered, regardless of any short-term failures in any individual copy.

VOMS is likely to be able to provide a common authentication/ authorisation platform

CVMFS is likely to be able to automate the distribution and curation of software suites for DRP.

Various other underlying components to implement what is known as a "Storage Element" (SE). Typical SEs are dCache or xRootD , but may be other things.

Various other components to implement what is known as a "Compute Element" (CE). Typical are ARC-CE, HTCondor-CE.

A likely advantage of these components is that they are known to work ,as they have been in full scale production for decades at the LHC. The question is whether in detail they are appropriate for the specific LSST data context and anticipated ways of working.

Complementing this, a number of pre-existing Rubin Observatory technologies are considered fundamental to DRP and hence need to be incorporated into the three-site DRP, including:

LSST Stack as noted above

Data Butler as noted above

Qserv a bespoke, distributed-memory, relational database into which the science catalogues are ingested for presentation to science users. It is understood that Qserv is not integral to DRP, though the outputs of DRP need to be ingested into Qserv, which is itself a challenging process.

Rubin Science Platform (RSP) a mix of client software and services which is used to interrogate raw, intermediate, and processed data. Again RSP is not integral to DRP. However, it is understood that Science Validation teams may wish to use RSP to interrogate candidate DR products and even in-progress DRP campaign data. The primary intermediaries between the RSP and a DR are Qserv and the Data Butler.

Kafka the stream processing platform that is used to serve a complement to DRP, called Alert Processing, which will serve pseudo-real-time information on transient astronomical activity to the science community. The Kafka platform is independent of DRP and likely only to involve the USDF. Therefore any overlap between AP and DRP should be limited to the USDF.

Where these technologies have been developed in-house, by Rubin staff, it is likely they were not envisaged to be employed for a multi-site campaign.

4.2.2 Assumptions

Data previews DP0.1 and DP0.2 have been completed at time of writing. However, the DP0.2 inputs (simulations from DESC Data Challenge 2) are a suitable collection of inputs with which to mature a design and implementation for the three-data-facility processing model. The DP0.2 inputs are readily available; are small enough to be tractable on modest computing resources while being large enough to exercise the scaling of the infrastructure; and we have a validated output with which to compare performance of each trial campaign.

Nothing new is needed between Data Preview 2 (DP2) and Data Release 1 (DR1), except more hardware.

Each Data Preview is built on the previous one's functionality.

4.2.3 Three Data Facility Rehearsal

- start date: Q2 Financial Year 23 (FY23) (that is, early 2023)
- end date: Q4 FY23
- Target: Successfully reprocess Data Challenge 2 (Dark Energy Science Collaboration (DESC)) (DC2) data across the three data facilities as a first rehearsal of operational data processing.
- Functionality:
  - DRP; Difference Image Analysis (DIA); RSP
  - Production ANd Distributed Analysis system (PanDA) and Processing team driving processing
  - some campaign management
  - run as Ops Rehearsal

Using DC2 and as a precursor to DP1, a rehearsal involving all three Data Facilities is envisaged to demonstrate the ability to do multi-site processing:

- cooperative development of the Data Butler contents across the three data facilities,

- Rucio clients for triggering data replication among the facilities,
- mechanism to accept jobs submitted by the PanDA central instance for local execution.

Here is a rough description of the rehearsal steps:

- define a suitable subset of DC2 as input data to be processed
- ingest input data into the distributed data management system (e.g., Rucio) and stage required data to each data facility.
- progress data processing campaign, step by step, using PanDA (or alternative) to action, monitor, and individual data processing jobs across the three sites
- Ensure a Data Butler is assembled as the processing progresses
- Ensure required outputs are staged back to US Data Facility and French Data Facility.

4.2.4 DP1 - ComCam

- start date: Q3 FY23
- end date: Q2 Financial Year 24 (FY24)
- Target: processing across USDF, FrDF, UKDF (need a backup infrastructure, if DP0.3 identifies significant issues that jeopardise this campaign).
- Released Products:
  - full DRP
  - Alert Production (AP), but no live alerts. Canned alerts are planned to allow interactions with brokers and the Minor Planet Center (MPC).
- Functionality:
  - Summit to USDF - dual path with NCSA
    * transfer images
    * transfer calibrations - bidirectional

    - * transfer Engineering and Facility Database (EFD) contents
  - DF production
    - * set up as processing site with sufficient storage and Central Processing Unit (CPU); configurable clusters for AP vs DRP. PanDA server at SLAC National Accelerator Laboratory (SLAC).
    - * Qserv + ingest mechanism
    - * Butler + ingest
    - * RSP
    - * connection to brokers for canned alerts. (MPC?)
    - * Data movement among DFs
    - * campaign management, monitoring and quality assessment
    - * IDACs may be under test at this phase

- Resources needed
  - Databases installed
  - EFD, butler, Prompt Products DataBase (PPDB), APDB...
  - PanDA server
  - Rucio instance (?)
  - Qserv size
  - CPU & storage

4.2.5 DP2 - LSSTCam

- start date: Q1 FY23
- end date: Q1 FY24
- Target: processing at USDF; FrDF, UKDF
- Released Products:
  - full DRP
  - AP, with canned and live alerts.
- Functionality in addition to DP1:

  - IDACs
  - Bulk downloads - including policy
  - Resources needed
  - CPU + storage increases

4.2.6 DR1 - Survey

This is already operations.

- start date: Q3 FY24
- end date: Q1 Financial Year 25 (FY25)
- Released products
  - full DRP
  - AP where templates are available, with some live alerts.
- Functionality
  - no additional functionality to DP2

5 Infrastructure

The bottom layer of the Data Facilities is the hardware on which the platforms run. These require a scalable architecture with sufficient storage and CPU to support the Data Preview/Release timeline. This element requires a design for the architecture followed by an acquisition and installation plan.

The middle layer includes the infrastructure to support deployment of hardware and tools for data movement and workflow management.

The top layer involves the applications: science platform, Qserv and pipelines.

In response to the timeline **??** the plan is as follows.

5.1 Hardware architecture and technology

See also DMTN-189 (scope) and DMTN-135 (sizing).

5.2 Key initial services

These initial services/resources are planned to support DP1. Support for DP2 and DR1 are largely by increments in hardware per the sizing model.

- Hardware
  - file systems: An architecture choice must be made for object store, likely between ceph and minIO. This may affect the hardware choice (Just a Bunch of Disks (JBOD) vs appliance). 3.5 PB of object store and 1.5 PetaByte (PB) of POSIX disk are envisaged.
  - CPU allocation: The bulk of the CPU is in batch (1000 cores) and staff RSP instances (500 cores).
  - Qserv: depending on the ultimate location of Qserv, we expect to do scale testing in the cloud and at National Center for Supercomputing Applications (NCSA) prior to deciding on an implementation.
-
- Science Platform
  - Kubernetes provisioning system (K8S) is the standard for deploying applications and resources. The Science Platform is built on top of it. Additionally, Continuous Integration (CI) activities are run via K8S.
  - RSP has been installed in multiple locations and architectures. For DP1, we expect science users to go to the IDF for data access, while the USDF provides staff access.
- Workflow and Data Movement Tools
  - PanDA is under serious consideration as the toolkit for at-scale workflow. It will get its first load testing in Data Preview 0 (DP0).2. It is expected that there will be worked needed to customize PanDA to Rubin's situation. We also anticipate a layer on top of PanDA to orchestrate campaign management.

  - Rucio is under consideration for data movement. It works with policies to schedule data movement and integrates with a transport layer (most commonly File Transfer Service (FTS)).

5.3 Enclave deployment

The USDF depends on an expansion of SLAC's SRCF-I data center ("SRCF-II"), which is scheduled for completion in March 2023, with an estimated 6 months needed to be ready for hardware installations.

5.3.1 Prompt

DP1 drives much of the USDF implementation. A difference from DP2 is that DP1 will feature only canned alerts.

5.3.1.1 DP1

- Prompt processing requires a cluster of compute nodes, of relatively fixed size.
- Kubernetes
- Alert Production DataBase (APDB) - Cassandra database
- butler repository
- Kafka database for alert distribution
- transfer mechanism for summit images to USDF
- PanDA server
- Data BackBone services

5.3.1.2 DP2

- connection to Minor Planet Center

- prompt processing cluster - 1200 cores

5.3.1.3 DR1

- increase cluster and storage sizes appropriate to DR1 sizing

5.3.2 Archive

5.3.2.1 DP1

- Data BackBone services
- Prompt Products database (Postgres)
- sufficient storage (1500 cores; 1.5 PB POSIX, 3.5 PB object store)

5.3.2.2 DR1

- increased storage (2000 cores; 60 PB disk and tape)

5.3.3 US Data Access

The DAC relies almost entirely on the RSP, which draws data from Qserv and the butler. Any A&A issues will need to have been addressed for the target users. These are all needed to be in place for DP1.

There may be distinct RSPs for staff and science users.

5.3.4 Developer and Integration

This enclave requires a staff RSP coupled to sufficient batch and storage resources. The install required for DP1 should satisfy these needs.

5.3.5 Offline Production

5.3.5.1 USDF All the services needed for Offline Production have been described above as needed in other enclaves: here, the USDF needs to add to its hardware base to satisfy each phase.

DP1

- sufficient cores and storage (using 1000 cores; 3.5 PB object store)

6 Implementation at Individual Data Facilities

6.1 USDF

6.1.1 US Data Facility as Hybrid Cloud-on prem

The USDF is envisaged as being a hybrid model: science users will interact with the RSP, hosted in the cloud, while processing and storage archive will be on-prem at SLAC. The cloud portion will be a continuation of the successful IDF demonstrated in DP0 and act as the US DAC.

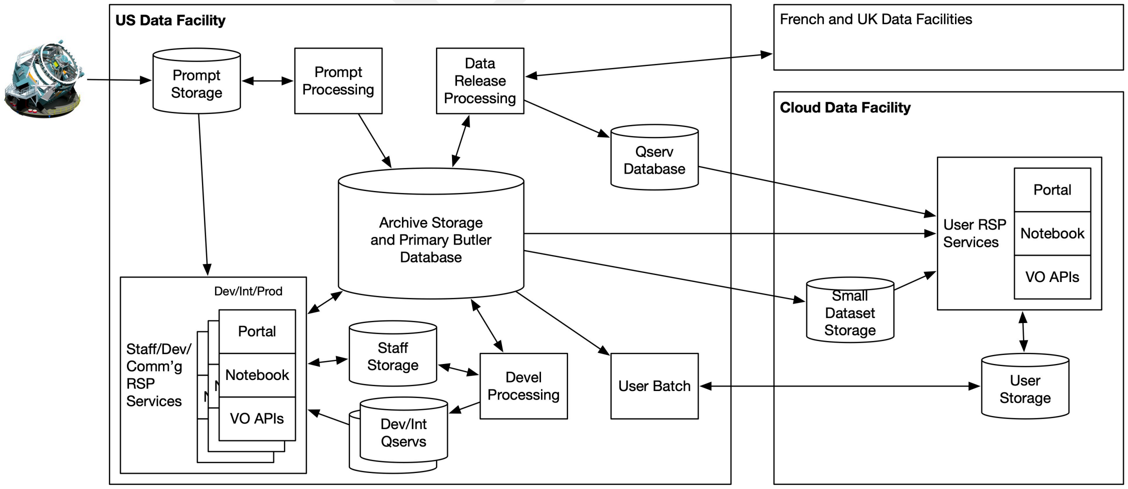


FIGURE 2: USDF as hybrid Cloud-on prem

Needed services will be to populate the cloud storage with a data cache from the on-prem archive, and to provide additional user batch resources outside of the cloud resources.

6.1.2 Replacing NCSA as US Data Facility

NCSA shut down for Rubin at the end of August 2022. In order to take over, the USDF:

- Transferred some 5 PB of historical data (raw data, processing output and user data) to SLAC
- Onboarded some 200 staff and in-kind commissioners, providing a similar environment to NCSA:
  - developer nodes
  - batch system with 2000 cores (also usable via HTCondor and Parsl)
  - instance of the Rubin Science Platform
  - Butler instance
  - at the time of transfer, workflow (PanDA) was still using the CERN instance used by the IDF for DP0.2, and a small Rucio instance for testing multi-site transfers.
  - live sync'ed EFD
  - data transfers from the summit to SLAC (eg for AuxTel and ComCam)

The USDF is hosted by the SLAC Shared Science Data Facility (S3DF). It provides:

- slurm batch system
- bastion host login nodes (to jump to Rubin developer nodes)
- weka file system, which provides a tiered flash/spindle storage system with POSIX and S3 direct access.
- tape silo using HPSS for backup
- data transfer nodes (DTNs)

User home directories, group space and software are stored on flash; each SLAC project provides its own storage for data in the weka tiered system.

6.1.3 Supporting Commissioning

During commissioning, the data from the EFD, rapid analysis ran on site, and the single-frame measurement pipeline ran at the USDF (as discussed in section 4.1) needs to be rendezvoused into a common database, at least as viewed by the user. This includes the ability to have data sorted per visit, average time, or for a given calibration set. The requirements for this database (or databases) are currently being refined.

Also undergoing definition, but will require support, is for "Processing and Analysis Management." This is where datasets are selected and processed based on data quality and/or other criteria. Identifiers are then attached (e.g. quality flags) to dataIds which indicate units of data (e.g. (visit, ccd)) which can be used in future processing.

- Need consolidated database - Process and analysis management Backed by a database

6.1.4 Supporting DRP Workflow and Data Movement

Production servers will be instanciated at SLAC for PanDA and Rucio, both deployed on kubernetes, both backed by postgres databases.

6.1.5 Resources Projection

The USDF follows the DMTN-135 sizing model for survey operation, essentially adding 2000-2500 compute cores and 30 PB each of disk and tape for each survey year. Buildup for survey is expected to start in FY24.

Resources for ComCam and LSSTCam support in commissioning will provide about 5000 cores and 15 PB of storage total.

6.1.6 Alert and Prompt Processing, and Alerts Distribution

A prompt processing harness is being developed on the IDF and expected to be available for port to the USDF in the last quarter of 2022. This will provide the kafka feed mechanism that brokers will be connecting to.

6.2 FrDF

6.2.1 Overview

The Rubin French Data Facility (FrDF) is hosted and operated by IN2P3 / CNRS computing center (CC-IN2P3), located in Lyon, France. This is a scientific data processing center which serves several major international projects using a pool of shared computing resources.

The compute and storage resources allocated to Rubin are progressively deployed as need arises, typically matching the calendar year funding cycle.

Documentation specific to Rubin users is available at doc.lsst.eu.

6.2.2 Data release processing

For Data Preview 0.2, which involves processing of the DESC DC2 simulated images, the following resources are deployed, operational and routinely used:

- a Slurm-powered batch processing farm with compute nodes equiped with CPUs of x86 architecture (64 bits). The allocation for DP0.2 is equivalent to 3600 reference CPU cores (Intel Xeon E5-2680v3 @ 2.5GHz, see DMTN-135),
- a POSIX-compliant file storage system implemented by CephFS with 5 PB available for image data and processing products,
- a webDAV-compliant object storage system implemented by dCache with 2.5 PB available for image data and processing products,
- two dedicated instances of PostgreSQL RBDMS with flash storage for butler registry databases, one devoted for user's private registries and another for data release processing registries,
- a set of 4 dedicated data transfer nodes, each with 10 Gbps network interface for exchanging data with the other data facilities.

Specifically for DP0.2, at the time of this writing a subset of the DESC DC2 simulated images is being processed locally and independently of the other data facilities. The purpose of this is to verify that the facility's infrastructure is well configured and to run the LSST pipelines at scale as well as to exercise the tools for comparing the products of the local processing are compatible with those produced by the Interim Data Facility.

6.2.3 Catalog database

A local production instance of Qserv composed of 15 physical, well-configured database server nodes is in production, populated with several catalog databases, including the DP0.2 catalog produced at the Interim Data Facility.

An integration instance deployed on the on-prem OpenStack cloud for experimenting with new releases of Qserv and Kubernetes.

6.2.4 Science platform

A 5-node evaluation instance of the Rubin Science Platform reachable at data-dev.lsst.eu is deployed and is being integrated with CC-IN2P3's specific services (e.g. identity and access management, user storage services, etc.) and with the local Qserv instance. The intended use of that instance is to serve local users.

The Kubernetes cluster used by this science platform instance is shared with the Qserv production instance.

6.2.5 Software distribution

The Rubin FrDF operates the source of truth software repository which distributes the LSST Science Pipelines to the other Rubin data facilities via CERN's CernVM file system (see sw.lsst.eu).

Stable and weekly releases of the LSST Science Pipelines are made available via this distribution mechanism and transparently accessed by users at the USDF and UKDF as well as from their personnal computers. The purpose of this mechanism is to ensure strict compatibility of the software releases used for data release production at the 3 facilities.

6.2.6 Fink alert broker

The Fink alert broker will be hosted in FrDF's on-prem cloud infrastructure on top of OpenStack. For year 2022, 250 CPU cores and 250 TB of storage on CephFS were allocated for deploying the initial production-level instance of Fink. This project is ongoing at the time of writing.

6.3 UKDF

The UKDF will provide an Offline Production Enclave capable of processing 25% of telescope observations and contributing these to the annual Data Release cycle.

The UKDF will also host a full Independent Data Access Centre and a Community Alert Broker.

The infrastructure that will constitute the UKDF will be sourced from the UK IRIS programme (http://www.iris.ac.uk/). IRIS is also expected to provide infrastructure to peer *experiments*, including the Euclid space telescope, the Square Kilometre Array, and the next generation of LHC experiments.

IRIS provides a mix of on-premises infrastructures hosted at partner institutions that can, at a high-level, be categorised as:

Cloud IaaS-style compute and storage resources, typically provided as OpenStack virtualised compute and Ceph clusters.

HPC Tier 1-scale high-performance computing resources and working storage, accessible via batch processing.

Grid Very high-throughput computing and storage based on LHC *grid* technologies.

IRIS provider institutions are connected via the UK academic network (Janet), which typically provides 100 Gbps potential bandwidth, as well as connections into the European academic network, Géant.

IRIS also operates cross-cutting services to enable full exploitation of the infrastructure, including authentication and authorisation services (based on IAM), usage accounting, security, policy development, and a helpdesk function. Some effort is likely to be required to harmonise these into a three-DF resource.

6.3.1 Resource Project

The UKDF team maintain a resource sizing mode, based on DMTN-135 for DRP and IDAC, and augmented by resource estimates for the Community Alert Broker and User-generated Products. This model is used to inform IRIS of the resource requirements for the UK in-kind contributions to telescope operations.

IRIS operates an annual infrastructure expansion and refresh, which coincides with the UK financial year (April–March). Each year, during October–December, experiments such as LSST:UK advise the IRIS Resource Scrutiny and Allocation Panel (RSAP) of their requirements for the following year (that is, April–March), plus give advanced notice of expected longer-term requirements for the subsequent four years. These requirements inform IRIS procurement plans for the subsequent year, as well as longer term funding requirements, to enable the experiments to continue to operate.

LSST:UK expect to utilise all three forms of IRIS infrastructure, loosely organised as follows:

- Data Release Processing will primarily be completed on Grid infrastructure, hosted at one of several UK provider sites. At the time of writing, LSST:UK has active allocations of grid resources at Lancaster University and Rutherford Appleton Laboratories. Storage will be provisioned to hold the UK's share of raw images (from the telescope), topical data products, and working storage for intermediate products and pipeline scratch space.
- The UK IDAC will primarily be hosted on cloud infrastructure, provisioned using Kubernetes or equivalent container-orchestration technology.
- Lasair will also be hosted on cloud infrastructure, alongside the UK IDAC to enable use cases spanning both Lasair and the RSP.
- Various User-generated Products are planned, to extend the science potential of the Data Release products. HPC resources will be employed for those User-generated Products that require significant computational resources to produce (for example, running part of the LSST Science Pipeline).

Table 1:

Infrastructure Type	Key Services	FY22	FY23	FY24
Compute (Millions of Core Hours)	DRP, RSP	2	11	21
Storage (Petabytes)	/work, Butler, Qserv	1.5	9.0	16.0

6.3.2 Offline Production

The UKDF team will provision compute and storage resources on Grid infrastructure, exposed to the Rubin Rucio service instance for distributed-date management and the Rubin PanDA service for job management.

These services will be supported by a team of expert staff – currently, at 2.0 FTE, though rising to 5.0 FTE by the beginning of Rubin Operations.

6.3.3 Independent Data Access Center

A prototype IDAC is operational, running on IRIS cloud resources at the University of Edinburgh. It is expected that this prototype IDAC will mature into the production UK IDAC in advance of telescope operations.

At the time of writing, the UK prototype IDAC is running an instance of the Rubin Science Platform, including a 10-node Qserv implementation that is hosting catalogues from DP0.1 and early runs of HSC-VISTA fused catalogues, as a preview of the planned in-kind contribution of LSST/near-IR User-generated Product.

6.3.4 Lasair Alert Community Broker

A preview of the Lasair alert community broker is running on IRIS cloud resources, at the University of Edinburgh, processing and serving alerts from the Zwicky Transient Facility.

7 Verification of the USDF

Certain tests from LDM-503 will have to be repeated at the USDF in one or more of the enclaves. This needs to be properly specified. These test relate to requirement verification and tend to be functional not scale oriented.

8 Construction tasks influenced by USDF

- Forwarders (from Chile to USDF)
- Bulk transfer and bulk download
- Workflow
- Campaign Management
- ...

A Services and enclaves

Put the list here including dependancies

B References

[DMTN-189], Lim, K.T., 2021, *Data Facility Specifications*, DMTN-189, URL `https://dmtn-189.lsst.io/`,
Vera C. Rubin Observatory Data Management Technical Note

[RTN-001], O'Mullane, W., 2021, *Data Preview 0: Definition and planning.*, RTN-001, URL `https://rtn-001.lsst.io/`,
Vera C. Rubin Observatory Technical Note

[LDM-503], O'Mullane, W., Swinbank, J., Juric, M., et al., 2022, *Data Management Test Plan*, LDM-503, URL https://ldm-503.lsst.io/,
Vera C. Rubin Observatory Data Management Controlled Document

[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

C Glossary

Alert A packet of information for each source detected with signal-to-noise ratio > 5 in a difference image by Alert Production, containing measurement and characterization parameters based on the past 12 months of LSST observations plus small cutouts of the single-visit, template, and difference images, distributed via the internet.

Alert Production Executing on the Prompt Processing system, the Alert Production payload processes and calibrates incoming images, performs Difference Image Analysis to identify DIASources and DIAObjects, and then packages the resulting alerts for distribution..

Alert Production DataBase A dedicated, internal database system used to support LSST Alert Production. Does not support end-user access..

AP Alert Production.

APDB Alert Production DataBase.

AVS Alert Vetting System.

Butler A middleware component for persisting and retrieving image datasets (raw or processed), calibration reference data, and catalogs.

CCD Charge-Coupled Device.

Center An entity managed by AURA that is responsible for execution of a federally funded project.

Charge-Coupled Device a particular kind of solid-state sensor for detecting optical-band photons. It is composed of a 2-D array of pixels, and one or more read-out amplifiers.

CI Continuous Integration.

CPU Central Processing Unit.

Data Management The LSST Subsystem responsible for the Data Management System (DMS), which will capture, store, catalog, and serve the LSST dataset to the scientific community and public. The DM team is responsible for the DMS architecture, applications, middleware, infrastructure, algorithms, and Observatory Network Design. DM is a distributed team working at LSST and partner institutions, with the DM Subsystem Manager located at LSST headquarters in Tucson.

Data Release Production An episode of (re)processing all of the accumulated LSST images, during which all output DR data products are generated. These episodes are planned to occur annually during the LSST survey, and the processing will be executed at the Archive Center. This includes Difference Imaging Analysis, generating deep Coadd Images, Source detection and association, creating Object and Solar System Object catalogs, and related metadata.

DC2 Data Challenge 2 (DESC).

DESC Dark Energy Science Collaboration.

DF Data Facility.

DIA Difference Image Analysis.

Difference Image Analysis The detection and characterization of sources in the Difference Image that are above a configurable threshold, done as part of Alert Generation Pipeline.

DP0 Data Preview 0.

DP1 Data Preview 1.

DP2 Data Preview 2.

DR1 Data Release 1.

DRP Data Release Production.

EFD Engineering and Facility Database.

FTS File Transfer Service.

FY23 Financial Year 23.

FY24 Financial Year 24.

FY25 Financial Year 25.

JBOD Just a Bunch of Disks.

K8S Kubernetes provisioning system.

monitoring In DM QA, this refers to the process of collecting, storing, aggregating and visualizing metrics.

MPC Minor Planet Center.

NCSA National Center for Supercomputing Applications.

PanDA Production ANd Distributed Analysis system.

PB PetaByte.

PPDB Prompt Products DataBase.

Prompt Products DataBase Data products within LSST data releases relating to LSST Alert Production.

Qserv LSST's distributed parallel database. This database system is used for collecting, storing, and serving LSST Data Release Catalogs and Project metadata, and is part of the Software Stack.

RDM Rubin Data Management.

RDP Rubin Data Production(Obsolete use RDM).

Release Publication of a new version of a document, software, or data product. Depending on context, releases may require approval from Project- or DM-level change control boards, and then form part of the formal project baseline.

RSP Rubin Science Platform.

Rucio Rucio is a project that provides services and associated libraries for allowing scientific collaborations to manage large volumes of data spread across facilities at multiple institutions and organizations. Rucio has been developed by the ATLAS experiment.

Science Collaboration An autonomous body of scientists interested in a particular area of science enabled by the LSST dataset, which through precursor studies, simulations, and algorithm development lays the groundwork for the large-scale science projects the LSST will enable. In addition to preparing their members to take full advantage of LSST early in its operations phase, the science collaborations have helped to define the system's science requirements, refine and promote the science case, and quality check design and development work.

Science Platform A set of integrated web applications and services deployed at the LSST Data Access Centers (DACs) through which the scientific community will access, visualize, and perform next-to-the-data analysis of the LSST data products.

SLAC SLAC National Accelerator Laboratory.

SLAC National Accelerator Laboratory A national laboratory funded by the US Department of Energy (DOE); SLAC leads a consortium of DOE laboratories that has assumed responsibility for providing the LSST camera. Although the Camera project manages its own schedule and budget, including contingency, the Camera team's schedule and requirements are integrated with the larger Project. The camera effort is accountable to the LSSTPO..

Summit The site on the Cerro Pachón, Chile mountaintop where the LSST observatory, support facilities, and infrastructure will be built.

USDF United States Data Facility.