Vera C. Rubin Observatory
Data Management

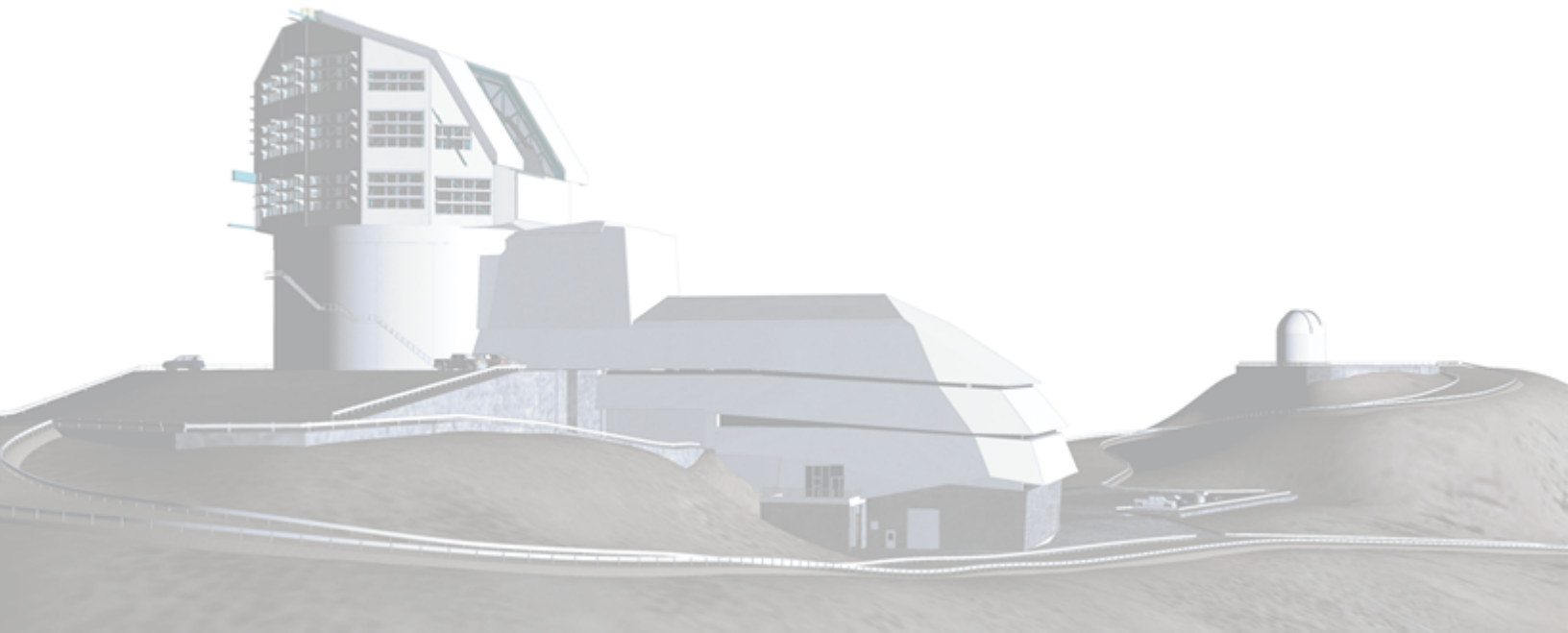
Data Facilities Transition Plan

Richard Dubois, William O'Mullane

RTN-021

Latest Revision: 2022-04-01

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

USDF operations are covered in the operations plan. However we need USDF in place ahead of Rubin operations so it can be fully functional on day one. This will require some initial setup and running some services in parallel with National Center for Supercomputing Applications (NCSA). Some tests run at NCSA will have to be rerun to verify requirements at the USDF.

This document captures the time line, structure and plan for getting the Data Facilities implemented.

Success in the setup up of the USDF is a shared responsibility between SLAC and Rubin Observatory, It is a preops activity but it has influence on some remaining construction tasks (see Section 6). Similarly, the annual Data Release Processings are a joint function of the three Data Facilities (France, UK, US).

1.1 Enclaves and functionality

A complete service list and how the services relate to enclaves is provided in Appendix A. In brief, these enclaves include: Prompt US Enclave; Offline Production Enclave; Archive US Enclave; US Data Access Center; and Development and Integration Enclave. Only the Offline Production enclave is required for the French and UK facilities.

2 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 (RDP) A&P alerts team.

3 Timeline for Data Facilities Implementation

The timeline is derived from functionality required by the Data Previews and then the first Data Release, DR1. Analysis indicates that the bulk of the service implementation is needed for DP1, with the exception of live alerts.

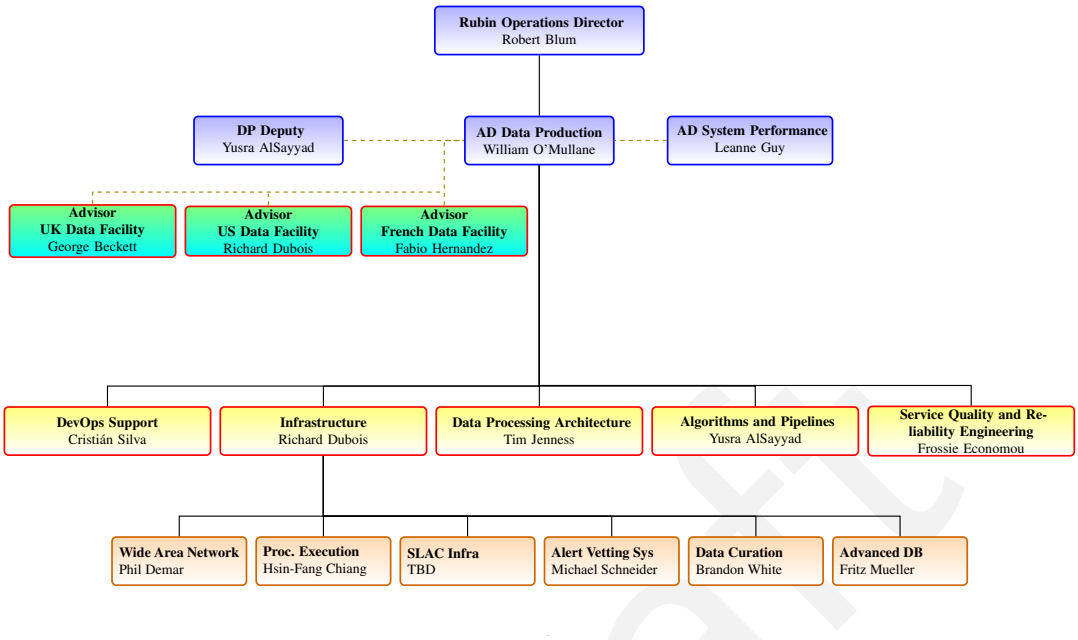


FIGURE 1: Rubin Data Production Org Chart from O'Mullane (RTN-001)

We need to develop a timeline for what needs done by when in the DFs, assuming something like mid 2023 to show everything is ready, if not scaled up already. This would mean that everything in the scope document and derived services list would be deployed.

With DP0.1 in progress, involvement with the Data Facilities will begin with DP0.2 [RTN-001] following through to Data Preview 2 (DP2). See also Rubin Directors Office (RDO)-011 for detail on the various Data Previews. Figure **??** shows the activities and events involved with preparation for ComCam and DP1.

The Data Production Milestones are listed in Table 1 - we may need to add some more specifically for USDF.

Table 1: Milestones for Rubin Observatory Data Production and System Performance

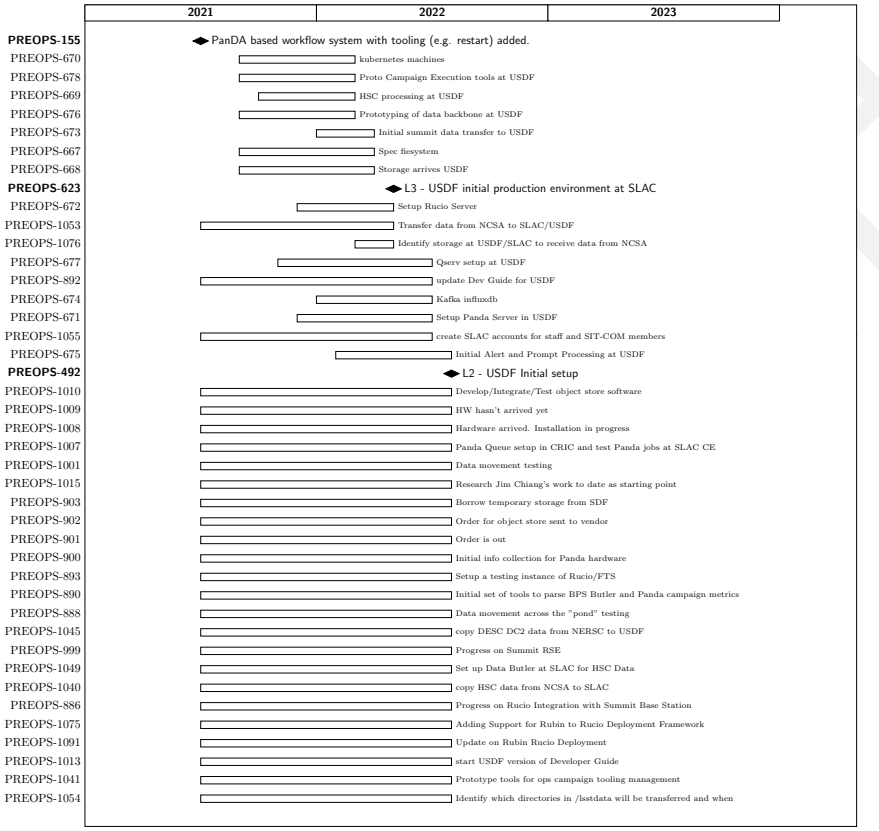
Milestone	Jira ID	Rubin ID	Due Date	Level	Status	Team
PanDA based workflow system with tooling (e.g. restart) added.	PREOPS-155	L3-MW-0060	2021-06-30	3	In Progress	Data and Processing Architecture
DP0.2 Qserv ingest preliminary run data	PREOPS-643	not set	2021-12-23	3	To Do	None
Plan for how to use IN2P3 in DP0.2	PREOPS-160	L3-EX-0010	2021-12-30	3	To Do	Infrastructure and Support
Deliver preliminary implementation plan for real-time and daily monitoring	PREOPS-515	L3-SC-0020	2021-12-31	3	To Do	Survey Scheduling
Deliver preliminary list of science measurements for quarterly monitoring	PREOPS-517	L3-SC-0040	2021-12-31	3	To Do	Survey Scheduling
Deliver preliminary list of measurements for real-time and daily monitoring	PREOPS-514	L3-SC-0010	2021-12-31	3	To Do	Survey Scheduling

Science Platform ready on for DP0.2	PREOPS-157	L3-PR-0040	2022-02-15	3	To Do	Service Quality and Reliability Engineering
Demonstrate EPO interface with DP0	PREOPS-152	L3-PR-0030	2022-02-18	3	To Do	Service Quality and Reliability Engineering
QA sign-off to begin DP0.2 single frame processing	PREOPS-662	L3-VV-0030	2022-03-01	3	In Progress	Verification and Validation
QA sign-off to begin DP0.2 coadd processing and beyond	PREOPS-661	L3-VV-0020	2022-03-01	3	To Do	Verification and Validation
DP0.2 Processing End	PREOPS-640	not set	2022-04-15	3	To Do	Infrastructure and Support
L3 - USDF initial production environment at SLAC	PREOPS-623	L3-IS-0010	2022-04-15	3	In Progress	Infrastructure and Support
QA sign-off on DP0.2 final run products	PREOPS-659	L3-VV-0010	2022-05-01	3	To Do	Verification and Validation
DP0.2: Commence Qserv Ingest	PREOPS-639	not set	2022-05-01	3	To Do	Infrastructure and Support
L2 - DP0.2 Internal Access: Staff access to RSP services and scale testing starts.	PREOPS-159	L2-DP-0040	2022-05-15	3	To Do	Service Quality and Reliability Engineering
Validation of DP0.2 products in RSP complete.	PREOPS-658	not set	2022-06-15	3	To Do	Verification and Validation
Documentation and Notebooks ready for DP0.2	PREOPS-638	not set	2022-06-15	3	Invalid	Community Engagement
DP systems stable and all teams done for DP0.2 (Go/ No go)	PREOPS-637	not set	2022-06-15	3	To Do	Service Quality and Reliability Engineering
Deliver LSST Data Products Documentation (DP0)	PREOPS-149	L3-CE-0010	2022-06-30	3	In Progress	Community Engagement
L2 - Prepare resources for DP0.2	PREOPS-651	L2-PF-0053	2022-06-30	2	In Progress	Community Engagement
L2 - DP0.2 Data Release: science-ready catalogs from reprocessed DP0 images released from the IDF	PREOPS-484	L2-PF-0052	2022-06-30	2	In Progress	System Performance Management
L2 - DP0.2 Public release to delegates	PREOPS-483	L2-DP-0051	2022-06-30	2	In Progress	Data Production Management
L2 - USDF Initial setup	PREOPS-492	L2-DP-0081	2022-07-31	2	In Progress	Infrastructure and Support
Deliver implementation of real-time and daily monitoring system	PREOPS-516	L3-SC-0030	2022-08-31	3	To Do	Survey Scheduling
Deliver initial Quality Assessment and Assurance (QA) plan for ComCam Data.	PREOPS-293	FY20-0010	2022-10-28	2	To Do	Verification and Validation
Deliver implementation of quarterly metric monitoring	PREOPS-518	L3-SC-0050	2022-12-30	None	To Do	Survey Scheduling

3.1 Assumptions

Nothing new is needed between DP2 and Data Release 1 (DR1), except more hardware. Each Data Preview is built on the previous one's functionality.

3.2 Data Release Productions



]

FIGURE 2: Timeline of DF Activities Supporting DP1 - from Jira

3.2.1 DP0.2 - DESC DC2 reprocessing

See RTN-001.

- start date: Q4 Financial Year 21 (FY21)
- end date: Q4 Financial Year 22 (FY22)
- Target: reprocess Data Challenge 2 (DESC) (DC2) data on Interim Data Facility (IDF); run as first Ops rehearsal
- Functionality:
  - DRP; Difference Image Analysis (DIA); RSP
  - Production ANd Distributed Analysis system (PanDA) and Processing team driving processing
  - some campaign management
  - French Data Facility (FrDF) planning parallel reprocessing and RSP; Probably not using PanDA. May attempt some at USDF:
  - run as Ops Rehearsal

3.2.2 Three Data Facility Rehearsal

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

- local butlers which would contain the inputs and products of the processing done at each facility,
- 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:

- decide on a small dataset to start

- have Rucio send it to CC-IN2P3 and then the UK and register with their butlers
- process with PanDA, submitted centrally. Divide up the dataset among the sites. Use campaign mgmt tooling on top of bps. Processing team monitors production.
- Local butlers updated
- Rucio transfers data products to SLAC and CC (and UK? IDF?); local butlers updated.

3.2.3 DP1 - ComCam

- start date: Q3 FY22
- end date: Q2 Financial Year 23 (FY23)
- Target: at NCSA; parallel processing at USDF; reprocessing at USDF, FrDF, UKDF
- 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

3.2.4 DP2 - LSSTCam

- start date: Q1 FY23
- end date: Q1 Financial Year 24 (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

3.2.5 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

4 Planning

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 Section 3 the plan is as follows.

4.1 Hardware architecture and technology

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

4.2 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 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 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)).

4.3 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.

4.3.1 Prompt

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

4.3.1.1 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

4.3.1.2 DP2

- connection to Minor Planet Center
- prompt processing cluster - 1200 cores

4.3.1.3 DR1

- increase cluster and storage sizes appropriate to DR1 sizing

4.3.2 Archive

4.3.2.1 DP1

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

4.3.2.2 DR1

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

4.3.3 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.

4.3.4 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.

4.3.5 Offline Production

4.3.5.1 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)

4.3.5.2 FrDF 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 for Rubin's DP0.2 through DP1 will be progressively deployed as need arises. For DP0.2, which involves processing of the DESC DC2 simulated images, the following resources are deployed and operational:

- a GridEngine-powered batch processing farm with compute nodes equiped with CPUs of x86 architecture (64 bits). The allocation for DP0.2 is equivalent to 1600 reference CPU cores (Intel Xeon E5-2680v3 @ 2.5GHz, see DMTN-135),
- a POSIX-compliant file storage system implemented by CephFS with 1 PB available for image data and processing products,
- a webDAV-compliant object storage system implemented by dCache with 1 PB available for image data and processing products,
- a dedicated instance of PostgreSQL RBDMS with flash storage for butler registry databases,
- 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, a subset of the DESC DC2 simulated images will be processed locally and independently of the other data facilities. The specific subset of input data will be selected so that the resulting products and the processing performance can be compared to the ones of the other data facilities.

4.3.5.3 UKDF The UK Data Facility plans to deliver infrastructure that will be expanded each year (from April) to meet Offline Production Requirements (from October).

UK DF will deploy, onto this infrastructure, the services required for each escalation in Data Release Processing experiments, towards operational readiness, with key infrastructure elements, as follows:

- Compute nodes, accessed via Kubernetes, Slurm or PanDA, as is appropriate, to host the Pipeline Stack, Qserv, the Data Butler, the Rubin Science Platform, and supporting services.
- (Disk-based) storage, accessible via Rucio, S3, or database connectors, as is appropriate, to host Pipeline working storage, raw and processed images, science catalogues, end-user storage, and supporting data.
- Network bandwidth to support bulk distribution and retrieval of campaign data via the Data Backbone, user access, plus monitoring and logging services, as required.
- Staff to maintain the infrastructure and the services that will be deployed on that infrastructure, along with necessary interfaces to connect the UK Data Facility to the US and French Data Facilities.

The high-level infrastructure capacity that is envisaged, in the UK, to support the boot-strapping of Data Facilities operations is captured in Table ???.

Table 2:

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

5 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.

6 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/

[RTN-001], O'Mullane, W., 2021, *Data Preview 0: Definition and planning.*, RTN-001, URL https://rtn-001.lsst.io/

[DMTN-135], O'Mullane, W., Dubois, R., Butler, M., Lim, K.T., 2021, *DM sizing model and cost plan for construction and operations.*, DMTN-135, URL https://dmtn-135.lsst.io/

[LDM-503], O'Mullane, W., Swinbank, J., Juric, M., et al., 2021, *Data Management Test Plan*, LDM-503, URL https://ldm-503.lsst.io/

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 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.

FrDF French Data Facility.

FTS File Transfer Service.

FY21 Financial Year 21.

FY22 Financial Year 22.

FY23 Financial Year 23.

FY24 Financial Year 24.

FY25 Financial Year 25.

IDF Interim Data Facility.

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.

RDO Rubin Directors Office.

RDP Rubin Data Production.

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.