Rubin Observatory

Data Preview 0: Definition and planning.

William O'Mullane

2020-12-18

1 Introduction

Table 1 shows the FY21 milestones for the Rubin Observatory, many of which concern, or relate to, data previews. Section 2 defines what Data Preview 0 is about and covers possible risks and mitigations to that definition. Section 3 Sets out the planning for achieving DP0.

Table 1: Milestones for Rubin Observatory Data Production and System Perfomrance FY21

Milestone	Jira ID	Rubin ID	Due Date	Level	Team
Develop a first model for community engagement for DP0.1	PREOPS-151	L3-CE-0020	2021-01-31	3	Community Engagement
IDF DP0-Ready: Complete IDF installation and IDF staff preparations for DP0.	PREOPS-140	L2-DP-0010	2021-01-31	2	Infrastructure and Support
Read only Gen3 butler for DP0 at IDF	PREOPS-143	L3-MW-0030	2021-03-31	3	Science Users Middleware
Qserv installation on IDF	PREOPS-142	L3-MW-0010	2021-03-31	3	Science Users Middleware
Science Platform Available on IDF	PREOPS-141	L3-PR-0010	2021-03-31	3	Science Platform and Reliability Engineering
PanDA based workflow system in place	PREOPS-154	L3-MW-0050	2021-03-31	3	Science Users Middleware
DP0.1 data loaded into Qserv on IDF	PREOPS-144	L3-MW-0020	2021-04-30	3	Science Users Middleware
DP0.1 QA Access: Provide access to processed images and catalogs from the IDF	PREOPS-146	L2-DP-0020	2021-05-03	2	Science Platform and Reliability Engineering
Gen3 butler and pipeline task ready for DP0 production	PREOPS-156	L3-MW-0070	2021-06-10	3	Science Users Middleware
Science Platform ready on for DP0.2	PREOPS-157	L3-PR-0040	2021-06-30	3	Science Platform and Reliability Engineering
Pipeline release for DP0.2	PREOPS-145	L3-AP-0010	2021-06-30	3	Algorithms and Pipelines
PanDA based workflow system with tooling (e.g. restart) added.	PREOPS-155	L3-MW-0060	2021-06-30	3	Science Users Middleware
Engage with the community to support shared-risk simulated data distribution to community for science with DP0	PREOPS-150	L2-SP-0020	2021-06-30	2	Community Engagement
Deliver beta LSST Data Products Documentation (DP0)	PREOPS-149	L3-CE-0010	2021-06-30	3	Community Engagement
DP0.1 Data Release: science-ready catalogs released from the IDF	PREOPS-148	L2-SP-0010	2021-06-30	2	Verification and Validation
Evaluate Batch Production System	PREOPS-153	L3-MW-0040	2021-07-31	3	Science Users Middleware
Plan for how to use IN2P3 in DP0.2	PREOPS-160	L3-EX-0010	2021-09-30	3	Execution
DP0.2 Early Access: Provide access to reprocessed images and visit level catalogs from the IDF	PREOPS-159	L2-DP-0040	2021-09-30	2	Science Platform and Reliability Engineering
DP0.2 Reprocessing Start: Begin early DRP-like reprocessing of DP0 simulated image data, at the IDF.	PREOPS-158	L2-DP-0030	2021-09-30	3	Execution
Demonstrate EPO interface with DP0	PREOPS-152	L3-PR-0030	2021-09-30	3	Science Platform and Reliability Engineering

Deploy early instantiation of service desk providing second-tier technical support for community	PREOPS-147	L3-PR-0020	2021-09-30	3	Science Platform and Reliability Engineering

2 Data Preview 0

In LSO-011 we outlined a number of scenarios for early releases of Rubin Observatory data. The purpose of the these releases are not only to prepare the community for LSST data, but also to serve as an early integration test of existing elements of the Data Management systems and to familiarize the community with our access mechanisms.

Two major new developments have occurred since LSO-011 was drafted:

- There have since been delays in construction such that we are now planning on making Data Previews with Rubin Observatory simulated data or on-sky data from other observatories (see Section 2.1.1) which would still allow us to meet some of the goals of the early releases.
- We are planning on carrying these activities at the Interim Data Facility, which is is dedicated to Pre-Ops activities infrastructure needs such as serving data and training operations staff. (Commissioning actives will continue at NCSA and in Chile.)

In this document we outline notable elements of DP0, the first of these planned data previews, from the Data Management and Pre-Operations perspective.

Data Preview 0 itself is broken down in several parts: 0.1 servings existing data products, 0.2 reprocessing that data and publishing new catalogs.

2.1 Elements of Data Preview 0.1

In this section we discuss the following key topics:

- Dataset choice considerations
- Data products offered

- Services offered
- Audience considerations

2.1.1 Dataset choice considerations

The Construction Project has been working for some time now with a number of pre-cursor datasets and simulated data. There are two leading candidates for forming the basis of DP0:

- The Subaru Hyper Suprime-Cam PDR2 dataset, provided permission can be secured from our HSC colleagues. As real (on-sky) data it is likely that users will interact with it in more realistic ways. It is a well understood dataset, and it is regularly re-processed with software that shares a common codebase with the LSST Science Pipelines.
- The simulated precursor to LSST data produced by the Dark Energy Science Collaboration, DESC DC2, provided permission can be secured. This is a very large dataset and putting DC2 catalogs in Qserv would be an excellent demonstration of its abilities.

There is interest from the science collaborations in working with data products from both of these datasets. DC2 was emphasized at the 2019 PCW, and at least one (AGN) has contributed to the simulation inputs since then. A comment at the PCW discussion was that without DC2 in DP0, the science collaborations would not see full frame LSST data until the year before the survey, too late for the needed analysis development.

Data Management is currently in transition between its 2nd and 3rd generation data abstraction layer (aka “Butler”). For DP0 to fulfill its aim as an early deployment/integration exercise, Gen 3 Butler must be used, preferably (stretch goal) using an S3 compliant Object Store as is the intent in production. This has bearing on the choice of dataset.

HSC PDR2 can either be converted from Gen 2 to Gen 3 or (stretch goal but ideally) reprocessed naively with Gen3. A smaller subset may be necessary to avoid production scaling issues. This is the preferred choice in the short term from an engineering point of view.

DC2 is available through Gen2 Butler and as we do not process that data with the Science Pipelines, the only option is conversion to Gen3. Estimates are that this is such a time-consuming

process that it cannot be done in time to meet milestone L2-DP-0020. Therefore if DC2 is to be involved in the short term, a significantly smaller subset would have to be selected.

Questions:

- Which dataset has the broader scientific interest? This question could be answered via a community survey: indeed, the possibility of such a survey was discussed at the 2019 PCW.
- For either dataset if we take a subset to avoid the Gen2-Gen3 conversion issues or production scaling issues, will that reduce the usefulness of the datasets or affect the choice? What would be the smallest data size that is still scientifically interesting?
- Are there HiPS maps available for either of these ?
- Given the delayed construction/commissioning schedule, could we consider including both of these datasets in DP0 over the course of FY21–FY22?

2.1.2 Data Products Offered

We will offer access to images and catalogs, though in more limited ways that will be available in Operations. Images will be stored in read-only Butler Gen3 repo. Catalogs will be stored in Qserv.

We may provide images and catalogs from different production runs based on the same dataset. For example, in the stretch goal of reprocessing the dataset in Gen 3, catalogs may not be available for Qserv to start ingesting in time. In such a scenario, we may choose to provide existing catalogs from the old run.

The exact science data products depend on what exist in the provided data repository (if serving existing data products) or what pipelines are ready for our reprocessing.

Questions:

- Are we offering parquet files? — No promise. Currently our SDMified parquet-generating pipelines are HSC only and Gen2 only. If parquet files are offered the access will be via the read-only Butler Gen3 repo.

- We should presumably explicitly rule out bulk download — YES. However, this (*was* discussed at the 2019 PCW, as a potential mitigation against there not being batch compute available in DP0. If a particular group requested bulk download, it could be an opportunity to start developing that capability. We will also need to know whether to allow DESC bulk download access as part of the MOU to gain access to DC2: they may well want to download all the re-processed products, for their own purposes (and to develop their capability to ingest and work with bulk downloads).
- When does ingest into Qserv has to start to be ready by DP0?

2.1.3 Services Offered

Although DP0 as a milestone described LSO-011 can be fulfilled with simple data distribution, we intend to offer limited Science Platform functionality as part of DP0. This includes:

- Provided the data is stored in Qserv or a Postgres database, catalogue access through TAP
- Access to the Science Platform's notebook-based analysis environment (Nublado); images can be accessed pragmatically via the Butler.
- Catalogue access only (no VO image services) via the Portal
- Authentication via Github (new self-service Identity Management system offering Federated Authentication will be offered subsequently to DP 0.1)

Shell access (except through Nublado) will not be offered.

The science platform will be reachable as data.lsst.cloud ("data" is specified by the Product Owner, "lsst" represents the eventual access to the Legacy Survey of Space and Time, and ".cloud" represents the GCP-deployed IDF, allowing us to bring up the USDF in parallel under a different TLD such as data.lsst.us.

2.1.4 Audience Considerations

Care should be taken to limit the target audience for the data previews; it is most critical that this is done for DP0.

- We have limited capacity to divert resources to support users.
- We will not have performed scaling tests on the Science Platform services by that point; current Science Platform usage is under 100 users, and any intent to exceed that should be communicated well in advance
- We will not yet have the ability to throttle excessive IDF usage

Authorization will be provided in an all-in basis (users will have the same level of access as project members currently have) since finer access control mechanisms will not be available by DP0; care should be taken in selecting them.

Questions:

- What is the authorization constraints for this data? For example, are DC2 data products only available to DESC science collaboration members? If so, if DC2 is chosen, does only DESC participate in DP0? **No: When agreed, DC2 would be available to all data rights holders.**
- How do we handle access? First come first served? Do we need a sign-up process?

2.2 DP0.2 - processing

The Milestone L2-DP-0040 includes re processing on IDF of the data set previously served as part of L2-DP-0020. This requires a workflow system and associated tools to preferable make this quite automated. Demonstrating a portable set of cloud enabled tools based on Butler Gen 3 and HTCondor would help to allay the main risk of moving to a new Data Facility in operations. As of today, processing based on Butler Gen3 has been limited to a very small scale, and no scalability testing has been performed. For L2-DP-0040 we may reprocess only a subset of the dataset constrained by scaling issues.

2.3 Risks and mitigation

The biggest schedule risk is not getting an interim data facility in place in time. This would delay the entire schedule and there is not much mitigation.

In the long run costs may be higher than expected in a cloud based IDF. This will be due to storage. An mitigation to this would be to store data on our own systems (NCSA or Chile) and expose it through S3. NCSA already have this in place and we should consider testing this for lesser used data sets.

There is some risk that Butler over S3 and Postgres might not be at production grade by DP0. We are working hard on that in construction. There is the possibility to run Gen 3 over a filesystem which would not be ideal on the cloud. If Gen3 does not work at all we will have to have a major rethink and build a much simpler butler. Similarly, the workflow system and associated tools may not be mature enough for large-scale production. Scalability in production is also not understood. We may need to limit the size of DP0 and rethink the system.

3 Planning and team(s) fro DP0

Planning epics have been (and are) being created in the PREOPS Jira project. On the dashboard you can see links to the tickets labeled DP0.1 and DP0.2.

We will have regular (every other week for now) DP0 meetings (see `https://confluence.lsstcorp.org/display/LSSTOps/Data+Production+Meetings`).

3.1 Teams

The Operations era org chart is shown in Figure 1.

The main departments involved in DP0 are Data Production and System Performance. With in those departments various people will be involved from the underlying teams but in small numbers. It makes most sense to approach DP0 with a task force approach. This might best be seen as two teams:

- Data production - with a focus on middleware and execution (Section 3.2);
- System Performance - with a focus on quality assurance and community support (Section 3.3).

Rubin Observatory

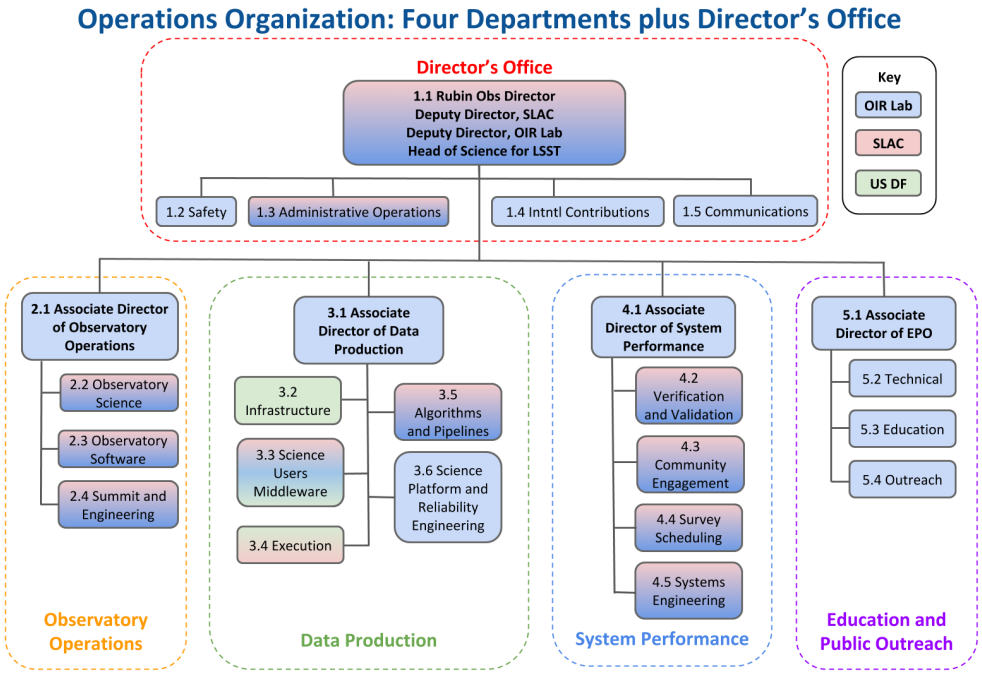

FIGURE 1: Organization of departments and teams for operations of Rubin Observatory.

As we advance the teams grow and we will transition to the an organization as in Figure 1 with team leads for each team as in Figure 2.

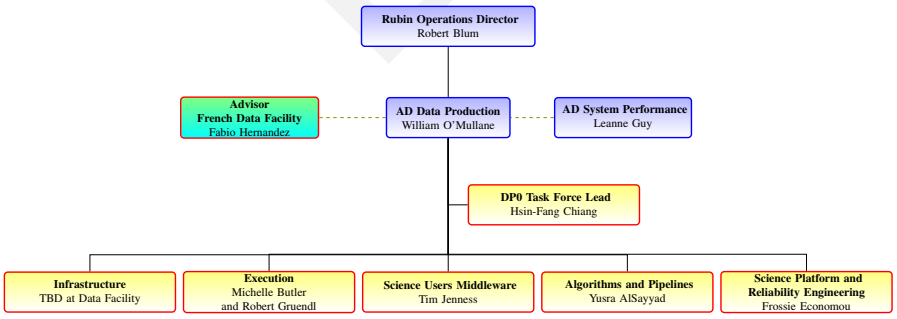

FIGURE 2: Data Production team structure

3.1.1 Task force lead

For DP0 on the IDF a task force approach seems most appropriate given the partial efforts in all teams. Hsin-Fang Chiang shall fulfill this role and coordinate Data Production activities for DP0. Responsibilities of this role include:

- Being point of contact for the IDF provider.
- Setting priorities for all work at the IDF until DP0 is fully complete.
- Evaluate stage-wise operational readiness wrt. to requirements.
- Make all components of the IDF work together (Science Platform, Middleware, Workflow ..)

The task force lead reports directly to the Data Production Associate Director and carries delegated authority for the above responsibilities.

3.2 DP Middleware and Execution

There is preops effort (fractional FTE) available in Execution and Pipelines as well as Middleware teams. The roles etc need some clean up from the ops proposal but the DP Roles are listed in Table 3 though the exact mix of roles is still under discussion.

3.3 SP Quality and Community Support

Note: DP0.1 and DP0.2 Early Access described in this document do not leave time for full-scale quality analysis. The provided data will not be science-ready; system performance milestones are succeeding.

Leanne ..

- How do we intend to do support? Slack? JIRA? CLO?

3.4 Planning

Table 2: Internal timeline

Date	Description	Reference
Jul 2020	Small test datasets identified to help dataset choice	Sec 2.1.1
Aug 2020	Decision on DP0.1 dataset	Sec 2.1.1
	Software freeze for repo conversion to Gen3 read-only Butler	L3-MW-0030
Dec 2020	Qserv installed and configured on IDF	L3-MW-0010
Jan 2021	Qserv ingestion starts on IDF	
Feb 2021	Qserv scale test	
Feb 2021	TAP service scale test	
Nov 2020	First workflow tools software release	
Jan 2021	Small test datasets available on IDF	
Jan 2021	Batch system configured on IDF	L3-MW-0050
Feb 2021	Test batch processing of the small dataset on IDF	
Apr 2021	Tract size verification run on stack candidate	
Jun 2021	Software freeze on DP0.2 pipeline stack	

Table 2 lists internal timeline.

3.4.1 Middleware

There are obvious middleware milestones such as L3-MW-0030 read only Gen3 Butler which are needed from the construction project. There is still installation work needed for the that on Google which includes the need for a Postgress (like) database for the registry. The DAX team are on the hook for this. For DP0.2 we need Butler to handle processing, not just locating files (L3-MW-0070).

3.4.1.1 Qserv should be installed and configured. Though we have some prior art for this we still will need some experimentation to get it correct. Getting DC2 loaded in Qserv is also a DAX activity we will have to do on IDF.

3.4.1.2 Workflow needs to be functioning at scale for DP0.2, ideally we should have basic workflow early on (milestone L3-MW-0050). Then more tooling such as restarting failed jobs (L3-MW-0060).

From the construction side we have BPS as a deliverable which may be useful on IDF also. We shall evaluate BPS as an option later in 2020 (L3-MW-0040). See LDM-636, LDM-633, DMTN-123. BPS translates the quantum graph to DAGMan for execution on HTCondor and submits the jobs. Most work has gone into the graph and execution.

As part of our march toward a potential more DOE oriented Data Facility, BNL will be part of the pre operations team to experiment with PanDA as an environment to monitor and control our processing jobs. This is a slightly parallel effort to construction attempting to take advantage of an existing set of tools for large scale job execution. In an ideal world the quantum graph translation of BPS would feed into a PanDA system to execute (retry etc) our jobs, this is still to be investigated. This may go through CWL.

See also Section 3.4.4.

3.4.2 Science Platform

The science platform and web services need to be deployed. In principle this is reasonable straight forward, an open issue may be configuring of the Portal aspect for the chosen dataset(s).

3.4.3 Pipelines

For DP0.2 we need a Gen3 version of the pipelines to process the dataset. This will have to run at scale for PDR2 or DC2. There may be several runs for quality purposes. Fractional FTE from the Pipelines will provide help in pipeline configuration, data repo preparation, workflow consulting, science verification, data model documenting, troubleshooting, and liaising. **Yusra will provide more info here.**

3.4.4 IN2P3

IN2P3 will contribute in Qserv and pipelines. **Fabio will provide more information here.** They bring experience running Gen3 workflows. The real interest with IN2P3 is to run remote jobs thus emulating the eventual operational DRP runs. This may be difficult to achieve in FY21 but we should make it a milestone for FY22.[1]. A more achievable goal for FY21 would be to duplicate the IDF processing at IN2P3.

Remote execution requires some features in Gen3 to be implemented. We will probably wish to execute jobs with a local registry then merge the results and registries.

[1] Tim, Fabio we should set a date for this

4 Other experiments

Apart from the milestones and planning in Section 3 there are some other activities it may be good to experiment with.

4.1 S3 access to NCSA

Storage remains the cost driver for cloud. We have an S3 interface exposing data a t NCSA, we could attempt some processing on the cloud accessing image data at NCSA.

4.2 Qserv 75% scaling

Qserv scale tests should go to 75% of DR1. This requires a lot of nodes for a short time, we do not need to necessarily keep all those nodes once the test is done. This is an ideal cloud scenario if we have Qserv working in an understood manner on the cloud. DMTN-125 would suggest we can at least do this in principle.

A References

[DMTN-123], Gower, M., Lim, K.T., 2019, *Batch Production Services Design*, DMTN-123, URL http://DMTN-123.lsst.io

[LDM-633], Kowalik, M., Gower, M., Kooper, R., 2019, *Offline Batch Production Services Use Cases*, LDM-633, URL https://ls.st/LDM-633

[LDM-636], Kowalik, M., Gower, M., Kooper, R., 2019, *Batch Production Service Requirements*, LDM-636, URL https://ls.st/LDM-636

[DMTN-125], Lim, K.T., 2019, *Google Cloud Engagement Results*, DMTN-125, URL http://dmtn-125.lsst.io

[LSO-011], William O'Mullane, L.G., Phil Marshall, 2019, *Release Scenarios for LSST Data*, LSO-011, URL https://lso-011.lsst.io

B Acronyms

Acronym	Description
AGN	active galactic nuclei
AP	Alert Production
BNL	Brookhaven National Laboratory
BPS	Batch Production Service
CWL	Common Workflow Language
DAGMan	Directed Acyclic Graph Manager
DAX	Data Access Services
DC2	Data Challenge 2 (DESC)
DESC	Dark Energy Science Collaboration
DMTN	DM Technical Note
DOE	Department of Energy
DP	Data Production
DP0	Data Preview 0
DR1	Data Release 1
DRP	Data Release Production
EPO	Education and Public Outreach
FTE	Full-Time Equivalent
FY21	Financial Year 21
GCP	Google Cloud Platform
HSC	Hyper Suprime-Cam
IDF	Interim Data Facility
IN2P3	Institut National de Physique Nucléaire et de Physique des Particules
L2	Lens 2
L3	Lens 3
LDM	LSST Data Management (Document Handle)
LSST	Legacy Survey of Space and Time (formerly Large Synoptic Survey Telescope)
MOU	Memo Of Understanding
NCSA	National Center for Supercomputing Applications
OPS	Operations
PCW	Project Community Workshop
PDR2	Public Data Release 2 (HSC)

PR	Pull Request
PanDA	Production ANd Distributed Analysis system
QA	Quality Assurance
RTN	Rubin Technical Note
S3	(Amazon) Simple Storage Service
SP	Story Point
TAP	Table Access Protocol
TLD	Top Level Domain
USDF	United States Data Facility
VO	Virtual Observatory

C Roles in Data Production FY21

These are the roles and individuals becoming active in FY21. More roles activate later as we approach operations.

Table 3: Team members for Data Production for Rubin Observatory FY21

WBS	Team	Role Title	Role Description	Institution	FY21 FTE	Staff
3.1a	Data Production Management	Associate Director for Data Production	The AD of the Data Production Department is one of the principal leaders of the Rubin Observatory operations phase. This position requires a Ph.D. level astronomer with extensive astronomical survey and science management experience, and reports directly to the Rubin Observatory Director. The primary responsibilities of this position include the management of the Data Production Department, participation in the leadership of Rubin Observatory survey, and coordination with other Rubin Observatory Departments. The AD for Data Production also has overall responsibility and authority for safely running the Rubin Observatory Data Facilities (LDF) including the generation of prompt data products (alerts) and the annual data release processing. This person will supervise a technical staff that will be responsible for all aspects of data processing, preparation of data products, archiving, and operation of the Chilean, French and other DACs, as well as the US LDF. They will be responsible for coordinating with project level Contract Management and Supplier Management when dealing with issues of business impact, and accountable for ensuring a disaster recovery plan is effective and able to be invoked. The AD of Data Production is also responsible for supervising the data flow from the Recinto to the LDF.	AURA	0.50	O'Mullane, William
3.1c	Data Production Management	Data Production Advisor - US DF	Each Rubin Observatory Data Facility (currently USA and France) have an advisory role to the AD for Data Production in terms of execution across the data facilities.	NCSA	0.35	Butler, Michelle
3.1d	Data Production Management	Data Production Advisor - IN2P3	Each Rubin Observatory Data Facility (currently USA and France) have and advisory role to the AD for Data Production in terms of execution across the data facilities. This has both a logistic and managerial element - there are local dtaff to manage but we need the entire organisation to work for prociessing.	IN2P3	0.25	Fabio Hernandez

3.2j	Infrastructure and Support	IT Network Engineer - US DF	Provides network hardware and operational functionality used from a site's border router to Rubin Observatory end equipment. Collaborates with the security engineer and also IT services related to dynamic reallocation of US DF enclaves to support these functions with network features. Supplies higher-level network services as needed at each site, such as DNS, NTP, domain name registrations, netflows, and support for security.	NCSA	0.25	Kollross, Matt
3.2k	Infrastructure and Support	Wide Area Network Technical Manager	Responsible for providing coordination amongst and managing relationships with the four independent WAN operators. Acts as the interface for services provided to the US DF in the context of the WAN. Responsible for managing the risk associated with each WAN operator, including developing mitigation strategies and proposed project responses to credible risks. Leads the Joint Wide Area Network Working Group. Well connected to DOE ESNet.	Fermilab	0.38	Demar, Phil
3.2l	Infrastructure and Support	Wide Area Network Architect	Familiar with WAN implementation technologies generally available in the networks supporting the Rubin Observatory. Familiar with technology roadmap of the ESNet WAN provider. Synthesizes and evolves network techniques and provisioning supporting the Rubin Observatory mission, as network technology evolves. Drawn from staff of WAN groups but explicitly supported by and work in the context of the Rubin Observatory.	Fermilab	0.38	Bobyshev, Andrey
3.3a	Science Users Middleware	Middleware Lead	Organizes the software maintenance effort and assigns work in a way that provides for continuity of maintenance for all Rubin Observatory maintained software. Is primarily responsible for further defining and enforcing software engineering rules related to maintenance, including maintenance of documentation, correct security practices, testing, and other aspects of delivery of a complete change set. Ensures that software tasks are consistent with authorized changes. Carries share of maintenance load. Participates in reviews.	AURA	0.25	Jenness, Tim
3.3b	Science Users Middleware	Database Engineer (Qserv) - SLAC	Develops, maintains, and implements the science Databases e.g. QSERV database, data butler, Prompt Products Database. May also work on other middleware as needed.	SLAC	0.75	Mueller, Fritz
3.3b	Science Users Middleware	Database Engineer (Qserv) - SLAC	Develops, maintains, and implements the science Databases e.g. QSERV database, data butler, Prompt Products Database. May also work on other middleware as needed.	SLAC	0.75	Gaponenko, Igor
3.3b	Science Users Middleware	Database Engineer (Qserv) - SLAC	Develops, maintains, and implements the science Databases e.g. QSERV database, data butler, Prompt Products Database. May also work on other middleware as needed.	SLAC	0.75	Pease, Nate
3.3c	Science Users Middleware	Service Software Engineer - SLAC	Develops, maintains, and implements DF software, including: Data Butler, Alert Filtering Service, orchestration software, workflow software, data backbone software, integration testing framework, authentication services, pipeline construction tools, operational fabric codes, logging, messaging, monitoring and health and status software, hosting environment for Rubin Observatory Data Space, Data Space batching services, and bulk export to other sites.	SLAC	0.75	Gates, John
3.3c	Science Users Middleware	Service Software Engineer - SLAC	Develops, maintains, and implements DF software, including: Data Butler, Alert Filtering Service, orchestration software, workflow software, data backbone software, integration testing framework, authentication services, pipeline construction tools, operational fabric codes, logging, messaging, monitoring and health and status software, hosting environment for Rubin Observatory Data Space, Data Space batching services, and bulk export to other sites.	SLAC	0.25	Hanushevsky, Andy

3.3c	Science Users Middleware	Service Software Engineer - SLAC	Develops, maintains, and implements DF software, including: Data Butler, Alert Filtering Service, orchestration software, workflow software, data backbone software, integration testing framework, authentication services, pipeline construction tools, operational fabric codes, logging, messaging, monitoring and health and status software, hosting environment for Rubin Observatory Data Space, Data Space batching services, and bulk export to other sites.	SLAC	0.25	Salnikov, Andy
3.3c	Science Users Middleware	Service Software Engineer - SLAC	Develops, maintains, and implements DF software, including: Data Butler, Alert Filtering Service, orchestration software, workflow software, data backbone software, integration testing framework, authentication services, pipeline construction tools, operational fabric codes, logging, messaging, monitoring and health and status software, hosting environment for Rubin Observatory Data Space, Data Space batching services, and bulk export to other sites.	SLAC	0.25	Lim, K-T
3.3c	Science Users Middleware	Service Software Engineer - SLAC	Develops, maintains, and implements DF software, including: Data Butler, Alert Filtering Service, orchestration software, workflow software, data backbone software, integration testing framework, authentication services, pipeline construction tools, operational fabric codes, logging, messaging, monitoring and health and status software, hosting environment for Rubin Observatory Data Space, Data Space batching services, and bulk export to other sites.	BNL	0.50	TBD
3.3c	Science Users Middleware	Service Software Engineer - SLAC	Develops, maintains, and implements DF software, including: Data Butler, Alert Filtering Service, orchestration software, workflow software, data backbone software, integration testing framework, authentication services, pipeline construction tools, operational fabric codes, logging, messaging, monitoring and health and status software, hosting environment for Rubin Observatory Data Space, Data Space batching services, and bulk export to other sites.	BNL	0.50	Padolski, Sergey
3.3d	Science Users Middleware	Data Management Software Engineer - SLAC	Maintain Rucio system which will be involved in the tracking and moving of data between multiple sites. This is in close conjunction with the Storage Engineers in the Data Facilities. Rucio is an open source HEP product which we have adopted on Rubin Observatory.	Fermilab	0.70	White, Brandon
3.3e	Science Users Middleware	Dev/ Ops Software Engineer - US DF	Maintain and improve the batch processing and data backbone services at the US DF. As we enter operations a set of tools (Pegasus, Condor, GPFS, Rucio) are used and some glue (middleware) sits between them to make the systems work. As these are upgraded the glue will need to be reshaped.	AURA	0.25	Chiang, Hsin-Fang
3.3e	Science Users Middleware	Dev/ Ops Software Engineer - US DF	Maintain and improve the batch processing and data backbone services at the US DF. As we enter operations a set of tools (Pegasus, Condor, GPFS, Rucio) are used and some glue (middleware) sits between them to make the systems work. As these are upgraded the glue will need to be reshaped.	NCSA	0.30	Gower, Michelle
3.3e	Science Users Middleware	Dev/ Ops Software Engineer - US DF	Maintain and improve the batch processing and data backbone services at the US DF. As we enter operations a set of tools (Pegasus, Condor, GPFS, Rucio) are used and some glue (middleware) sits between them to make the systems work. As these are upgraded the glue will need to be reshaped.	NCSA	0.25	Kowlik, Nic
3.3g	Science Users Middleware	Dev/ Ops Software Engineer - IN2P3	Develops, maintains, and implements DF software, including: QSERV database, data butler, DAX, Alert Filtering Service, orchestration software, workflow software, data backbone software, integration testing framework, authentication services, pipeline construction tools, operational fabric codes, logging, messaging, monitoring and health and status software, hosting environment for Rubin Observatory Data Space, Data Space batching services, and bulk export to other sites.	IN2P3	0.13	Fabrice James

3.3g	Science Users Middleware	Dev/ Ops Software Engineer - IN2P3	Develops, maintains, and implements DF software, including: QSERV database, data butler, DAX, Alert Filtering Service, orchestration software, workflow software, data backbone software, integration testing framework, authentication services, pipeline construction tools, operational fabric codes, logging, messaging, monitoring and health and status software, hosting environment for Rubin Observatory Data Space, Data Space batching services, and bulk export to other sites.	IN2P3	0.13	Sabine Elles
3.4a	Execution	Lead Production Scientist - US DF	Responsible for leading the Execution Team. This includes responsibility for managing the activities of the team members, planning work, and reporting on progress and issues to the next level of management. Additionally, the Lead Production Scientist must possess all of the skills and qualifications of a Production Scientist.	NCSA	0.50	Gruendl, Robert
3.4b	Execution	Production Scientist - US DF	Responsible for processing and database ingestion of Prompt (Alert) and Batch (Annual Data Release) Data Products. This includes responsibility for: acting as the Scientific Code Liaison, hardware and software deployment, oversight and responsibility for processing execution, prompt SDQA and response, alert filtering service operations, and external (community) broker operations.	NCSA	0.25	Adamow, Monika
3.4c	Execution	Production Scientist - SLAC	Responsible for processing and database ingestion of Prompt (Alert) and Batch (Annual Data Release) Data Products. This includes responsibility for: acting as the Scientific Code Liaison, hardware and software deployment, oversight and responsibility for processing execution, prompt SDQA and response, alert filtering service operations, and external (community) broker operations.	Fermilab	0.25	Yanny, Brian
3.4c	Execution	Production Scientist - SLAC	Responsible for processing and database ingestion of Prompt (Alert) and Batch (Annual Data Release) Data Products. This includes responsibility for: acting as the Scientific Code Liaison, hardware and software deployment, oversight and responsibility for processing execution, prompt SDQA and response, alert filtering service operations, and external (community) broker operations.	Fermilab	0.25	Lin, Huan
3.4e	Execution	Computation Facility Scientist - SLAC	Improve performance of the Data Production codes on the specific hardware of the day. This means improving computational performance and data throughput of the different pipelines. This spans all parts of the software and fits well in the middleware team, between all parts of the system.	Fermilab	0.25	Kuropatkin, Nikolay
3.4e	Execution	Computation Facility Scientist - SLAC	Improve performance of the Data Production codes on the specific hardware of the day. This means improving computational performance and data throughput of the different pipelines. This spans all parts of the software and fits well in the middleware team, between all parts of the system.	Fermilab	0.10	Tucker, Douglas
3.4e	Execution	Computation Facility Scientist - SLAC	Improve performance of the Data Production codes on the specific hardware of the day. This means improving computational performance and data throughput of the different pipelines. This spans all parts of the software and fits well in the middleware team, between all parts of the system.	Fermilab	0.15	Neilsen, Eric
3.4f	Execution	Computation Facility Scientist - IN2P3	Improve performance of the Data Production codes on the specific hardware of the day. This means improving computational performance and data throughput of the different pipelines. This spans all parts of the software and fits well in the middleware team, between all parts of the system.	IN2P3	0.50	Dominique Boutigny
3.4h	Execution	Workload Manager - US DF	Provides workload management service (a batch service and data access methods upon a hardware cluster provided by the ITC function). This work is resident at the US Rubin Observatory Data Facility and used to provision the various clusters with uniform provisioning and administrative methods. Interfaces with security policy to ensure access by authorized users and supports workflows deployed on the system and data transfers to and from the corresponding batch system.	US DF	0.25	
3.4i	Execution	Environment Manager - US DF	Maintains data management policy/ environment for release builds/ computation/ distribution.	US DF	0.25	

3.5a	Algorithms and Pipelines	Lead of Algorithms and Pipelines	Responsible for the leadership and coordination of the Algorithms and Pipelines Team, the scientific integrity of Alert Production and Data Releases, and interaction and coordination with the Lead Community Scientist, Lead Scheduler Scientist, and the Lead Production Scientist. This position requires a Ph.D. level astronomer with extensive astronomical survey and software experience, or a software engineer with extensive astronomical experience.	Princeton	0.25	Alsayyad, Yusra
3.5b	Algorithms and Pipelines	Alert Production Pipeline Group Leader	Applying extensive astronomical knowledge, including solar system, explosive transients, and time-domain surveys in general, and Rubin Observatory software experience, this role acts as product owner for the prompt processing pipelines and oversees the day-to-day work of the Alert Production Pipeline Scientists. Recommends changes to Alert Production Pipelines, and provides support to accept or reject software changes based on a scientific validation of new algorithms and an understanding of their impact on required computational resources. This position requires a Ph.D. level astronomer with extensive astronomical survey and software experience, or a software engineer with extensive astronomical experience.	UW	0.25	Bellm, Eric
3.5c	Algorithms and Pipelines	Alert Production Pipeline Scientist - NOIRLab	This role combines an understanding of one or more specific prompt processing science use cases with software engineering expertise and an understanding of the Rubin Observatory Science Pipelines to work in conjunction with the Science Software Engineering Group to modify, extend, and update the Prompt Processing Pipelines in response to emergent scientific needs, community requests, and bug reports. This role requires a Ph.D. level astronomer with extensive time-domain survey and software development experience, or a software engineer with extensive astronomical experience. Reports to the Alert Production Pipeline Group Leader.	UW	0.25	Sullivan, Ian
3.5c	Algorithms and Pipelines	Alert Production Pipeline Scientist - NOIRLab	This role combines an understanding of one or more specific prompt processing science use cases with software engineering expertise and an understanding of the Rubin Observatory Science Pipelines to work in conjunction with the Science Software Engineering Group to modify, extend, and update the Prompt Processing Pipelines in response to emergent scientific needs, community requests, and bug reports. This role requires a Ph.D. level astronomer with extensive time-domain survey and software development experience, or a software engineer with extensive astronomical experience. Reports to the Alert Production Pipeline Group Leader.	UW	0.25	Morrison, Chris
3.5c	Algorithms and Pipelines	Alert Production Pipeline Scientist - NOIRLab	This role combines an understanding of one or more specific prompt processing science use cases with software engineering expertise and an understanding of the Rubin Observatory Science Pipelines to work in conjunction with the Science Software Engineering Group to modify, extend, and update the Prompt Processing Pipelines in response to emergent scientific needs, community requests, and bug reports. This role requires a Ph.D. level astronomer with extensive time-domain survey and software development experience, or a software engineer with extensive astronomical experience. Reports to the Alert Production Pipeline Group Leader.	UW	0.25	Parejko, John

3.5d	Algorithms and Pipelines	Alert Production Pipeline Scientist - SLAC	This role combines an understanding of one or more specific prompt processing science use cases with software engineering expertise and an understanding of the Rubin Observatory Science Pipelines to work in conjunction with the Science Software Engineering Group to modify, extend, and update the Prompt Processing Pipelines in response to emergent scientific needs, community requests, and bug reports. This role requires a Ph.D. level astronomer with extensive time-domain survey and software development experience, or a software engineer with extensive astronomical experience. Reports to the Alert Production Pipeline Group Leader.	Fermilab	0.50	Herner, Ken
3.5e	Algorithms and Pipelines	Data Release Pipeline Group Leader	Applying extensive astronomical knowledge of all key Rubin Observatory science cases, including dark energy, galaxies, and stars; and of wide-field astronomical surveys in general, and Rubin Observatory software experience, this role acts as product owner for the data release processing pipelines and oversees the day-to-day work of the Data Release Pipeline Scientists. Recommends changes to Data Release Production Pipelines, and provides support to accept or reject software changes based on a scientific validation of new algorithms and an understanding of their impact on required computational resources. This position requires a Ph.D. level astronomer with extensive astronomical survey and software experience, or a software engineer with extensive astronomical experience.	Princeton	0.25	AlSayyad, Yusra
3.5f	Algorithms and Pipelines	Data Release Pipeline Scientist - NOIRLab	This role combines an understanding of one or more specific data release processing processing science use cases with software engineering expertise and an understanding of the Rubin Observatory Science Pipelines to work in conjunction with the Science Software Engineering Group to modify, extend, and update the Data Release Pipelines in response to emergent scientific needs, community requests, and bug reports. This role requires a Ph.D. level astronomer with extensive astronomical survey and software development experience, or a software engineer with extensive astronomical experience. Reports to the Data Release Pipeline Group Leader.	Princeton	0.25	Saunders, Clare
3.5f	Algorithms and Pipelines	Data Release Pipeline Scientist - NOIRLab	This role combines an understanding of one or more specific data release processing processing science use cases with software engineering expertise and an understanding of the Rubin Observatory Science Pipelines to work in conjunction with the Science Software Engineering Group to modify, extend, and update the Data Release Pipelines in response to emergent scientific needs, community requests, and bug reports. This role requires a Ph.D. level astronomer with extensive astronomical survey and software development experience, or a software engineer with extensive astronomical experience. Reports to the Data Release Pipeline Group Leader.	Princeton	0.25	Waters, Chris
3.5f	Algorithms and Pipelines	Data Release Pipeline Scientist - NOIRLab	This role combines an understanding of one or more specific data release processing processing science use cases with software engineering expertise and an understanding of the Rubin Observatory Science Pipelines to work in conjunction with the Science Software Engineering Group to modify, extend, and update the Data Release Pipelines in response to emergent scientific needs, community requests, and bug reports. This role requires a Ph.D. level astronomer with extensive astronomical survey and software development experience, or a software engineer with extensive astronomical experience. Reports to the Data Release Pipeline Group Leader.	Princeton	0.25	Taranu, Dan

3.5f	Algorithms and Pipelines	Data Release Pipeline Scientist - NOIRLab	This role combines an understanding of one or more specific data release processing processing science use cases with software engineering expertise and an understanding of the Rubin Observatory Science Pipelines to work in conjunction with the Science Software Engineering Group to modify, extend, and update the Data Release Pipelines in response to emergent scientific needs, community requests, and bug reports. This role requires a Ph.D. level astronomer with extensive astronomical survey and software development experience, or a software engineer with extensive astronomical experience. Reports to the Data Release Pipeline Group Leader.	Princeton	0.25	Kannawadi, Arun
3.5g	Algorithms and Pipelines	Data Release Pipeline Scientist - SLAC	This role combines an understanding of one or more specific data release processing processing science use cases with software engineering expertise and an understanding of the Rubin Observatory Science Pipelines to work in conjunction with the Science Software Engineering Group to modify, extend, and update the Data Release Pipelines in response to emergent scientific needs, community requests, and bug reports. This role requires a Ph.D. level astronomer with extensive astronomical survey and software development experience, or a software engineer with extensive astronomical experience. Reports to the Data Release Pipeline Group Leader.	SLAC	0.50	Meyers, Josh
3.5g	Algorithms and Pipelines	Data Release Pipeline Scientist - SLAC	This role combines an understanding of one or more specific data release processing processing science use cases with software engineering expertise and an understanding of the Rubin Observatory Science Pipelines to work in conjunction with the Science Software Engineering Group to modify, extend, and update the Data Release Pipelines in response to emergent scientific needs, community requests, and bug reports. This role requires a Ph.D. level astronomer with extensive astronomical survey and software development experience, or a software engineer with extensive astronomical experience. Reports to the Data Release Pipeline Group Leader.	SLAC	0.50	
3.5h	Algorithms and Pipelines	Lead Calibration Scientist	Together with the Calibration Support Scientist at the summit, ensures that data is available to enable proper astrometric and photometric calibration of Rubin Observatory data as part of regular pipeline processing. The Lead Calibration Scientist reports to the Lead of the Algorithms and Pipelines Team. The Lead Calibration Scientist position requires a Ph.D. level astronomer with extensive astronomical survey and Rubin Observatory software experience: it is desirable for this position to be filled by a person who contributed to the construction of the Calibration Products Production pipeline.	D4D Ltd	0.25	Fisher-Levine, Merlin
3.5i	Algorithms and Pipelines	Lead Science Software Engineer	Defines, develops and maintains the overall architecture of the Rubin Observatory scientific processing pipelines, and provides advice and to the Pipeline and Calibration Scientists to ensure that the overall Rubin Observatory Science Pipelines form a coherent whole. Provides leadership to the Science Software Engineering Group, and reports to the Lead of Algorithms and Pipelines. Requires an expert in software architecture with extensive expertise in scientific software design in general and in the Rubin Observatory software system in particular.	Universities	0.25	Jim Bosch
3.5j	Algorithms and Pipelines	Science Software Engineers - NOIRLab	Provides software engineering support to the Alert Production, Data Release, and Calibration Groups. Works with the pipeline scientists on general code development, to help keep code maintainable and optimized, and reports to the Lead Science Software Engineer.	Princeton	0.25	Lust, Nate
3.5k	Algorithms and Pipelines	Science Software Engineer - SLAC	Provides software engineering support to the Alert Production, Data Release, and Calibration Groups. Works with the pipeline scientists on general code development, to help keep code maintainable and optimized, and reports to the Lead Science Software Engineer.	SLAC	0.50	Wittgen, Matthias

3.5l	Algorithms and Pipelines	Science Software Consultant	Supports and reports to the Lead Science Software Engineer. The Science Software Consultant position is a senior software position filled by fractions of individuals with deep and intimate knowledge of the Rubin Observatory Data Management system, presumably from the construction period.	Princeton	0.25	Lupton, Robert
3.6a	Science Platform and Reliability Engineering	Technical Lead/ Manager	Responsible for technical leadership and management of the Science Platform and Reliability Engineering Team. This includes running stand ups and looking after budgets and staff issues as well as making technical calls where decisions are needed.	AURA	0.50	Economou, Frossie
3.6b	Science Platform and Reliability Engineering	DevOps Infrastructure Engineer - NOIRLab	Generalist software engineers who work through the entire software stack. A DevOps engineer must be able to understand the software and infrastructure enough to know it is working well. They must also be able to improve the infrastructure and debug problems which can span hardware, network and operating system all the way to the end user delivered service.	AURA	0.25	Sick, Jonathan
3.6b	Science Platform and Reliability Engineering	DevOps Infrastructure Engineer - NOIRLab	Generalist software engineers who work through the entire software stack. A DevOps engineer must be able to understand the software and infrastructure enough to know it is working well. They must also be able to improve the infrastructure and debug problems which can span hardware, network and operating system all the way to the end user delivered service.	AURA	0.25	Allbery, Russ
3.6b	Science Platform and Reliability Engineering	DevOps Infrastructure Engineer - NOIRLab	Generalist software engineers who work through the entire software stack. A DevOps engineer must be able to understand the software and infrastructure enough to know it is working well. They must also be able to improve the infrastructure and debug problems which can span hardware, network and operating system all the way to the end user delivered service.	AURA	0.25	
3.6c	Science Platform and Reliability Engineering	DevOps Infrastructure Engineer - US DF	Generalist software engineers who work through the entire software stack. A DevOps engineer must be able to understand the software and infrastructure enough to know it is working well. They must also be able to improve the infrastructure and debug problems which can span hardware, network and operating system all the way to the end user delivered service. At least one of these engineers will have special competence in cybersecurity issues, and will ensure that Data Production services are developed and managed in accordance with Rubin Observatory cybersecurity policy.	US DF	0.38	
3.6c	Science Platform and Reliability Engineering	DevOps Infrastructure Engineer - US DF	Generalist software engineers who work through the entire software stack. A DevOps engineer must be able to understand the software and infrastructure enough to know it is working well. They must also be able to improve the infrastructure and debug problems which can span hardware, network and operating system all the way to the end user delivered service. At least one of these engineers will have special competence in cybersecurity issues, and will ensure that Data Production services are developed and managed in accordance with Rubin Observatory cybersecurity policy.	US DF	0.38	
3.6d	Science Platform and Reliability Engineering	Science Platform Engineer	Maintains and develops the Science Platform (User Services) software [e.g. Jupyter notebooks, etc].	AURA	0.25	Banek, Christine
3.6d	Science Platform and Reliability Engineering	Science Platform Engineer	Maintains and develops the Science Platform (User Services) software [e.g. Jupyter notebooks, etc].	AURA	0.25	Thornton, Adam
3.6e	Science Platform and Reliability Engineering	Data Exploration Developer	Communicates data exploration needs to the engineers, documents and develops tools, demonstrates how to achieve scientific goals with the tools provided. This would explicitly include technical support to the Community and EPO scientists. These are specialist software engineers with science backgrounds who can cater to the scientific needs of the users.	AURA	0.25	Krughoff, Simon

3.6e	Science Platform and Reliability Engineering	Data Exploration Developer	Communicates data exploration needs to the engineers, documents and develops tools, demonstrates how to achieve scientific goals with the tools provided. This would explicitly include technical support to the Community and EPO scientists. These are specialist software engineers with science backgrounds who can cater to the scientific needs of the users.	AURA	0.25	Fausti, Angelo
3.6f	Science Platform and Reliability Engineering	Data Visualization Engineer	Documents and develops data visualization tools, demonstrates how to achieve scientific goals with the tools provided, produces example visualizations. Includes technical support in data visualization to the Community and EPO scientists. These are specialist software engineers with science backgrounds who can cater to the scientific needs of the users.	SLAC	0.25	Kaehler, Ralf