Rubin Observatory

Interim Data Facility & Data Preview Zero

William O'Mullane and Hsin-Fang Chiang

NSF NOIRLab AURA SLAC U.S. DEPARTMENT OF ENERGY

Two topics in this session

Data Preview Zero

Working in Interim Data Facility

Acronym Definitions Available Here

Rubin Observatory

Data Preview Zero

Data Previews ..

Rubin Observatory

- Original idea - ship some commissioning data for the community to look at.
  - Potentially files on a server
- But several things happened over time
  - We really wanted to provide more of a platform than just files
    - LSO-011 outlined a number of scenarios for early releases of commissioning data.
  - ComCam and Sumit delays have hit
  - USDF location became uncertain
- Hence we changed ideas a little and we got an Interim Data Facility
- RTN-001 provides more details on the Data Preview 0 (DP0)

Data Preview 0 - Purpose

- Early integration test of existing elements of the Data Management systems
  - Familiarization of new Rubin staff with operation of Rubin software
  - Still time for feedback to development

- Familiarize the community with our access mechanisms.
  - This is NOT a finished system more like a construction site ..
  - DP0 access for a limited number of science community delegates
    - More on this later from Leanne, Community Engagement session

- Prepare the community for Rubin data
  - Again restricted set

Data Preview 0 - Dataset

- We selected DESC DC2 (MOU pending completion)
  - This is a large dataset
  - Putting DC2 catalogs in Qserv will be an excellent demonstration of its abilities.
  - We may use a subset of the 300deg^2 5-yr WFD
  - What science data products will be included is TBD
  - Douglas Tucker (FNAL) is the main contact

- DP0.1 will serve the existing products
  - Bulk download would not be available
  - All experimentation would be via the Science Platform
- DP0.2 will serve reprocessed products.
  - Gen3 DM pipelines will be used
  - Similarly, the access would be via the Science Platform

Data Preview 0 - Services

- Catalogue will be stored in Qserv and accessed through TAP.
- Users will have access to the Science Platform's notebook-based analysis environment (Nublado)
- Images can be accessed via a read-only Butler.
- Federated Authentication - though may be GitHub Org based

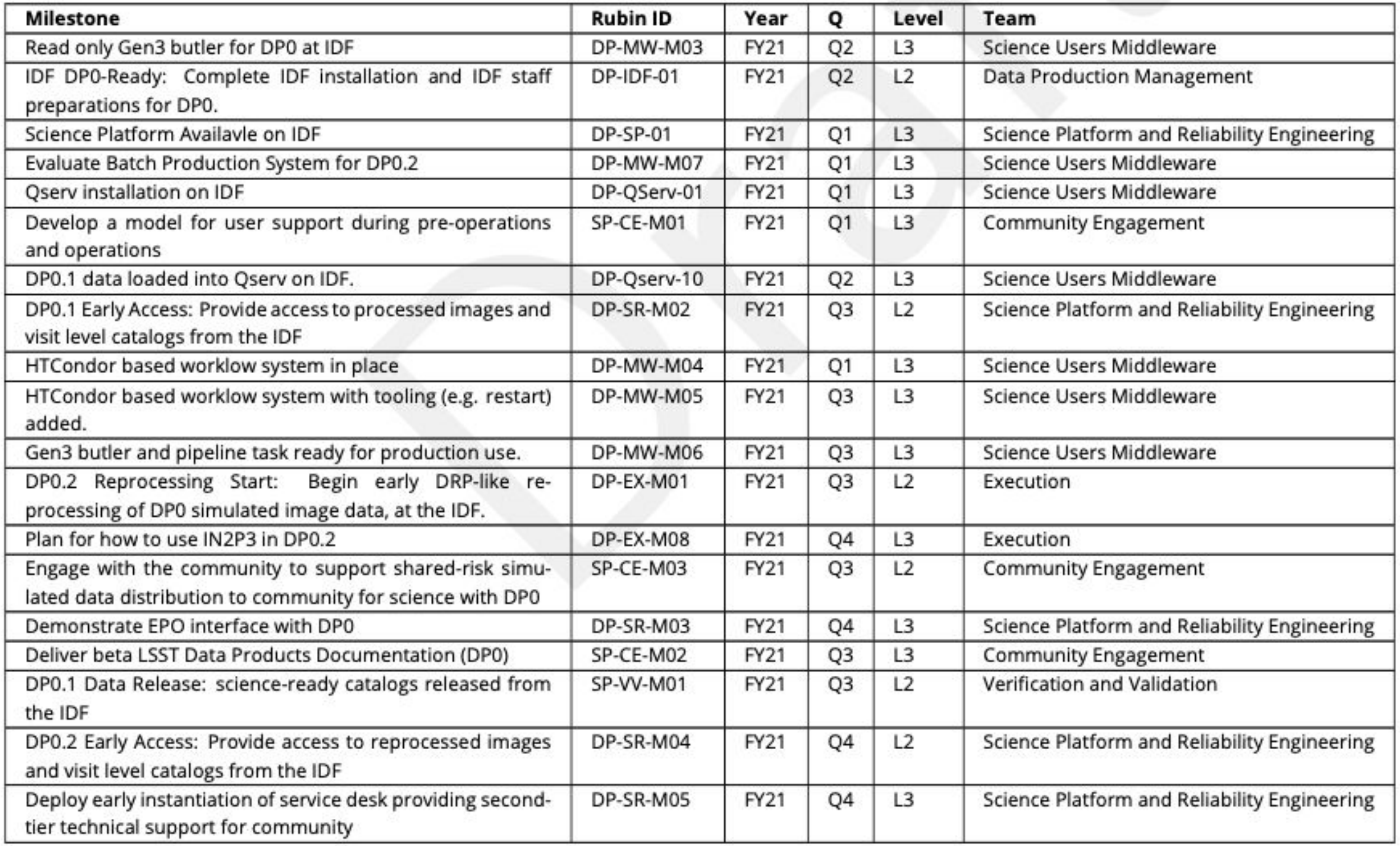
- Some stretch goals not promised in DP0 include: Portal Aspect, user batch compute

Data Preview 0 - Timeline

IDF

DP0.1

DP0.2



Milestone	Rubin ID	Year	Q	Level	Team
Read only Gen3 butler for DP0 at IDF	DP-MW-M03	FY21	Q2	L3	Science Users Middleware
IDF DP0-Ready: Complete IDF installation and IDF staff preparations for DP0.	DP-IDF-01	FY21	Q2	L2	Data Production Management
Science Platform Availavle on IDF	DP-SP-01	FY21	Q1	L3	Science Platform and Reliability Engineering
Evaluate Batch Production System for DP0.2	DP-MW-M07	FY21	Q1	L3	Science Users Middleware
Qserv installation on IDF	DP-QServ-01	FY21	Q1	L3	Science Users Middleware
Develop a model for user support during pre-operations and operations	SP-CE-M01	FY21	Q1	L3	Community Engagement
DP0.1 data loaded into Qserv on IDF.	DP-Qserv-10	FY21	Q2	L3	Science Users Middleware
DP0.1 Early Access: Provide access to processed images and visit level catalogs from the IDF	DP-SR-M02	FY21	Q3	L2	Science Platform and Reliability Engineering
HTCondor based worklow system in place	DP-MW-M04	FY21	Q1	L3	Science Users Middleware
HTCondor based worklow system with tooling (e.g. restart) added.	DP-MW-M05	FY21	Q3	L3	Science Users Middleware
Gen3 butler and pipeline task ready for production use.	DP-MW-M06	FY21	Q3	L3	Science Users Middleware
DP0.2 Reprocessing Start: Begin early DRP-like re-processing of DP0 simulated image data, at the IDF.	DP-EX-M01	FY21	Q3	L2	Execution
Plan for how to use IN2P3 in DP0.2	DP-EX-M08	FY21	Q4	L3	Execution
Engage with the community to support shared-risk simulated data distribution to community for science with DP0	SP-CE-M03	FY21	Q3	L2	Community Engagement
Demonstrate EPO interface with DP0	DP-SR-M03	FY21	Q4	L3	Science Platform and Reliability Engineering
Deliver beta LSST Data Products Documentation (DP0)	SP-CE-M02	FY21	Q3	L3	Community Engagement
DP0.1 Data Release: science-ready catalogs released from the IDF	SP-VV-M01	FY21	Q3	L2	Verification and Validation
DP0.2 Early Access: Provide access to reprocessed images and visit level catalogs from the IDF	DP-SR-M04	FY21	Q4	L2	Science Platform and Reliability Engineering
Deploy early instantiation of service desk providing second-tier technical support for community	DP-SR-M05	FY21	Q4	L3	Science Platform and Reliability Engineering

Reminder FY21 starts Oct 2020.

Reference: rtn-001.lsst.io

Rubin Observatory

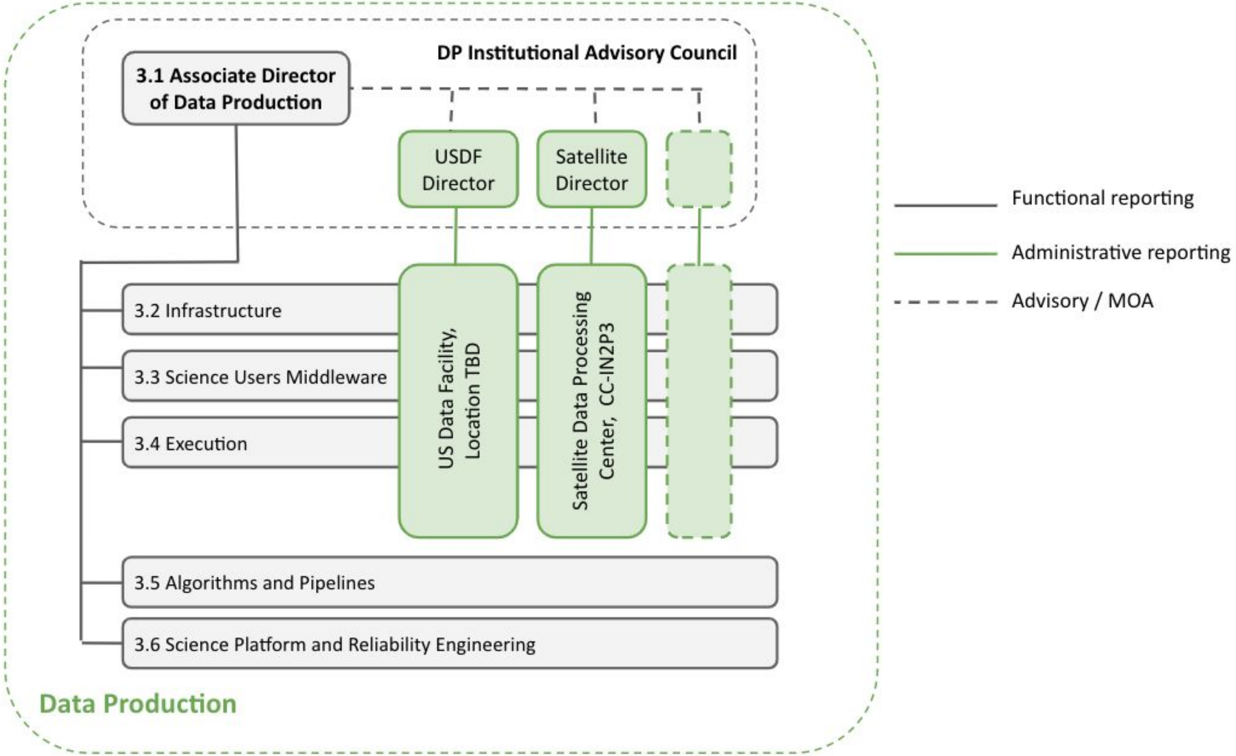
Interim Data Facility

Rubin Operations Data Facilities

Rubin Observatory

Multiple data facilities, including the USDF, will form constituent parts of one integrated Data Production Department.

- Satellite Data Processing Center at CC-IN2P3 will perform 50% of annual data release processing.
- We are discussing a UK DF taking on up to 25% of DRP


DP Institutional Advisory Council
3.1 Associate Director of Data Production
USDF Director
Satellite Director
3.2 Infrastructure
3.3 Science Users Middleware
3.4 Execution
US Data Facility, Location TBD
Satellite Data Processing Center, CC-IN2P3
3.5 Algorithms and Pipelines
3.6 Science Platform and Reliability Engineering
Data Production
Functional reporting
Administrative reporting
Advisory / MOA

Cloud Experiments

Google

- Initial cost estimates based on a simplified sizing model (DMTN-072) formed basis. Report in DMTN-125 — highlights include:
- Deployed Qserv on Google with reasonable performance (80% or better of in-house)
- Data transfer adequate for Prompt Processing demonstrated, within the limits of the available network. Prompt Product Database stood up and tested.
- Science Platform deployed and users simulated.
- A second POC with HTcondor processing and networking is concluding - DMTN-157

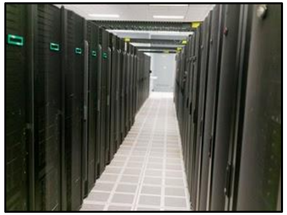
Amazon

- Processing with Amazon Web Services / Elastic Compute Cloud and HTCondor. Led by Hsin-Fang Chiang (AURA). See DMTN-114 for setup. Report in DMTN-135.
- Demonstrated HSC data processing on Amazon, integrated with their S3 object storage system;
- Dino Bektesevic (UW) continued to refine this showing higher efficiency
- Offered tutorial at Data Inclusion Revolution meeting in Boston, November 2019

Interim Data Facility

Rubin Observatory

- To mitigate the risks of — and the delay imposed by — the USDF selection process, we have set up an Interim Data Facility (IDF).
- Cloud hosting seen as best option, on the basis of maximising flexibility and minimising investment in hardware and new staff.
- Cloud contract has been tendered, Google were selected as a provider for 3 years of service.
- Now we have an IDF. When we know the US DF (Ops facility) we will plan transition, then ramp to full ops readiness there (FY23).
- IDF will provide data management services primarily in support of Data Preview releases and other pre-operations activities.
- Constructions & Commissioning activities continue at NCSA and in Chile.



IDF Current Status

- Our collaboration with Google is transitioned from POC (constructions) to IDF (operations)
- A working group in Data Production department has been discussing services for early handover to operations and for defining resource organization, security access model, quota, billing, etc.
- Next week we will select partners.
- With help from Google team and partners, we will then do onboarding, training and migration.
- Timeline: Nov

Google Team

Jess Masciarelli - Rubin Observatory Account Manager
jmasciarelli@

Responsible for account health, assisting with strategic direction, evaluations and coordinating Google resources.

Google Cloud

Dr. Ross Thomson - Solutions Architect
jrossthomson@

Flora Huang - Sr. Customer Engineer
huangflora@

Sarosh Naseem
EDU W. Region Manager
snaseem@

Len Zheleznyak - Cloud Value Advisor
lenzheleznyak@

Dr. Karan Bhatia - Sr. Cloud Specialist
karanbhatia@

Jesus Trujillo Gomez - Strategic Business Advisor
jesusgomez@

Google Cloud Platform 101

- GCP provides a wide range of resources, services, and tools.
- Resources are hosted in different regions/zones, e.g. us-west1/us-west1-a
- Some quotas are regional.

Compute

- Compute Engine
- Kubernetes Engine
- Cloud Function

Databases

- Cloud SQL
- Cloud Bigtable
- Cloud Spanner

Storage

- Cloud Storage
- Persistent Disk

Tools

- Cloud Logging
- Cloud Monitoring

Working in the IDF

Rubin Observatory

We expect *MOST* IDF users to interact with the Science Platform

- This provides some level of command line access
- Most work/experiments can be done in notebooks.
- Tomorrow 12:00 Science Platform Overview by Frossie

A limited set of developers will work directly in GCP

- Only those working on the services require GCP access.
- Care should be taken we shall be billed for all usage.
- Egress is not free. Stop unused resources. Rightsizing.
- Resources will be organized hierarchically.
- The right "project" should be used. Ask your team leads.

Google Cloud Platform demo

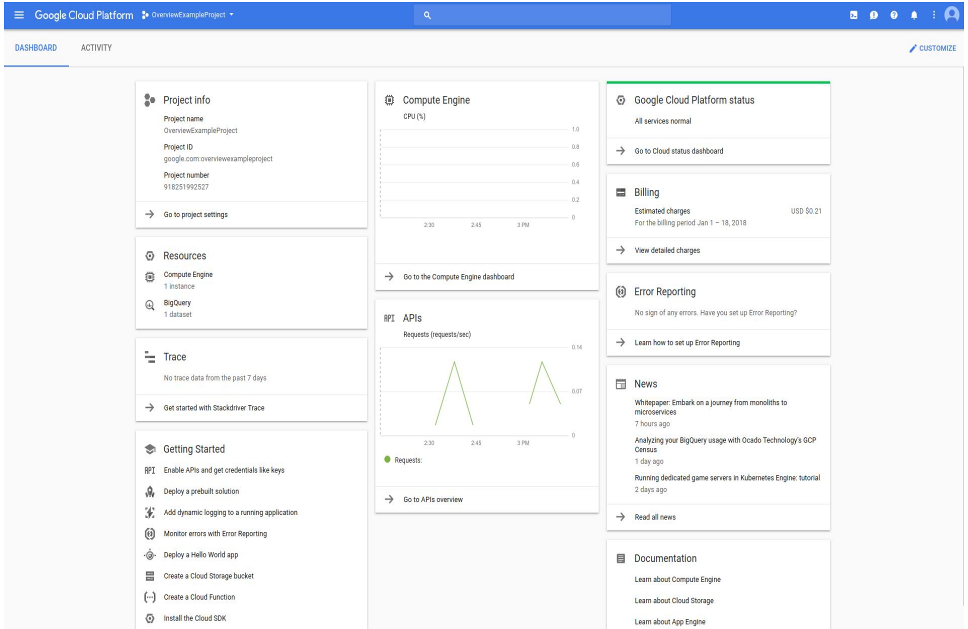
Rubin
Observatory

Google Cloud can be accessed via

- `gcloud` command line interface
- Web-based Google Cloud console.

In this demo we will

- Show basic navigation in the console
- Create a Virtual Machine instance
- ssh into the VM
- Send a message to system log and find it in Google Logging
- Delete the VM when done



Project info

Project name
OverviewExampleProject

Project ID
google.com:overviewexampleproject

Project number
918251992527

Go to project settings

Resources

Compute Engine
1 instance

BigQuery
1 dataset

Trace

No trace data from the past 7 days

Get started with Stackdriver Trace

Getting Started

Enable APIs and get credentials like keys
Deploy a prebuilt solution
Add dynamic logging to a running application
Monitor errors with Error Reporting
Deploy a Hello World app
Create a Cloud Storage bucket
Create a Cloud Function
Install the Cloud SDK

Compute Engine

CPU (%)

Go to the Compute Engine dashboard

APIs

Requests (requests/sec)

Requests:

Go to APIs overview

Google Cloud Platform status

All services normal

Go to Cloud status dashboard

Billing

Estimated charges USD $0.21
For the billing period Jan 1 – 18, 2018

View detailed charges

Error Reporting

No sign of any errors. Have you set up Error Reporting?

Learn how to set up Error Reporting

News

Whitepaper: Embark on a journey from monoliths to microservices
7 hours ago

Analyzing your BigQuery usage with Ocado Technology's GCP Census
1 day ago

Running dedicated game servers in Kubernetes Engine: tutorial
2 days ago

Read all news

Documentation

Learn about Compute Engine
Learn about Cloud Storage
Learn about App Engine