# Rubin Science Platform on Google: the story so far.

William O'Mullane,[1] Frossie Economou,[1] Flora Huang,[2] Dan Speck,[3] Hsin-Fang Chiang,[1] Melissa L. Graham,[4] Russ Allbery,[1] Christine Banek,[1] Jonathan Sick,[1] Adam J. Thornton,[1] Jess Masciarelli,[2] Kian-Tat Lim,[5] Fritz Mueller,[5] Sergey Padolski,[6] Tim Jenness,[1] K. Simon Krughoff,[1] Michelle Gower,[7] Leanne P. Guy,[1] and Gregory P. Dubois-Felsmann[8]

[1] *Rubin Observatory Project Office, 950 N. Cherry Ave., Tucson, AZ 85719*

[2] *Google Inc., 1600 Amphitheatre Parkway, Mountain View, CA 94043*

[3] *Burwood Group, 125 South Wacker Drive Suite 2950, Chicago, IL 60606*

[4] *University of Washington, Dept. of Astronomy, Box 351580, Seattle, WA 98195*

[5] *SLAC National Accelerator Laboratory, 2575 Sand Hill Rd., Menlo Park, CA 94025*

[6] *Brookhaven National Laboratory, Upton, NY 11973*

[7] *NCSA, University of Illinois at Urbana-Champaign, 1205 W. Clark St., Urbana, IL 61801*

[8] *IPAC, California Institute of Technology, MS 100-22, Pasadena, CA 91125*

**Abstract.**
We describe Rubin Observatory's experience with offering a data access facility (and associated services including our Science Platform) deployed on Google Cloud infrastructure as part of our pre-Operations Data Preview program.

## 1. Introduction

The Legacy Survey of Space and Time (Ivezić et al. 2019) is "deep wide fast" optical/near-IR survey of half the sky in ugrizy bands to r 27.5 (36 nJy) based on 825 visits over a 10-year period. Carried out by Rubin observatory on Cerro Pachón Chile, the survey will produce around 100 PB of data consisting of about a billion 16 Mpix images, enabling measurements for 40 billion objects!

As a big data project, Rubin is providing a "bring-your-compute-to-the-data" model relying on a project-provided Rubin Science Platform (hereafter RSP) to allow science users to access, visualize and analyze survey data products. This paper describes our experience with an early release of the RSP to users through our pre-operations program.

## 2.  The Rubin Interim Data Facility on Google Cloud

We decided to deploy services involved in our pre-operations program on a commodity cloud services provider as an Interim Data Facility (IDF) to mitigate uncertainty in the final location of our operations-era US Data Facility (USDF). Following a competitive tender, we selected Google Cloud for this three-year activity to prepare our team for operational readiness and our community for interacting with Rubin data and services. The IDF is a full stand-alone facility (without dependance on on-premises computing) and is a real high-availability production environment for all services currently oriented to external users. These include the Rubin Science Platform as well as underlying services such as Qserv (our high performance database system) and Butler (data abstraction middleware).

The architecture is depicted in Figure 1 and leverages the Google Kubernetes Engine (GKE) and Google Cloud Storage (GCS) to scale up and down as needed depending on demand. Kubernetes is our reference deployment platform both on-premises and in the Cloud as our services make heavy use of container orchestration patterns and consume containers produced by our build systems based on Jenkins and GitHub Actions (see e.g., Jenness et al. 2018).

Management of Google Cloud resources is fully captured in code (Terraform) developed by the Burwood Group and no production services rely on ad-hoc interactive use of the Google console. We use continuous deployment techniques to manage our services including tools such as ArgoCD[1] even though we don't practice pure continuous deployment, having instead settled on a gated release model (Economou & Allbery 2021).
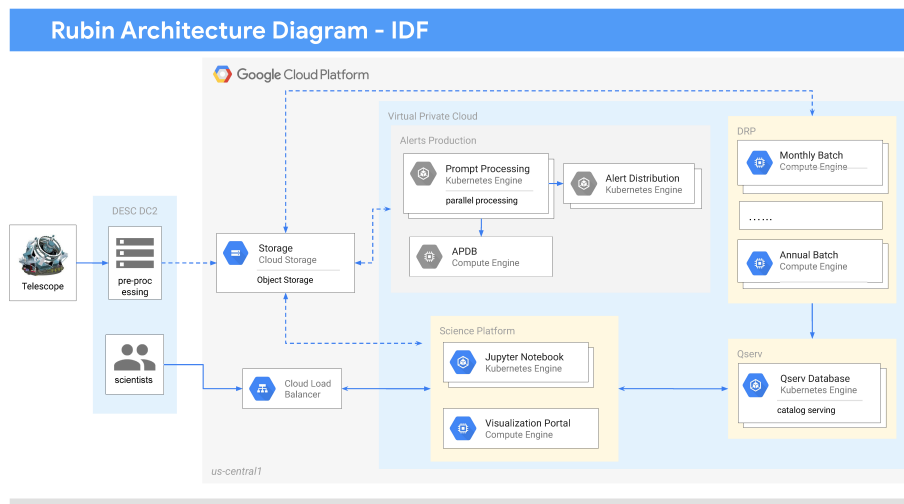


Figure 1.     Architecture of the Interim Data Facility on Google, gray Items (alerts) are not yet implemented but shall be in 2022.

---

[1]`https://argoproj.github.io/cd/`

### 3.   The Pre-operations Data Preview programme

The IDF is hosting what is the first of a series of Data Previews (DP). DP0 is using simulated data (as the telescope has not yet achieved First Light), DP1 will offer commissioning camera data, and DP2 LSSTCam commissioning data. DP0 currently has two phases(O'Mullane 2020) with one focussing on data access and the other on data processing readiness. In June 2021 data.lsst.cloud was opened up to the first 200 delegates selected as part of Data Preview 0.1 with data from the DC2 simulated dataset provided by the DESC collaboration (Abolfathi et al. 2021). For Data Preview 0.2 we shall reprocess the DESC data published "as is" in DP0.1 with the LSST Science Pipelines v23, to generate a fully self-consistent data release that matches the LSST data model. This will entail producing many types of the final data products and catalogs as well as loading the generated catalogs in Qserv and will demonsrate readiness (and on-prem/cloud portability) for a number of systems including our PanDA-based processing system and the Generation 3 version of the Butler (Jenness et al. 2019).

### 4.   Community Engagement

Besides the technical and operational goals, a key aim of deployment of the Interim Data Facility was the engagement of the future community of scientists and students by providing early access to "DP0 delegates",a group of users representing the broad science community as learners, testers, and providers of feedback, and tasked with sharing the benefits of their DP0 participation with their communities.

Besides extensive documentation at dp0-1.lsst.io supporting delegates in a number of ways, delegate resources included biweekly virtual "Delegate Assemblies" with hands-on demonstrations and breakout rooms to facilitate co-working; a dedicated category in the Rubin Community Forum (`https://community.lsst.org`) (based on the Discourse web forum platform) where delegates can ask questions and discuss their DP0 work; and a GitHub repository where delegates can contribute to shared code and notebooks. GitHub Issues were also used as the primary means for delegates to request technical assistance. All of the delegate pedagogical resources are publicly accessible.

Some aspects of the Community Engagement model as applied to DP0 are designed to scale to thousands of users for the future Data Previews, such as the extensive documentation, the tutorials, and the use of the Community Forum. Other aspects, such as the frequent live virtual sessions, may not scale well to thousands of users and were specifically designed to ramp up skill levels from novice to intermediate in order to seed expertise with Rubin software in the community. For future Data Previews, the Community Engagement strategy will focus on building infrastructure to foster a vibrant science community, that enables self-help, peer-to-peer help and crowdsourced support.

### 5.   Cloud as a primary platform for astronomical data services

The successful launch of our Data Preview program with our cloud-based Science Platform reinforced our view that commodity infrastructure offers unparalleled technical and programmatic advantages in delivering large-scale data services to astronomers. Some key points:

- Architecting systems using cloud-native techniques (such as targetting Kubernetes as the deployment platform) allows painless transition between commodity and on-premises infrastructure.

- There are tangible benefits to working with a highly popular toolchain both for a project and its developers, including the ease of contracting short-term bootstrapping help as Rubin did with the Burwood Group.

- Developers love the self-serve aspect of cloud infrastructures (and their velocity reflects this).

- Use of commercial cloud environments offers an ideal way to mitigate schedule and technical risk in the deliver of on-premises computing.

- There are significant security advantages in placing public users in a separate security domain from on-premises resources.

- We are seriously evaluating whether an on-prem/cloud hybrid model is actually the best way forward permanently for Rubin Operations.

While use of these services are not cheap (despite academic and government discounts that can be had for the asking), they offer great value for money (and are frequently disadvantaged in comparisons with on-premises cost by the lack of total cost accounting of in-house facilities). Management buy-in is a big part of any cloud adventure; in our case we were helped by our proof of concept studies with both Amazon (Bektesevic et al. 2020; Chiang et al. 2020) and Google (Lim 2019). Each ran for six months and demonstrated we could deploy and operate on either platform. This allayed concerns about vendor lock-in and improved our cost models.

We believe use of commodity cloud infrastructure will allow scientific facilities to focus on fertile domain-relevant middleware development instead of on-premises compute facilities and we hope that the funding landscape can evolve to facilitate this paradigm shift.

**References**

Abolfathi, B., et al. 2021, The Astrophysical Journal Supplement Series, 253, 31. URL `http://dx.doi.org/10.3847/1538-4365/abd62c`

Bektesevic, D., Chiang, H.-F., Lim, K.-T., Miller, T. L., Thain, G., Jenness, T., Bosch, J., Salnikov, A., & Connolly, A. 2020, arXiv e-prints, arXiv:2011.06044. `2011.06044`

Chiang, H.-F., Bektesevic, D., & the AWS-PoC team 2020, AWS Proof of Concept Project Report. URL `http://DMTN-137.lsst.io`

Economou, F., & Allbery, R. 2021, Guidelines for gated updates for SQuaRE services. URL `http://sqr-056.lsst.io`

Ivezić, Ž., et al. 2019, ApJ, 873, 111. `0805.2366`

Jenness, T., et al. 2018, in Software and Cyberinfrastructure for Astronomy V, vol. 10707 of Proc. SPIE, 1070709

Jenness, T., et al. 2019, in ADASS XXVII, edited by P. J. Teuben, M. W. Pound, B. A. Thomas, & E. M. Warner, vol. 523 of ASP Conf. Ser., 653. `1812.08085`

Lim, K.-T. 2019, Google Cloud Engagement Results. URL `http://dmtn-125.lsst.io`

O'Mullane, W. 2020, Data Preview 0: Definition and Planning. URL `http://RTN-001.lsst.io`