Alternative science operations approach for LSST

William O'Mullane

2019-04-29

1 Introduction

We are currently experimenting with Google and Amazon services for science platform and processing. These services are priced to deliver compute and storage - our current model at the LDF is also service oriented but is not proceed in the same manner making comparisons difficult. It has been difficult to get an alternative cost model together. An initial approach to a cloud costing was outlined in DMTN-072, this approach was to try to cost the hardware and compare to cloud pricing.

In this document a radical restructuring of LSST operations is explored - a technology stack underpinned by commodity services which could be provided by commercial providers or computing centers. Here we look first at how we would run something like this -we can then leave one free variable which is the cost of the underlying compute and storage services. This will both help to sanity check the LDF costing and potentially allow us to have a ball park for assessing commodity provider offers.

2 Plugable service oriented architecture

In the Kavli workshop in Vegas (Feb 2019) we took a long term view to astronomy archives and data processing. We suggested a layered service model as depicted in Figure 1[1]. Our requirements are no longer unique and we have access to a wealth of open source software, commodity hardware, and managed cloud services (offered by commercial providers and federally-funded institutions) that are well positioned to meet the needs of LSST Momcheva et al. (2019); Bektesevic et al. (2019).

We took Figure 1 and made a more LSST oriented version in Figure 2. This is pretty close to how we are currently but we do not treat the compute and storage as pure services.

[1]The full document is here `https://petabytestoscience.github.io/PetaBytes-2019-04-26.pdf`

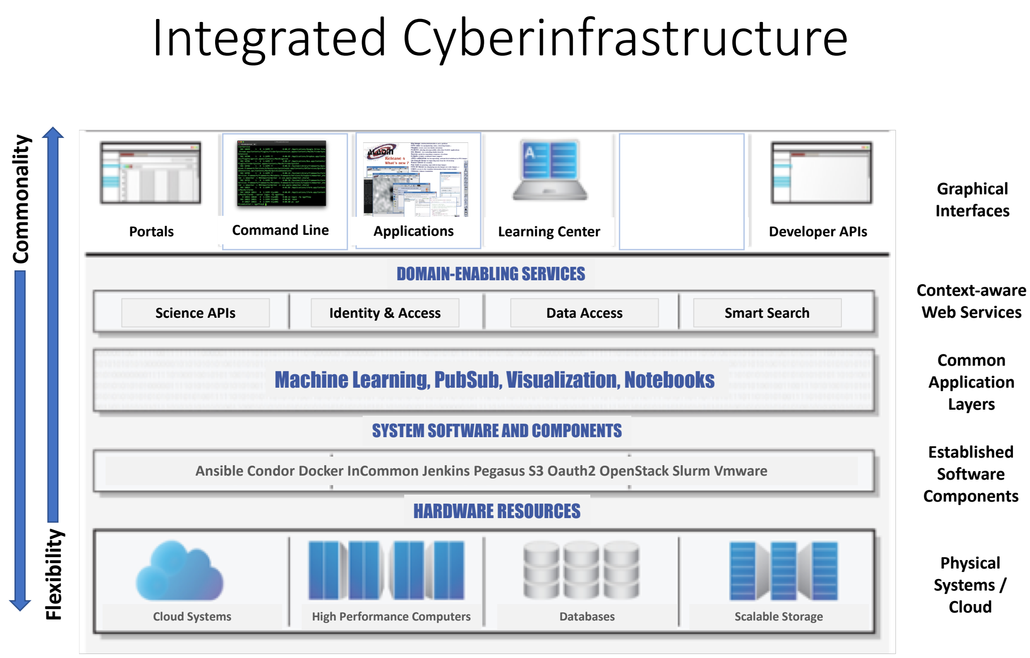

FIGURE 1: An example a cyberinfrastructure built on an Infrastructure as Code design model. Note that while this example does not have astronomy-specific tooling, our recommendations highlight the importance of developing astro-specific layers that are fully accessible to scientists in both the application and the graphical interface layers.

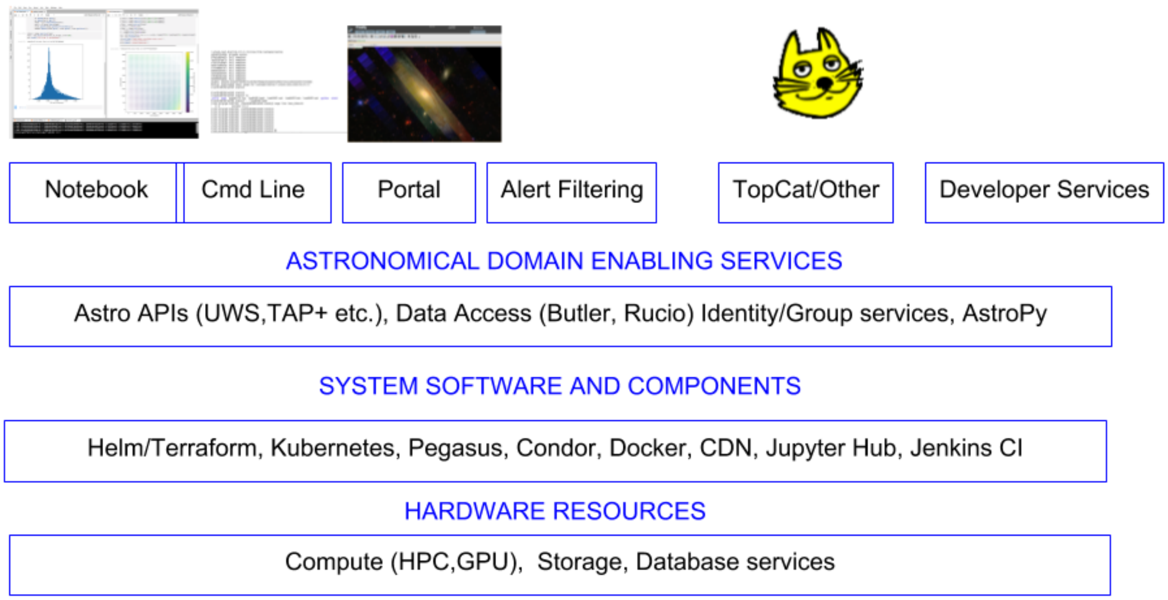

FIGURE 2: An example LSST cyberinfrastructure built analogous to the CI model shown in Figure 1.

3 Data Production Department

The role of data production within LSST is to deliver LSST's science products: the science images, the alert stream, the annual data releases, the science software, and the Science Platform. In the current ops proposal not all groups required to do this are under control of Science operations.

Figure **??** gives a view of the Data Production teams which combine some of the old science operations and LDF departments. This is far more analogous to Data Management moving into operations than in the current proposal and would make for a smoother transition.

The FTE counts are estimated in Table **??**.

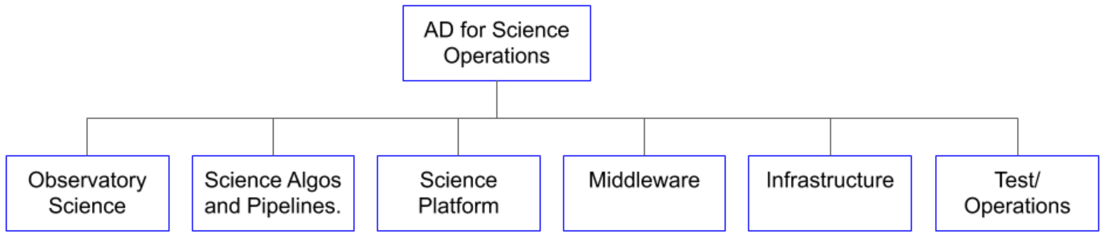

FIGURE 3: Possible configuration of Science Operations Department for operations of LSST

3.1 Teams

Figure 3 introduces several teams some of which were not in the original ops proposal. I little detail is given here about each.

3.1.1 Observatory Science

As in the ops proposal the primary responsibility of this team is to understand the end-to-end impact of the Observatory hardware and environment on the science images and to work with the Observatory Operations department to ensure that the image quality meets requirements.

This team may be better in Observatory operations department

3.1.2 Science algorithms and pipelines

This team is responsible to assess and assure the alert stream and annual data releases. This includes extensive QA to compare the data products against requirements, and responsibility for the underlying software pipelines themselves. Monitoring and updating the calibration plan and algorithmic implementation is also a responsibility of this team. The Calibration Support Scientist on the Observatory Science team will be responsible for monitoring the physical implementation of the calibration plan at the summit.

3.1.3 Science platform

This team will be responsible for maintaining and evolving LSST's user access portal, the Science Platform. This will include keeping up with evolving technologies and computing infrastructure, as well as providing basic code-base maintenance, bug fixes, and low-level response to science community and internal LSST requests for new features.

3.1.4 Middleware

In a service oriented model with a layered architecture as outlined in Section 2 it is essential to have a cross cutting team who compose and debug services. Software such as the `butler` is not part of the pipeline but the pipeline needs it. This would also cover the builds and how the code interacts with the infrastructure (Section **??**.

3.1.5 Infrastructure

This is for deployment of of various systems and pipelines. Configuration is included in this. There needs to be a couple of people who manage keys/secrets for access to commodity services. We would need a security resource as well as database expertise. This then implies using tooling for system management as provided by e.g. AWS console. This team would include paying for a liaison at any service provider e.g. Amazon professional Services or a Service Manager at NCSA. (2FTE calculated)

3.1.6 Verification/Operations

This team will take and verify new releases for operations before they are deployed to the operations system. They will monitor the operational system to make sure it is functioning - they should have some science knowledge to know it is actually working properly as opposed to not just giving errors. A team of 4 should be able to handle this. Some support for this is assumed from IN2P3 .

3.2 Other implied changes to the current operations proposal

Notably missing from Figure 3 is QA. Currently QA is spread across three departments - the suggestion here is to place all QA activities under the survey performance department. Consolidation of the QA activities in one department may allow for some personnel saving.

The data release team in science operations would require a verification scientist (this may be 0.5FTE) while the SDQA and Semantic scientists may move to QA in survey science.

All data facility work, be it with a partner or in commercial cloud should be firmly under science operations - hence there is no LDF department and no associate director for LDF as depicted in Figure **??**.[2]

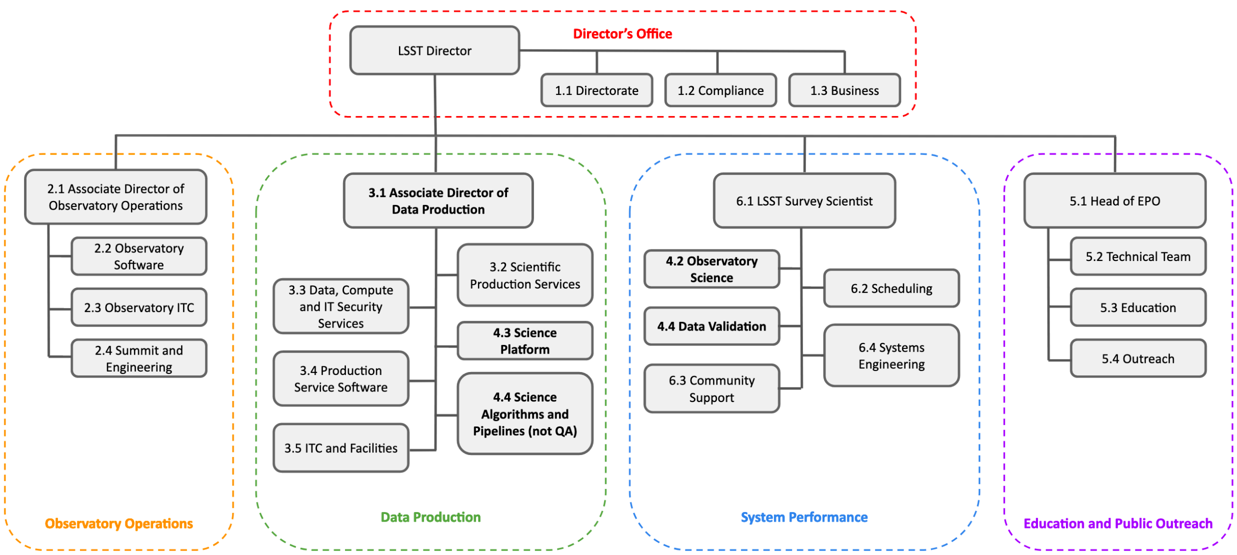

FIGURE 4: Possible new organisation chart for LSST operations

NOTE: I have said before communications should report directly to the director - in Figure ?? there is NO communications.

ITC and facilities (Figure **??**) in this model should come from NCOA logically also this does not belong in data production but at a higher level its for the entire org, the observatory already has its own ITC group so they are good.

[2]This is in line with AMCL recommendations

4 Conclusion

A restructuring of operations would give more transparent cost, allow for a better comparison to commodity pricing for many services and would yield considerable savings.

Team	FTE
Management (AD)	1
Observatory Science	5.5
Science Platform	6
Science Algos and Pipelines	16
Middleware	5
Infrastructure	7
Verification/ Operations	4
Total	**44.5**

A References

References

Bektesevic, D., Mehta, P., Juric, M., et al., 2019, In: American Astronomical Society Meeting Abstracts #233, vol. 233 of American Astronomical Society Meeting Abstracts, 245.05, ADS Link

Momcheva, I., Smith, A.M., Fox, M., 2019, In: American Astronomical Society Meeting Abstracts #233, vol. 233 of American Astronomical Society Meeting Abstracts, 457.06, ADS Link

[DMTN-072], O'Mullane, W., Swinbank, J., 2018, *Cloud technical assesment*, DMTN-072, URL `https://dmtn-072.lsst.io`,
LSST Data Management Technical Note

B Acronyms and glossary items

Acronym	Description
AMCL	Aura Management Council for LSST
AWS	Amazon Web Services, one of the largest cloud computing providers.
CI	cyberinfrastructure
DM	Data Management
DMTN	DM Technical Note
FTE	Full Time Equivalent
ITC	Information Technology Center
LDF	LSST Data Facility
LSST	Large Synoptic Survey Telescope
NCOA	National Center for Optical-Infrared Astronomy
NCSA	National Center for Supercomputing Applications
NSF	National Science Foundation
OPS	Operations
QA	Quality Assurance
SDQA	Science Data Quality Assurance
US	United States
cloud	A visible mass of condensed water vapor floating in the atmosphere, typically high above the ground or in interstellar space acting as the birthplace for stars. Also a way of computing (on other peoples computers leveraging their services and availability).
cyberinfrastructure	Sometimes denoted CI, A term first used by the US National Science Foundation (NSF), and it typically is used to refer to information technology systems that provide particularly powerful and advanced capabilities.
software	The programs and other operating information used by a computer.