Alternative science operations approach for Large Synoptic Survey Telescope (LSST)

William O'Mullane

2019-05-29

1 Introduction

We are currently experimenting with Google and Amazon Web Services for science platform and processing. These services are priced to deliver compute and storage - our current model at the LSST Data Facility (LDF) is also service oriented but is not priced in the same manner making comparisons difficult. An initial approach to a cloud costing was outlined in DMTN-072; this approach was an attempt to try to cost the hardware and compare to cloud pricing.

In this document a 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 estimate for assessing commodity provider offers.

2 Plugable service oriented architecture

At the Kavli workshop held in Las Vegas (Feb 2019 **?**) we took a long term view of astronomy archives and data processing. We suggested a layered service model as depicted in Figure 1[1]. Our astronomy 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 planning to operate LSST, but we do not yet treat the compute and storage as pure services.

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

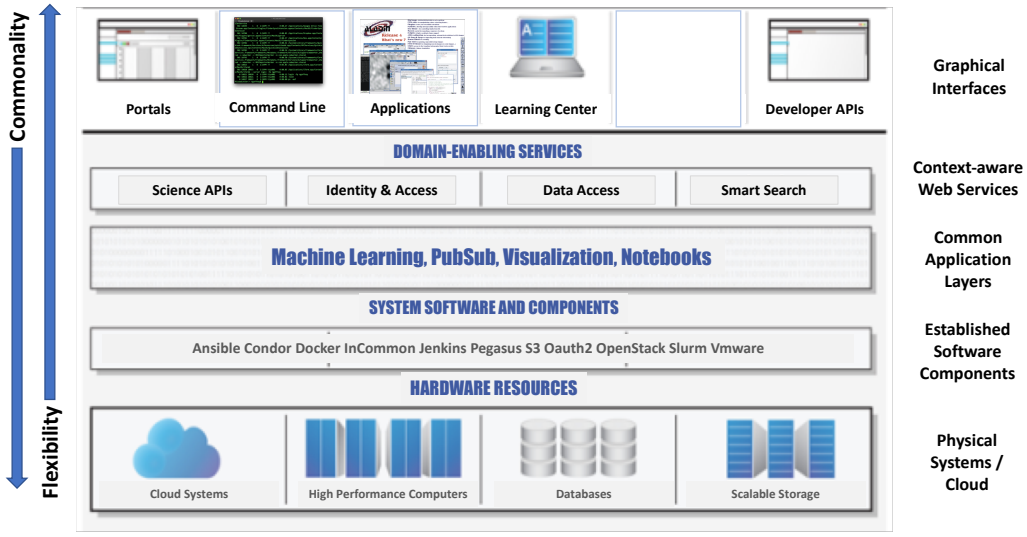

FIGURE 1: An example of a cyberinfrastructure built on an Infrastructure as Code design model. Note that while this example does not have astronomy-specific tooling, our recommendations highlight the importance of developing astro-specific layers that are fully accessible to scientists in both the application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

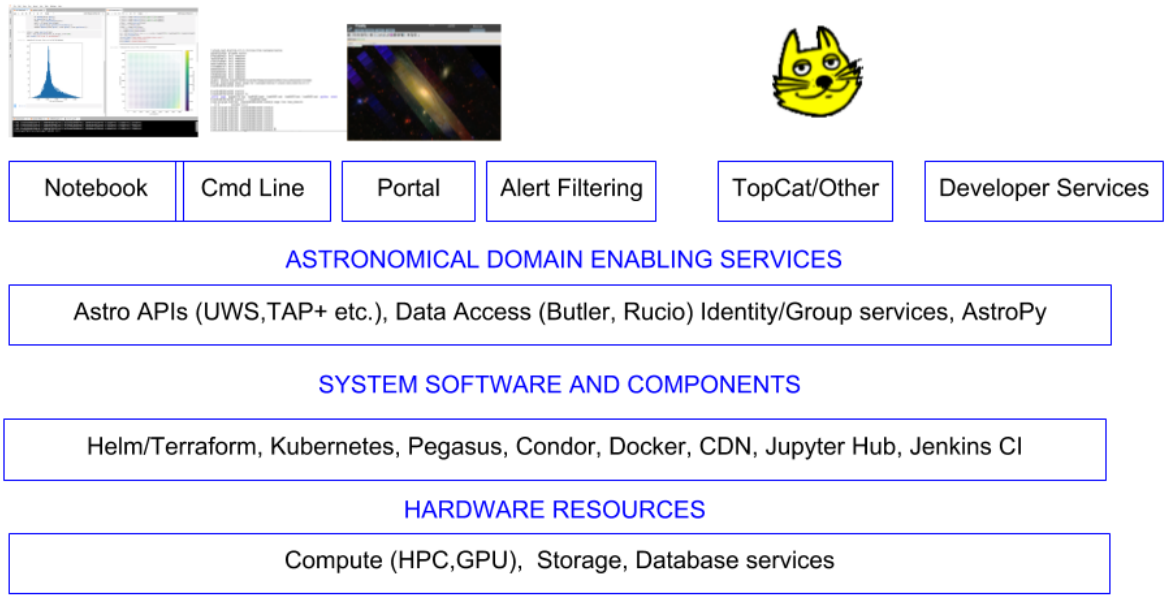

FIGURE 2: An example LSST cyberinfrastructure built analogous to the CI model shown in Figure 1.

An important part of following this architecture design is the ability to choose best in class components. In the science platform, for example, we have probably the best notebook implementation, but we could possibly pick up a better portal. In processing NCSA insist on a shared nothing approach - a move to an Object Storage could profoundly change that. It may raise other questions though on replication and redundancy. Then the shared nothing approach brings its own problems for deployment and Quality Assurance (QA).

A more service oriented approach should allow us to move between service providers to use the best in class for our underlying services as well. A clear model, understood by many, will make QA an easier task as well.

Getting to operations in this model will require some rethinking in construction – construction is a big ship which is already steaming ahead, so a change in course will take some effort. It is absolutely worth pursuing though.

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 the Science Operations Associate Director (AD).

Figure 3 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.

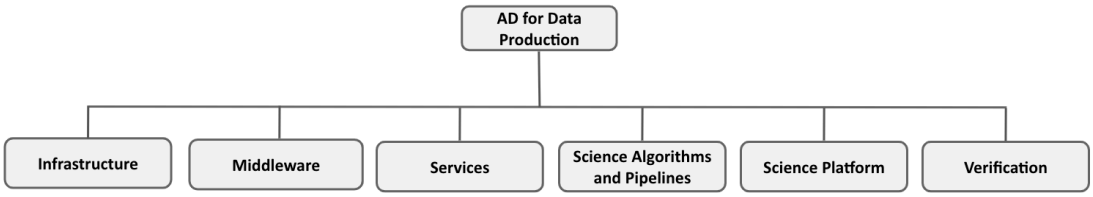

FIGURE 3: Possible configuration of Science Operations Department for operations of LSST

The Full Time Equivalent (FTE) counts are estimated in Table 1, which also gives the team sizes from the operations proposal for comparison. A brief description of the teams is given in Section 3.1.

Table 1: Size (in FTE) of the various teams in data production department with sizes including LDF teams from the proposal in the fourth column - a zero imples the team did not exist inthe proposal.

Team	2023	2026	2023(P)	Note
Management (AD)	1.5	1.5	1	
Infrastructure	8	7	0	
Middleware	7.5	6.5	0	
Science Alg and Pipelines	22	16	21	QA to system performance?
Science Platform	6	6	6.5	
Verification/ Operations	5	4	0	
LDF Management (AD)	0	0	3	In infrastruture
LDF Scientific Prod. Services	0	0	6.75	In Verification/ operations

LDF Information Technology Center (ITC) Security	0	0	8.5	Some in infrastructure/ middleware
LDF Prod. Services Soft.	0	0	7.6	Some in infrastructure
LDF ITC and Facilities	0	0	13	Should be in Services
Total FTE	**55.5**	**46.5**	**72.85**	

Note: The numbers in Table 1 assume we choose the people in the roles based on experience and effectiveness - in the current plan there are a number of Department of Energy (DOE) provided FTEs where the reason in some case seems to be availability rather than suitability. In the submitted proposal there was probably a certain number of duplicated roles (certainly in LDF) to cover this. We must consider this aspect carefully.

3.1 Teams

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

3.1.1 Infrastructure, Site Reliability Engineering

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.

In general an Site Reliability Engineering (SRE) team is responsible for the availability, latency, performance, efficiency, change management, monitoring, emergency response and capacity planning of their services Beyer et al. (2016).

This team would include paying for a liaison at any service provider e.g. Google Professional Services or a Service Manager at NCSA. (2FTE calculated)

3.1.2 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. In house developments such as Qserv should be covered here (1.5FTE has been included for this those are DOE/SLAC personnel, it could be 2FTE). This would also cover the builds and how the code interacts with the infrastructure (Section 3.1.1.

3.1.3 Services

Services contains the provision of compute and storage services for LSST independent of the source of those services in a computing center or via commercial cloud services.

3.1.4 Science algorithms and pipelines

This team is responsible to assess and assure the alert stream and annual data releases. In the submitted proposal this includes extensive QA to compare the data products against requirements - this may be be better merged with System Performance/Verification.

The main responsibility of this team would then be the underlying software pipelines themselves. That would include monitoring and updating the calibration plan and algorithmic implementation. The Calibration Support Scientist on the Observatory Science team will be responsible for monitoring the physical implementation of the calibration plan at the summit. In Table 1 this team is initially sized similarly to the AP/DRP teams in construction. There will be significant maintenance in the first two or three years of operations. As mentioned above there may be some consolidation with QA activities in System Performance.

3.1.5 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.

3.1.6 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.

3.2 Service budget

If we assume we do not reduce the operations budget then the total cost of the LDF services is the difference in FTE and the non labor costs for computer purchases. This is calculated in Table 2. One must also bear in mind that the data volume etc. increases each year, initial costs could be a little lower with final costs being a little higher. Service prices though are

TABLE 3: Size summary based on LDM-141

Table	Bytes/row	Rows (DR1 -> DR11)	DR1 (TB)	× Growth	DR10 (PB)
Object_Lite	1840	$2.26^{10} -> 4.74^{10}$	42	2.1	0.08
Object_Extra	20393	$2.26^{10} -> 4.74^{10}$	461	2.1	0.9
Source	453	$4.51^{11} -> 9.01^{12}$	204	20.0	4.0
ForcedSrc	41	$1.20^{12} -> 5.01^{13}$	49	42	2.0
DiaObject	1405	$7.94^{08} -> 1.54^{10}$	1.1	19.4	0.002
DiaSource	417	$2.26^{09} -> 4.52^{10}$	0.9	20	0.002
DiaForcedSource	49	$1.50^{10} -> 3.01^{11}$	0.7	20	0.001
Year 1 raw images:$3PB$, tables:$\sim 1PB$, half for Object_Extra,$0.2PB$ Sources					
Year 10 raw images:$30PB$, tables:$\sim 7PB$,$4PB$ Sources,$2.0PB$ Forced ,$1PB$ Object_Extra					

dropping every year so do we add significant data each year our costs should not in crease by the same fraction.

Table 2: Estimate of service budget/cost using FTE and non-labor costs from the proposal.

	FTE	Cost K$
Annual labour diff from Table 1	17.35	$3,081
Non labour hardware (average of all years)		$7,600
Total		**$10,681**

We would still require some hardware on the mountain and the base in Chile where we would potentially still keep a copy of the raw data. We should consider a hybrid model services at the LSST base facility for more realtime work and services in the cloud for even short term processing but also long term data processing and releases. An integrated model here can be provided by at least Amazon - analogous to how your mobile phone does some local processing but goes off to big brother for more heavy processing.

The LSST data volumes are in Table 3.

The compute estimates are a little more difficult to extract in Table 4 from DMTN-072 an estimate is made in terms of FLOPs.

Table 4: Various inputs for deriving costs

Year	2017	2018	2019	2020	2021	2022
FLOPs Needed Total (no Alerts)	9.48261E+19	1.00E+19	1.00E+19	9.48261E+19	1.00E+19	4.74131E+20
Time to Process days	252.0	365.0	365.0	252.0	365.0	252.0

Time to Process seconds	21772800.0	31536000.0	31536000.0	21772800.0	31536000.0	21772800.0
Instantaneous GFLOP/ s	4355.255691	3.17E+02	3.17E+02	4355.255691	3.17E+02	21776.27846
Instantaneous GFLOP/ s (inc Alerts)	4355.255691	3.17E+02	3.17E+02	30025.25569	2.60E+04	21776.27846
Disk Space TB	1000	1000	1000	10000	20000	30000
I/ O for year TB	10	100	3000	30000	60000	90000
Base numbers	Ecyc	FLOP	GFLOP			
LDM-138 DR1,2 Data Rel sheet row 1	155.17	4.26718E+20	426717500000			
LDM-138 DR3 Data Rel sheet row 2	348.76	9.5909E+20	959090000000			
LDM-138 Alert Instananeous	0.00023434	25670000000000	25670			
Alert Total, assuming 275k visits/ year	64.4435	1.7722E+20	177219625000			
Total Yr1 (inc DAC)		**4.74131E+20**	**474130555556**			
	Optimistic	Pessimistic				
Moore Factor Proc	0.7	0.9				
Kryder Factor Disk	0.8	0.9				

3.2.1 Potential next steps

We should publish the commissioning data - this will require compute and storage and a science platform. This could be as a pre ops task on pre ops money. The sizing for this is currently being worked on, the max data volume is about 0.5PBi, but is probably much less If we decide this might only have 40 or so users we have a good idea of the processor power required since we ran the platform on Google for LSST Europe in 2018 with 40+ users on a few nodes[2].

The correct thing to do here would be spec this up and ask for bids from interested parties to host it for the number of years we need it (up to DR1). With an option of course to expand to DR1. This would allow us to see how this model works on a dataset we were not planning to serve in construction. It would also allow us to become comfortable with our potentially commercial partner and vice versa before DR1.

4 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 Science Data Quality Assurance (SDQA) and Semantic scientists may move to QA in survey science.

[2]This cost is about $400 for a week - apparently for 3 nodes

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 4.[3]

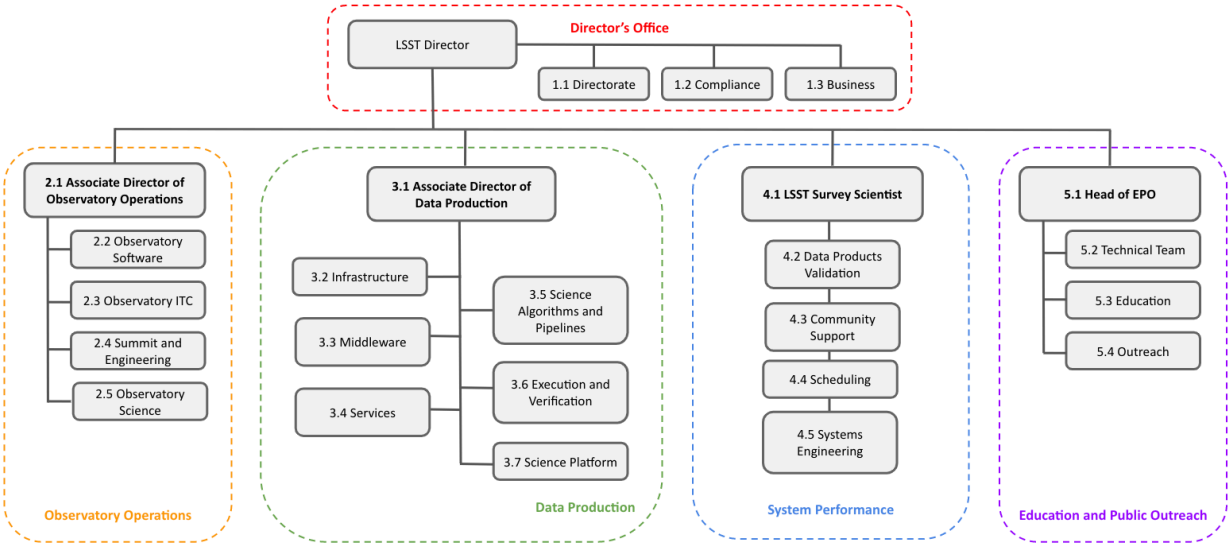

FIGURE 4: Possible new organisation chart for LSST operations

ITC and facilities (Figure 4) 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.

There are still some developments in NCSA such as Data Acquisition System (DAQ) forwarders which would need to be included in possibly the observatory software.

There are other developments Data Back Bone (DBB) which could be replaced by commodity services like Amazon S3 which includes replications and reliability.

5 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.

[3] This is in line with AMCL recommendations

A References

References

[LDM-141], Becla, J., Lim, K.T., 2013, *Data Management Storage Sizing and I/O Model*, LDM-141, URL https://ls.st/LDM-141

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

Beyer, B., Jones, C., Petoff, J., Murphy, N.R., 2016, *Site Reliability Engineering: How Google Runs Production Systems*, O'Reilly Media, Inc., 1st edn.

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 Glossary

AD Associate Director.

AWS Amazon Web Services, one of the largest cloud computing providers..

calibration The process of translating signals produced by a measuring instrument such as a telescope and camera into physical units such as flux, which are used for scientific analysis. Calibration removes most of the contributions to the signal from environmental and instrumental factors, such that only the astronomical component remains..

CI cyberinfrastructure.

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)..

configuration A task-specific set of configuration parameters, also called a 'config'. The config is read-only; once a task is constructed, the same configuration will be used to process all data. This makes the data processing more predictable: it does not depend on the

order in which items of data are processed. This is distinct from arguments or options, which are allowed to vary from one task invocation to the next..

cyberinfrastructure Sometimes denoted CI, A term first used by the US National Science Foundation (National Science Foundation (NSF)), and it typically is used to refer to information technology systems that provide particularly powerful and advanced capabilities..

DAQ Data Acquisition System.

Data Management The LSST Subsystem responsible for the Data Management System (DMS), which will capture, store, catalog, and serve the LSST dataset to the scientific community and public. The DM team is responsible for the DMS architecture, applications, middleware, infrastructure, algorithms, and Observatory Network Design. DM is a distributed team working at LSST and partner institutions, with the DM Subsystem Manager located at LSST headquarters in Tucson..

DBB Data Back Bone.

Department of Energy cabinet department of the United States federal government; the DOE has assumed technical and financial responsibility for providing the LSST camera. The DOE's responsibilities are executed by a collaboration led by SLAC National Accelerator Laboratory..

DOE Department of Energy.

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.

Object Storage A storage architecture that stores files as objects (opposed to a hierarchy etc.) commonly provided by cloud services, i.e AWS S3 (Simple Storage Service) or Google Compute Cloud "Cloud Storage"..

Operations The 10-year period following construction and commissioning during which the LSST Observatory conducts its survey.

pipeline A configured sequence of software tasks (Stages) to process data and generate data products. Example: Association Pipeline..

QA Quality Assurance.

Qserv Proprietary LSST Database system.

S3 Structured, imperative high level computer programming language, used as implemen-

tation language for the Virtual Machine Environment (Virtual Machine Environment (VME)) operating system.

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..

SDQA Science Data Quality Assurance.

software The programs and other operating information used by a computer..

SRE Site Reliability Engineering.

stack A record of all versions of a document uploaded to a particular DocuShare handle.

VME Virtual Machine Environment.