LSST Science Platform Design Review
Final Report

Review Committee:

LSST Science Collaborations:
Rachel Street (Chair)
Transients and Variable Stars Science Collaboration (TVS) co-chair,
David Kirkby
Science Advisory Committee, Dark Energy Science Collaboration (DESC)
Michael Mommert
Solar System Science Collaboration (SSSC), expert on moving objects
Christian Johnson
Stars, Milky Way and Local Volume Science Collaboration (SMWLV)
Lee Steven Kelvin
Galaxies Science Collaboration (GSC)
Tina Peters
Active Galactic Nuclei Science Collaboration (AGNSC), DESC

LSST Project:
Keith Bechtol
LSST Commissioning, Science Verification, DESC
Richard Dubois
LSST Camera. DESC

External:
Ani Thakar
SciServer

Review Meeting

The committee and Project representatives met in the offices of the National Optical Astronomy Observatory in Tucson, AZ in April 10-12, 2019. The committee presented a verbal summary of this report at the end of that meeting. The agenda for that meeting can be found in Appendix A.

Review Charge

Objectives

The LSST Science Platform is a unified set of web applications and services made available to the scientific community to access, visualize, and perform 'next-to-the-data' analysis of the LSST data. The platform exposes the LSST data and services to the user through three primary user facing "Aspects" — the web Portal, the JupyterLab analysis environment, and a machine-accessible Web API interface, each providing different ways to access the data and analysis services provided by the LSST Data Access Centers (DACs).

Although primarily conceived as a platform for scientific analysis of the LSST data in the operations- era, the LSP will also be the major platform for integration and test activities during LSST commissioning, and as such, the relevant stakeholders comprise not only members of the scientific community but also the LSST Camera and Commissioning teams.

The objective of this review is to evaluate the vision and design of the LSP against the LSST science requirements, and to verify both the current implementation and the design for future planned operations-era aspects of the system. LSST Data Management want to explore how the LSP can best meet the needs of the LSST science user community by identifying issues now before committing to further implementation or reprioritization of already-planned work.

Scope of the Review

This is the Final Design Review of the LSST Science Platform (LSP), as defined and described in LSE-319: LSST Science Platform Vision Document.

The services provided by the LSP can be decomposed into the three different "Aspects":

- A web Portal designed to provide essential data access and visualization services through a simple-to-use website.
- A JupyterLab environment, that will provide a Jupyter Notebook-like interface enabling next-to-the-data analysis.
- An extensive set of Web APIs that the users will be able to use to remotely examine the LSST data with familiar tools.

The LSST teams developing the Science Platform — Science User Interface and Tools (SUIT) at IPAC, Software Quality and Reliability Engineering (SQuaRE) at LSST-Tucson, and Data Access (DAX) at SLAC — are requested to demonstrate the detailed product design to be realized as well as the realization process.

Science use cases addressing all four of the LSST key science themes — probing dark energy and dark matter, taking an inventory of the Solar System, exploring the transient optical sky, and mapping the Milky Way — should be addressed. These use cases should cover database-oriented data access, data access for large-scale analytics (possibly non-database), bulk data access for non data centers, and user interfaces.

It should be noted that the LSST database, Qserv, is considered to be outside the scope of this review; however, certain aspects of the database design can affect the performance of the LSP. Where pertinent, those aspects of the Qserv design and expected performance relevant to science users and the ability of the LSP to satisfy the LSST science requirements should be presented.

A representative group of LSST stakeholders will be asked to provide science-oriented feedback and recommendations on the design of the LSST Science Platform.

Charge to the review committee

The review committee is asked to assess the following items based on the material presented and made available:

1. Is the traceability of requirements from higher design documents, e.g., from LPM-17: The LSST Science Requirements Document, to LDM-554: The LSST Science Platform Requirements, complete and will it ensure coverage of the four key LSST science themes?
2. Are the stakeholders clearly identified and understood? Have the requirements been prioritized and communicated to a representative set of the stakeholders?
3. Does the design presented in LDM-542: LSST Science Platform Design capture the requirements for the LSP as detailed in LDM-554: The LSST Science Platform Requirements?
4. Are the verification, validation and software quality assurance plans adequate?
5. Does the performance of the current system and its development status inspire confidence that both the interim and operations-era functionality can be delivered?
6. How does the design of the LSP compare with that of other contemporary astronomical data archives and interfaces, or, more generally, other scientific data analysis environments? How well does the design and current implementation reflect trends in software engineering? Do the current design and technology choices give confidence that the LSP can evolve over time with the needs of 21st century astronomy?
7. Are there items of significance in the design that would unnecessarily limit the science harvest of LSST?
8. Are the risks associated with the design of the LSST Science Platform understood and adequately captured? Are there any overlooked areas of risk?
9. Are there appropriate scope options accompanying the plan? If cuts had to be made, are there areas of the plan that could be descoped with minimal impact on LSST science? What is the scope for use of third party-tooling in place of in-house development?

In addition, the committee is asked to provide actionable advice on addressing any issues raised during the course of the review as well as guidance based on experience that will ensure the success of the LSST Science Platform.

Preamble

The committee is excited to learn how scientists will be able to discover, explore, visualize and analyze data products from the Large Synoptic Survey Telescope (LSST). This Project is anticipated to bring about extraordinary changes in astronomy; not just as a result of the expected wealth of new discoveries revising our physical understanding of the universe but also in terms of the new technologies necessary to handle and exploit a data set of this volume and rate of delivery. As one of the landmark projects of the “Big Data” era, LSST presents an opportunity for science to benefit from technological developments in other industries.

Table 1 shows LSST data products in the context of other modern optical astronomical surveys, comparing their overall size and rate of delivery. One of the distinctive characteristics of these projects is their exploration of *time-domain astrophysics* [LSST Science Book, v2.0] in ways that were technologically impractical until recently. The combination of wide-field (multi-degree) instruments with deep (>21 mag) limiting magnitudes with rapid data processing (<minute) and massive (TB/night) data transfer and storage capabilities, makes it possible for the first time to explore astronomical variability (temporal and spatial) over timescales ranging from seconds to years across a large section of the sky.

Project	Etendue [m^2deg^2]	Total # objects	Alerts/night	Expected final data release size
LSST	319	~37 billion	~10 million	512 PB
ZTF	8.85	~1.8 billion	~1 million	~3.2 PB
Gaia		~1 billion	10-20	>550 GB

Table 1: The size and rate of LSST data products in comparison with other alert-issuing optical surveys. Sources: LSST [http://lsst.org], Zwicky Transient Facility [ZTF, Masci et al. 2018], Gaia [**https://www.cosmos.esa.int/web/gaia/dr2**]

The size of this dataset demands a paradigm shift in the way astronomers perform their analyses. It will not be possible for thousands of individuals or even institutions to download a copy of the entire dataset to local storage. Instead, astronomers must adopt the currently-unfamiliar model of “moving their analysis to the data”, a capability which is expected to be hosted at Data Access Centers (DACs). The efficacy, suitability and utility of software tools used by the astronomical community to access the DACs will therefore be critically important to the scientific yield of LSST as a whole.

The LSST Science Platform (LSP) is a multi-faceted software platform designed to facilitate several different modes of accessing astronomical data which have become expected cornerstones of current astronomical analysis. Since LSST is one of the largest international projects in astronomy in the 2020s, it will have profound influence on astronomical software going forward, and the LSP can expect to set the standard for the next decade.

Response to the Charge to the Committee

1. Is the traceability of requirements from higher design documents, e.g., from LPM-17: The LSST Science Requirements Document, to LDM-554: The LSST Science Platform Requirements, complete and will it ensure coverage of the four key LSST science themes?

Great effort has been made to thoroughly detail the LSP from its top level vision, inclusive of scientific motivations, through to appropriate functional requirements. However, the committee found that the requirements documentation supplied as part of the review remains incomplete in places (Appendix B lists a few example cases), so it is not possible to fully trace requirements between them. That said, the committee acknowledges the rapidly-evolving nature of software and the need - and great benefits - of the developers remaining flexible in their approach, in order to take maximum advantage of new technologies as they emerge. For example, it may prove advantageous for the LSP to become a Cloud-based system (discussed further below), which may require corresponding revision of the baseline design. An ongoing review and revision of requirements and design elements is highly desirable to ensure that the LSP provides cutting-edge tools. The committee therefore endorses the development process undertaken by the team to date, and stresses that this charge should not delay the LSP development. Instead, to address the wording of this particular charge, the requirements documentation should be completed, and regularly updated, in order to provide criteria against which progress may be evaluated.

It was the committee's impression during the review that the toolset for the LSP is being developed by the LSST Construction Team, with the potential implication that no additional tools will be developed once operations start. Given the lifetime of LSST, and the rapid evolution of computing resources, this could result in the LSP becoming outdated. The Team should clarify if there will be a point at which the design of the LSP is frozen.

It was evident from presentations made at the review that there has been some prioritization of the requirements; however, these priorities are not made clear in the documentation itself and should be added. The committee notes that the 3rd level of priority effectively indicates that a requirement is optional, and suggests that these items be rebaselined as "goals" rather than "requirements".

The committee expressed surprise that the scientific requirements outlined in LPM-17 were not accompanied by scientific use-cases, which could help the Project to specify and to test necessarily functions, and the reviewers to establish whether these use-cases were adequately covered. In particular, demonstrating that the LSP can answer a representative set of science questions (the "20 Queries") - that are based on the four main science themes listed in LPM-17 - would have been very valuable in ascertaining the ability of the LSP to meet the science requirements and the expectations of the science collaborations.

2. Are the stakeholders clearly identified and understood? Have the requirements been prioritized and communicated to a representative set of the stakeholders?

The stakeholders for the LSP include the scientific community as well as the LSST camera and commissioning teams. The committee found that these groups were clearly identified, and that development timelines for all three LSP aspects have been designed with commissioning use-cases in mind.

The committee acknowledges a recent presentation by Leanne Guy to the Project Science Team/Science Collaboration (PST) Telecon which outlined the design and functions of the LSP. While substantial documentation on the LSP exists, it was not clear whether the Science Collaborations have had the opportunity to provide feedback on the requirements (aside from this committee) or on the prioritizations made. While the presentation to the PST Telecon was welcomed, it could not be as detailed as a full review, and gave limited opportunity for in-depth feedback.

As noted above, while the presentations made at the review indicate that priorities have been drawn up, the committee is unaware of these being communicated to the Science Collaborations in general. Similarly, the understanding of the committee is that the scientific community has been unaware of the rescoping decisions that were implied by several of the review's presentations. This does not imply the committee necessarily disagrees with the decisions made.

The LSST Stack Club was recognized as the best forum for science collaboration members to gain familiarity with Project software tools. The committee recognizes the important role of the Stack Club, and appreciates the efforts of Project personnel in supporting it. To date, the Stack Club has focused primarily on the LSST Stack software rather than the LSP, and not all Science Collaborations are represented in its membership (though all are welcome). The committee recommends that Science Collaborations identify representatives to participate in the Stack Club, acknowledging that the ease of collaboration in the shared, highly-functional, environment provided by the LSP has been essential to the success of the Club. The committee emphasizes the importance of making the LSP available to a wider science user group at the earliest opportunity, to facilitate astronomers a) gaining familiarity with new modes of analysis, giving greater opportunity for existing analysis software to be updated as appropriate, b) disseminating information about the LSP to a wider audience, c) providing ongoing feedback on tools and priorities and d) acting as ambassadors to encourage uptake of the LSP among the community. The committee further recommends that the LSP team provide detailed presentations regarding the LSP to each of the science collaborations, which could be tailored to the needs of those communities. Such presentations could be developed in cooperation with science collaboration members that are active in the LSST Stack Club. Additionally, the committee notes that the HSC Team are currently using the LSST Stack and could act as well-informed alpha testers for the LSP as well as benefitting from its tools.

The LSP Team are encouraged to consider and to document their goals and (where possible) plans for the long-term role of, functionality of and support for, the LSP. For example, will early data releases continue to be accessible through the LSP towards the end of the main survey? How will users continue to exploit the LSST data after the end of the main survey? Who will the stakeholders be in this era - for example, will it still be aimed at science users, or overlap or merge with the Education and Public Outreach portal?

3. Does the design presented in LDM-542: LSST Science Platform Design capture the requirements for the LSP as detailed in LDM-554: The LSST Science Platform Requirements?

The LSP design does seem to capture all necessary functionality as identified in the completed sections of the Science Requirements Document. The committee recognise the importance of all three aspects of the LSP as serving distinct workflows that are expected to become increasingly common science use-cases. However it was not clear whether all elements of this design will be implemented following the rescoping of the LSP, most notably its portal aspect.

The committee stresses the importance of the portal functionality as outlined in the requirements, as being essential tools for some users, for example those who are less comfortable coding in Python. It was felt all users would expect that a time-domain survey such as LSST should support fundamental time-domain functions such as the ability to visualize and analyze light curve data. In addition, further steps should be made to support and document a bulk image cutout service, providing trivial access not only to the image data itself, but also the associated image metadata as stored in a FITS header (e.g., gains, exposure times, WCS information).

4. Are the verification, validation and software quality assurance plans adequate?

The LSP developers presented a detailed plan for the verification of the software functionality. Verification is centered around a Jira Test Plan where pre-written scripts designed to test specific functions can run systematically, providing the developers with detailed output which will alert them to rapidly identify any failures. The Test Cases upon which these scripts are based are derived from the requirements documents. Overall, this is a powerful and highly repeatable testing framework which should enable thorough verification of the software.

The team reports that all requirements in LDM-554 have been mapped to verification elements within this system. This is supplemented by the implementation of Continuous Integration processes to execute a suite of unit tests for the software itself every time updates are made to the repositories.

However, the committee draws the team's attention to the sections of LDM-554 that remain to be completed, and appropriate tests remain to be written. The documentation provided includes the specification of step-by-step verification procedures for ~20% of the requirements, including a combination of automated and manual execution steps (which the committee notes may be tested by a simulator package such as Selenium). The full documentation of verification procedures should be completed.

The LSP Team described a validation plan centered around making scientifically interesting datasets available through the portal during the Project's commissioning phases, which the committee strongly endorses. To date, the team have performed one formal testing exercise of the Portal (report DMTR-22), though none for the Notebook or API aspects, and further validation has been carried out in an informal way by the LSP developers and commissioning team members. It is envisaged that over the next 1-2 years a limited number of science users can provide validation by participating in the Stack Club. The committee strongly encourages the team to recruit some community "alpha-testers" through the Stack Club or appropriate members of the scientific community (e.g., the HSC team) as soon as possible, to give science users as long as possible to learn about the platform, adapt their analysis process and to advocate for it within the community. Such an approach would furthermore enable the community to contribute functionality to the Portal and other aspects. The committee recommends the development of a (potentially phased) plan to increase the level of involvement of science collaboration members in validation activities, and to share this with the community with an appropriate timeline.

The committee notes that the LSP is already supporting a level of routine operations demand, in parallel with ongoing development work, and this is likely to increase as the Project nears commissioning. Personnel must therefore strike a balance between development and operations.

5. Does the performance of the current system and its development status inspire confidence that both the interim and operations-era functionality can be delivered?

The committee unanimously agrees that the overall development of the LSP is impressive, and inspires confidence that the required functionality can be delivered. In particular, it was felt that the notebook aspect may prove to be revolutionary and set the standard for all future astronomical programs. The provision of a 'birthright' allocation of resources to all users (storage, CPU, GPU, RAM) was recognized as important and a valuable way to 'level the playing field' and ensure that LSST science is open to all.

Cloud computing has undergone significant development, led by industry, since the LSP design was first outlined, and now offers a credible alternative deployment option for the LSP compared with the local-storage model originally envisaged, and several potential benefits. Most notable of these is the Cloud's ability to scale up or down both storage, CPU and even GPU allocations in real-time response to user demand. It was recognized by both the LSP Team and the committee that demand is likely to fluctuate strongly, and undergo spikes (e.g., immediately following data releases). The committee recognizes the insightful and timely work of the LSP Team in exploring Cloud-based technologies and in developing the LSP in such a way that moving to a fully Cloud-based platform would be relatively easy to accomplish. The LSP Team have made good progress in their discussions with commercial vendors, but a concrete plan (that also includes cost comparisons) for the Cloud-based deployment of the LSP is not yet available. In the context of this part of the charge, the committee acknowledges that this may result in a revision of the requirements, and recommends that this decision be made as soon as possible.

The committee feels that the versatility, power and ease of use of the tools offered by the LSP are highly likely to be extremely popular among astronomers. So much so, that it was felt that the performance requirements for user storage allocation (DMS-LSP-REQ-0012, LDM-554), and the volume of simultaneous queries to be handled (DMS-LSP-REQ-0028, DMS-LSP-REQ-0029) were likely to underestimate community demand, potentially excessively so. It was unclear how the thresholds for these requirements were set. The committee recognises that, at the time of writing, it would have been very difficult to accurately predict demand in the LSST era; indeed, since there are few comparable projects in astronomy, this is still challenging. Nevertheless, more credible performance criteria could be set by investigating user demand for current projects of similar breadth and scale, in particular usage of the Gaia Mission Data Releases. The committee strongly recommends that the LSP Team review, revise and document these performance requirements. It does recognize that resources provided by the Project for community science meet the mandated requirement of 10% of their computing resources, and that the Project may not be able to single-handedly meet the full demand. Therefore, if user demand is found to be likely to exceed the available resources, the Project should urgently explore the development of iDACs funded external to itself, and facilitate their providing similar tools and facilities.

The uncertainty and time-variability of user demand led the committee to recommend that, once a revised estimate of user demand is available, they perform a cost/benefit analysis of Cloud deployment and, if adopted, draw up detailed plans as soon as possible, since some revision of the requirements (and verification/validation) may be required, particularly with respect to resource allocation and management.

The importance of a tool to provide a quick-look visualization of the full focal plane was highlighted in particular. This will be particularly valuable in the context of real-time data quality assessment during commissioning, where a tool to quickly assemble mosaic images is needed for rapid inspection, ideally with interactive pan and zoom features, but is likely to be of interest to the science community also. This was felt to be most useful as a notebook aspect function, and could be made available through the portal aspect at lower priority. Although suitable functionality appears to be under development, the Team should ensure this capability is adequately developed in time for commissioning.

6. How does the design of the LSP compare with that of other contemporary astronomical data archives and interfaces, or, more generally, other scientific data analysis environments? How well does the design and current implementation reflect trends in software engineering? Do the current design and technology choices give confidence that the LSP can evolve over time with the needs of 21st century astronomy?

The current design of the LSP, as defined by LDM-554, describes a very modern, versatile and user-friendly approach to serving large astronomical datasets to the community, which takes advantage of modern software engineering and data science technologies. The choice to adopt VO-standards will facilitate analysis of LSST data in conjunction with other survey datasets as well as the use of commonly-used tools such as `astropy` and `TopCat`. The LSP enables next-to-the-data analysis in several ways that are considerably more technologically advanced than other data archives (such as Vizier, IRSA/IPAC, MAST, NASA PDS). Most notable of these is the implementation of easy parallelization of tasks through Dask, which, although presented as an exploratory technology, has the potential to be transformative, and the committee endorses its further evaluation for possible inclusion in the final LSP.

The committee strongly feels that these technologies will be essential to LSST, for meaningful and timely exploitation of such a large dataset, to provide the community with tools that exploit cutting edge technologies at the start of commissioning, and to allow the LSP to evolve over the lifetime of LSST to take advantage of new developments.

The LSP incorporates a number of elements that are well aligned with current (and foreseeable) trends in astronomical data analysis, notably the JupyterLab notebook and API aspects, which will allow for a smooth evolution over time (for example to support new kernels/languages, addition of further explanatory demonstration notebooks).

The LSP Team have made excellent use of containerized runtime environments, which will be a key element of reproducible science, as well as enabling users to share their analysis and/or computing resources and environments, facilitating current modes of collaboration.

The committee recommends that a regular review process be instigated, throughout the lifetime of LSST, to provide continual feedback on trends in astronomical analysis and industry-developed technologies going forward. LSST should plan to enable these technologies to be incorporated within the LSP.

The committee felt that while astronomers are increasingly turning to Notebooks as a means of data access and discovery, many users would find the data difficult to explore if a Portal aspect were not also included. In particular, the Python-centric approach currently inherent within the design of the Notebook aspect may result in some users preferring to use the Portal aspect to download small subsets of the data where they can continue data analysis locally using familiar languages and software packages. The absence of a browser based Portal with the exploration and visualization functions would put the LSP at odds with the current standard of service delivered by existing archives such as the ESO Archive Science Portal, Vizier, IRSA/IPAC, MAST, SDSS etc. Some associated Portal tools, such as batch modes of operation and a bulk image cutout service providing access to imaging and associated metadata, should also continue to be developed and documented. Furthermore, the lack of a light curve plotting tool in the Portal (functionality already successfully trialed in the prototype) was strikingly absent from the current design, since LSST is first and foremost a time-domain survey, and light curve based analysis is likely to be a leading use-case. Lastly but by no means least, concern was expressed that the implementation of functionality relevant to Solar System research has been stopped due to the current re-scoping plan applied to the Portal.

The committee therefore recommends that consideration be given to upscaling the Portal aspect, or, at a minimum, a commitment to launch it as-is and maintain it at its current level (e.g., bug fixes, ingestion of new data).

The committee recommends that the FTEs allocated to the long-term maintenance of the LSP be reviewed in the light of the proposed re-evaluation of user demand. It seems likely that greater effort may be required, including for ongoing support of the 3rd-party tools employed.

The committee recognized the advantages offered by the powerful and efficient Qserv database system, and acknowledged the impressive upstream contributions made by LSP developers to commercial tools (such as those developed by Facebook, for example). However, the committee notes that this custom solution (a layer above the MariaDB base) brings with it development and support overheads for the lifetime of the project. The committee notes that this has places some restrictions on the query language that Qserv can support, in the interests of providing faster queries and a simpler codebase.

7. Are there items of significance in the design that would unnecessarily limit the science harvest of LSST?

It was felt that the most significant risk presented regarding the LSP was that of user demand exceeding the available resources, and the committee discussed a number of ways in which this concern could be mitigated. A Cloud-based deployment offers the advantage that fees scale up and down in real-time in direct proportion to demand, as discussed above. The committee recommended encouraging the development of multiple additional DACs, perhaps with some providing access to data subsets of interest to specific user communities, sponsored by partner institutions, consortia and, particularly, countries. It was felt that this would also help to offset issues with network latency that might otherwise restrict the useable functionality of the LSP in some areas.

The LSP design is Python-centric, which seems to be a defensible choice given the language's widespread adoption in astronomy and its use in many scientific pipelines. It should be noted, however, that it still presents a significant learning curve for some users. The committee strongly supports the provision of a terminal within the Notebook aspect, which makes it possible for the user to install code written in other languages (with the exception of any which require a license server, a choice the committee endorsed). Given the rapid evolution of programming languages, it is possible that another language may develop in preference, tools and functionality over the lifetime of LSST. While this is not foreseeable at this stage, the LSP maintenance and review program should consider this. The committee advocates providing additional introductory-level training materials (e.g. websites, demonstration notebooks) to assist users who are unfamiliar with Python and/or Notebooks to adopt the LSP.

The absence of a Portal was felt to limit the scientific productivity of the LSP, since users will face additional hurdles to data discovery and visualization, as discussed above. At a more detailed level, the committee raised the question of whether the use of ADQL exclusively (as opposed to also supporting fuller versions of SQL) could limit the use-cases that may be implemented through the LSP.

A large section of the scientific community (e.g., extragalactic) will require access to the imaging data produced during the lifetime of the LSP and LSST. This may include single-epoch snapshots, co-added multi-epoch imaging, point spread functions and LSST stack pipeline imaging such as subtracted sky maps. To that end, further efforts should be made to provide fully featured and documented tools which provide bulk image cutout and download capabilities. In addition to the image data itself, such tools should strive to also provide the associated metadata typically found in a FITS header for each image (e.g., gains, exposure times, WCS information).

8. Are the risks associated with the design of the LSST Science Platform understood and adequately captured? Are there any overlooked areas of risk?

The primary risk associated with the LSP was felt to lie in the uncertainty regarding user demand, which may outstrip available resources. For example, the "birthright" storage space allocation per user remains unclear but is expected to be of the order of ~50GB. This is already small in comparison with the products of many typical analyses. This may be further complicated by the ongoing and future evolution in typical user workflows, as astronomers become more familiar with Notebook-style and API interfaces. At the present time, the level of user demand for the Notebook aspect remains unclear.

The rescoping of the Portal aspect carries the risk that several science cases will not be serviced via this aspect (e.g. light curve analysis, bulk image access, and functionality for Solar System science such as visualization of orbit predictions). This will force users to use other aspects or perhaps DACs to achieve their science goals, or alternatively to attempt to download large data subsets in order to conduct an analysis on local servers. This raises related risks such as increased bandwidth requirements (or reliance on download-throttling), leading to decreased performance and utility for the LSP overall, and a likely increase in user frustrations and corresponding Helpdesk inquiries.

A related concern is the impact of network latency on the user experience of the LSP, particularly if it is hosted from only one physical location. The committee recommends exploring ways to encourage the provision of multiple DACs hosting the LSP, in different continents as described above, to mitigate this concern.

The committee felt that the plan for scaling the current Qserv database up and out to the full LSST cluster is insufficiently quantified, and that this should be clarified. How many nodes is the 100% scaled out version of Qserv expected to have, and what are the performance ramifications of the full deployment? Are the query performance figures (that were presented) expected to apply to the full scale database as well? What will be the database ingest and index rebuild times be?

It appears that the LSP Team are on track to deliver the tools necessary for commissioning in a timely manner. The committee emphasized the need for science users to be allowed adequate lead time, prior to the start of operations, to become familiar with all aspects of the LSP and to adapt their analysis workflow and existing software to the new tools. To that end, the committee reiterated the benefits of a phased program to gradually increase science community usage of the LSP. It was generally accepted that a relatively small fraction of users will be willing and able to participate in this way, but the `Astropy` example has demonstrated what community contributions can accomplish.

LSST data will become even more powerful when it is combined with data from other major surveys and facilities, in particular the Gaia catalog. To achieve this, it will be necessary to cross-match sources between LSST and external catalogs, a CPU-intensive and time-consuming task. It is likely that a high number of users (including Project commissioning teams) will attempt to do this, potentially leading to significant resources being repeatedly dedicated to effectively the same operation. The committee recommends providing pre-computed cross-matched tables of objects/sources (referred to as "neighbors tables") to selected external catalogs, along with documentation regarding the method used. It is acknowledged that this will not satisfy every user, but could well be sufficient for most purposes (i.e., meet the 80/20 rule).

Although the LSP Team outlined a thorough plan for data security in the LSP, the committee discussed a number of concerns, chiefly focused on the management of legitimate users. Questions were raised over the mechanism to prevent the LSP being used for non-LSST science, for example, and to identify and manage (and perhaps charge) users who repeatedly and heavily use CPU or storage facilities, or who attempt large downloads. To a certain extent, the impact of this risk may be adequately mitigated via the provision of appropriate end-user information, e.g., informing the user why they might choose to spawn a smaller notebook over the largest available. The committee recommends implementing systems to identify and manage heavy usage, and that a user's resource allocation be made clear to them when their account is first allocated. Displaying the current usage of the granted CPU or storage resources in a notebook would be a useful asset to identify the actual needs of the user and enable the user to make a potentially more appropriate resource allocation selection in the future. The option to allocate additional resources via competitive proposal was also discussed. This was felt to be unduly onerous for relatively small and/or short-term increases in storage/processing limits, and that this use-case would be better addressed by the provision of a scratch space. However, the committee supports a resource-proposal process for users with longer-term and/or large-scale or specialized processing requirements.

9. Are there appropriate scope options accompanying the plan? If cuts had to be made, are there areas of the plan that could be descoped with minimal impact on LSST science? What is the scope for use of third party-tooling in place of in-house development?

It was evident from the presentations that some rescoping decisions have already been made, including some that the committee felt were detrimental (e.g., light curve plotting, Portal development) as discussed above.

While the presentations included a number of rescoping options, the committee felt that very few of them could be implemented without significant impact on LSST science, commissioning, and the overall utility of the LSP, and that few of them would result in any substantial saving for the Project. It was unclear whether any of the options would produce better performance. The LSP Team is already making judicious and effective use of open-source 3rd party tooling. The committee agreed with some of the scoping options and priorities that were presented, for example, the reduction of contextual mouse-over help tips or information buttons which would aid the user in finding any relevant data cross-linkages (e.g., in finding the uncertainties column associated with a particular parameter column).

The committee discussed the rescoping potential of the Project providing a “bare-bones” terminal environment, with fewer pre-installed common-user packages, since users are enabled to install the packages they require. The Project may wish to evaluate what savings in maintenance effort that this might produce. However, a cost/benefit analysis should be performed, since many of the same packages will be required by the LSP itself and/or commissioning teams, so the potential gains could be minimal, and might be eliminated by a corresponding increase in user-support enquiries to the Helpdesk.

Recommendations:

In addition, the committee is asked to provide actionable advice on addressing any issues raised during the course of the review as well as guidance based on experience that will ensure the success of the LSST Science Platform.

1) While the LSP represents an extremely powerful and groundbreaking new approach to astronomical data analysis, the level of demand for its services remains unclear and there is a significant risk that demand could exceed available resources, potentially by a large margin. The committee recommends that the Project:
   a) Reevaluate potential user demand, based on recent usage of the Gaia archive and clearly document the assumptions.
   b) Reassess the performance requirements for user storage and simultaneous query load in light of the new estimates of demand.
   c) Review the FTEs allocated to the long-term maintenance of the LSP in the light of the proposed re-evaluation of user demand.
   d) Perform a cost/benefit analysis of a Cloud-based deployment of the LSP and determine whether to include this in the baseline design as soon as possible.
   e) Explore ways to encourage and support the development of other DACs, ideally distributed across the world offering similar functionality, as well as DACs which may wish to serve a subset of the LSST data with an interface customized to a specific user-base.
   f) Provide baseline “neighbors tables” cross-identifying LSST sources/objects against other major contemporary catalogs, such as the Gaia final data release. These are likely to be required by the commissioning teams, and this will avoid the unnecessary overhead of many users attempting to perform the cross-matching independently. Engage the SCs in defining a suitable baseline.
2) Some elements of the LSP requirements and verification/validation documents remain incomplete. The committee acknowledges legitimate reasons why the development process needed to be able to explore the best available tools, which evolve rapidly. Nevertheless, the specification of the requirements should be completed, which will enable the project’s verification process to be completed and thoroughly applied. The committee feels that this should not delay LSP development.
   a) If a Cloud-based model is adopted for the LSP, the baseline plan should be revised accordingly.
   b) If there will be a point at which no further tools can be added, this should be clearly documented.
   c) The priority of each requirement should also be documented.
   d) The committee recommends that a program of regular reviews be adopted for the LSP throughout the lifetime of LSST, to ensure that the LSP remains responsive to users needs and software trends. At minimum, the committee recommends an additional design review in ~1 year from the date of this report.

3) The Portal aspect is likely to remain an important mechanism for many users to discover, explore and exploit LSST data. Its descoping is likely to be seriously detrimental to a number of science cases (notably Solar System and time-domain science) and significantly hamper the ability of users who might prefer to use a language other than python to access and analyse LSST data. Regardless of language, some tasks benefit from GUI-like visualization tools.
    a) The committee recommends that the Project explore the feasibility of up-scoping the Portal aspect, ideally including time-series plotting functions, or at a minimum, a commitment to launch and maintain it at the current level of functionality. Time-series plotting functionality should be provided within the notebook aspect, even if it has to be descoped from the Portal.
4) The Project should clearly document their plan to ensure fair allocation of resources for a large and diverse community of scientific users, including mechanisms to enable users to temporarily expand their storage/CPU limits for clearly defined periods and a proposal mechanism for users with long-term high-usage requirements. It does not seem reasonable to ask users to submit a proposal when they need a temporary increase in resources; this should be handled via a temporary resource availability option (e.g., at least some scratch space that is cleaned up periodically) which is clearly documented. For longer term, larger or more specialized needs, this could include users purchasing additional computing resources and/or disk volumes for inclusion in the LSP resource pool, either in a local DAC or in the Cloud, and the option to apply for additional resources via competitive proposal.
5) Science users should be introduced to the LSP at the earliest possible opportunity, to afford them time to adapt their workflow and existing analysis software to the new interfaces.
    a) This could be achieved by gradually increasing user-engagement, starting with the LSST Stack Club, in a phased program, and/or reaching out to the HSC science community as with the LSST Stack. The Science Collaborations and Project should be encouraged to ensure that each Science Collaboration is represented by at least one member in the Stack Club.
    b) The LSP Team should engage with each Science Collaboration directly to ensure their users are aware of the LSP capabilities and that their needs are accommodated.
    c) A clear plan and timeline for user engagement should be drawn up and publicized.
    d) If descoping is necessary, the Project should explore ways to enable a limited number of community developers to contribute to LSP functionality.
    e) The resources allocated to users should be clearly explained when an account is allocated to them. A mechanism to identify and manage heavy users should be planned and implemented.
6) The Project should proceed with development of an LSST data (and software) archival plan consistent with requirements in LSR and OSS to allow for long-term scientific reproducibility.

7) The committee supports the VO-first stances adopted by the project, since it ensures compatibility of the LSP with external services such as Vizier and with many commonly-used astronomical analysis packages. This should be continued.
8) The committee make the following recommendations regarding specific technologies included in the LSP:
    a) The Dask software to enable parallelization of tasks appears to be an extremely powerful way to use the available resources with maximum efficiency and minimize computational overhead. A number of alternative parallelization packages (e.g. Parsl) are also available. These should be explored further and, if suitable, should be formally incorporated with the LSP design.
    b) The Project should formalise procedures for supporting a batch processing / bulk image cutout service, providing trivial access to imaging and relevant metadata as stored in a FITS header (e.g., gains, exposure times, WCS information).
    c) The Project should ensure a tool exists to quickly visualize the full focal plane in time for commissioning.
9) Although the qserv database was outside the scope of this review, it is tightly integrated with the LSP functionality and was covered in the Team's presentations. The committee have no concerns about the impressive qserv development effort, but note two science use cases that it would be useful for the project to track:
    a) The Project should consider including a priority-2 requirement for a mechanism to allow the random sampling of database tables in a reproducible way.
    b) The Project should consult with stakeholders (DESC, and others) to ensure that Qserv can efficiently support standard 2- and 3-pt correlation function estimators, especially in light of the 1 arcmin margin parameter presented.

Closeout Comments

The LSP represents the next major evolution in astronomical data analysis, providing a set of powerful tools which can convincingly meet the needs of the Project and commissioning teams and make “next-to-the-data analysis” a credible reality for the wider science community. Without platforms such as the LSP, much of the science from LSST would not be viable owing to the sheer size of the dataset, and its user-base would be restricted to a privileged few. The committee congratulates the team on its development of a set of innovative tools which will not only enable the project to fulfill its enormous scientific potential, but also set a new standard for all future astronomical data services.

Appendix A:

Agenda for the Design Review Meeting
NOAO Main Conference Room, Tucson, AZ,
April 10-12, 2019

SESSION	TOPIC	SCHEDULING

Wednesday April 10th

10.1 Wed AM 1 9:00-10:30	Welcome and goals/charge LSST & DM project context LSP vision & high level requirements	15+5 min 15+5 min 30+10 min
	Break	
10.2 Wed AM 2 10:50-12:30	LSP functional architecture and alignment with IVOA standards; cross-Aspect integration LSST Science Platform Demo Overall development status and timeline, including milestones for integration and for provision of services to upcoming project phases; risks	30+10 min 15 min 30+10 min
12:30-13:30	Lunch	
10.3 Wed PM 1 13:30-15:10	API Aspect detailed requirements and design: overall picture and web services API Aspect project management status API Aspect technical details: ADQL and Qserv (query capabilities)	20+10 min 15+5 min 15+5 min
	Break	
10.4 Wed PM 2 15:30-17:15	Science Data Model and associated tooling Next-to-data processing architecture Authentication and authorization; cyber-security considerations Single Sign On demo Dinner on your own	15+5 min 15+5 min 20+10 min 15 min

SESSION	TOPIC	SCHEDULING

Thursday April 11th

11.1 Thu AM 1 09:00-10:30	Q and A Portal Aspect detailed requirements and design (supplemental PDF) Portal Aspect project management status	15+5 min 20+10 min 15+5 min
	Break	
11.2 Thu AM 2 10:50-12:30	Notebook Aspect detailed requirements and design Notebook Aspect project management status	20+10 min 15+5 min
	Lunch	
11.3 Thu PM 1 13:30-15:10	Infrastructure and Operational considerations Integration and test (verification) planning Scientific testing (validation) planning	30 min 15+5 min 15+5 min
	Break	
11.4 Thu PM2 15:30-17:15	Unallocated time Informal drinks & dinner with project personnel	

SESSION	TOPIC	SCHEDULING

Friday April 12th

	Q and A time	30 min
	Supplement: Commissioning visualization use cases	
	Committee Time	
	Closeout	30 min

Appendix B

Incompletely specified requirements

Requirements that were missing or incompletely specified in the documentation include (but may not be limited to):

LDM-554 (LSST Science Platform Design):

1.6 Control and Management

1.6.1 Alert Subscription

1.7 Documentation

1.7.1 Documentation

2.11 General

2.11.1 Identity and Security

4.5 Performance

4.6 Control and Management

4.7 Documentation

References

Masci, F.J. et al., 2018, PASP, 131, 995.
LSST Science Collaborations, 2009, LSST Science Book, v2.0, arXiv:0912.0201