What to Expect of the LSST Archive: The LSST Science Platform

Mario Juric,
University of Washington
LSST Data Management Subsystem Scientist

for the LSST Data Management Team.
<table>
<thead>
<tr>
<th>Level 1</th>
<th>Level 2</th>
<th>Level 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>LSST Data Products: Level 1, 2, and 3</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- A stream of ~10 million time-domain events per night, detected and transmitted to event distribution networks within 60 seconds of observation.
- A catalog of orbits for ~6 million bodies in the Solar System.
- A catalog of ~37 billion objects (20B galaxies, 17B stars): ~7 trillion observations (“sources”), and ~30 trillion measurements (“forced sources”), produced annually, accessible through online databases.
- Reduced single-epoch, deep co-added images.
- User-produced added-value data products (deep KBO/NEO catalogs, variable star classifications, shear maps, …)

*For more details, see the “Data Products Definition Document”, [http://ls.st/lse-163](http://ls.st/lse-163)*
LSST Mission Goals and System Vision

The LSST will be a facility whose primary mission is to acquire, process, and make available to the data-rights holders the data collected by its telescope and camera. Our primary products are the stream of events alerts (Level 1) and Data Release data products (Level 2).

To make those products available and useful to the community, we’re building Data Access Centers (one in the U.S., and one in Chile). These will expose the LSST data to the data rights holders through a number of well-integrated data access center services.

We call those services the “LSST Science Platform”.
The **LSST Science Platform** is a set of integrated web applications and services deployed at the LSST Data Access Centers (DACs) through which the scientific community will access, visualize, subset and perform next-to-the-data analysis of the data.
The LSST Science Platform exposes the underlying DAC services services through three user facing aspects: the Portal (novice), the JupyterLab (intermediate), and the Web APIs (expert and remote tools).

Through these, we enable access to the Data Releases and Alert Streams, and support next-to-the data analysis and Level 3 product creation using the computing resources available at the DAC.
The Web Portal to the archive will enable browsing and visualization of the available datasets in ways the users are accustomed to at archives such as IRSA, MAST, or the SDSS archive, with an added level of interactivity.

Through the Portal, the users will be able to view the LSST images, request subsets of data (via simple forms or SQL queries), construct simple plots, and generally explore the LSST dataset in a way that allows them to identify and access (subsets of) data required by their science case.

This will all be backed by a petascale-capable RDBMS.
LSST Portal: The Web Window into the LSST Archive

We currently have an initial version of the Portal running at NCSA.

Datasets:
- SDSS Stripe 82
- NEOWISE

Soon:
- HSC (LSST-reprocessed)

*The Firefly Web Science User Interface (Wu et al, 2016; ADASS)*
JupyterLab: Next-to-the-data Analysis

The tools exposed through the Web Portal will permit simple exploration, subsetting, and visualization of LSST data. They may not, however, be suitable for more complex data selection or analysis tasks.

To enable that next level of next-to-the-data work, we plan to enable the users to launch their own Jupyter notebooks at our computing resources at the DAC. These will have fast access to the LSST database and files. They will come with commonly used and useful tools preinstalled (e.g., AstroPy, LSST data processing software stack).

This service is similar in nature to efforts such as SciServer at JHU, or the JupyterHub deployment for DES at NCSA.
JupyterLab: Next-to-the-data Analysis

YouTube demo of the LSST JupyterLab Aspect Demo: [http://ls.st/bgt](http://ls.st/bgt)
Web APIs: Integrating With Existing Tools

Backend Platform services – such as access to databases, images, and other files – will be exposed through machine-accessible web APIs.

We have a preference for industry standard and/or VO APIs (e.g., WebDAV, TAP, SIA, etc.) – the goal is to support what’s broadly accepted within the community. This will allow the discoverability of LSST data products from within the Virtual Observatory, federation of the LSST data set to other archives, and the use of widely utilized tools (eg., TOPCAT or others).
Computing, file storage, and personal databases (the “user workspace”) will be made available to support the work via the Portal and within the Notebooks.

An important feature is that no matter how the user accesses the DAC (Portal, Notebook, or VO APIs) they always “see” the same workspace.
How big is the “LSST Science Cloud” (@ DR2)?

- **Computing:**
  - ~2,400 cores
  - ~18 TFLOPs

- **File storage:**
  - ~4 PB

- **Database storage**
  - ~3 PB

This is shared by all users. We’re estimating the number of potential DAC users not to exceed 7500 (relevant for file and database storage).

Not all users will be accessing the computing cluster concurrently. **We are estimating on order of a ~100.**

Though this is a relatively small cluster by 2020-era standards, it will be **sufficient to enable preliminary end-user science analyses** (working on catalogs, smaller number of images) and creation of some added-value (Level 3) data products.

**Think of this as having your own server with a few TB of disk and database storage, right next to the LSST data, with a chance to use tens to hundreds of cores for analysis.**

For larger endeavors (e.g., pixel-level reprocessing of the entire LSST dataset), the users will want to use resources beyond the LSST DAC (more later).
Level 3: Added-value Data Products

- Level 3 Data Products: Added-value products created by the community

- These may enable science use-cases not fully covered by what we’ll generate in Level 1 and 2:
  - Custom processing of deep drilling fields
  - SNe photometry (e.g. CFHT-LS type forward modeling)
  - Extremely crowded field photometry (e.g., globular clusters)
  - Characterization of diffuse structures (e.g., ISM)
  - Custom measurement algorithms
  - Catalogs of SNe light echos

- The user computing and storage present in the LSST Science Platform are meant to enable next-to-the-data realization of use cases like the ones above.

- Level 3 software/data products may be migrated to Level 2 (with owners’ permission); this is one of the ways how Level 2 products will evolve.
How we (think) we will work with LSST data?

**LSST Science Platform**

- **Portal**
- **JupyterLab**
- **Web APIs**

- **Data Releases**
- **Alert Streams**
- **User Databases**
- **User Files**
- **User Computing**
- **Software Tools**

**LSST Users**
How we (think) we will work with LSST data?

− Most users are likely to begin with the Web Portal, to become familiar with the LSST data set and query smaller subsets of data for “at home” analysis. Some may use the tools they’re accustomed to (e.g., TOPCAT, Aladin, AstroPy, etc.) to grab the data using LSST’s VO-compatible APIs.

− Some users may choose to continue their analysis by utilizing resources available to them at the DAC. They’ll access these through Jupyter notebook-type remote interfaces, with access to a mid-sized computing cluster. It’s quite possible that a large fraction of end-user (“single PI”) science may be achievable this way.

− For users who need larger resources, they may be able to apply for more resources at adjacent computing facilities. For example, U.S. computing is located in the National Petascale Computing Facility at the National Center for Supercomputing Applications (NCSA). Significant additional supercomputing is expected to be available at the same site (e.g., NPCF currently hosts the Blue Waters supercomputer).

− Finally, rights-holders may utilize their own computing facilities to support larger-scale processing or even put up their own Data Access Centers. As they’re open source, they may re-use our software (pipelines, middleware, databases) to the extent possible.
Putting it all together: the LSST Science Platform

For more details, see the “LSST Science Platform Vision Document”, http://ls.st/lsp