LSST Data Management Overview

DM Project Manager

LSST Rio
25th September 2018
Outline

Introduction

LSST status

Data Management Overview

Data Management Recent Achievements

Conclusion
Milky way analogue

Can we unravel the formation, composition, and evolution of the Galaxy?

Our view is severely obstructed by the dust in the disk and relatively little is known about the origin, history, and structure of our own Galaxy.
Origin of the Milky Way

Simulation frame: Amina Helmi

Image credit: R. Jay GaBany
Beyond our galaxy

The modern cosmological models can explain all observations, but need to postulate dark matter and dark energy (though gravity model could be wrong, too).
Killer asteroids: the impact probability is not 0!

LSST is the only survey capable of delivering completeness specified in the 2005 USA Congressional NEO mandate to NASA (to find 90% NEOs larger than 140m).

The Barringer Crater, Arizona: a 40m object 50,000 yr. ago
Outline

Introduction

LSST status

Data Management Overview

Data Management Recent Achievements

Conclusion
LSST: uniform sky survey

An 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: *deep wide fast.*

- 90% of time spent on uniform survey: every 3-4 nights, the whole observable sky scanned twice per night
- 100 PB of data: about a billion 16 Mpix images, enabling measurements for 40 billion objects!

see also http://www.lsst.org and Ivezic et al. (2008)-arXiv:0805.2366

Call for white papers - https://www.lsst.org/call-whitepaper-2018

10-year simulation of LSST survey: number of visits in u,g,r band

(Aitoff projection of eq. coordinates)
LSST Camera

The largest astronomical camera:
- 2800 kg
- 3.2 Gpix
Modular design: $3200 \text{ Megapix} = 189 \times 16 \text{ Megapix CCD}$

9 CCDs share electronics: raft (=camera 144 Megapix)

**First of 21 rafts**

About to accept final raft.
Site shaping up (July 2018)

http://ls.st/8p0
Marine tracking and logistics (1st Sept 2018)

- 9 Pallets of M1M3 lift fixture hardware is MIA! Found in Phoenix warehouse Chapter 7 Trustee Custody

- Valparaiso Express 5 containers held up in Peru, transshipment required to MN Callao Express (already had 10 Containers)

- BBC Arizona Late to Antwerp for Coating vessel & 7 crates

- K&N warehouse Ready in Chile for 15 containers

- Los Angeles Warehouse ready for M2

- 6 EIE Containers in Italy scheduled for Packing in Sept.
Coating Chamber now en Route
LSST Project Schedule
Outline

Introduction

LSST status

Data Management Overview

Data Management Recent Achievements

Conclusion
LSST Project is large and dispersed.

↑ Data Management is just one of five subsystems.
DM’s mission: 
*Stand up operable, maintainable, quality services to deliver high-quality LSST data products for science, all on time and within reasonable cost.*

Development is distributed across the Americas.  
*Plus we have partners like IN2P3.*  

See Management Plan LDM-294.
DM with System Engineering Joint All Hands Meeting IPAC 2018
Leanne Guy replaces Mario Jurić (who stays with the project) as Subsystem Scientist.

Michelle Butler replaces Don Petravick as Infrastructure Product Owner.

New Release Manager Gabriele Comoretto.

Deputies John Swinbank (PM), Colin Slater (PS), Yusra AlSayyad (Pipelines) and Vaikunth Thukral (DAX).

Toughest thing in any project is communication.
DM must build everything to get LSST data products—as described in LSE-163—to end users.

- Large data sets (20 TB/night)
- Complex analysis with small systematics
- Science alerts issued within one minute

$\sim \frac{3}{4}$ million lines of code/comments

(C++/Python/Java/JavaScript/Kotlin)

See SPIE paper by Jenness et al. (2018)


Concept of Operations LDM-230.

Upper diagram courtesy K-T Lim, LDM-148.

Lower diagram by Tim Jenness; covers only the Science Pipelines codebase.
Data Access

– Construction includes US and Chilean (LDM-572) Data Access Centers
  - Data rights and access details being worked on LPM-261
  - US 2400 cores, 3PB DB, 4PB disk, Chile 10% of that

– LSSTC (Corporation) considering International DACs
  - International agencies interested in establishing international DACs (intDACs) - potential added value to LSST science that
    - May prefer national investments in local facilities using funding programs targeting infrastructure investments
    - Such national investments are not part of the scope of the LSST project and operations planning
  - Data access fees are a critical component of provisioning and funding LSST operations
  - Goal: need to structure an approach to intDACs in a way that is practical, seamless and beneficial to the full LSST science community while not undermining the operations funding
    - LSSTC (D. MacFarlane) have setup a working group to consider criteria for intDAC
All Comes Down to Images

Simulated image based on three filters
From just one of 189 CCDs

Processed through the
*LSST Science Pipelines*
https://pipelines.lsst.io/
The Pipelines are already in use with other facilities, e.g. Hyper Suprime-Cam.

Still working on performance, algorithmic enhancements, orchestration, etc.

Design LDM-151; Test Specs LDM-533;
LDM-534; Test Reports DMTR-52; DMTR-53
Fast start with pipelines

Find release you want at https://hub.docker.com/r/lsstsqre/centos/tags/

```bash
>docker run -ti lsstsqre/centos:7-stack-lsst_distrib-v16_0
```

You may see.. Unable to find image ‘lsstsqre/centos:7-stack-lsst_distrib-v16_0’ locally

Setup up the environment ..

```bash
>source
/opt/lsst/software/stack/loadLSST.bash
```
Alerts: identify time varying object

Figure from Eric Bellm
Brokers

- LSST Mini Broker - Users can create filters that return
  - a subset of LSST alerts based only on data in the alert packet
  - can use lightcurve, variability parameters, colors, etc.,
  - no crossmatch to external catalogs
  - Runs in the LSST Data Access Center (-> users must have data rights)

- Community Alert Brokers - further enabling science on alerts e.g.:
  - Provide public access to alerts
  - Classification and Crossmatch to other catalogs or data streams
  - Provide filtering, visualization, and search
  - Coordinate scientific activity and/or followup observations
  - Aggregate alert annotations (community classifications, etc.)

- Mini Broker capacity/Number of community brokers limited by bandwidth
  - see LDM-612
  - Working out selection process for 2019
End-User Documentation on the Web

See lsst.io, pipelines.lsst.io, nb.lsst.io
DM adopted a test driven schedule to address verification and progress.

In line with the approach taken by project-level Systems Engineering — adapting our test specs to Jira test manager system.

LDM-503 (test plan) and LDM-564 (releases) now automatically generated from P6.
High level features of all releases are listed in LDM-564. These coincide with the test milestones.

Example **LDM-503-7**: Camera Data Processing

– Camera package supporting the LSST Camera. (DM-DRP-5)

– LSSTCam data display and visualization. (DM-SUIT-6)

– Mapping between SUIT systems & NCSA auth system. (DM-SUIT-7)

– SUIT portal integrated with workspace. (DM-SUIT-8)

– Basic instrument signature removal (ISR) capability. (DM-AP-7)

– Camera package supporting the Commissioning Camera. (DM-DRP-38)

– Calibration product generation in support of basic ISR. (DM-DRP-4)

The test specs provide further refinement on the delivered product. The bracketed tag is a level 3 milestone in P6.
Verification and Validation

**Verification**: Have we built everything we are supposed to build?
- In line with the Project’s System Engineering approach
- Demonstrate that we cover all requirements on DM
- LDM-503 shows the DM verification matrix

**Validation**: Have we built the right thing and does it work as expected?
- Must tackle *both* Scientific and Operational Validation
- Talking with Commissioning Team: some *rehearsals* will be joint
- **There is no difference between DM Science Validation and Commissioning Science Validation** — DM will work closely with Commissioning Team.

**LDM-503** addresses DM’s plans for verification & validation.
High Level Goals

- 2018: Prototypes for various processes and databases - “Minimum Viable System”
  - Aug: Mountain base network up
  - Oct: “Generation 3” pipeline execution middleware
  - Nov: Ready for spectrograph data acquisition
  - Dec: Prototype QA/Commissioning Environment
- Dec 2019: ComCam L1, L2 Production
- Dec 2019: Base Center Integration Complete
- Jun 2020: Camera L1, L2 Production
- Jul 2021: US Data Access Center Integrated

LDM-564 describes functionality provided by each milestone.
Test specs to confirm milestone completion under development.

LDM-532; LDM-533; LDM-534; LDM-535; LDM-536
LDM-537; LDM-538; LDM-539; LDM-540; LDM-541
November 2019: DM for Commissioning (minimum required for start of commissioning with ComCam):

- Pipeline: single-frame measurement including ISR, ghost masking, cosmic ray detection, PSF estimation, astrometric and photometric calibration, background estimation, single-frame deblending, master calibration image generation, atmospheric characterization

- Services: archiving, EFD transformation, Data Backbone for files (Base/NCSA), telemetry gateway, OCS-controlled batch, offline processing

- LSST Science Platform on Commissioning Cluster: Notebook Aspect, image access, user file storage, batch computing

Milestones:

- LDM-503-09 – 2018-11-30: Demonstrate and rehearse commissioning capabilities

- LDM-503-11a – 2019-10-21: Verify that we are ready for commissioning ComCam

(See LSE-79 §3.3 and table 8)
Early commissioning is coming soon!!!

- Auxiliary Telescope is in Chile already — the spectrograph will arrive after summer.
- We plan to be taking and processing data early 2019!
  - Official operations late 2019 — milestone LDM-503-08 in preparation for that.
- We will have a pipeline for processing AuxTel data on the commissioning cluster.
- Initially *selected* data will be transferred to NCSA and made available in the Science Platform.
- By summer 2019 we will have an automated processing system for AuxTel data.
LDM-503-2: HSC (Hyper Suprime Cam) reprocessing milestone

- First (equal) post-replan, NSF-visible milestone hit by the DM project.
- Joint effort to reprocess (Data Facility team) and analyze (DRP team) HSC data under operations-like conditions
- Milestone successful: DMTR-51!
LDM-503-2: HSC (Hyper Suprime Cam) reprocessing milestone

- First (equal) post-replan, NSF-visible milestone hit by the DM project.
- Joint effort to reprocess (Data Facility team) and analyze (DRP team) HSC data under operations-like conditions
- Milestone successful: DMTR-51!

“Warp Compare” coadds

- New algorithm to robustly reject artifacts when coadding images.
- Now default for HSC processing; stack-wide default to be RFCed soon.
– **LDM-503-3: Alert generation milestone**
  - First(equal) post-replan milestone hit by the DM project DMTR-53!
  - Demonstrating a *end-to-end* alert production pipeline.

– Prototype alert distribution system using Kafka & AVRO; benchmark results on DMTN-028.

– New MOPS (Moving Object) linking algorithm under development and approach to Minor Planet Center DMTN-087.

– **Jointcal replaces meas_mosaic**
  - Simultaneous astro- and photometric fitting to source lists derived from multiple images.
  - The all new, much improved, more generic replacement for the HSC-specific meas_mosaic.

Figure shows the variation in photometric calibration not captured by single frame processing, normalized to 1. This demonstrates fine structure in photometry which Jointcal picks up but per-CCD processing doesn’t catch.

*Figure: Parejko.*
Successful transfer of digital data over LSST/AURA fiber optic networks from the Summit Site on Cerro Pachon to NCSA. A set of 6 x 10 Gbps Network Interface cards on Data Transfer Nodes (DTN) configured with iPerf3 generated a sustained data rate of approximately 44 gigabits per second, over a period of 24 hours, exceeding the target of 40 gigabits per second.
Data Access Services

- Catalog Database (Qserv) to 100 TB range
  - Three 30-node clusters operating:
    - NCSA (PDAC): science dataset (Stripe 82 + AllWISE + NEOWISE)
    - CC-IN2P3 (2 x dev): synthetic dataset
  - 30% DR1 KPM measurements DMTR-17
  - Deployment under Kubernetes
- Gen 3 Data Butler and Supertask in progress
LSST Data Facility (NCSA)

– Observatory Operations Support (Level 1) Services

- Working within the LSST Systems Engineering Early Pathfinder group, developing and testing integration of T&S, Camera, and DM service software via a series of early integration activities.
- Initial header service developed and configured for Camera subsystem and AuxTel use cases, ability to acquire pixel data and write FITS files, all commandable by OCS. Demonstrated on Level 1 Test Stand (DMTR-61) and with Spectrograph in Tucson.

EIA-341 reading image from the spectrograph in Tucson July 2018
Science Platform Vision to Reality (1.1)


Portal/Browser
Notebooks
Web API
(Batch)
Outline

Introduction

LSST status

Data Management Overview

Data Management Recent Achievements

Conclusion
Conclusion

- LSST is getting very real!
- A lot of work on verification for DM
- AuxTel coming very rapidly.
- Looking forward to first data next year.
Questions?

~ 3.5' SDSS image

http://www.lsst.org http://community.lsst.org

HSC image (COSMOS)

$g, r(1.5 \text{ hrs}), i(3 \text{ hrs})$ PSF matched co-add ($\approx 27.5$)

Images: Lupton and HSC collaboration see also Lupton et al. (2004)
Outline

Backup Slides

Reference material
Outline

Backup Slides

Reference material
<table>
<thead>
<tr>
<th>Acronym</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>AI</td>
<td>Action Item</td>
</tr>
<tr>
<td>AP</td>
<td>Alerts Production</td>
</tr>
<tr>
<td>API</td>
<td>Application Programming Interface</td>
</tr>
<tr>
<td>AURA</td>
<td>Association of Universities for Research in Astronomy</td>
</tr>
<tr>
<td>AVRO</td>
<td>Apache data serialization system</td>
</tr>
<tr>
<td>BBC</td>
<td>German shipping company</td>
</tr>
<tr>
<td>C</td>
<td>Specific programming language (also called ANSI-C)</td>
</tr>
<tr>
<td>CCD</td>
<td>Charge-Coupled Device</td>
</tr>
<tr>
<td>D</td>
<td>Deutschland (Germany)</td>
</tr>
<tr>
<td>D</td>
<td>Specific project phase (production; concluded by QR and FAR)</td>
</tr>
<tr>
<td>DAX</td>
<td>Data Access Services</td>
</tr>
<tr>
<td>DB</td>
<td>DataBase</td>
</tr>
<tr>
<td>DM</td>
<td>Data Management</td>
</tr>
<tr>
<td>DMTN</td>
<td>DM Technical Note</td>
</tr>
<tr>
<td>DMTR</td>
<td>Data Management Test Report</td>
</tr>
<tr>
<td>DRP</td>
<td>Data Release Production</td>
</tr>
<tr>
<td>DTN</td>
<td>Data Transfer Node</td>
</tr>
<tr>
<td>EFD</td>
<td>Engineering Facilities Database</td>
</tr>
<tr>
<td>EIA</td>
<td>Early Integration Activity</td>
</tr>
<tr>
<td>EIE</td>
<td>European Industrial Engineering - Italian engineering company (Dome)</td>
</tr>
<tr>
<td>FITS</td>
<td>Flexible Image Transport System</td>
</tr>
<tr>
<td>Acronym</td>
<td>Definition</td>
</tr>
<tr>
<td>---------</td>
<td>------------</td>
</tr>
<tr>
<td>HSC</td>
<td>Hyper Suprime-Cam</td>
</tr>
<tr>
<td>IPAC</td>
<td>Infrared Processing and Analysis Center</td>
</tr>
<tr>
<td>IR</td>
<td>Infra Red</td>
</tr>
<tr>
<td>ISR</td>
<td>Instrument Signal Removal</td>
</tr>
<tr>
<td>K</td>
<td>Kelvin; SI unit of temperature</td>
</tr>
<tr>
<td>KPM</td>
<td>Key Performance Metric</td>
</tr>
<tr>
<td>L1</td>
<td>Level 1 (ambiguous could mean milestone or processing)</td>
</tr>
<tr>
<td>L2</td>
<td>Level 2 (ambiguous could mean milestone or processing)</td>
</tr>
<tr>
<td>LDM</td>
<td>LSST Data Management (handle for controlled documents)</td>
</tr>
<tr>
<td>LPM</td>
<td>LSST Project Management (Document Handle)</td>
</tr>
<tr>
<td>LSE</td>
<td>LSST Systems Engineering (Document Handle)</td>
</tr>
<tr>
<td>LSST</td>
<td>Large Synoptic Survey Telescope</td>
</tr>
<tr>
<td>M2</td>
<td>Second mirror</td>
</tr>
<tr>
<td>MIA</td>
<td>Missing In Action</td>
</tr>
<tr>
<td>MN</td>
<td>Meeting Minutes</td>
</tr>
<tr>
<td>MOPS</td>
<td>Moving Object Pipeline System</td>
</tr>
<tr>
<td>N</td>
<td>Newton; SI unit of force</td>
</tr>
<tr>
<td>NASA</td>
<td>National Aeronautics and Space Administration (USA)</td>
</tr>
<tr>
<td>NCSA</td>
<td>National Center for Supercomputing Applications</td>
</tr>
<tr>
<td>NEO</td>
<td>Near-Earth Object</td>
</tr>
<tr>
<td>NSF</td>
<td>National Science Foundation</td>
</tr>
<tr>
<td>OCS</td>
<td>Observatory Control System</td>
</tr>
<tr>
<td>PB</td>
<td>PetaByte</td>
</tr>
<tr>
<td>PDAC</td>
<td>Prototype Data Access Center</td>
</tr>
<tr>
<td>Acronym</td>
<td>Definition</td>
</tr>
<tr>
<td>---------</td>
<td>------------</td>
</tr>
<tr>
<td>PM</td>
<td>Project Manager</td>
</tr>
<tr>
<td>PS</td>
<td>Project Scientist</td>
</tr>
<tr>
<td>PSF</td>
<td>Point Spread Function</td>
</tr>
<tr>
<td>QA</td>
<td>Quality Assurance</td>
</tr>
<tr>
<td>Qserv</td>
<td>Query Service, Proprietary LSST Database system</td>
</tr>
<tr>
<td>S</td>
<td>Strip (CCD chip along-scan coordinate identifier in focal plane)</td>
</tr>
<tr>
<td>SDSS</td>
<td>Sloan Digital Sky Survey</td>
</tr>
<tr>
<td>SPIE</td>
<td>the international society for optics and photonics</td>
</tr>
<tr>
<td>SUIT</td>
<td>Science User Interface and Tools</td>
</tr>
<tr>
<td>T&amp;S</td>
<td>Telescope and Site</td>
</tr>
<tr>
<td>TB</td>
<td>TeraByte</td>
</tr>
<tr>
<td>US</td>
<td>United States</td>
</tr>
<tr>
<td>USA</td>
<td>United States of America</td>
</tr>
<tr>
<td>arcmin</td>
<td>arcminute, minute of arc (unit of angle)</td>
</tr>
<tr>
<td>kg</td>
<td>kilogram; SI unit of mass</td>
</tr>
<tr>
<td>s</td>
<td>second; SI unit of time</td>
</tr>
</tbody>
</table>
References


References II


[DMTN-028], Patterson, M.T., 2018, Benchmarking a distribution system for LSST alerts, DMTN-028, URL https://dmtn-028.lsst.io, LSST Data Management Technical Note


[DMTR-17], Thukral, V., 2018, Qserv Fall 17 Large Scale Tests/KPMs, DMTR-17, URL https://ls.st/DMTR-17


