LARGE SYNOPTIC SURVEY TELESCOPE

Large Synoptic Survey Telescope (LSST)
LSST Document

Data Management Database Requirements

Jacek Becla

LDM-555

Latest Revision: 2017-06-30

Change Record

Version	Date	Description	Owner name
1.0	2013-10-10	Requirements included in LDM-135	J. Becla
2.0	2017-06-29	Initial version of standalone requirements extracted directly from LDM-135.	T. Jenness
2.1	2017-06-30	Correct requirements prefix DM- to DMS- for consistency with LSE-61	T. Jenness
2.2	2017-06-30	Remove draft status	T. Jenness

Document source location: MagicDraw SysML
Version from source repository: 23

Contents

1 General Requirements **1**

1.1 Reliability 1

1.2 New Technologies 1

1.3 Incremental Scaling 1

2 Data Production **2**

2.1 Level 1 Database Public Updates 2

2.2 Level 1 Schema Stability 2

2.3 Level 1 Database Reliability 2

2.4 Engineering and Facility Database 3

2.5 Level 1 Database Performance 3

2.6 Data Release Schema Stability 3

2.7 Data Release Production ingest 3

2.8 Calibration Database 4

2.9 Alert Database 4

3 Query Access **4**

3.1 Query Complexity 4

3.2 Flexibility 4

3.3 Reproducibility 5

3.4 Cross-matching with external/user data 5

Data Management Database Requirements

The key requirements driving the LSST database architecture include: incremental scaling, near-real-time response time for ad-hoc simple user queries, fast turnaround for full-sky scans/correlations, reliability, and low cost, all at multi-petabyte scale. These requirements are primarily driven by the ad-hoc user query access.

Database performance requirements are specified in the DM Requirements Document (LSE-61).

1 General Requirements

1.1 Reliability

ID: DMS-DB-REQ-0003

Specification: The system, including disaster recovery backup storage, must not lose data, and it must provide at least **dbSystemUpTime** up time in the face of hardware failures, software failures, system maintenance, and upgrades.

Description	Value	Unit	Name
Minimum uptime percentage for all database systems.	98	percent	dbSystemUpTime

1.2 New Technologies

ID: DMS-DB-REQ-0002

Specification: New technologies that become available during the life of the system must be able to be incorporated easily.

1.3 Incremental Scaling

ID: DMS-DB-REQ-0001

Specification: The system must scale to tens of petabytes and trillions of rows. It must grow

as the data grows and as the access requirements grow. Database sizes are described in LDM-141.

2 Data Production

2.1 Level 1 Database Public Updates

ID: DMS-DB-REQ-0007

Specification: The public-facing view of the L1 database must be updated within at least **L1PublicT** of the visit being observed.

Description	Value	Unit	Name
Maximum time from the acquisition of science data to the public release of associated Level 1 Data Products (except alerts)	24	hour	L1PublicT

2.2 Level 1 Schema Stability

ID: DMS-DB-REQ-0008

Specification: The Level 1 database shall contain the entire history of Level 1 data products. It shall be designed such that the schema can be modified during the lifetime of the survey so long as query results are not altered.

Discussion: The internal database can be rebuilt when not observing.

2.3 Level 1 Database Reliability

ID: DMS-DB-REQ-0004

Specification: The live Level 1 Database shall be implemented such that external access shall not be disabled for extended periods of time for maintenance.

Discussion: Alerts need to be generated in under a minute after data has been taken, data has to be ingested/updated in almost-real time. The number of row updates/ingested is modest: 40K new rows and updates occur every 39 sec.

2.4 Engineering and Facility Database

ID: DMS-DB-REQ-0012

Specification: The Engineering and Facility Database (EFD) shall be made available at the Archive Center in a form optimized for queries with latency less than **efdArchiveCenterLatency**.

Description	Value	Unit	Name
Maximum delay between an item appearing in the EFD and it being accessible for querying from the Archive Center.	3600	second	efdArchiveCenterLatency

2.5 Level 1 Database Performance

ID: DMS-DB-REQ-0005

Specification: The Alert Production system shall have access to all previous Level 1 catalog data when performing source association.

2.6 Data Release Schema Stability

ID: DMS-DB-REQ-0009

Specification: It shall be possible to modify the schema for a data release after the release has been made, so long as query results do not change.

Discussion: Example of non-altering changes including adding/removing/resorting indexes, adding a new column with derived information, changing type of a column without loosing information, (eg., FLOAT to DOUBLE would be always allowed, DOUBLE to FLOAT would only be allowed if all values can be expressed using FLOAT without loosing any information). Each data release can have a different schema to the previous data release.

2.7 Data Release Production ingest

ID: DMS-DB-REQ-0010

Specification: Ingestion of Data Release catalogs into the Data Release database shall be

incremental and asynchronous.

Discussion: The ingestion needs to be independent of ongoing calculations in the Data Release Production, and it cannot wait until all Data Release Production calculations have completed.

2.8 Calibration Database

ID: DMS-DB-REQ-0011

Specification: A database shall contain queryable results from the Calibration Products Pipeline. This database shall be bitemporal, providing information on when a particular value was valid, and not solely the most recently calculated value.

2.9 Alert Database

ID: DMS-DB-REQ-0006

Specification: The contents of all alerts shall be maintained in a database, searchable by alert ID.

3 Query Access

3.1 Query Complexity

ID: DMS-DB-REQ-0015

Specification: The system shall handle complex queries, including spatial correlations, and time series comparisons. Spatial correlations are required for the Object catalog only.

Discussion: Spatial queries require highly specialized, 2-level partitioning with overlaps.

3.2 Flexibility

ID: DMS-DB-REQ-0016

Specification: Catalogs in relational databases shall be queried using a dialect of ADQL supporting as many features of the standard as feasible.

Discussion: ADQL is a superset of SQL92, and can therefore be used to query the alert database and EFD as well as spatial catalogs. Native SQL supported by the underlying database system, with LSST-specific extensions, might also be desirable, but without the promise of longevity if the backend database technology changes during the survey lifetime.

3.3 Reproducibility

ID: DMS-DB-REQ-0013

Specification: It must be possible to issue queries on any Level 1 and Level 2 data products that will produce the same results when re-issued at any time before the next Data Release and when re-issued with at most minor, well-documented modifications at any time thereafter.

3.4 Cross-matching with external/user data

ID: DMS-DB-REQ-0014

Specification: Users shall be able to cross-match the LSST catalogs with external catalogs. Some catalogs shall be provided by LSST, whilst other catalogs can be uploaded by the user. Results from these cross-matches can be used in subsequent queries.

Discussion: Example catalogs hosted by LSST will be those from SDSS, SKA or Gaia.

References

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

[2] **[LSE-61]**, Dubois-Felsmann, G., Jenness, T., 2018, *LSST Data Management Subsystem Requirements*, LSE-61, URL `https://ls.st/LSE-61`