Astronomical Data Analysis Software and Systems XVI
ASP Conference Series, Vol. XXX, 2007
R. A. Shaw, F. Hill and D. J. Bell, eds. **O1.1**

Designing for Peta-Scale in the LSST Database

Jeffrey Kantor

LSST Corporation, Tucson, AZ, USA

Tim Axelrod

Steward Observatory, University of Arizona, Tucson, AZ, USA

Jacek Becla

Stanford Linear Accelerator Center, Stanford University, Menlo Park, CA, USA

Kem Cook, Sergei Nikolaev

Lawrence Livermore National Laboratory, Livermore, CA, USA

Jim Gray

Microsoft Research, San Francisco, CA, USA

Ray Plante

National Center for Supercomputing Applications, University of Illinois, Urbana, IL, USA

Maria Nieto-Santisteban, Alex Szalay, Ani Thakar

Johns Hopkins University, Baltimore, MD, USA

Abstract. The *Large Synoptic Survey Telescope* (LSST), a proposed ground-based 8.4 m telescope with a 10 deg^2 field of view, will generate 15 TB of raw images every observing night. When calibration and processed data are added, the image archive, catalogs, and meta-data will grow 15 PB yr^{-1} on average. The LSST Data Management System (DMS) must capture, process, store, index, replicate, and provide open access to this data. Alerts must be triggered within 30 s of data acquisition. To do this in real-time at these data volumes will require advances in data management, database, and file system techniques.

This paper describes the design of the LSST DMS and emphasizes features for peta-scale data. The LSST DMS will employ a combination of distributed database and file systems, with schema, partitioning, and indexing oriented for parallel operations. Image files are stored in a distributed file system with references to, and meta-data from, each file stored in the databases. The schema design supports pipeline processing, rapid ingest, and efficient query. Vertical partitioning reduces disk input/output requirements, horizontal partitioning allows parallel data access using arrays of servers and disks. Indexing is extensive, utilizing both conventional RAM-resident indexes and column-narrow, row-deep tag tables/covering indices that are extracted from tables that contain many more attributes. The DMS Data Access Framework is encapsulated in a middleware framework to provide a uniform service interface to all framework capabilities. This framework will provide the automated work-flow, replication, and data analysis capabilities necessary to make data processing and data quality analysis feasible at this scale.

1. Introduction

The *Large Synoptic Survey Telescope* (LSST) is a proposed ground-based 8.4 m telescope with a 10 deg^2 field of view that will capture digital images of faint astronomical objects across the entire sky, night after night, in a ten-year survey. In a relentless campaign of 15 s exposures, LSST will cover the available sky every three nights, opening a movie-like window on objects that change or move on rapid timescales: exploding supernovae, potentially hazardous near-Earth asteroids, and distant Kuiper Belt Objects. The superb images from the LSST will also be used to trace billions of remote galaxies and measure the distortions in their shapes produced by lumps of Dark Matter, providing multiple tests of the nature of the mysterious dark energy.

The LSST Data Management System (DMS) will produce:

- An image data archive of millions of raw and calibrated scientific images, each image being available within 24 hr of capture
- Thousands of verified alerts of transient and moving objects detected within the last 60 s, every time a new pair of images is captured
- Astronomical catalogs containing billions of stars and galaxies, richly attributed in both time and space dimensions and setting a new standard in uniformity of astrometric and photometric precision.

The archive will be openly accessible via direct query or for fusion with other astronomical surveys. The user community will vary widely, from students having only personal computers and commodity internet access to researchers processing terabyte-sized sections of the entire catalog on a dedicated supercomputer cluster or across a scientific grid. The workload placed on the system by these users will be actively managed to ensure equitable access to all segments of the user community. The system that produces and manages this archive will be robust enough to keep up with the LSST's prodigious data rates, with an operational staff typical of today's much smaller surveys. It will never lose a byte of data and will improve over time in terms of both the quantity and quality of the data. This system will be initially constructed and subsequently refreshed using commodity hardware rather than "bleeding edge" technology to ensure affordability, even as technology evolves.

The design, implementation, and operation of the LSST DMS pose a number of challenges:

- The system must, of necessity, routinely collect large volumes of data at high rates. In order to produce the required data products, extensive

computations must be performed on the data at high throughput and, in the case of transient alerts, with low latency.
- Performance of these functions must extend over a wide range of time scales—from roughly 15 s between images, to over a decade for the generation of some catalog data products, and to multiple decades for the curation of all data products.
- High reliability and availability are essential; routine human intervention to correct data processing problems would be impractical, inefficient, and cost-prohibitive.
- Given the rate of technological change in computing, we estimate that hardware will change generations four times over the life of the LSST survey. This implies that the DMS must be capable of operating in a heterogeneous hardware environment. The software will be more stable but it, too, will evolve multiple times over the survey life.

These challenges have driven the LSST DMS designers to incorporate a number of significant architectural features. First, the proposed system is distributed across four specialized types of LSST-managed facilities: the Mountain Summit, Base Facility, Archive Center, and Data Access Centers. The distributed structure is essential to maintaining balance between near real-time vs. non real-time requirements, and between storage and network capacity for data reduction/data product generation vs. scientific/community access and analysis.

This vertical partitioning is, however, only half of the story. The LSST requirements and challenges for scalability, reliability, modularity, and evolutionary design are best addressed by what is commonly known as a *layered* architecture, a concept that is widely employed in information systems with these requirements. This horizontal layering separates the primarily custom application software from the physical infrastructure (i.e., hardware and system software) with middleware software technology. This layering is ubiquitous in the DMS and incorporated into the design of each of the individual Centers of the DMS:

- The *Application Layer* embodies the fundamental roles and responsibilities of the LSST DMS; i.e. it is where the scientific algorithms, pipelines, and data products implementations reside. Notable in this layer is the inclusion of highly automated data quality analysis features, a key to making the system maintainable with an operational staff similar to existing observatories. This layer is implemented as custom-developed, open-source software.
- The *Middleware Layer* enables portability by providing a thin, abstract, and uniform interface to the hardware and system software; it also provides standard services to the application layer for security, reliability, plug-in pipeline algorithms and stages, and extendable data types. This layer is implemented by custom integration of off-the-shelf, open-source software.
- The *Infrastructure Layer* provides all the hardware and system software in the DMS computing, storage, and network resources that are the environment in which the Application and Middleware layers are hosted and execute. This layer is implemented as off-the-shelf, commercial hardware and system software.

In order to estimate the size of the DMS, we developed a comprehensive analytical model driven by input from the requirements specifications. Key input parameters include the processing operations per data element, the data transfer rates between and within processing locations, the data ingest and query rates, the alert generation rates, and latency requirements. The resultant performance and sizing requirements show the DMS to be a *supercomputing-class* system, with over 100 TFLOPS of aggregate processing power and correspondingly large data input/output and network bandwidth rates.

To define communications requirements, we developed a model of the data transfers and user query/response load, extrapolated from existing surveys, and adjusted to LSST scale. With the exception of a new fiber to be installed from the Mountain Summit on Cerro Pachon to the Base Facility in La Serena, Chile, existing fiber and research networks and their successors will carry LSST data communications.

To derive processing requirements, we extrapolated from the functional model of operations/element and from existing precursor survey pipelines (SDSS, DLS, SuperMACHO, MACHO, and Raptor) adjusted to LSST scale. To derive storage and input/output requirements, we extrapolated from the data model of LSST data products, pre-cursor schemas (SDSS, 2MASS), and existing DBMS overhead factors in precursor surveys (SDSS, 2MASS, BaBar) adjusted to LSST scale. Working in the petabyte regime frequently requires non-conventional techniques and innovative solutions. Optimizing queries is extremely important in LSST's world of multi-billion-row tables. As an example, the Source table is expected to reach over 2×10^{11} rows in the first year of data taking. Triggering a single scan through this table (which can be done through an innocent looking "SELECT * FROM Source" query) would result in fetching 100 TB of data from disks. Further, having to support simultaneous, multi-dimensional (spatial and temporal) searches significantly complicates the database design.

The set of queries used for sizing and costing the system is representative enough to drive the schema design and cover both the public access as well as the professional astronomers' access. They were generated based on experience from other astronomical experiments and archives (SDSS, MACHO, IRSA), carefully selected such that both spatial and temporal aspects are well covered. They are aligned with the precursor schema that was used for sizing and testing and were optimized using rewriting, clustered index preferred over non-clustered, and other well-known techniques. The key schema optimizations include:

- Pre-calculating the most frequently used fields to minimize full table and/or full index scans, e.g., storing colors in addition to bandwidths.
- Extracting the most frequently used information into summary (tag) tables to de-randomize disk I/O, e.g., extracting information about variable sources and replicating it in a much smaller table than the original table containing all sources.
- Preparing summaries to reduce size of searched data set, e.g., extracting time-dependent information like wavelets from the Source table and storing it in the smaller Object table.
- Maintaining and using optimal indexes. This includes preparing covering indexes, and building highly specialized spatial indices (based on R-trees and the Hierarchical Triangular Mesh developed by the SDSS team).

- Employing query plans and estimates and row/byte/time limits to queries to ensure that a few users do not dominate the available resources to the exclusion of others.

The key layout optimizations include appropriate data clustering, table partitioning and choosing optimal data and index block sizes (based on extensive database size and input/output models). Finally, based on the survey reference design, we established an overall availability percentage for the DMS, as a constraint for the design effort, since this drives the amount of spare capacity and redundancy that must be included in the design. We defined the availability in terms of data preservation and the ability to keep up with the data volume.

2. LSST Facilities and Data Flows

As described above, the infrastructure layer provides the total hardware and system software for DMS computing, storage, and networking. Processing and storage activities are distributed across four DMS facilities: the *Mountain Summit*, *Base Facility*, *Archive Center*, and *Data Access Centers*. Many features to support reliability and scalability have been designed into this layer, including spare computing and storage capacity, and redundant communications links. Each DMS area operates 24/7 and is staffed for both nighttime (observing) and daytime (non-observing) shifts. Distribution of processing and storage activities among separate facilities promotes efficient performance of the system via a separation of real-time and non-real-time resources.

The Mountain Summit site will be on Cerro Pachon in Chile. Its sole DMS responsibility is to capture data and move it down to Base Facility for nightly pipeline processing. The Data Acquisition System (part of the Camera Subsystem) and the Observatory Control System (OCS, part of the Telescope Subsystem) interface to the DMS on the summit, with image read-out in 2 s and immediate data transfer to the Base at 10 Gb s^{-1} on fiber optic lines dedicated to this traffic. Meta-data from the Engineering and Facility database in the OCS moves to the Base on a dedicated 1 Gb s^{-1} line on a nightly basis. The Mountain Summit also has a data buffer sufficient for 4 nights of data in case of communications failure to the Base Facility. There is a redundant link to the Base Facility for fail-over and in the extremely unlikely event of simultaneous failure in the primary and backup links it is possible to transport one set of the backup data drives to the base.

The Base Facility will be located at the Cerro Tololo Inter-American Observatory in La Serena, Chile. The Base Facility's primary role is processing the image data to generate transient alerts within the required latency. The Nightly Data Pipelines and the most recent previously processed co-added sky templates and catalogs are hosted here on a 25 teraflops-class computing cluster to provide primary data reduction and transient alert generation within the required 60 s. Similar to the Mountain Summit, the Base Facility has sufficient capacity to store 4 nights worth of data in case of a communications failure to the Archive Center.

The Base to Archive Center network is a 2.5 Gb s^{-1} full duplex, protected, clear channel fiber optic circuit, with protocols optimized for bulk data transfer. The term *protected* means that the link is a ring with traffic moving in both directions; only a dual outage on both sides of the ring between source and destination will prevent data delivery. *Clear channel* means that LSST will have a dedicated portion of the network bandwidth guaranteed by an Indefeasible Right to Use (IRU). This link is used both for high priority command and alert traffic, as well as to trickle the raw image data over a 24 hr period and to send processed data back to the Base Facility. In the event of an outage in this link, once service is restored the link supports a 400% on-demand burst increase in capacity for transmission of buffered data.

The Archive Center will be co-located at the National Center for Supercomputing Applications (NCSA) at the University of Illinois in Champaign, Illinois. This center publishes the alerts via public alerting protocols based on the Virtual Observatory *VOEvent* standard. It also repeats the Base Facility processing (recall that only the raw image and meta-data are transferred from the Base in order to keep the data rates feasible) and merges the processed data into the Science Data Archive. The Archive Center computing resources handle the processing needed to generate Nightly and Data Release data products, as well as the planned reprocessing required. Next, the Archive Center acts as the source of the replication operations that deliver the data to the Data Access Centers, and along with the latter supports end user data queries and data access. Finally, the data releases are published from the Archive Center. These responsibilities require supercomputers capable of 100 TFLOPS (year 1 of the survey) interfaced to a 15 PB yr^{-1} data archive.

The Data Access Centers will be located, at a minimum, in the United States and Chile. Together, they provide a backup copy of the Science Data Archive and additional access capacity for external users. No data reduction processing beyond servicing data access requests is performed at the Data Access Centers, except for on-demand capacity available for end users. The Centers are connected to Archive Center via a high-speed 10 Gb s^{-1} network (TeraGrid in the United States, REUNA in Chile). We anticipate that the scientific collaborations may wish to establish additional Data Access Centers to host selected data products based on relevance to the communities served. Funding for these additional data centers will be provided by the host institutions.

With its huge repository of data, LSST will greatly accelerate the advance of "e-Science" for professional astronomers through the use of the Virtual Observatory tools and interfaces combined with the advanced computational resources available both at large, end-user sites and, more generally, through the Grid. In addition to the data, the data processing and analysis software will be openly available for deployment not only on proprietary resources (for groups who have funds to build and manage significant computational centers) but also for deployment onto NSF-funded supercomputing and Grid resources. While the primary data processing will be executed on dedicated computing resources, and additional computing resources are budgeted for query support, the DMS is being designed to allow expansion onto the Grid and other available resources as needed to provide scalability as the user community grows.

Managing access to core LSST resources will be important in order for the project to achieve its mission-critical goals, both in day-to-day operations and in the support of the core science programs. The concept of *service levels* will be used to manage resources effectively as well as distribute access loads over a variety of shared or collaborative sites. While this hierarchy is somewhat new in astronomical circles, the concept has been quite successful in the high-energy physics context where *tiers* of access and shared responsibilities are common. Level 1 sites will be relatively rare and typically have their own supercomputers dedicated to limited groups, while Level 2 through Level 5 sites will have more limited resources but will be more common. Combining these provides flexible access to cover the gamut of scientific and Education/Public Outreach users.

3. LSST Pipelines

The LSST algorithms are executed by a set of pipelines that are structured to partition the overall functionality of the DMS cleanly. Each pipeline has sole responsibility for producing one or more data products. Several pipelines are run on a nightly basis, first at the Base Facility and later at the Archive Center, producing the Nightly Data Products. The Image Processing Pipeline transforms the raw science image from the sensor into a calibrated science image, which has had the instrumental signature removed and is both astrometrically and photometrically calibrated. This calibration will later be improved at the Archive Center. From the calibrated science image and a template image obtained from the Archive Center, this pipeline produces the subtracted science image. The Detection Pipeline is responsible for detecting and measuring sources in the subtracted science image, producing the Source Catalog. The Association Pipeline is responsible for associating new entries in the Source Catalog with the information on previously known astrophysical objects contained in the Object Catalog. Sources that do not correspond to a previously known object cause a new entry to be made in the Object Catalog. Alerts are generated for those objects whose behavior meets the alert criteria, and these are forwarded to the Archive Center for distribution to the community. The Moving Object Pipeline is responsible for the Orbit Catalog. It takes as input the entries in the Source Catalog that were not matched with known static objects by the Association Pipeline. From these, and the existing Orbit Catalog of solar system objects, it determines which of these unmatched sources correspond to already cataloged orbits, and whether new orbits need to be formed. In the latter case, it can send an alert to the Archive Center.

The remaining pipelines are run on a less frequent cadence (as needed in the case of the Calibration pipeline, roughly every six months for Classification and Deep Detection) only at the Archive Center and produce the Data Release Data Products. The Calibration Pipeline is responsible for producing calibration products such as flats and darks that are needed for the nightly data processing. The Deep Detection Pipeline is responsible for optimally combining information from multiple images to produce the best possible measures of object properties, including shapes, photometry, proper motions, and parallaxes. This information is stored in the Object Catalog. Additionally, the Image Processing Pipeline produces image stacks for use as template images. The Classification Pipeline utilizes the multi-epoch, multi-filter information produced by the Deep Detection Pipeline to classify objects into astrophysically meaningful categories and summarize properties of their time history, such as the waveform and period of periodic variable stars.

4. LSST DMS Data Access Framework

The Data Access Framework (DAF) is the middleware that provides uniform open interfaces to all LSST data, copies and organizes data in preparation for pipeline execution, and incorporates features for data and meta-data extensions.

The Archive Services incorporate data into the Science Data Archive from transfers or pipelines. Thus, this service is used primarily by internal applications although external use is also possible. Primary drivers for this service are high performance and scalability to keep up with LSST data volumes. Underlying this service is a distributed file system (DFS) for image and other files, and a database management system (DBMS) for record-based catalog and meta-data. Both are selected from off-the-shelf solutions and integrated with the rest of the middleware layer. The DAF uses a DFS for:

- Staging input data for pipeline processing
- Staging output data for ingest
- Storing, replicating, and serving image files

Current distributed file systems under evaluation include GPFS, Google File System, Lustre, and IBRIX.

The VO Compliant Interfaces provide support for responding to VO protocols to provide standard formats in response to queries for data. The VO Architecture Overview (IVOA 2006b) provides detailed information on VO protocols and architecture. LSST members from NCSA, Johns Hopkins University, Lawrence Livermore National Laboratory, and NOAO both lead and participate in VO workshops and summer schools, and contribute and review VO standards. By ensuring VO-compliance, LSST enables access via the many VO-community based tools that are being created.

The DAF employs advanced software engineering techniques to support extendable data and meta-data to evolve the system rapidly when the LSST needs to enable astronomical analyses that have not been previously conducted. These techniques have been utilized by LSST designers in previous scientific applications to achieve data extendability and include:

- Providing services that expose logical paths to data and map to physical paths transparently
- Employing context-specific, externally visible labels that are distinct from internal identifiers
- Defining pre-allocated spare attributes, variant records, and pre-defined extension tables in the database
- Utilizing data type handlers that map data types into tools and visualizations dynamically
- Providing deep and shallow cloning operations
- Employing abstract data types and object-oriented data access objects that encapsulate the physical implementation of the data type.

5. Database System Architecture

Systems for managing very large data sets are complicated by nature. The LSST database system must support multi-petabyte scalability, good transactional insert performance, ad-hoc queries with ability to scan data at tens of GB s^{-1}, and high data availability—all that at a reasonable cost. The LSST data ingest rate at the Archive Center will be 30 MB s^{-1} from pipelines, or 1 TB within a 10 hr observing window. One ingest server will be sufficient from a performance standpoint, but there is nothing in the design that would prohibit introducing more ingest servers if needed. Ingest server(s) will be mirrored to provide fault tolerance. Newly ingested data has to be queried together with previously taken (archived) data in real time. To minimize competition between readers and writers, and to facilitate recovery process, ingested data will be transferred to archive data servers in real time, e.g. through replication. Archived data will be partitioned to reduce the amount of searched data. In addition, ingest for data indexing and re-processing brings the total ingest rate across all centers to 6 GB s^{-1}, which will be accomplished by parallelizing the ingest across servers.

The reference design features several independent server farms, each dedicated to serving specific task or a specific group of users and tuned accordingly for the expected access patterns. Two or more data centers will serve data for astronomers and general public. Key design features providing high performance for query access include:

- Caching major indexes in RAM
- Partitioning tables, indexes and queries
- Executing queries at the best place
- Combining disk arrays with a virtual file system to balance the load
- Replicating most frequently used data

The proposed design includes relational database management system (RDBMS) technology, which is much more widely used than object-oriented database management systems (ODBMS) technology. An open source system, MySQL, is currently baselined, but at least one commercial DBMS will be fully evaluated during the research and development phase. A technical “fly-off” and a cost trade-off analysis between the open-source and commercial system(s) will be completed prior to the start of construction.