Astronomical Data Analysis Software and Systems XVIII **O7A.2**
ASP Conference Series, Vol. XXX, 2009
J. Lewis, R. Argyle, P. Bunclarck, D. Evans, and E. Gonzales-Solares, eds.

LSST Science Data Quality Analysis Subsystem Design

Russ R. Laher[1], Deborah Levine, Vince Mannings, Peregrine McGehee, Jeonghee Rho

Infrared Processing and Analysis Center, California Institute of Technology, M/S 100-22, Pasadena, CA 91125

Richard A. Shaw

National Optical Astronomy Observatory, 950 N. Cherry Avenue, Tucson, AZ 85719

Jeff Kantor

LSST Corporation, 933 N. Cherry Avenue, Tucson, AZ 85721

Abstract. The Large Synoptic Survey Telescope (LSST) will have a Science Data Quality Analysis (SDQA) subsystem for vetting its unprecedented volume of astronomical image data. The SDQA subsystem inhabits three basic realms: image processing, graphical-user-interface (GUI) tools, and alarms/reporting. During pipeline image processing, SDQA data are computed for the images and astronomical sources extracted from the images, and utilized to grade the images and sources. Alarms are automatically sent, if necessary, to initiate swift response to problems found. Both SDQA data and machine-determined grades are stored in a database. At the end of a data-processing interval, e.g., nightly processing or data-release reprocessing, automatic SDQA reports are generated from SDQA data and grades queried from the database. The SDQA reports summarize the science data quality and provide feedback to telescope, camera, facility, observation-scheduling and data-processing personnel. During operations, GUI tools facilitate visualization of image and SDQA data in a variety of ways that allow a small SDQA-operations team of humans to quickly and easily perform manual SDQA on a substantial fraction of LSST data products, and possibly reassign SDQA grades as a result of the visual inspection.

1. Introduction

About a year ago, an experienced team of engineers, scientists, and managers at IPAC began working on defining the role for and operational purpose of the LSST SDQA subsystem, as driven by LSST science and functional requirements. The following definition emerged: "An assessment subsystem comprised of substantially automated processes that examines and reports on the quality of LSST science data and derived science data products. The subsystem will enable the project to verify that the data meet LSST science requirements." Shortly thereafter, work on the high-level design and prototyping commenced.

[1]E-mail: laher@ipac.caltech.edu

LSST data processing will be done both nightly and episodically. The latter will be annually or semiannually, and all image data acquired up to date will be reprocessed and its products issued as a "data release". The SDQA subsystem must be able to operate in both processing scenarios, and take full advantage of all information available over different temporal/spatial/color scales.

The SDQA development team has been planning incorporation of SDQA-metric calculational stages into LSST image-processing pipelines, which will be programmed in C++ with "thin" Python wrappers. The SDQA calculational results will be thresholded to yield hard decisions for grading the quality of the observed images and derived products. Real-time alarms will be raised for human intervention if problems are uncovered. SDQA data and grades will be stored in a database for historical trending and visualization. Nightly and data-release SDQA summary reports will be automatically generated.

Humans will be part of the SDQA subsystem during operations. They will be primarily responsible for manually spot-checking and grading the data products shortly after their creation. GUI software will facilitate human analysis of the images, derived products, and SDQA information.

The goal for the first round of development is to focus on just a few SDQA metrics, such as delivered seeing, as test-particles in end-to-end SDQA subsystem development. Other contemplated metrics for prototyping work are critical factors from the difference-image pipeline's convolution-kernel generation.

The purpose of this paper is to give some insight into key aspects of the LSST SDQA subsystem that have evolved in the design and planning thus far.

2. Very Brief LSST Overview

LSST's 10-year mission is to survey the entire visible sky deeply in multiple colors every week. The project is currently in the R & D phase. It will go into construction phase in year 2010, experience camera first-light in year 2015, and begin normal operations several months later.

The mountain-top observatory will be constructed on Cerro Pachón in Chile. The telescope will have an 8.4-m primary mirror and a 9.6 square degree field-of-view. The camera will have: 189 science CCDs in its focal plane, where each CCD has 4K×4K pixels (a 3200-MP digital camera!); 3024 amplifiers/channels to read out the CCD-image data; and 6 optical filters (*ugrizy*), modeled after the Sloan Digital Sky Survey.

During observing, pairs of 15-s exposures will be taken on the same sky position, totaling approximately 3.5×10^5 CCD images per night. A data-acquisition rate of ≈ 15 PB yr^{-1} is projected (where "PB" stands for petabytes).

Kantor *et al.* (2007) give further information on the project and plans for peta-scale data processing and management.

3. SDQA Database Schema & C++/Python Classes

Figure 1 shows our SDQA database-schema design. The schema's tables closely parallel the SDQA classes that we have identified in our UML (universal modeling language) design, which reveals an interesting connection between the

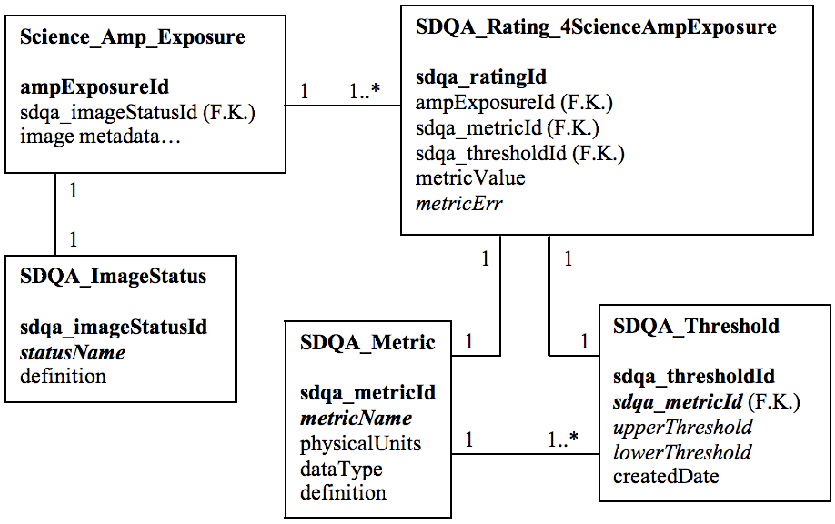

Figure 1. SDQA database-schema design ("F.K." stands for foreign key, and "1 1*.." stands for one record to many records).

database-table structure and related software classes, seemingly a lesson in database-driven software design.

The Science_Amp_Exposure database table stores metadata for processed images associated with independently read-out 512×2048-pixel portions of raw images (called amplifier segments). The sdqa_imageStatusId field in this table points to the grade category determined for the image by the SDQA subsystem. The SDQA_Rating_4ScienceAmpExposure database table is associated with the Science_Amp_Exposure database table in a one-to-many relationship, and stores multiple, what we refer to as image "SDQA ratings".

An SDQA rating is basically the computed value of an SDQA metric and its uncertainty. SDQA metrics are diverse, predefined measures that characterize image quality; e.g., image statistics, astrometric and photometric figures of merit and associated errors, counts of various things, like extracted sources, etc.

A processed image, in general, has multiple SDQA ratings, which are computed at various pipeline stages and stored in pipeline-shared memory. The SDQA pipeline will be executed after the data-processing pipelines have computed the required SDQA ratings. It will threshold SDQA ratings for hard decisions about the image's quality, in order to determine a setting for the sdqa_imageStatusId field, ultimately the image's SDQA grade.

Bulk loading of SDQA_Rating_4ScienceAmpExposure database records will reduce impact of the SDQA subsystem on pipeline throughput.

4. SDQA GUI Tools

GUI and visual-display software tools will be developed in Java, primarily for its platform-independent multi-threading capabilities. Both thick (downloadable applications) and thin (web-browser-based) clients are envisaged, each leveraging a common code base. Applications are generally faster and more powerful,

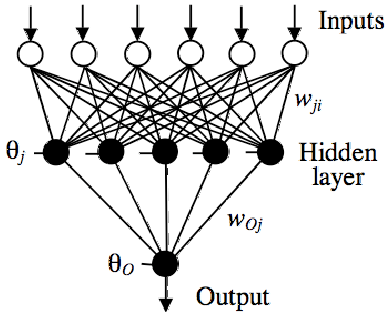

Figure 2. Artificial neural network with single-hidden-layer architecture.

but have to be installed on the host machine. Through advances over time, the very convenient web-browser solution is expected to become a formidable contender. The Google toolkit could be used to convert the Java code into Javascript for relatively trouble-free SDQA subsystem access via popular web browsers. The SDQA GUI will have drill-down capability to selectively obtain additional information about the SDQA metrics reported in nightly summaries.

5. Novel Approaches

In future versions of the SDQA subsystem, we will apply machine-learning techniques to a broad range of metrics (e.g., source counts, photometric depth, environmental factors, and telescope-engineering data) to identify more subtle data quality problems, as well as likely causes of those problems. Artificial neural networks are particularly promising for this application (see Figure 2).

Completeness and reliability assessments of SDQA subsystem performance will be made, for quantitative comparison between both traditional and novel approaches. C & R determination is discussed by Laher, Grant, and Fang (2008).

6. Conclusions

This paper covers some important facets of the current design of, and thinking about, the LSST SDQA subsystem. Its development is incipient, and future work involves defining and effectively utilizing meaningful metrics, and exploring approaches to visualization and assimilation of the information. Progressively more of the SDQA system will be built and tested in a series of project-wide "data challenges" over the next few years. Novel approaches for maximizing SDQA-subsystem completeness and reliability, such as the employment of artificial neural networks, will be investigated.

References

Kantor, J. *et al.* 2007, in ASP Conf. Ser. 376, ADASS XVI, ed. R. A. Shaw, F. Hill, & D. J. Bell (San Francisco: ASP), 376, 3.

Laher, R., Grant, A., Fang, F. 2008, PASP, 120, 1325.