Astronomical Data Analysis Software and Systems XII
ASP Conference Series, Vol. 295, 2003
H. E. Payne, R. I. Jedrzejewski, and R. N. Hook, eds.

The SuperMacho+SuperNova Survey Database Design: Supporting Time Domain Analysis of GB to TB Astronomical Datasets

R. Hiriart and C. Smith

NOAO/CTIO, Casilla 603, La Serena, Chile

Andy Becker

Lucent Technologies, Bell Labs, 600 Mountain Ave., Murray Hill, NJ

C. Stubbs, A. Rest and G. Miknaitis

University of Washington, Dept. of Astronomy, Seattle, WA, USA

Gabe Prochter

Lawrence Livermore Nat. Lab., Livermore, CA 94550, USA

SuperMacho project and ESSENCE project collaborations

Abstract. Two large scale survey projects to discover transient events have recently been initiated at NOAO's Cerro Tololo Inter-American Observatory (CTIO) in Chile. The SuperMacho project seeks to detect and follow microlensing events toward the Large Magellanic Cloud (LMC) and the ESSENCE Supernova project seeks to detect and follow intermediate to high-redshift supernovae. Together, these projects (SM+SN) present challenging data management needs both due to the large size of the datasets and the kind of analysis to be performed on the data.

The database requirements of the SM+SN projects can be divided into three broad categories: support for the survey operation, storage and analysis of the data that comes from the image reduction and transient detection pipeline, and communication of the results to the project users and the astronomical community.

Current relational database technologies are being applied to address these requirements. The open source database PostgreSQL has been selected for the implementation of the system. This work presents the design of the database, along with some performance considerations that are necessary for the fast retrieval of information, thus allowing the development of data mining applications to take full advantage of the database.

1. Introduction

The combination of wide-field CCD detectors and increased computing power has opened the opportunity for the study of astronomical events over large spatial scales **in the time domain**. Over the past few years, interest in such studies has grown, as most recently demonstrated by the growing support for the Large Synoptic Survey Telescope (LSST). At CTIO, we have recently begun two large scale synoptic surveys which serve as precursors to LSST, exploring the parameter space of managing large (GB to TB) datasets in **real-time** in order to achieve the scientific goals of the projects. The SuperMacho project is designed to study microlensing events due to MACHOs (MAssive Compact Halo Objects) passing in front of the background of stars in the LMC. The goal of this survey is to monitor millions of stars in the densest portions of the LMC in order to detect ≈12 microlensing events per year over a five year period. The ESSENCE project aims to constrain the equation of state of dark energy through the study of intermediate redshift ($0.15 < z < 0.75$) supernovae (SNe). In order to accomplish this goal, this survey must discover ≈200 type Ia SNe distributed evenly over the redshift range during the five year lifetime of the project.

These two survey projects have many common data management requirements. Both are designed around an observing cadence of a half night every other night during dark and grey time on the CTIO Blanco 4m telescope. The observing period lasts for three months (October through December). Each produces 10 to 15GB of data per half-night of observing (for a total of ≈25GB per night). Both projects must process this data in near-real-time and automatically detect faint transient sources. These transients must be cataloged, matched against previously known variable sources, and if new, classified and announced to the astronomical community to allow for rapid follow-up on other telescopes around the world. A modern database optimized for time-domain science is necessary to support the combined SM+SN projects and manage the millions of transient detections which are produced.

2. Database Schema

In general terms, the database requirements of the SM+SN projects are:

- Support for the survey operations
- Storage and analysis of the pipeline output data
- Communication with users and the astronomical community

As the observations take place, a considerable amount of information is generated. This information includes: observation (coordinates, exposure time, airmass, seeing sky conditions, etc.), area of sky covered, images produced, and related calibration frames.

The entity **observation** represents the action of pointing a given **optical_system** to a point in the sky, at a given time, and performing an astronomical observation of an **obs_object**. Since the SM+SN survey projects have divided the sky into predetermined *fields*, it makes sense to attach to the **obs_object** entity the attributes that define these fields. As a result of the **observation**, an observation file (**obs_file**) or image is generated and stored in

a given physical location. Associated calibration frames may be taken before or after the **observation** to which they are linked.

The SM+SN pipeline reduces the images creating **reduced_images** entities, after applying cross-talk correction, WCS calibration, bias subtraction and flat fielding. The pipeline then proceeds through image subtraction and transient detection. A set of entities have been added to allow the definition of pipelines in the database. A **pipeline** is composed of individual **stages**. The execution of the pipeline at a given time is represented by the entity **pipeline_run**, which is configured by **pipeline_parameter**s and a set of **stage_parameter**s. The input of a **pipeline_run** is an image and the output is a set of **var_detection**s and **abs_detection**s, which represent the potential transient objects and absolute (non-transient) objects identified by the pipeline. This structure allows us to track the parameters that were used to generate any result stored in the database.

As a result of the transient analysis, a set of detections (**var_detections**) is generated for each processed image. These detections, including positions and all additional information produced by the analysis, are stored in the database, together with the relationships between the detections and the images, observations, and pipeline runs they come from. Because of the inherent precision of the observations and numerical analysis performed by the pipeline, two detections which were presumably generated by the same source at different epochs do not necessarily share the same exact spatial coordinates, although they are very close. As a result, the detections need to be aggregated into *clusters* (**diff_clusters** entity) around the same spatial coordinates before the extraction of lightcurves and object classification. A number of entities have been defined to take into account the different kinds of classifications for an object. As one object can be classified in different ways depending on the classifier, a **diff_classifier** entity along with a ternary relationships have been defined.

To effectively communicate the results of the transient analysis, a web site is being developed which integrates the PostgreSQL database through PHP, Perl, and Python. The delivery of data products to the community will eventually make use of XML-based technologies as these evolve to meet the needs of the surveys and the astronomical community.

3. Performance Considerations

Because of the large sizes of some tables (for year 2001 SuperMacho observations, the size of the **var_detection** table is over 20 million tuples), it is necessary to define convenient access methods over these tables in order to get answers for queries in a practical time. Currently, PostgreSQL implements Hash, B-tree and R-tree indices into its database management system.

One critical query over the database was the search for detections in a box in the sky. An R-tree was created for this purpose, but because it is not possible to define R-trees over more than one column in the current PostgreSQL release, it was necessary to create an additional column in the **var_detection** table of type "box". This type is not part of the SQL standard, but is one of the object-oriented extensions of PostgreSQL. These boxes were initially generated as a function of the Full Width Half Maximum values of the detections, although

we will probably redefine this definition to base the box on the astrometric uncertainty in the detection's position.

4. Data Mining Application: Cluster Analysis

The essential product of the pipeline analysis of the SM+SN is not a list of variable objects, but rather a list of individual (single-epoch) **detections**. Before validation and/or classification of the detections identified by the pipeline can begin, the detections must be grouped over multiple epochs into objects. For stationary objects this is done on the basis of the distances between detections.

To solve this *cluster analysis* problem, a state of the art clustering algorithm, OPTICS (Ankerst et al. 1999) has been implemented to be applied over the detections, creating additional entries and relationships in the database. This algorithm was selected because of its scalability with the number of detections as well as its ability to find subclusters in a given cluster of detections.

The OPTICS algorithm does not produce an explicit clustering for the data, but instead creates two additional attributes per detection: an order index and a *reachability distance*. Roughly speaking, the reachability distance is a measure of density at the location of a given detection, and it is the distance of a point to the set of its neighbors. Plotting the reachability distance for each one of the detections in the order generated by the algorithm, it is possible to reveal the clustering structure of the dataset.

5. Summary

Over the past ten years, the use of modern relational databases has become more common in astronomical contexts, due largely to the growing datasets and the complexity of the information astronomers are trying to track. The time-domain represents a new challenge in astronomical database design and use, especially in the face of the TB datasets of today and the PB datasets of tomorrow (e.g., LSST). Through support of the SuperMacho and ESSENCE projects, and based upon experience from previous surveys such as MACHO and the High-z SN searches, we have begun to explore the application of modern relational database technologies in support of time-domain astronomical research. While the details of the SM+SN database may be specific to the support of these projects, the general flow and large-scale structure of the database should be instructive for future time-domain database support, and many of the data mining tools and applications we are developing (clustering being but one example) will be applicable to future projects such as LSST.

References

Ankerst, M., Kriegel, H. P., & Sander, J. 1999. Proc. 1999 ACM-SIGMOD Int. Conf. Management of Data (SIGMOD'99), pp. 49–60