Astronomical Databases Challenges[♣]

© Oleg Bartunov

Sternberg Astronomical Institute, Moscow University
PostreSQL Global Development Group

oleg@sai.msu.su

Annotation

Modern astronomy undergoes a big change due to a new possibilities enabled by technology development. Large scale survey projects produce huge amount of data, which needs to be processed and organized in databases to provide access by astronomical community. There are many problems in the current state of art of accessing large astronomical databases and organizing programmatic access to the distributed and diverse data, which Virtual Observatory initiative should eventually to solve. One of the most important problem is the effective execution of queries in the very big databases. We expect petabyte databases in the 5 years from several big projects, like Large Synoptic Survey Telescope (lsst.org), PAN-STARRS (pan-starrs.ifa.hawaii.edu), which planned to produce Petabyte/year data size.

Nowadays, it's not unusual to work with billions of objects in terabyte-sized database. Astronomy is the only science which has so many objects. These objects are intrinsically 2-dimensional, and what is more, they are located on the celestial sphere, which makes even basic queries like find objects near some point with fixed radius difficult and crossmath queries using standard algorithms useless. The challenge is to provide execution time about several seconds for easy queries like spatial query and several minutes for catalogs crossmatch.

Huge databases change patterns of data access - it's impossible to download data and do science locally. Users will query databases via VO (Virtual Observatory) services, so we need flexible access policy to the system resources (disk, memory, cpu usage) and handle users quotas in databases.

Clustering algorithms, which tend to be N^2 or N^3 complex, are need to improved to be applicable for petabyte databases.

Astronomical data are not static, the scale and rate of changes are different. We need version management to be able to reproduce scientific results. Current practice is to work with monolithic releases for big catalogs, but there are many rapidly changed catalogs which require version management on the row level.

SAI RVO development group was organized in Sternberg Astronomical Institute, Moscow University, in summer 2005, to meet the requirements of modern astronomy to develop unified access to astronomical data using generally adopted standards. Primary goal of the group is to develop fully functional node of Virtual Observatory in Russia and facilitate solution of typical astrophysical problems using VO technology.

We realized original spatial algorithm for 2-dimensional data with spherical attributes in open-source database PostgreSQL, which allow us to work with several terabytes databases. Our sky-indexing scheme Q3C is available for download from q3c.sourceforge.net. The total number of objects in our database is about 4 billion (10^9) objects. Our hardware, which is HP rx1620 entry-level server, dual Itanium2, 8Gb RAM and MSA 20 storage, was kindly provided by HP Russia. We provide conesearch and crossmatch query via standard web-based interface for interactive work and webservices for programmatic access (vo.astronet.ru) . We developed uniform access to the diversed catalogs with the help of metadata catalog.

We're working on developing of VO registry - a searchable directory of VO services, with additional full-text search of astronomical papers archive (arxiv.org) to find information about astronomical objects. We developed full-text search engine in PostgreSQL, which supports online index update and users pluggable methods for document parsing and lexemes processing.

We participate in the creation of scalable data processing and storage data center of Moscow University. We'll use data center to store scans of SAI Glass Library - photos of sky for more than hundred years. The largest plate is 30x30 cm with scan size about 4Gb. There are about 60,000 plates of different sizes and we estimate the total size in about 20 Tb. Images will be accessed using SIAP (Simple Image Access Protocol), all image metadata will be stored in PostgreSQL and indexed using our Q3C sky-indexing scheme.

[♣] The SAI RVO project is being developed in the framework of the Astronet project, supported by RFBR (Russian Foundation for Basic Research), grant 05-07-90225.

Proceedings of the Spring Young Researcher's Colloquium On Database and Information Systems SYRCoDIS, St.-Petersburg, Russia, 2007