Astronomical Data Analysis Software and Systems XIII
ASP Conference Series, Vol. 314, 2004
F. Ochsenbein, M. Allen, and D. Egret, eds.

PostgreSQL: the Suitable DBMS Solution for Astronomy and Astrophysics

Igor Chilingarian[1,2,5], Oleg Bartunov[2,3,6], Janko Richter[4], Teodor Sigaev[3,6]

[1] *Sternberg Astronomical Institute, MSU*

[2] *Special Astrophysical Observatory, RAS*

[3] *Delta-Soft LLC*

[4] *International Meteor Organization*

[5] *A participant of MIGALE (Prugniel et al., this conference) project, responsible for DBMS subsystem*

[6] *Official members of PostgreSQL development team*

Abstract. PostgreSQL, the open-source ORDBMS, is one of the best solutions for science. Compared to several available commercial and non-commercial database engines, it appears to be the most versatile. Extensibility is the most remarkable feature of PostgreSQL – it allows to develop custom data types, queries and indexed access methods, optimized for specific tasks. We present two contribution modules for PostgreSQL: pgSphere, offering the capability for dealing with geometrical objects in spherical coordinates, and pgAstro, based on the pgSphere, providing astronomy-specific functions and methods

1. What is PostgreSQL and why we're using it?

PostgreSQL is an object-relational database management system (ORDBMS) based on Postgres v.4.2, developed at the University of California at Berkeley CS Department. PostgreSQL provides SQL92/SQL99 language support and numerous powerful features making it well-suited for different scientific and technological tasks. A lot of object-relational concepts assisting in modern commercials were pioneered in Postgres.

The main concepts and features of PostgreSQL are:

- rapidly developing open source freely distributed DBMS
- SQL support and object-relational concepts
- simple and handy front-end interfaces for different software platforms
- extensibility of the DB server functions, i.e. developing of custom data types and data access methods

These features allow PostgreSQL to be used in different scientific projects. At present it is being used in the following projects related to astronomy:

1) HyperLEDA database, http://leda.univ-lyon1.fr/, a part of MIGALE project.
2) SAI Astronomical Databases, http://www.sai.msu.su/database.html
3) TASS, The Amateur Sky Survey, http://www.tass-survey.org/
4) MAPS, Minessota Automated Plate Scanner, http://aps.umn.edu/

Usually scientific objectives imply the dealing with data types different from integer and floating point numbers, strings, timestamps and money, provided by the standard SQL. For example, many of astronomical and astrophysical tasks require effective operation with celestial coordinates. This implies the 2D indexing of the positions on sphere to achieve high performance on large datasets. Unfortunately, no standard solution exists for this problem in modern DBMSs and there is no standard data types even for 2D objects on cartesian plane. So, the extensibility of the DBMS becomes the most valuable feature. Let's consider several database solutions available on the market to compare them and conclude about their suitability for astronomy:

- Oracle is a market leader. It is full featured database solution, it is extensible and high-performance, it has support for user-written data types and access methods and for GiST access methods, but it is very expensive.
- MS SQL Server has similar feature set, but it can be used for Windows only. SDSS project successfully uses this DBMS. It is also quite expensive.
- DB2 (Informix). It is another example of full featured solution, and it is used in several astronomical projects, such as NED. Again, it is expensive.
- Sybase is similar to DB2 by the abilities, it is traditionally used in many astronomical applications.
- MySQL is open source RDMBS with a reputation for efficiency. But is not extensible and feature set is rather poor. Anyway it is quite cheap solution for static datasets.
- PostgreSQL is open source and easily extensible, has extremely rich feature set, supports GiST access methods. Unfortunately, many of its features are poorly documented

So, PostgreSQL is the only extensible free open source DBMS solution.

2. Extensibility of PostgreSQL

As noted before, the extensibility becomes the most important feature of the DBMS to be used in science. PostgreSQL provides very wide possibilities for extending the database and adopting it to the raised objective.

1) PostgreSQL allows to create user-defined functions and aggregates in the upper layer using SQL or one of the available procedure languages. This feature is quite common for the most of the DBMSs. Also it is possible to create custom data types and use these high level functions for dealing with them.

2) PostgreSQL provides a powerful functionality for so called back-end programming. This allows developer to create functions in a low-level language (i.e. C), compile them and load dynamically into the running database server as shared objects. Binary code usage increases the performance dramatically. Moreover, the standard interface to GiST (Generalized Search Tree) is provided to create custom data types with indexed access methods and extensible set of queries for specific domain experts not a database one.

GiST was implemented in an early version of PostgreSQL by J. Hellerstein and P.Aoki, more details is available from "The GiST Indexing Project" (http://gist.cs.berkeley.edu/) at Berkeley. As an "university" project it has a limited number of features and was in rare use. Since version 7.1 of PostgreSQL the GiST was taken up by Oleg Bartunov and Teodor Sigaev. Current implementation of GiST supports:

- Variable length keys
- Composite keys (multi-key)
- provides NULL-safe interface to GiST core

But GiST cannot be used to implement such well known multi-dimensional indexing methods as Hierarchical Triangular Mesh, because HTM is a kind of Space Partitioning Trees. More general index structure called SP-GiST (Aref et al.) exists for dealing with SP-Tree algorithms. It also can be implemented as extension to PostgreSQL.

Several extensions to PostgreSQL based on GiST interface exist. They're described here: http://www.sai.msu.su/˜ megera/postgres/gist/ We'll emphasize the pgSphere extension, useful for astronomy more then the others.

3. pgSphere project and concepts of pgAstro

We have developed pgSphere contribution module, http://www.pgastro.org/cgi-bin/wiki.pl?pgSphere for PostgreSQL using backend programming and GiST interface. It is distributed under BSD license. It introduces data types for geometrical objects on a sphere and access methods for them. The project is hosted by Gborg, http://gborg.postgresql.org/projects/pgsphere
pgSphere provides the following functionality:

- input and output of spherical data (points, circles, polygons, ellipses, boxes) in several formats (radians, degrees, DMS, HMS)
- containing, overlapping and other operations for spherical objects
- various input and converting functions and operators
- calculation of circumference and area of spherical objects
- spherical transformations
- indexed data access methods for spherical data types

Hence it is possible to do a fast search and analysis for objects with spherical attributes, using PostgreSQL. For instance it is possible to manage data for geographical objects on the Earth or astronomical catalogs conveniently using a SQL interface. The main goal of pgSphere is to provide an uniformed access to spherical data.

Several performance tests were made with different datasets. We used Tycho catalog and its parts to compare the performance of GiST R-tree based algorithm implemented in pgSphere to 2-column B-tree index on celestial coordinates. The selection of objects within 6 by 6 degrees area from 10^6-record dataset takes about 2 ms using pgSphere and 17 ms using 2-column B-tree. More details about the benchmarks are available in the full electronic version of this paper at http://www.sai.msu.su/˜chil/ADASSXIII_poster.pdf

PgSphere is close to the first stable release now, and we hope to finish it available before January 2004. Now it can be downloaded from CVS repository.

Using pgSphere module it becomes possible to solve some astronomical tasks using SQL queries.

We are introducing pgAstro contribution module, distributed under GPL2 license. It will be a set of tools on SQL-layer and backend layer devoted to astronomical tasks. Two possible applications are clear now:

1) Positional astronomy. Some astrometric functionality will be included, for instance, it will be possible to do cone search for a given epoch and equinox taking into account proper motions to calculate precession and nutation on the fly, to check if the given object belongs to the given constellation etc.

2) Coordinate based cross-correlation. This task is important for identifying objects in different catalogs.

4. Conclusions

From the given examples PostgreSQL appears to be the most versatile DBMS solution for astronomy and astrophysics. It is easily extensible, has powerful set of features well comparable to leading commercial database solutions. The fact that PostgreSQL is freely distributed open source software indicates a very important advantage. Many people can create contributions useful for scientists, which is hardly possible with any commercial database solutions.

The further features of PostgreSQL will include XML support. It may be very useful for many VO applications and tools.

Acknowledgments. Our development is supported by the Russian Foundation for Basic Research, projects #02-07-90222 and #03-07-06116. Also we greatly appreciate PostgreSQL community, TASS Amateur Sky Survey working group, especially Robert Creager and Chris Albertson. Great thanks to ADASS-XIII organizing committee for financial support, provided to complement our attendance the conference.

References

Aref, W. et al., http://www.cs.purdue.edu/homes/aref/dbsystems_files/SP-GiST/

Baruffolo, A., & Benacchio, L. 1998, in ASP Conf. Ser., Vol. 145, ADASS VII, ed. R. Albrecht, R. N. Hook, & H. A. Bushouse (San Francisco: ASP)

Baruffolo, A. 1999, in ASP Conf. Ser., Vol. 172, ADASS VIII, ed. D. M. Mehringer, R. L. Plante, & D. A. Roberts (San Francisco: ASP)

Page, C. 2003, in ASP Conf. Ser., Vol. 295, ADASS XII, ed. H. E. Payne, R. I. Jedrzejewski, & R. N. Hook (San Francisco: ASP), 39

Page, C. http://www.star.le.ac.uk/~ cgp/ag/skyindex.html