Geophysical Research Abstracts
Vol. 19, EGU2017-7203, 2017
EGU General Assembly 2017

Hosting and pulishing astronomical data in SQL databases

Anastasia Galkin, Jochen Klar, Kristin Riebe, Gal Matokevic, and Harry Enke
Leibniz Institute for Astrophysics Potsdam (AIP), Germany (agalkin@aip.de)

Abstract

In astronomy, terabytes and petabytes of data are produced by ground instruments, satellite missions and simulations. At Leibniz-Institute for Astrophysics Potsdam (AIP) we host and publish terabytes of cosmological simulation and observational data. The public archive at AIP has now reached a size of 60TB and growing and helps to produce numerous scientific papers.

The web framework Daiquiri offers a dedicated web interface for each of the hosted scientific databases. Scientists all around the world run SQL queries which include specific astrophysical functions and get their desired data in reasonable time.

Daiquiri supports the scientific projects by offering a number of administration tools such as database and user management, contact messages to the staff and support for organization of meetings and workshops. The webpages can be customized and the Wordpress integration supports the participating scientists in maintaining the documentation and the projects' news sections.

1 Astronomical data at AIP archive

To name a few scientific databases in use at AIP since 2013:

- AIP is one of the four data centers to publish the Gaia data starting with the Gaia Data Release 1 from September 2016. Gaia@AIP services include auxilliary catalogues for crossmatch purposes. Gaia@AIP on https://gaia.aip.de/
- MultiDark and Bolshoi simulations on CosmoSim https://cosmosim.org
- Digitized archive for astronomical photographic plates APPLAUSE https://www.plate-archive.org
- RAVE Survey archive https://www.rave-survey.org

All of the projects listed above rely on the SQL database technology for data retrieval.

2 The setup

Depending on the size of the database, the setup is based on either sharded or single MariaDB database nodes. The sharded nodes are orchestrated by a head node which runs the MariaDB Spider engine. ParallelQuery (PaQu) reformulates the SQL queries for the use in distributed environments.

The Daiquiri framework can be used with the parallel MariaDB database setup including PaQu and the Spider Engine or with a single SQL node.

All software we develop is published under an Open Source license.

- Daiquiri on GitHub: https://github.com/aipescience/daiquiri
- PaQu on GitHub https://github.com/adrpar/paqu
- DBingestor on Github https://github.com/aipescience/DBIngestor
- UWS client on GitHub https://github.com/aipescience/uws-client

Another essential part of the system besides the software and the hardware are the data curators. Their role is the most crucial to the publishing of the data. The job of a data curator begins well before the data is ingested in the public database and even with the public release there is always space for improvements suggested by the scientists or by the data curators themselves.

The current setup is in use since 2013. Since then we identified several areas that will be improved in the coming up version of the framework in Python relying on the Django framework. The deployment of the new system is planned for 2017 with the newest MariaDB version.