Enabling data discovery in big datasets

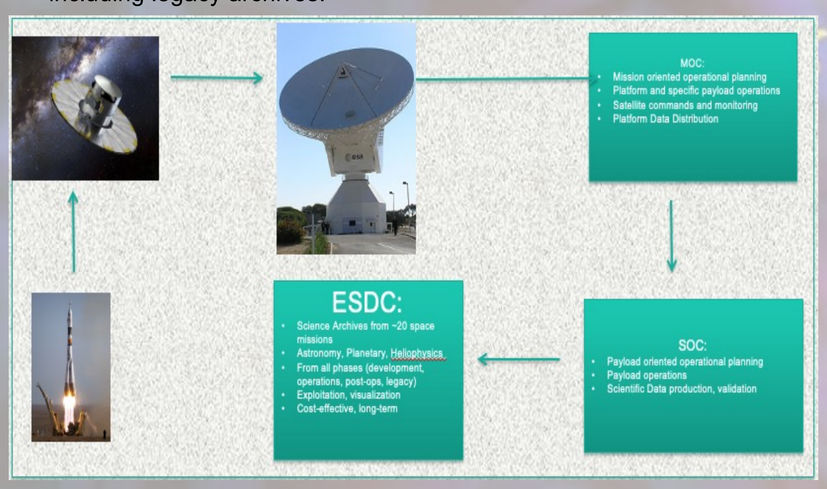
Pilar de Teodoro, AURORA BV for ESA
Sara Nieto, Rhea group for ESA
Monica Fernández, Rhea group for ESA
Hector Pérez, Rhea group for ESA
Christophe Arviset, ESA
ESAC Science Data Centre (ESDC), European Space Astronomy Centre (ESAC), Madrid, Spain

Abstract

The ESAC Science Data Centre (ESDC) is handling the archive data for several astronomy, solar and planetary missions. We started with some gigabytes of information, currently in the hundreds of terabytes and not so far in the future we will handle petabytes. An important fraction of them reside in database systems which allows to analyse the data, structured, semi and unstructured, directly in the ESDC systems using VO protocols. How to store this data in a database to give the users the ability to query easily the contents of a space mission?. How do we choose a solution that will handle some small data to one that scales better for big data?. May this be a nightmare? One does not fit all, but maybe in the future it well may happen. We will review the evolution of database solutions for big data space projects with special focus on the ones that we have already tested (PostgreSQL, CitusDB, PostgresXL, Greenplum) with specific implementation for the Gaia DR3 release, the European JWST archive, the Euclid Science Archive and the future PLATO Data Archive.

How ESDC gets the data

The ESDC is the library for ESA missions catalogues. It allows the searching on its data for all scientific missions in all its phases from development to operations phase. It is distributed into helio, human and robotic exploration, observatory, survey and planetary missions including legacy archives.

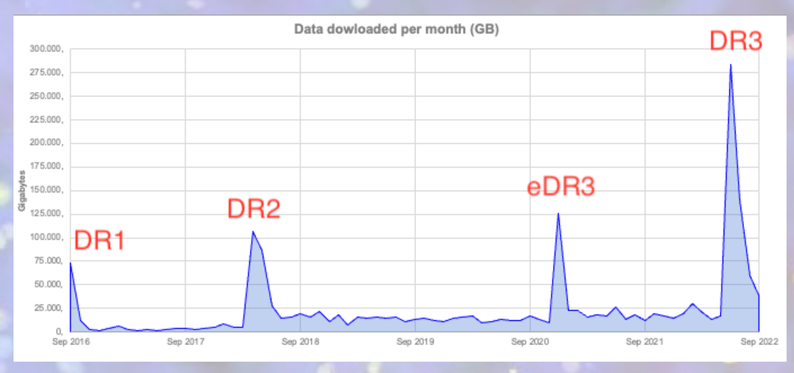


ESAC archives

The main webpage of the ESDC archives can be found in http://archives.esac.esa.int. From that page you can navigate to any ESAC archive.

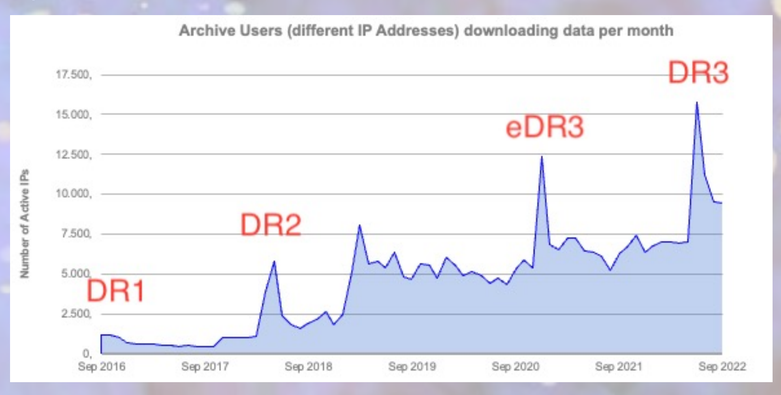
Gaia Data downloaded

The Gaia monthly statistics clearly show an increasing trend in the downloads associated to the data releases.



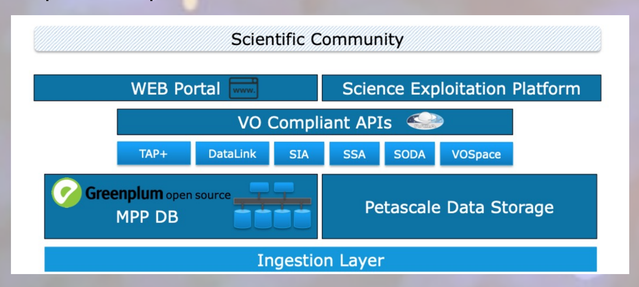
Gaia Archive Users

The amount of users is growing with each release for Gaia. We can analyze accesses to the archives, showing why and what was important to the users. Statistics are measured for all the archives.



System Stack

When the scientific data is ready it will be ingested in our databases(mainly metadata and/or catalogues) and if there is a data repository it will be stored in a petascale data storage. Then, using VO protocols, the data will be able to be retrieved via a web portal, Jupyter notebooks or under the science exploitation platform.



Distributed databases

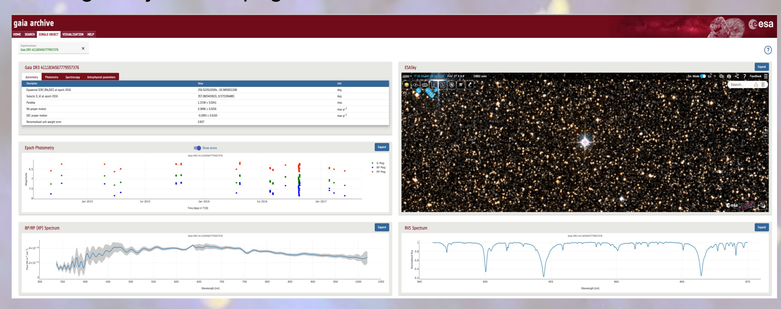
We have compared parallel databases based on PostgreSQL for typical data base server requirements and specific astronomical use cases:

- Scalability queries
- Aggregation queries
- Geometrical Queries
- Xmatch queries

We determined that Greenplum was the best solution to use with IVOA TAP but also Spark may be used for cut-off products.

Gaia Archive New features

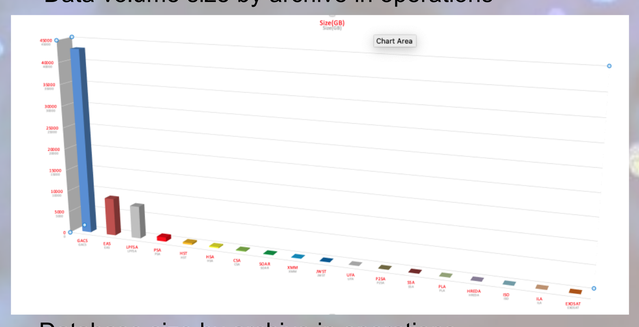
VO protocols are used in the forms to query the archives and also through astroquery. Gaia Archive is also adding new features such as the single object web page.



September 2022 archives size

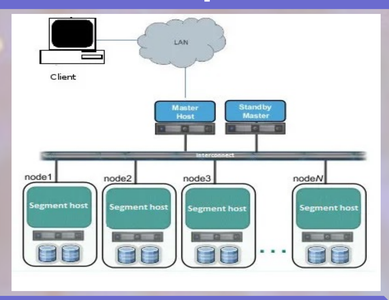
The data on the archives is stored directly in the database or in filesystem. The first option allows to run queries directly on the data using a database management system, while the second option, will store only the metadata into the database and will locate the data in the filesystem to retrieve it. Eamples of the first option is Gaia and Euclid where all the catalogues are stored inside the database.

Data volume size by archive in operations

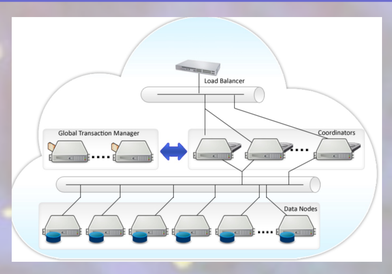


Database size by archive in operations

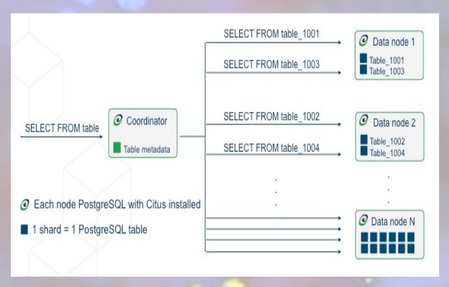
Greenplum



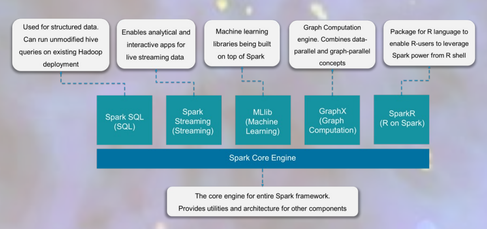
Postgres-XL



CitusDB

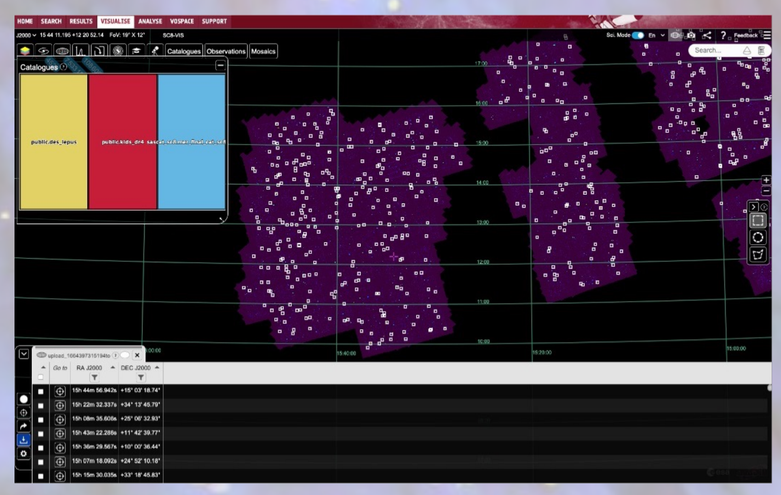


Spark



Euclid Archive Visualization

Euclid visualization is based on ESASky and will allow the users to search for specific data, postcards, run cutoffs on the images…



Pilar de Teodoro
pilar.teodoro@ext.esa.int

PV2023@CERN
02-04 May 2023