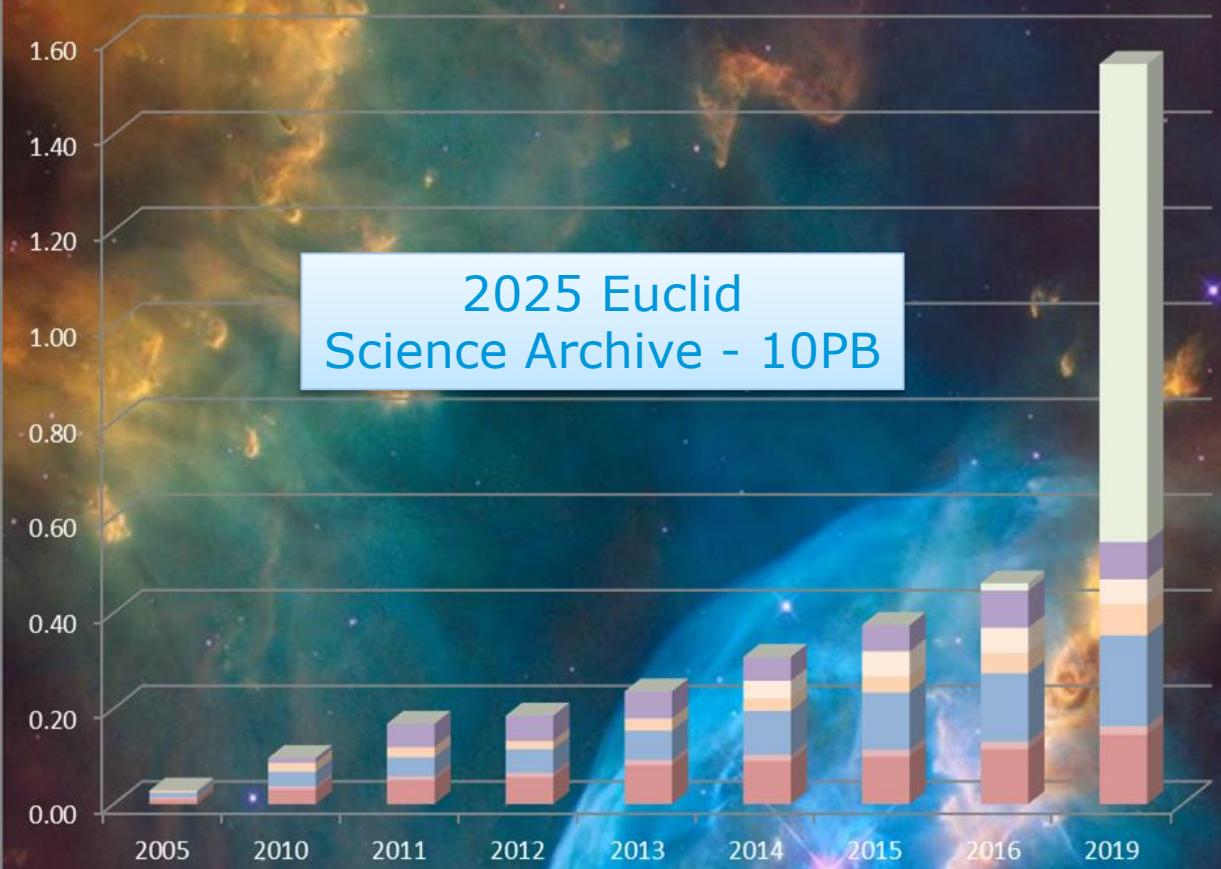
Data-driven Space Science at ESAC Science Data Centre (ESDC)

Beatriz Martinez
ESDC, European Space Astronomy Centre, ESA, Spain

ADASS XXVIII, 14/11/2018

Data-driven science

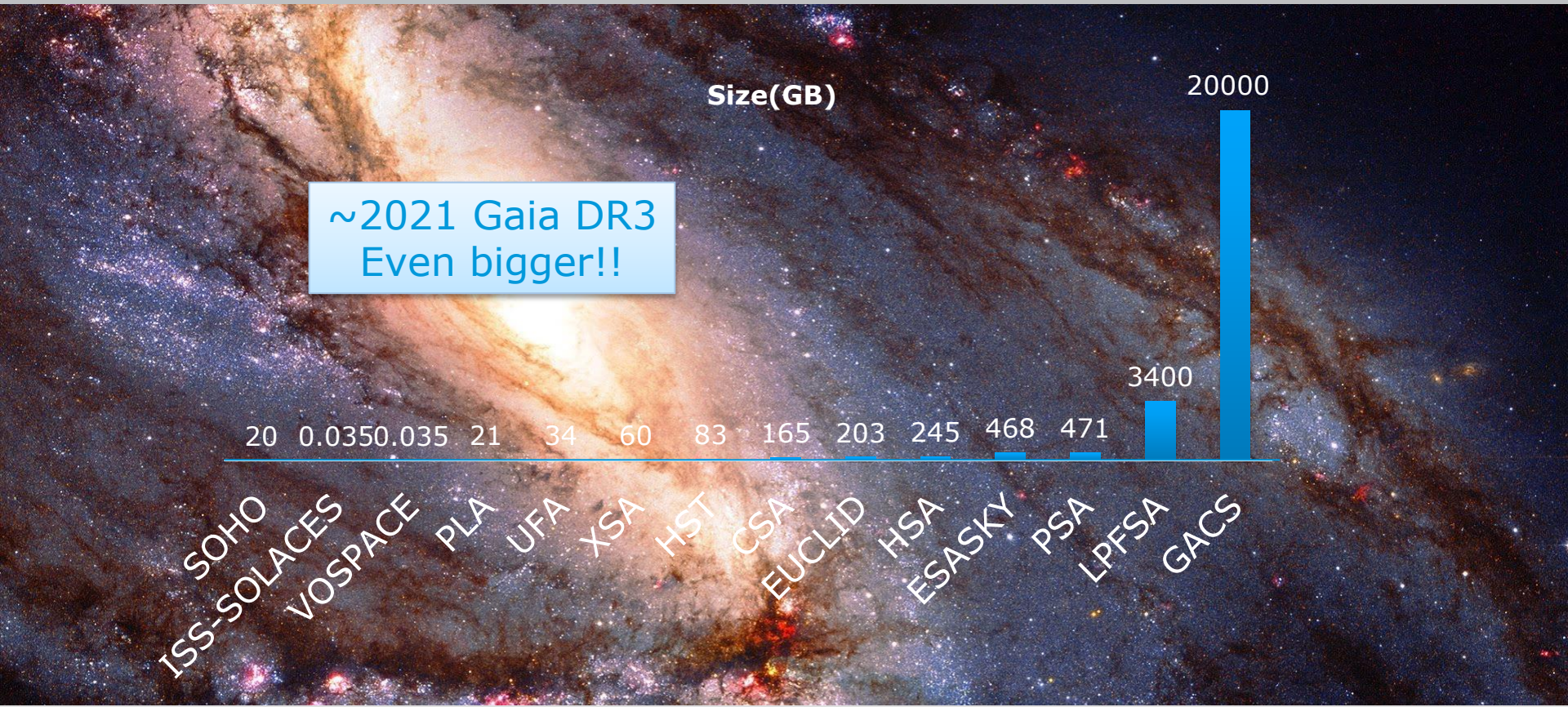
ESA Space Science Archives - Volume (PB)



2025 Euclid
Science Archive - 10PB

- Gaia - 2013
- Herschel - 2009
- Planck - 2009
- Planetary (MEX - 2003, Rosetta - 2004, VEX - 2005, BepiC - 2018)
- Cluster - 2000
- XMM-Newton - 1999
- Soho - 1995
- Hubble - 1990

Databases Size at ESDC (October 2018)



ESDC Challenges

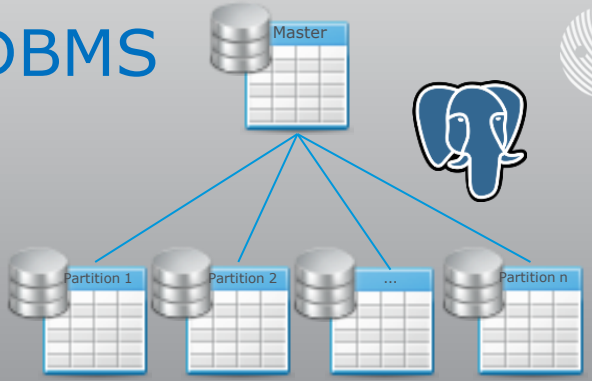
- Manage large volume of data and high heterogeneity
- Enable collaboration between scientists
- Provide tools for exploring and mining the data
- Integrate data (the value of data explodes when it can be linked with other data)
- Manage data in context (track provenance, handle uncertainty and error)

Solutions adopted / implemented

Scientific collaboration and code to data via Interoperability

- **TAP+** parametric search for metadata in catalogues based on ADQL
- **Universal Worker Service** (UWS) to manage sync/async queries
- **SAMP** to interoperate with other analysis applications (Aladin, Topcat, Autoplot,…)
- **EPN-TAP** to query planetary datasets in a standard way, based on TAP
- …

Handling of large datasets in RDBMS



- Table partitioning with Postgresql 10+
- Down-sampling algorithm(s)

As example, the Lisa Path-Finder Science Archive:

- Tables > 10 billion rows, 10 partitions
- ↑ query performance ➔synchronous queries & DB is scalable
- Interactive plots of telemetry parameter values (ex.: > 2 million points)

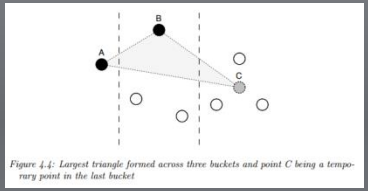
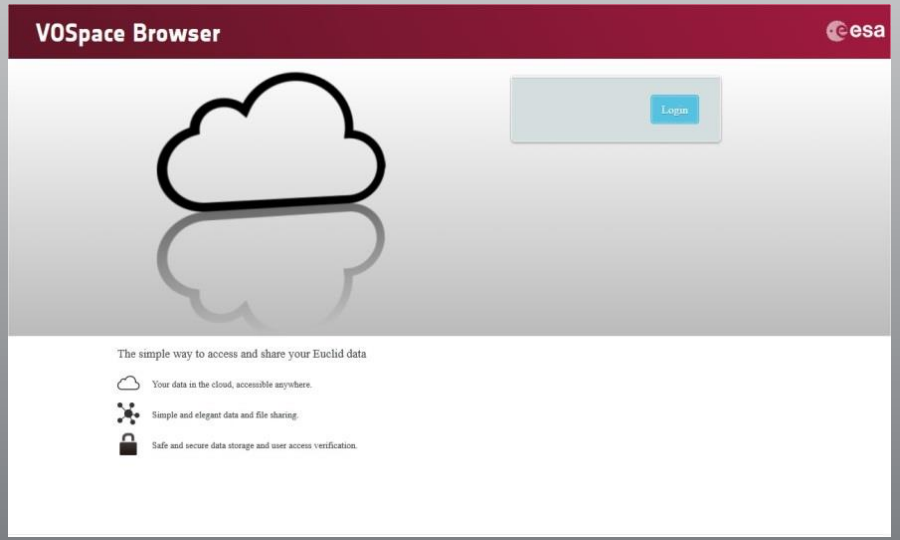


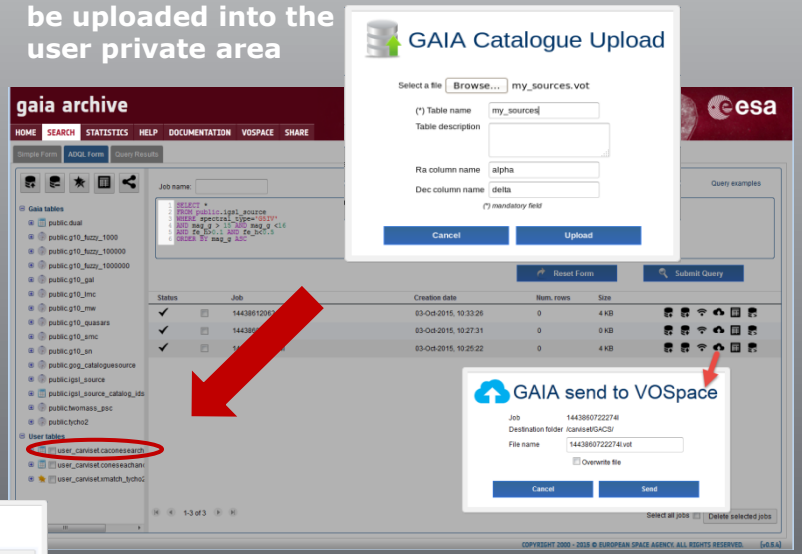
Figure 4.4: Largest triangle formed across three buckets and point C being a temporary point in the last bucket

Largest-Triangle-Three-Buckets algorithm
http://hdl.handle.net/1946/15343

Enabling scientific collaboration



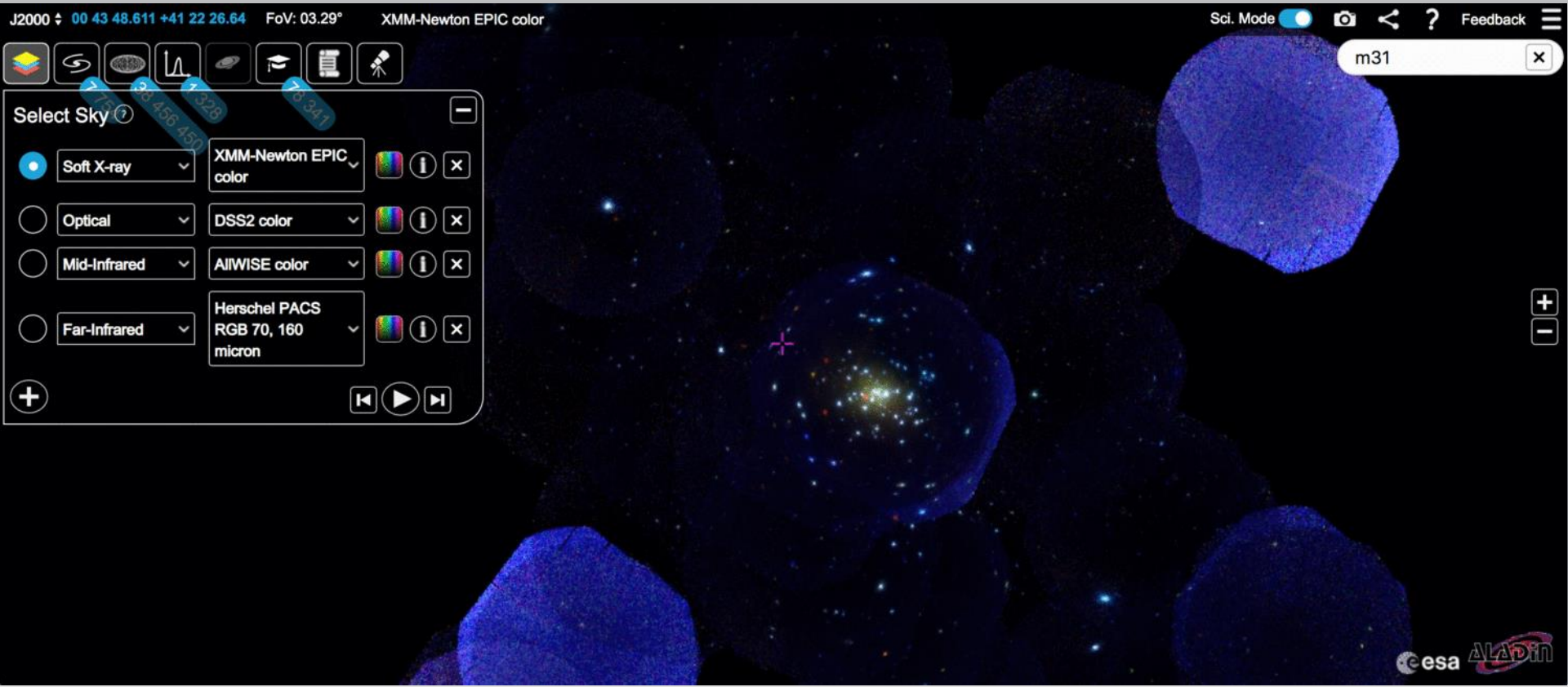
Upload: a table can be uploaded into the user private area



Sharing: any private table can be shared with other users

Crossmatch: an uploaded table can be crossmatched with any other table

Explore heterogeneous data: multi-mission, multi-wavelength



Towards “Archive 2.0” concept

Solutions under evaluation / prototyping

Massive Parallel Processing for big catalogues

As example, the Gaia archive:

- Stores in Postgres-XL time-series, spectra, etc... provided through Datalink service.

Exploring distributed relational DBs that scale-out

PostgreSQL:

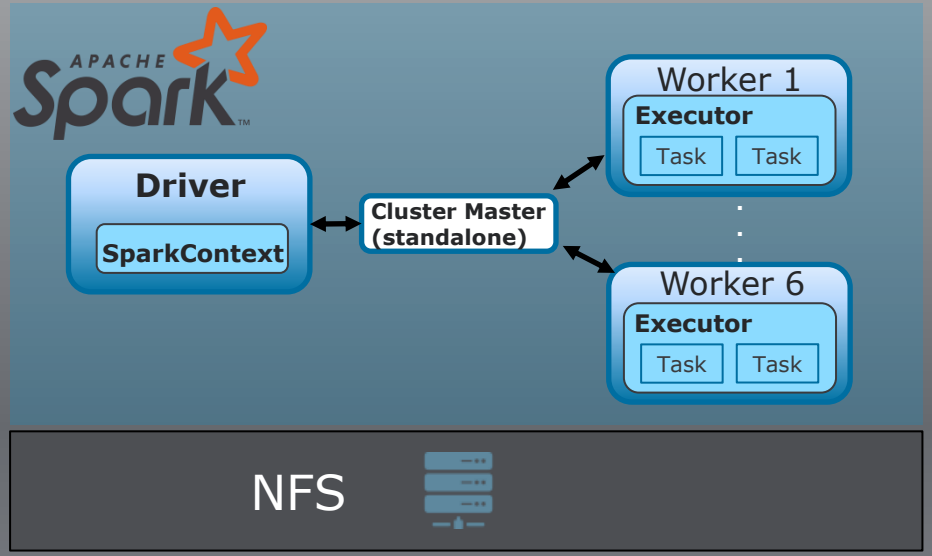
- Open source / Big community
- Specific extensions: Spherical queries (pg_sphere, q3c), pg_healpix, location queries (postgis)

Working on developing specific Query Profile per use case

Interactive/Batch Data Analysis

Prototyping Massive Parallel Processing over large scale datasets for:

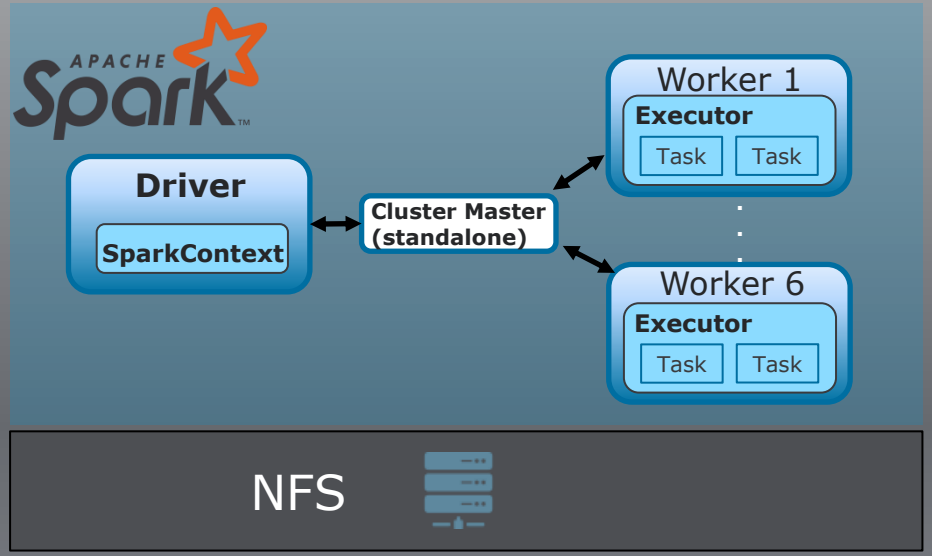
- Morphology analysis / classification of objects with machine learning, in batch mode
- Cutout service or customized source extraction, in interactive mode

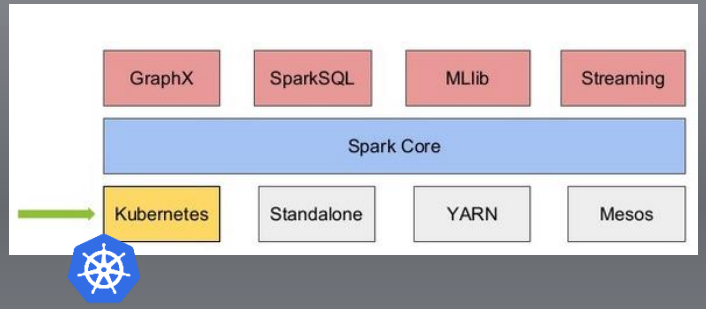


Interactive/Batch Data Analysis

Prototyping Massive Parallel Processing over large scale datasets for:

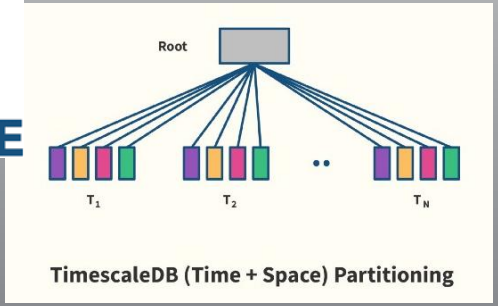
- Morphology analysis / classification of objects with machine learning, in batch mode
- Cutout service or customized source extraction, in interactive mode





Specific searches by data nature

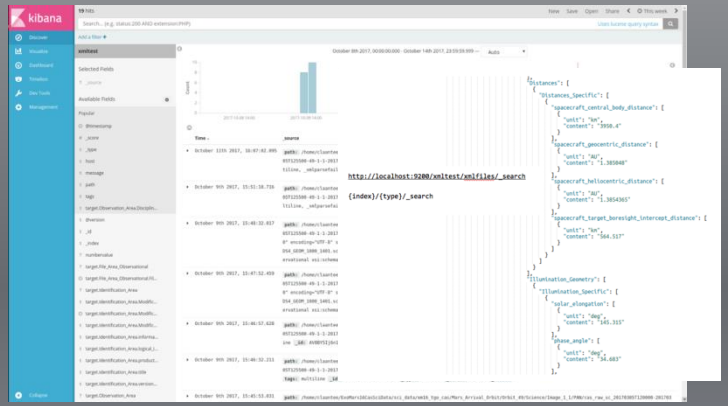
Exploring Time Series oriented databases for large **Time Series data**:



Relational TS DB,
extension of PostgreSQL

Non-SQL solutions *(pending of evaluation)*

Prototyping full text search on planetary data:



Code to the Data and Scientific Collaboration

Astropy: ESDC **open** contributed libraries

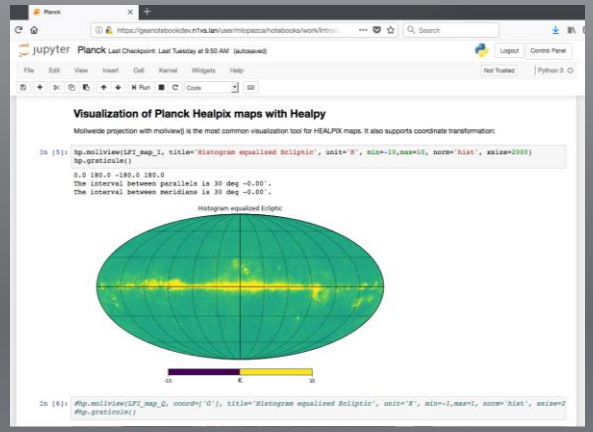
- **Gaia** module: TAP+ access to GACS
  - Reusable to build access to any TAP based archive
- **pyESASky** module
  - Visualization app to visualize data for any Astro archive
- **Hubble** module: TAP+AIO access to HST
  - Reusable to build access to any ABSI/legacy based archive

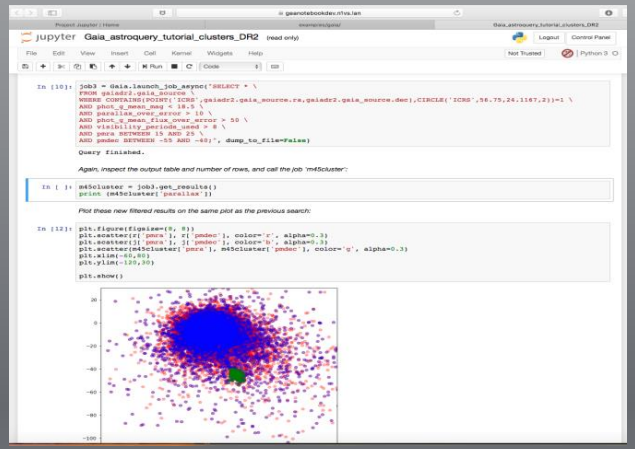
First step to provide Jupyter Notebook "code to the data" services

Prototype of a **JupyterHUB** environment at ESDC

Within a future **S**cience **E**xploitation and **P**reservation **P**latform (SEPP) a collaborative data analysis environment with Jupyter Notebooks will be available.

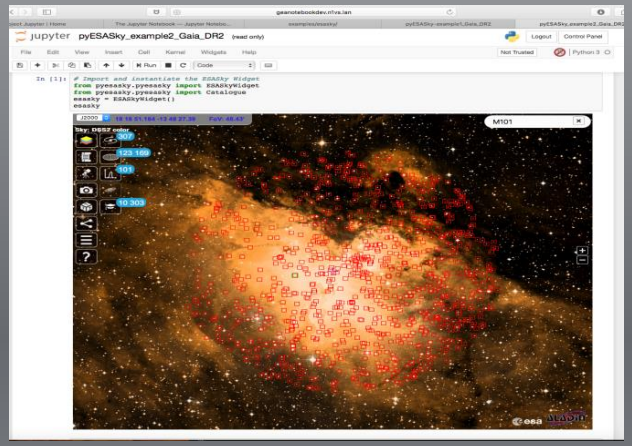
Access to Planck archive





Access to Gaia archive

Access to ESASky

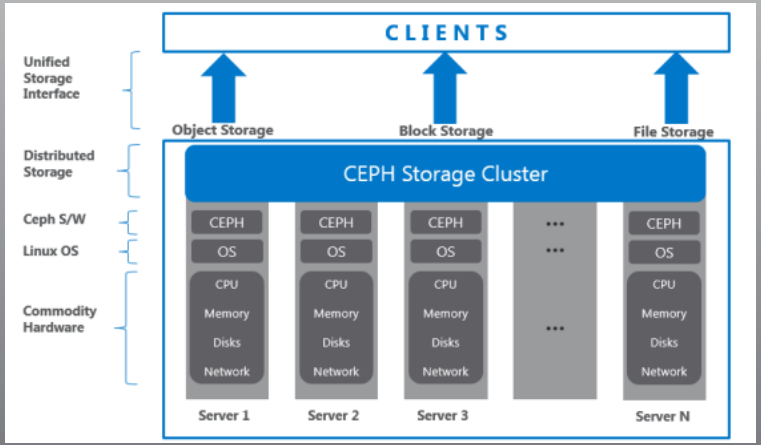


Improving scalability - studies

Scale-out ESDC VOSpace storage using **Ceph**.

Ceph is a software defined storage solution:

- Massive scalable (to Exa-Bytes)
- Highly reliable
- Easy to manage
- Open source



Increase Jupyter Notebooks data analysis using Spark clusters via **PySpark** library

ESDC proposed solutions

- storage of big catalogues through distributed databases,
- storage of long time series in high resolution via time series oriented databases,
- data search and processing via specialized analysis engines,
- and enabling scientific collaboration and closer access to data via JupyterLab, Python client libraries and integration with pipelines using containers.

Thank you

http://archives.esac.esa.int

I. Barbarisi, J. Gonzalez, M. Fernandez, C. Laantee, B. Martinez, B. Merin, H. Perez, S. Nieto, J. Salgado, P. de Teodoro

European Space Astronomy Centre, European Space Agency, Spain

@ESAesdc