SCIENCE EXPLOITATION IN A BIG DATA ARCHIVE: THE EUCLID SCIENTIFIC ARCHIVE SYSTEM

Sara Nieto[1], Pilar de Teodoro[1], Fabrizio Giordano[1], Elena Racero[1], Monica Fernandez[1], Damien Noiret[2], Jesus Salgado[1], Bruno Altieri[1], Bruno Merin[1] and Christophe Arviset[1]

[1]*European Space Astronomy Center, European Space Agency, Spain*

[2]*Paris University, TBD, France*

Abstract.

Euclid is an ESA M2 mission and a milestone in the understanding of the geometry of the Universe. The Euclid Archive System (EAS) is a joint development between ESA and the Euclid Consortium and is led by the Science Data Centres (SDC) of the Netherlands and the ESDC (ESAC Science Data Centre). Big-data technologies, driven by data nature and volume, are transforming the way of doing scientific research towards collaborative platforms whose first goal is to enable access and process large data sets in ways that could not be done downloading the data. Some examples of the main technologies explored as part of the Euclid scientific archive are: JupyterLab, Apache Spark, GreenPlum and PostgresXL.

1. Introduction

Euclid is the ESA M2 mission (Laureijs et al. 2011) which will map the sky in a single optical band and three near-infrared bands (H, J and Y). It will measure photometric and spectroscopic redshift of galaxies to understand the properties and nature of dark matter and dark energy. Euclid will be launched in 2021 and will complete a wide survey (covering 15000 deg^2) and a deep survey (covering 40 deg^2 and 2 magnitudes deeper than the wide survey) during 5 and a half years of observations.

During the nominal operations phase, Euclid will combine space surveys with ground-based surveys to achieve its scientific objectives. This will boost the data volume produced by Euclid SGS up to 26PB per year and a catalogue up to 10 billion objects (Pasian et al. 2014). The data in the Euclid Science Ground Segment (SGS) is a combination of images, spectra and catalogs in FITS files and an extensive metadata description of these data products. To manage such amount of information, the Euclid Archive System (EAS) will handle data and metadata within the SGS according to mission premises.

2. Euclid Archive System

The Euclid Science Ground Segment is a distributed data processing and data storage system, which is responsible for the delivery of the science-ready data to ESA. The SGS is formed by 9 national Science Data Centres (SDCs) and the Euclid Science

Operations Centre (SOC) at the European Space Astronomy Centre (ESAC). The task of the SGS is to process the data from ingested raw frames to science-ready images, spectra and catalogs and deliver them to ESA (Pasian et al. 2014).

According to requirements on the SGS and the EAS, the EAS design was established as a combination of 3 independent subsystems: the Data Processing System (DPS) consists of metadata storage and services which support the data processing inside the SGS; the Science Archive System (SAS) which is a gateway for end-users to Euclid data and supports the scientific use-cases, the release delivery of data to the wide astronomical community, and long-term data preservation; the Distributed Storage System (DSS) consists of data files storage for both the DPS and SAS.

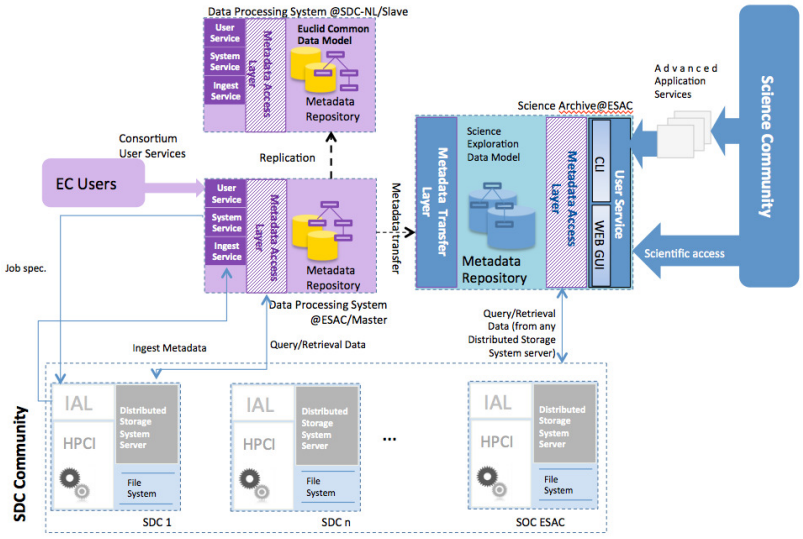

Figure 1. EAS Architecture: DPS, SAS and DSS

The DPS contains metadata describing the data processed as well as operational and orchestration metadata. The main objective of the DPS is to provide data and metadata to other SGS operations and to trace any operation on the data inside the SGS. The DPS keeps the full data lineage for each data product. The DPS will assist in the preparation of Euclid data releases as subset of the data processed by SGS. The DPS services allow the retrieval of metadata in the form of XML files, the ingestion of data products, and the browsing of available data products through web applications.

The DSS is a distributed file storage for both the DPS and SAS and implements the distribution of data files according to the Euclid processing plan. The DSS consists of a grid of DSS servers which utilize a simple interface to the different storage solutions implemented by each SDC. Each DSS server communicates with all other DSS servers and allows a set of operations on files: ingestion, retrieval, copy to designated SDC and registration of the file in DPS metadata database.

3. The Scientific Archive System

As part of the EAS, SAS aims to support the scientific exploitation of the best Euclid data for the Euclid Consortium and the wider astronomical community. The SAS

is currently under development at the ESAC Science Data Centre (ESDC), which is responsible for the development and maintenance of the scientific archives for the Astronomy, Planetary and Heliophysics missions of ESA. The design and architecture of the SAS follows the latest technology generation of archives developed by the ESDC, taking full advantage of the existing knowledge, expertise and software libraries.

The Euclid mission will generate a huge and increasing amount of scientific metadata and catalogues up to 10 billion of galaxies, providing the worldwide astronomical community with an extremely large source of targets for future missions. For these reasons, the SAS will face the challenge of providing the public community with access to this data through scientific interfaces and tools, while guaranteeing the long-term preservation of Euclid data.

The scientific requirements on the SAS cover three main exploitation areas: parametric search for metadata and catalogues, exploitation of spectra, visualization of Level 2 (L2) images and data retrieval. Therefore, the architecture of the SAS is built as a three layered architecture, where the core of the system is based on these modules: metadata, spectra and L2 images.

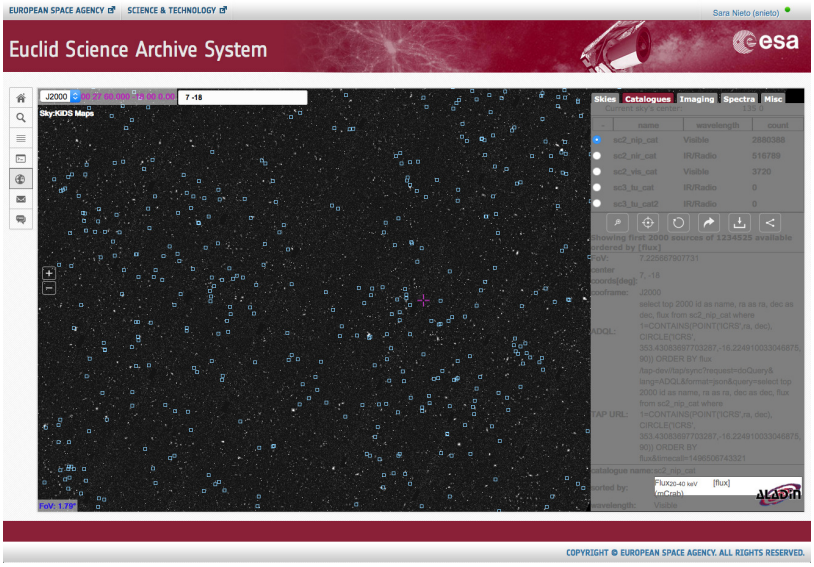

Figure 2. SAS Visualization Interface

4. Interactive Computing

Euclid will produce observations up to 26PBs per year. In this order of magnitude, to keep and process data locally is not an affordable option for most of the Euclid users, as it was not for Gaia users. Therefore Euclid will enable the Software-to-Data Paradigm and "bring the software to the data", providing a set of libraries to allow the users execute queries, retrieve results and share the output with the scientific community.

Interactive computing is a key element of the strategy to enable coding close to the data and avoiding local analysis for big data volumes. In this context, state-of-the-art interactive tools like Jupyter Notebooks are a way to achieve this goal, together with

the tools and VO interfaces needed to facilitate the access to the Euclid science stored in SAS.

Sharing scientific results with collaborators is an important point of interactive computing. VOSpace (**?**) is the VO protocol for distributed data storage that enables data sharing within the scientific community. Users could use their own VOSpace account to store and share query results with the rest of the community. Python libraries, currently under development, will enable to access VOSpace Rest API from interactive computing environments like Jupyter Notebooks.

5. References

This template has no bibtex file. Look for the larger template and Makefile how to do this. By default the Makefile will create an empty P10-10.bib. When you add references to this, uncomment the line `\bibliography` below, make use "make" to create your beautifully looking PDF.

References

Laureijs, R., et al. 2011, arXiv. `arxiv1110.3193`

Pasian, F., Hoar, J., Buenadicha, G., Dabin, C., Sauvage, M., Poncet, M., Noddle, K., Delouis, J., & Mansutti, O. 2014, in Astronomical Data Analysis Software and Systems XXIII, edited by N. Manset, & P. Forshay, vol. 485 of Astronomical Society of the Pacific Conference Series, 505