Towards Analytics-Hub: a data & computing environment for multi-model data analysis in the Earth System Grid Federation

D. Elia[1,3], C. Palazzo[1], P. Nassisi[1], A. D'Anca[1], E. Scoccimarro[1], T. Weigel[2], S. Bendoukha[2], S. Gualdi[1], S. Fiore[1], G. Aloisio[1,3]

[1] Fondazione Centro Euro-Mediterraneo sui Cambiamenti Climatici (CMCC), Lecce, Italy - [2] Deutsches Klimarechenzentrum (DKRZ), Hamburg, Germany - [3] Università del Salento, Lecce, Italy

ANALYTICS-HUB

The **Analytics-Hub** is a new component in the **Earth System Grid Federation** landscape joining data and computing aspects to provide a multi-model environment for CMIP-based analytics experiments. It represents a one-stop-shop for scientists focusing on 'variables' rather than 'models' as for the ESGF data nodes.

Such peculiarity allows scientists to find ready to use datasets from the **CMIP5** and soon **CMIP6** experiment for a specific variable without the need to download all the relevant datasets on the end-user side.

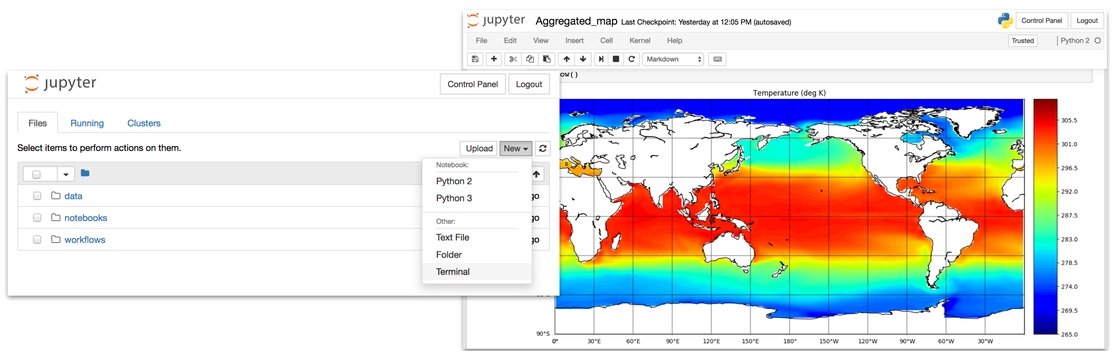
Specific components at the data management layer provide the data-backbone to synchronize the local repository with the larger **ESGF Data Archive**.

ECASLAB

ECASLab is an Analytics-Hub implementation. It provides a user-friendly scientific data analysis environment deployed at CMCC and DKRZ based on ENES Climate Analytics Service (ECAS). It integrates data and analysis tools to support scientists in their daily research activities. The environment joins the features of the *Ophidia data analytics framework* with a large set of Python libraries for running data manipulation, analysis, and visualization. ECASLab integrates the following services and features:

- an ECAS cluster hosting an instance of **Ophidia framework**, with WPS-enabled interface accessible through the Ophidia Terminal and any WPS-compliant client;
- a **JupyterHub** web-based instance enabling the user to create, execute and share **Jupyter notebooks** (Python-based) supporting live-coding and visualization;
- an instance of **Synda**, i.e. an efficient download manager well-suited for search&discovery, data transfer and synchronization from the ESGF Data Archive;
- a monitoring system based on **Grafana**.

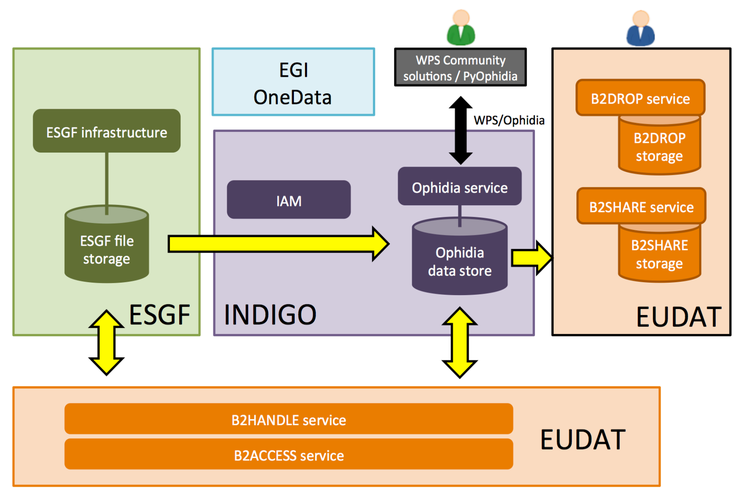
The environment also provides access to some datasets (by using **THREDDS Data Server**, TDS), a number of example Jupyter notebooks and real-world workflows describing indicators from several use cases. The experiment output can be either exported in the user space or published by means of TDS, whereas JupyterHub provides the features to update files and navigate the file system.



The features of the Ophidia framework can be directly exploited in the notebooks to run data analytics tasks on big datasets and plot the results on charts and maps using well-known Python libraries. **PyOphidia** - the Ophidia Python bindings - allows an easy interaction with Ophidia and other Python-based modules (e.g. Matplotlib, NumPy).

ENES CLIMATE ANALYTICS SERVICE (ECAS)

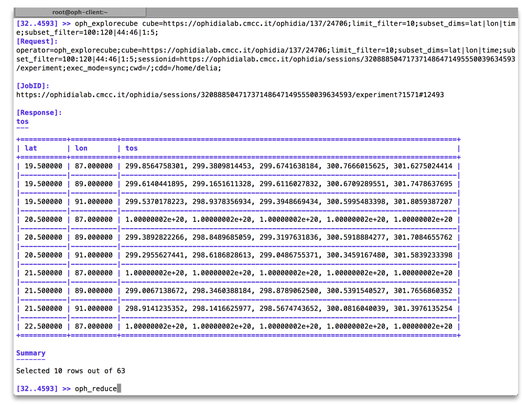
The **ENES Climate Analytics Service** is one of the EOSC-Hub Thematic Services. It builds on top of the Ophidia big data analytics framework with additional components and services from the **INDIGO-DataCloud** software stack, **EUDAT** and **EGI** e-infrastructures. ECAS has been ranked as the *1st out of 64 Thematic Service proposals.*



ECAS & THE OPHIDIA BIG DATA ANALYTICS FRAMEWORK

Ophidia is the core component/engine of ECAS. It represents a complete software stack developed by CMCC for data analytics in multiple eScience domains, such as climate change, astrophysics, etc. In terms of end-user features, the framework provides, among others:

- data **reduction** and **subsetting**;
- data intercomparison;
- metadata and **provenance** management;
- time series analysis with a wide set of array-based primitives (around 100);
- **interactive data analysis**;
- **workflows** of tasks;
- agile setup of operational chains.

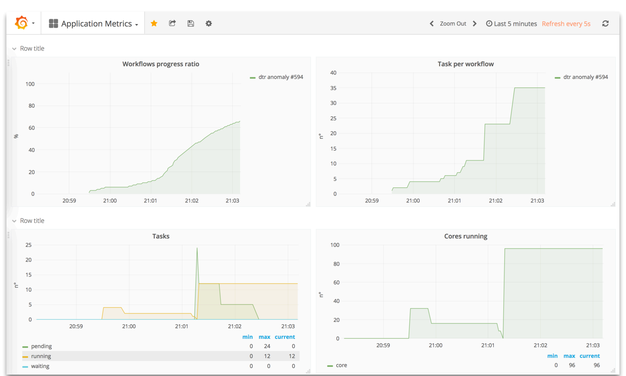


ECAS CLUSTER MONITORING AND ACCOUNTING

ECASLab relies on various types of nodes: **computing nodes**, a **server front-end** and **client/services machines**.

In this environment, ECAS allows the execution of **single operators**, **massive tasks** and **workflows of tasks**.

A **Grafana**-based monitoring and accounting system keeps track of resource usage and activity from an **infrastructural**, **application** and **user** point of view.

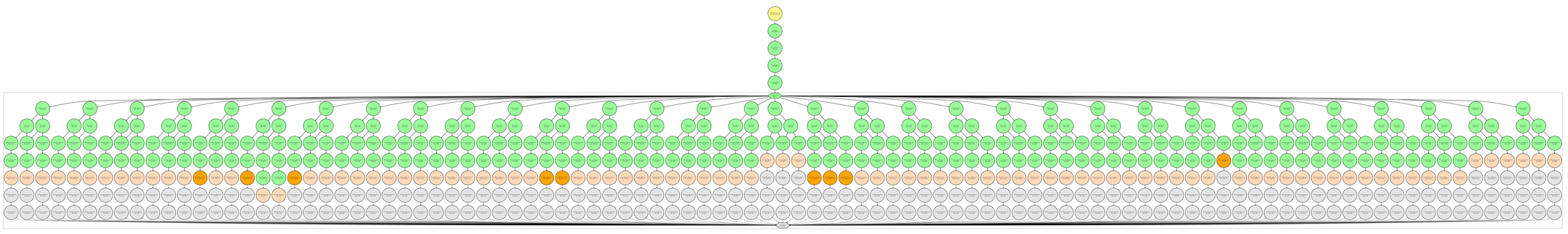


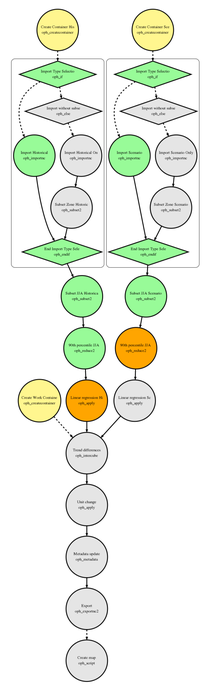
ECAS ANALYTICS WORKFLOWS CAPABILITIES FOR SCIENTIFIC INDICATORS

ECAS embeds an analytics **workflow manager** designed to make the platform more flexible, to help reduce the complexity of scientific experiments, to increase the re-usability, and to fully exploit the available computational resources.

In the climate change context, several workflows for real-world use cases have been defined. By writing down a **simple** task graph including the basic operations to be executed, the user is able to quickly process large input datasets and evaluate one or more **indicators** like *sea surface temperature anomaly*, *precipitation trend* (workflow on the right), *snow season statistics* (workflow at the bottom), *climatological averages, unusual warm events*. Furthermore, the parallel workflow interface allows an easy replication of the same set of operations over different datasets to compute complex analyses with no effort (see the picture below).

By using ECAS, a number of workflows have already been defined to perform experiments also in other scientific domains (e.g. astronomy, seismology, biology).





EOSC-hub receives funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 777536.