CEOS Analytics Lab

CEOS Systems Engineering Office and CSIRO Chile

April 2024
David Borges, CEOS SEO
Jonathan Hodge, CSIRO Chile

The Challenge

The increasing cadence and volume of new Earth observation data produced by evermore missions has challenged the traditional ways researchers and institutions ingest, archive, and process these data, requiring new solutions. The greater volume, velocity, and variety of data has changed foundational aspects of civil space-based Earth observation programmes on a global basis by putting an emphasis on scalability and data centric architectures. A combination of cloud-related technologies and capabilities represents a new paradigm for working with the vast collections of data now available on the public cloud and is referred to as cloud native geospatial. These technologies move users closer to these large archives of data enabling them to directly process these data in a scalable, parallel, and continuous way.

Many Committee on Earth Observation Satellites (CEOS) Agency members are already exploring new cloud-based data and analytics solutions. However, broad analytics access has historically been limited externally by various necessities including security and cost. To ensure international coordination and promote efficient exchange of data to optimize societal benefit across this new landscape, some CEOS activities may benefit from an open and collaborative analytics environment.

A Shared Platform

The CEOS Analytics Lab is a platform designed to meet the challenges of a cloud based, data centric approach to Earth observations. It is an evolution and response to the CEOS Earth Analytics Interoperability Lab proposal, developed by CEOS WGISS and SEO in 2020. From expanding CEOS-ARD activities, increasing WGISS Interoperability Framework maturity, to supporting commercial sector engagement, the CEOS Analytics Lab has been established to support the CEOS community. Current utilization includes the Ecosystem Extent Task Team, COAST Virtual Constellation, and WGDisasters Flood Pilot. The CEOS Analytics Lab is freely and openly available for use by all CEOS Members and Associate Members. This access is possible due to contributions from CSIRO, CSIRO Chile, Chilean Data Observatory, and the CEOS SEO.

CEOS Analytics Lab Capabilities

The CEOS Analytics Lab (CAL) is a cloud native multiuser gateway for spatial data science deployed on Amazon Web Service (AWS) infrastructure to enable a high degree of flexibility and scalability. CAL is similar to other Earth observation analytics platforms such as Digital Earth Australia and Digital Earth Africa and shares some similarities with the Microsoft Planetary Computer. CAL is specifically based on CSIRO's Earth Analytics Science and Innovation (EASI) platform. It knits together multiple open source projects to provide a range of capabilities and make the power of the cloud more accessible. A chief advantage of building on open architectures is a shared python ecosystem, where examples and analyses can be adapted to run across platforms and shared between systems. CAL serves as a testbed to further advance knowledge sharing for cloud native platforms.

Using JupyterHub (https://jupyter.org/hub), CAL is able to serve a large number of concurrent users economically, only using additional resources when a user is logged in. Users are then automatically provisioned with a customized JupyterLab user environment, preconfigured with a range of data science, machine learning, and geospatial python libraries. This ensures users have access to a reproducible environment with the same capabilities and no manual setup. Users may also customize their own environments if they have custom datasets or have requirements which are not included in the default environment. The user environment is also connected to an Open Data Cube (ODC) to simplify loading data from a wide variety of sources with a single call. Additionally, the environment is connected to a Dask GateWay allowing users to scale their analysis across multiple servers on demand. This allows users to process areas larger than a single server can handle, faster and more economically. By using Dask, an analysis task may be transparently broken up into multiple steps and spread across multiple powerful computational nodes called "workers" . These workers are spun up only at execution time and work in parallel to execute the analysis. CAL also includes a highly scalable workflow solution, the Argo Workflows platform (https://argoproj.github.io/workflows), for advanced data processing and ingestion. This platform is used to simplify and automate the process of incorporating new data into CAL and enables very large-scale parallel processing of data.
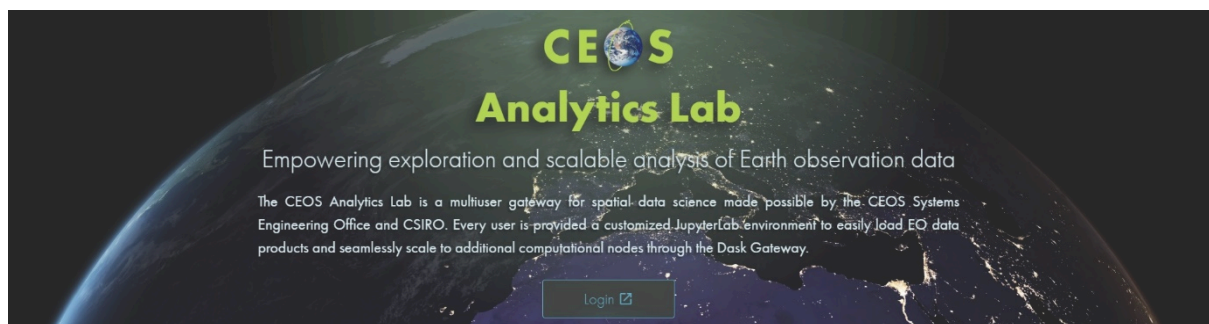
Utilization

Accounts are available to all Members or Associate members participating in CEOS activities. Simply request an account using the radio button on the homepage and fill out the form. We will manually approve accounts and you will receive a notification to the provided email when approved. Other requests may also be submitted using the Support Request and Data Request forms available under the Services dropdown. We are happy to support you and answer any questions you may have about CAL. Your feedback, questions, and requests are welcome.

CEOS Analytics Lab: https://ceos.org/cal/

David Borges: david.borges@nasa.gov

Jonathan Hodge: jonathan.hodge@csiro.au

SEO Team: brian.b.terry@ama-inc.com , james.d.hasselman@ama-inc.com, andrew.w.cherry@ama-inc.com