CSIRO EASI Hub data-pipelines – presentation notes

CEOS WGISS-50, 22-24 September 2020

Matt Paget, Jonathan Hodge, Peter Wang, Robert Woodcock

Slide 1 – Introduction

Hello.

This presentation introduces CSIRO's EASI Hub "data pipelines" software.

EASI Hub is CSIRO's cloud-based Earth Analytics and Science Innovation platform. The platform infrastructure is controlled through Terraform and Kubernetes and managed by CSIRO. Platform features include Open Data Cube, Jupyter Hub, and custom apps and services.

EASI Hub powers the CEOS Earth Analytics Interoperability Lab with SEO – see Rob Woodcock's talk.

Slide 2 – Overview and Intent

The intent of the Data-Pipelines software is to simplify and automate searching, ordering, downloading and preparing CEOS data for use in EASI Hub.

The software has recently been updated to include Argo workflows, which has proved very successful for us in running scalable and resilient workflows in the cloud.

The data storage model for EASI Hub is to hold only data that is required for current applications. This is intended to reduce costs, avoid data archive management overheads, and increase flexibility for deployments and data products (including responding to upstream reprocessing).

Slide 3 – Design

EASI Hub data-pipelines relies on machine-readable interfaces to CEOS agencies and data provider archives. Ideally "agency standard" ARD processing services or products, where available, also provide for consistent global data use and interpretation.

The data-pipelines software provides a common interface for requesting data:

- Platform
- Product
- Spatial and Temporal
- Additional API parameters such as filtering by tile code

After download, a simple per-file or object task workflow allows for unpacking, pre-processing (custom ARD), reformatting (COG, Zarr) and data-cube preparation. Data cube indexing is a separate step routinely run over a set of prepared data.

Request or order information and state are held in python dict structures. While simple, this forces the workflow components to be idempotent as there is no dependency on other system components.

Slide 4 – Ecosystem

The data pipelines software builds on existing APIs and available client libraries. There are quite a number of client libraries available for agency APIs. In some cases we use them directly, in other cases we use them as examples for interacting with the API.

We have current data-pipelines interfaces for ESPA, AppEEARS and Copernicus Hub, as well as direct datacube indexing of Geoscience Australia's public AWS buckets for their Australian products.

ESPA has been our primary test case as it requires a delay between order and download, with a variety of possible error states (e.g., scene unavailable). Our workflows also take into account any concurrent request or download limits set by each API.

We have more interfaces in development including NovaSAR, Himawari and CEOS OpenSearch.

Slide 5 – Trajectory

EASI Hub data-pipelines has been designed with the emerging trajectory of CEOS agencies towards cloud-based archives, granule or area-selection APIs, and "agency standard" ARD production.

With agency archives and ARD products available from cloud stores and services, consumers can efficiently access and use just the data they need. This reduces the need for local copies of agency archives, provided access to a regional archive hub is available and "fast enough".

Slide 6 – Towards more interoperability

Some observations to share and discuss:

ARD is great, and "agency standard" ARD gives great confidence. However, we see third-party distributors have a mixed approach to available and clear documentation on their ARD products. This becomes a challenge for consumers to compare, select and interpret data particularly for science applications. In our case poorly described ARD products are of limited use.

CEOS OpenSearch is great for discovering the wealth of products across agencies. We find the "Short name" field is simplest for searching for granules. Where granules are not searchable we look for another service hosted by the source agency.

Multiple sources for the same data products provide options for consumers, e.g. multiple cloud providers or mirrored copies across regions. Key information we suggest would help users make an informed choice are the update frequency, completeness and provenance (source and any structure or format changes).

Data stored in the cloud is not only accessible (outside agency firewalls) but can also be compute-ready if stored in selected formats. Compute-ready data allows applications to take advantage of the cloud provider's parallel or distributed network I/O. Mostly it means that data unpacking and cloud-formatting is done once by the agency rather than by each user.

Slide 7 – Chile coverage

We used data-pipelines to build a Landsat data cube for the Chilean Data Observatory initiative.

The first version occurred during and contributed to our test and development of the data-pipelines software. After finding some reformatting and prepare bugs, we threw away the data and re-ordered and built the full coverage in one week, with little human intervention.

This shows that its quite possible to request and prepare areas of interest on demand, given ARD APIs and scalable compute technology.

The graphic below the text shows an Argo workflow (left to right) waiting for sets of scenes within an ESPA order to become available (the horizontal trunk) and then scaling-out worker nodes to download each scene and run the post-download tasks (the vertical branches).

Slide 8 – Thank you

Thank you. Please ask or contact us if you have any questions.