Science Demonstrator 4: EuroArgo Data Subscription Service (Use Case TC_2)

Please provide your feedback on this Science Demonstrator using the questionnaire at https://survey2.icos-cp.eu/ENVRIplus-evaluator!

Overview

The EuroArgo Data Subscription Service (DSS) allows researchers to subscribe to customized views on Argo data, selecting specific regions and time-spans, and choosing the frequency of updates. Tailored updates are then provided on schedule to researchers' private storage.

Scientific Objectives

An extensive number of Research Infrastructures need to publish and give access to datasets that may accumulate over time and need to remain available for download for researchers. The accumulated datasets are queried and analyzed, leading to new data results. Keeping an eye on the accumulated and result datasets at the Argo data center is time consuming, thus a subscription model was adopted to facilitate researchers' needs. The subscription model does not require direct interactions between the researcher and possibly time-consuming analysis actions, thus allowing more flexible design and integration of the system components. In practice this means that even when time consuming actions can be optimized to operate within time-constrained requirements, the adopted model alleviates effects of long-running actions on the data, especially when the size of accumulated datasets increases.

The objective of this use case is to develop and integrate a system to access, download, and subscribe to EuroArgo DataSets. The EuroArgo community aggregates the marine domain datasets into a community repository from which the data is pushed to the EUDAT B2SAFE service. The new developed service allows data registration through data identification. Optionally the community registers the subscription actions and their parameters in B2SAFE. Data could be requested through interfaces provided by EUDAT and IFREMER web sites. Subscription is managed by EUDAT and actions are processed using EGI FedCloud. Users can select different attributes for their subscription like localization, stations, parameter, update frequency and more.

Description

Architecture

As shown in (Figure 1), the Data Subscription Service (DSS) involves the following basic components: 1) a data selection portal as frontend, 2) the Global Data Assembly Center (GDAC) of EuroArgo, 3) EUDAT B2SAFE storage, 4) DRIP, 5) EGI FedCloud resources, and 6) a subscription service component for managing the subscriptions registered via the data selection portal.

IFREMER provided accumulated monthly marine domain datasets to be pushed to EUDAT e-Infrastructure (B2SAFE). The datasets were then synchronized between EUDAT and EGI Cloud resources on demand. EUDAT provides services for storage and data transfer, while EGI FedCloud provides the services for computing data products for each subscription.

A Web Portal has been developed for users to select and subscribe to the interested data. Data selection is made through different criteria (platform type, measure type, parameter, platform Id, time period and more). A python script has been provided by IFREMER to extract data with user's criteria.

DRIP (the Dynamic Real-time Infrastructure Planner, developed by WP7, Task T7.2) is integrated to execute the data selection process using parallel computation. DRIP can dynamically deploy and manage as many Virtual Machines as required to cope with the load in order to be able to process the subscriptions in a timely manner. Once results were available, they were pushed to B2SAFE and the user was notified by email.

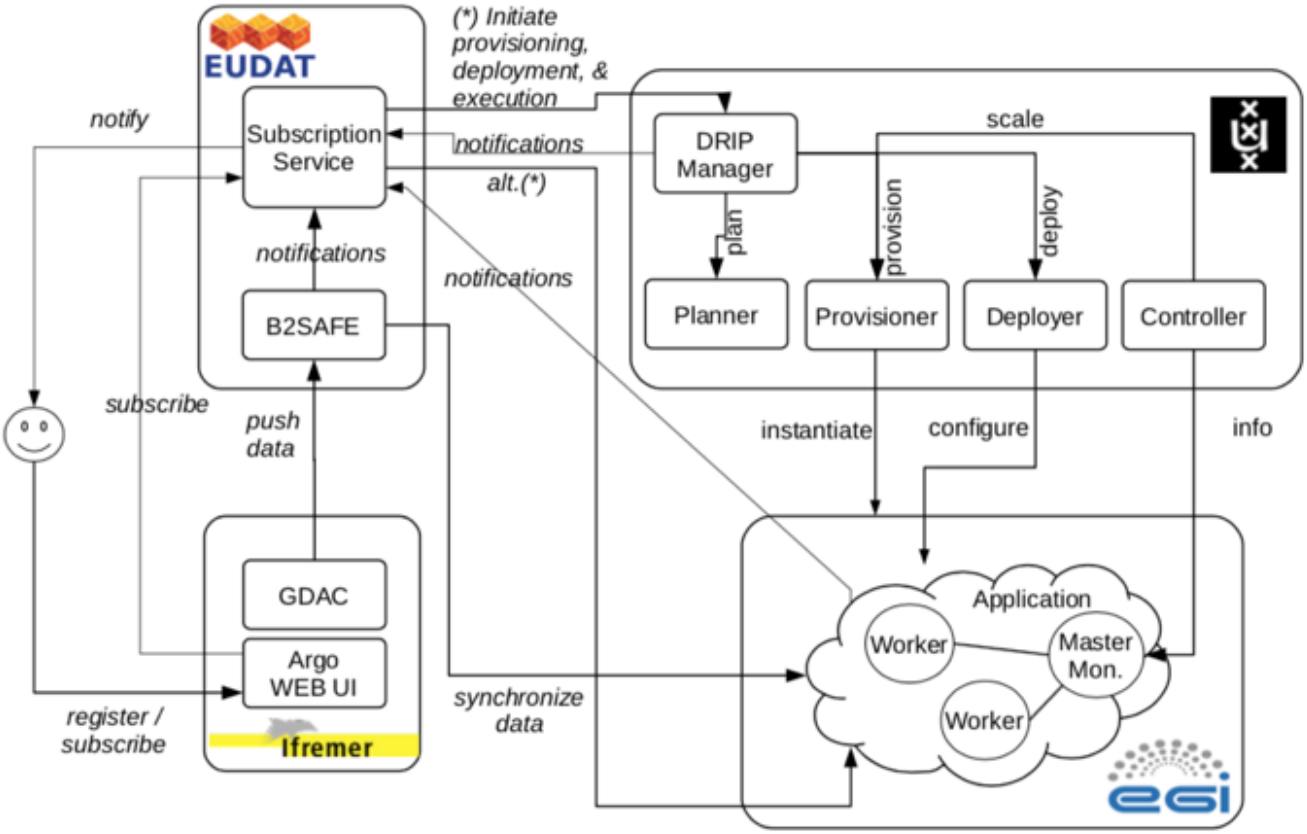


Figure 1. System architecture for EuroArgo Data Subscription Service.

Workflow

The typical workflow is as follows: users interact with the DSS via the portal, registering to receive updates for specific areas and time ranges for selected parameters such as temperature, salinity, and oxygen levels. The GDAC receives new datasets from regional centres and pushes them to the B2SAFE data service. The DSS maintains records of subscriptions including selected parameters and associated actions. DRIP plans, provisions, deploys, scales and controls the data filtering application. EGI FedCloud provides cloud resources to host the application. The application itself is composed of a master node and a set of worker nodes.

When new data is available to the GDAC, it pushes them to the B2SAFE service, triggering a notification to the DSS, which consequently initiates actions on the new data. If the application is not deployed to FedCloud then DRIP provisions the necessary VMs and network so that the application may be deployed. Next, the deployment agent installs all the necessary dependencies along with the application including configurations to access the Argo data. The DSS signals to the application master node the availability of the input parameters to be processed, whereupon it partitions the input tasks into sub-tasks and distributes them to the workers. If the input parameters include deadlines then the master will prioritise them accordingly. The monitoring process keeps track of each running task and passes that information to the DRIP controller. If the programmed threshold is passed, then the controller will request more resources from the provisioner. Finally, the results of each task are pushed back to the B2SAFE service triggering a notification to the subscription service, after which it notifies the user.

User Interface

Shown in Figure 2, users can use DSS web portal to subscribe to interested datasets. A typical subscription task is made up of a set of inputs: a) an area expressed as a bounding box; b) a time range; c) a list of parameters required in data products (e.g. temperature); and d) optionally, a deadline.

Figure 2. DSS Web portal

Advantages

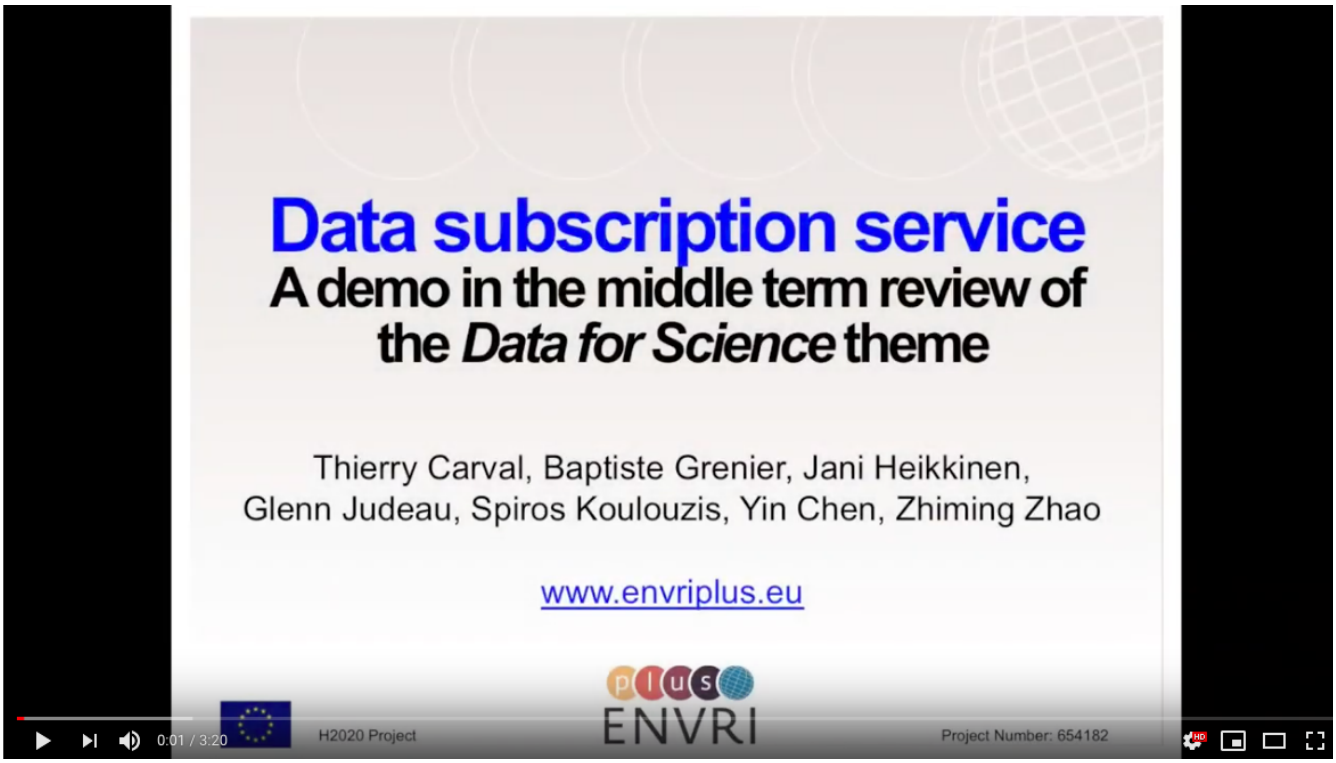
The pilot activity was initiated by the marine research community, however, the possibility to receive regular transmissions of data, especially in near-real time, directly from the organisation responsible for the data collection and (pre-)processing, is very important to many large initiatives. Generic initiatives will themselves be interested to operate subscription services for their outputs, based around a trusted repository hosting synchronised versions of their data collections. Such a service may also allow the provision of new features to end users, generating more visibility.

Research Infrastructures can benefit from subscription services in several ways. They may set up one to serve their own dedicated end user communities, either by pushing new or updated data sets to their "customers", or by configuring a system that automatically advertises the existence of new data e.g. to those who downloaded previous versions. RIs that require data from other sources, for example, to create more elaborated data products, can optimise their internal work flows by signing up to receive automatic updates.

The end users can certainly benefit from signing up to services that automatically advertises the existence of new versions or updates to data that they have downloaded previously – for example annually receiving the observational data from a station for which they have a long-standing scientific interest. Typically, the greatest interest here would be for accessing finalised and aggregated data products created by an RI at its data centre. However, subscription services can also be configured to allow customised processing, bringing an opportunity for research groups to benefit from large-scale cluster computations at the data centre side.

Data subscription services are expected to play an increasing role in the future, as the number of data producers and their respective output continues to increase rapidly. The mechanism tested and implemented by this demonstrator could contribute to the development of both a common standard for input streams to enhancements of digital collaborative spaces for researchers and data providers.

Link to the Demonstrator



Youtube video is at: https://youtu.be/PKU_JcmSskw

Contributors

- Dr Thierry Carval, IFREMER, Thierry.Carval@ifremer.fr
- Dr Glenn Judeau, IFREMER, Glenn.Judeau@ifremer.fr
- Dr Jani Heikkinen, CSC, jani.heikkinen@csc.fi
- Baptiste Grenier, EGI, baptiste.grenier@egi.eu
- Dr Zhiming Zhao, UvA, z.zhao@uva.nl
- Dr Paul Martin, UvA, p.w.martin@uva.nl
- Dr Spiros Koulouzis, UvA, S.Koulouzis@uva.nl