SeaDataNet OneData pilot 2nd phase

About the pilot

Description of work

In the SeaDataNet practice, MARIS is confronted many times with situations that data sets are stored at different locations while we want to undertake central processing. For instance, there is great interest in so-called BioGeoChemical (BGC) data sets as these provide input for determining indicators about the quality of marine waters and as such are very relevant for the Marine Strategy Framework Directive of the EU which aims at establishing Good Environmental Status (GES). Through its engagement with EMODnet Chemistry, SeaDataNet is actively supporting Regional Sea Conventions, EU DG Environment, and European Environment Agency (EEA) in compiling and providing harmonised and validated data collections for eutrophication and contaminants which are derived from the BGC data as gathered by the SeaDataNet data centres. Moreover, SeaDataNet has established cooperations with Copernicus CMEMS as well as with EuroArgo to work together on mutual data exchanges and on improving and innovating quality control and processing of large BGC data collections for various purposes, including MSFD. Access to the data, as well as controlling quality and processing the distributed datasets, currently have performance issues.

For this purpose, it is of great added value to set up a test configuration using OneData in combination with Cassandra and Elastic Search. OneData will be configured to give access to a number of data hubs on the cloud, each provided with BGC data collections in the SeaDataNet ODV format. Cassandra is an open source NoSQL database with wide column

store, which allows high searching performance on large data sets with many numbers. Elastic Search can be configured on top to optimize free text search on the metadata of the data sets to facilitate fast and precise subsetting of data collections from the master collection.

Cassandra is also being analysed by IFREMER for improving access to its large collection of NetCDF files as acquired through the EuroArgo monitoring programme. This analysis is partly done as part of EOSC-HUB in the Marine Competence Centre (MCC) activities. Therefore, MARIS will learn from the insights and best practices as gained by IFREMER in MCC. Moreover, MARIS will add the connection of Cassandra to OneData cloud and build up further experience and complement the EOSC-HUB knowledge base of OneData and Cassandra as both are very interesting tools for handling and federating big data coming from multiple locations.

Cassandra might be installed and configured at a local MARIS server; however, it will also be considered to install and configure it on the cloud. Moreover, the test configuration in a later stage might be expanded to include also OneData connection to the Cassandra instance of IFREMER for exchanging subsets of BGC data between the two installations as part of joint activities for innovating generation of high-quality data collections.

Participants

Participant	Name and Surname	Organization
peter@maris.nl	Peter Thijsse	MARIS
bert@maris.nl	Bert Broeren	MARIS
arko@maris.nl	Arko Rietdijk	MARIS
gergely.sipos@egi.eu	Gergely Sipos	EGI Foundation
andrea.manzi@egi.eu	Andrea Manzi	EGI Foundation
lukasz.dutka@cyfronet.pl	Lucasz Dutka	Cyfronet
aloga@ifca.unican.es	Alvaro Lopez	CSIC
jpina@lip.pt	Joao Pina	LIP
david@lip.pt	Mario David	LIP

Technical Plan

M1-2	Analyse the required architecture and consult IFREMER for existing experience
M3-4	Develop prototype components, declare datahubs at OneData
M5-10	Create integrated working prototype, testing various options (e.g. Cassandra local vs Cassandra in Docker in cloud).
M11	Create report of results to EOSC-HUB

Technical Info

EGI DataHub and OneData info

The main access point for the Pilot is the EGI DataHub : https://datahub.egi.eu based on OneData.

The hub is where the virtual data spaces are created , providers are assigned to support space with physical resources and end users can manage their files.

More info can be found on the EGI DataHub user docs and OneData docs.

EGI DataHub and MARIS space access

For this Pilot a **MARIS** space has been created in the EGI DataHub and the group **seadatanet-onedata** has been created to manage it.

To access the EGI DataHub, users need to use the **EGI Checkin service**. You can find the user guide at AAI_usage_guide in EGI Wiki in order to create an account.

Afterwards, you can add yourself to the **seadatanet-onedata** group and have therefore access to the **MARIS** space,

by accessing the Managing group membership page, searching for the group , ticking on the *Member* option and save the settings.

please contact Andrea Manzi in case of issues.

OneProvider installations

Organization	Hostname	HW details	Status	Version	Main Contacts
LIP	seadata.ncg.ingrid.pt	500 GB volume	installed	19.02.3	Mario David, Joao Pina
IFCA	seadata.ifca.es	500GB volume	installed	19.02.3	Mario David

Meetings

- **Pilot Kickoff meeting: 14 April**
- **2nd meeting : 8 July**
- **3rd meeting: 26 October**
- **minutes for all meetings**